Mechanistic Understanding and Mitigation of Language Model Non-Factual Hallucinations

Lei Yu[1,*], **Meng Cao**[2,3,*], **Jackie Chi Kit Cheung**[2,3], **Yue Dong**[4]
[1]Department of Computer Science, University of Toronto
[2]School of Computer Science, McGill University
[3]Mila – Québec AI Institute
[4]University of California, Riverside
jadeleiyu@cs.toronto.edu
{meng.cao@mail,jcheung@cs}.mcgill.ca
yue.dong@ucr.edu

Abstract

State-of-the-art language models (LMs) sometimes generate *non-factual hallucinations* that misalign with world knowledge. To explore the mechanistic causes of these hallucinations, we create diagnostic datasets with subject-relation queries and adapt interpretability methods to trace hallucinations through internal model representations. We discover two general and distinct mechanistic causes of hallucinations shared across LMs (Llama-2, Pythia, GPT-J): 1) **knowledge enrichment hallucinations**: insufficient subject attribute knowledge in lower layer MLPs, and 2) **answer extraction hallucinations**: failure to select the correct object attribute in upper layer attention heads. We also found these two internal mechanistic causes of hallucinations are reflected in external manifestations. Based on insights from our mechanistic analysis, we propose a novel hallucination mitigation method through targeted restoration of the LM's internal fact recall pipeline, demonstrating superior performance compared to baselines [1].

1 Introduction

Language models (LMs) serve as repositories of substantial knowledge (Petroni et al., 2019; Jiang et al., 2020; Srivastava et al., 2023) through their parametric knowledge gained from pre-training. However, they are susceptible to generating "hallucinations" that contain factual errors. At the level of logit predictions, these hallucinations often display a pattern similar to factual generations. For example, LMs have been observed to produce seemingly confident completions that are, in reality, hallucinations (Dong et al., 2022; Zhang et al., 2023b).

To understand how hallucinations differ from factual outputs and whether they are uniformly generated or equally challenging to fix, thorough analysis tools that monitoring the information flow are required, extending beyond merely last-layer predictions (Kaddour et al., 2023). However, research on understanding the internal mechanisms of hallucination generation is limited. Most efforts on detecting and mitigating hallucinations (Elaraby et al., 2023; Mündler et al., 2023; Manakul et al., 2023; Zhang et al., 2023a) treat the LM as a black box, devising methods based on external features like predictive uncertainty (Xiao and Wang, 2021; Varshney et al., 2023) and logical consistency (Cohen et al., 2023). These approaches provide little insight into the internal mechanisms of factual errors and have been shown to be unreliable with often contradictory signals (Turpin et al., 2023).

In contrast, interpretability research, which examines the internal mechanisms of transformers in white-box settings, enables the identification of components that contribute to accurate factual predictions. For example, existing work has identified several critical model "components" (e.g., attention heads, feedforward layers) related to knowledge flow that are essential for answering questions accurately (Lu et al., 2021; Dai et al., 2022; Meng et al., 2022a; Geva et al., 2023). However, it remains unclear whether the results of mechanistic interpretability on factual predictions can generalize to hallucinations. Specifically, it is *unknown which model components deviate from normal functioning to cause hallucinations*. Localizing the source of non-factual hallucination in LMs may help us design targeted and efficient methods to mitigate hallucinations without significantly impacting utility (e.g., by editing a small set of model weights identified as causing hallucinations, without affecting other parts that are important for information flow).

In this study, we employ mechanistic interpretability (Olah, 2022) to investigate the origins and manifestations of non-factual hallucinations in LMs. To address the lack of datasets for non-

[*]Equal contribution.
[1]Code and data are available at: https://github.com/jadeleiyu/lm_hallucination_mechanisms.

Findings of the Association for Computational Linguistics: EACL 2024, pages 7943–7956

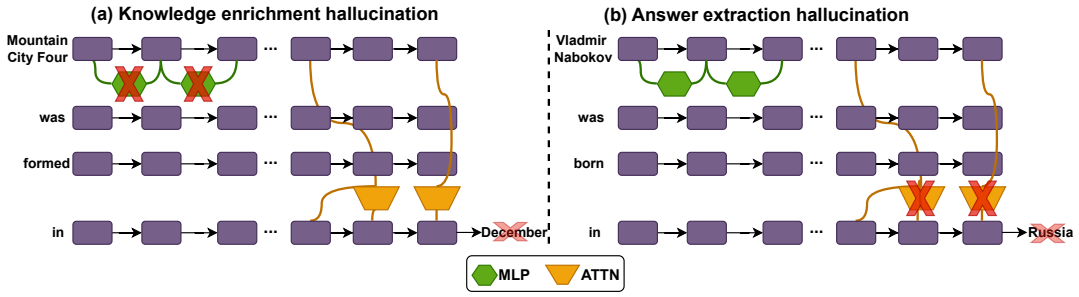

Figure 1: **Our main finding of two non-factual hallucination mechanisms. Left (a)**: The **knowledge enrichment hallucinations** are caused by lacking general knowledge of the subject retrieved from early and middle layer MLPs – in these cases, the subjects tend to be relatively unknown and the incorrect answer is often nonsensical. **Right (b)**: The **answer extraction hallucinations** are caused by the failure of middle and late layer self-attention heads to identify the most relevant object to the given subject and relation – in these cases, the subjects are often more strongly associated with the hallucinating answers than the with the true answers.

factual hallucinations, we constructed a diagnostic hallucination dataset from ParaRel (Elazar et al., 2021), which contains cloze-style factual knowledge queries. This enables the examination of information flow associated with non-factual hallucinations. Specifically, we adapt two established interpretability methods for hallucination analysis—logit lens (Geva et al., 2022b; Dar et al., 2023) and causal mediation analysis (Pearl, 2001; Vig et al., 2020)—aiming to assess the influence of model components on the generation of hallucinatory predictions. Through extensive analyses on LMs of various sizes and architectures (Llama-2, Pythia, GPT-J), we obtain converging evidence that there exist two groups of crucial components for factually incorrect predictions: 1) the multi-layer perceptrons (MLPs) in lower transformer layers, 2) the multi-head self-attentions in upper transformer layers.

Figure 1 illustrates two distinct scenarios where the identified hallucinating components exhibit different behaviors. In some instances, lower-layer MLPs function normally, successfully retrieving semantic attributes about queried entities, while upper-layer attention heads struggle to distinguish the most relevant attributes that lead to the correct answer. In other cases, 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, we demonstrate that the **mechanistic insights gained from our analyses can be leveraged to develop an effective method to reduce LM hallucinations** on multiple open-domain question answering datasets. Our research offers the first mechanistic explanation and mitigation of LM factual errors, fostering future research on model explainability and transparency.

2 Related Work and Background

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; Geva et al., 2021; 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) (Cao et al., 2020; Ji et al., 2023; Zhang et al., 2023b). Here, we focus on the latter type of hallucination. Existing studies aimed at detecting or mitigating hallucinations leverage

features such as internal activation patterns (Yuksekgonul et al., 2023; Li et al., 2023), predictive confidence (Cao et al., 2022a,b; Varshney et al., 2023), and generation consistency (Mündler et al., 2023; Manakul et al., 2023; Zhang et al., 2023a).

Mechanistic interpretability. Mechanistic interpretability (Olah, 2022; Nanda, 2023) is an evolving research area. Recent works employ projections to the vocabulary (Geva et al., 2022b,a; Nostalgebraist, 2020; Katz et al., 2024) and interventions in transformer computation (Finlayson et al., 2021; Haviv et al., 2022; Stolfo et al., 2023; Ghandeharioun et al., 2024) to study LM inner workings, 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 principled account and mitigation method for non-factual hallucinations.

Background and notations Our work builds on the inference pass of decoder-only, transformer-based LMs. An auto-regressive transformer (Vaswani et al., 2017), denoted as G, maps an input sequence of tokens $u = [w_1, ..., w_T]$ 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 [2], 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: $h_i^{(l)} = h_i^{(l-1)} + a_i^{(l)} + m_i^{(l)}$. 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 hidden representations $z \in Z = A \bigcup M$ are causing the model to generate a factually incorrect answer for an input question.

[2] $a_i^{(l,k)} = \text{softmax}\left(\frac{Q_i^k (K_i^k)^T}{\sqrt{d}}\right) \cdot V_i^k$ and Q_i^k, K_i^k, V_i^k are derived from $h_i^{(l-1)}$ with linear transformations.

3 Dataset for LM Hallucination

Dataset Construction We collect a set of questions from the ParaRel (Elazar et al., 2021) dataset of cloze-style factual knowledge queries. Each example in ParaRel consists of a subject-relation-object triple (s, r, o) (e.g., (*Paris*, CAPITAL_CITY, *France*)) and a set of prompts $u(s, r, o)$ generated from hand-curated templates that contains (s, r) and has o as its ground-truth next word continuation (e.g., "The capital city of France is "). To ensure the uniqueness of the true answer for each query, we only take prompts generated from triples in the "many-to-one" relational classes in ParaRel where each subject-relation has a single associated object entity that begins with a capitalized English letter. This yields a large set of approximately 80K factual knowledge queries.

We evaluated three widely used pretrained LMs on our constructed query dataset: 1) Llama-2 (32 layers, 7B parameters, fine-tuned on instruction following) (Touvron et al., 2023), 2) Pythia (Biderman et al., 2023) (32 layers, 6.9B parameters), and 3) GPT-J (28 layers, 6B parameters) (Wang and Komatsuzaki, 2021). For each prompt u, we compute the LM predicted conditional probability $p(t|u)$ of the next token continuation, where t is taken from the collection of all capitalized alphabetical tokens in the model vocabulary. We define the **non-factual hallucination set** as the queries for which a model predicted next token $\hat{t} = \underset{t}{\text{argmax}}\, p(t|u)$ is not a prefix of the true object answer o (i.e., the model makes a factual error), and otherwise the query is an example of the **factual set** (i.e., the model answers correctly). 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 and therefore not suitable for our subsequent analyses. Table 1 summarizes the dataset statistics.

4 Mechanistic Analysis of Hallucinations

We wish to know which "broken" LM components are causing the model to produce a factually incorrect answer. Recent studies have shown that given a subject-relation query, an LM predicts a factual object answer via two steps (Geva et al., 2023; Yu et al., 2023; Jin et al., 2024): 1) during the

	Llama-2	Pythia	GPT-J
No. of factual queries	25204	10277	8646
No. of hallucinating queries	25478	31110	23831
Model accuracy	0.497	0.248	0.266
% of enrichment hall.	22.1	30.2	67.3
% of extraction hall.	77.9	69.8	32.7

Table 1: Statistics of the ParaRel query datasets of three language models.

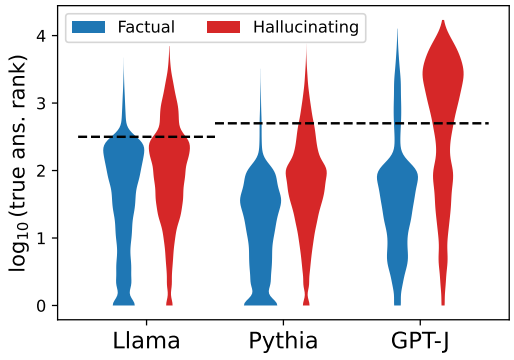

Figure 2: Minimum (over all transformer layers) true object token rankings in the logit lens distributions of intermediate MLP outputs (shown in log scale). Dashed lines denote the threshold $\rho_s^*(o) = 0.01|V|$ ranks to distinguish between knowledge enrichment and answer extraction hallucinations ($\rho_s^*(o) = 320$ for Llama-2 and $\rho_s^*(o) = 502$ for Pythia/GPT-J).

knowledge enrichment step, the model retrieves from MLP sublayers many relevant semantic attributes of the subject and propagates them to the last query token position; and 2) during the **answer extraction step**, the self-attention modules select the most relevant object entity among the previously retrieved attributes. We postulates that an LM hallucinates if any one of these two steps get compromised during inference, and perform a series of mechanistic interpretability analyses to pinpoint the malfunctioning model components.

4.1 Model inspection through logit lens

Method We inspect the semantic information encoded in the intermediate hidden representations within each transformer layer through logit lens (Nostalgebraist, 2020; Elhage et al., 2021; Dar et al., 2023). In particular, for each $z_i^{(l)} \in A \bigcup M$ produced by either the MLP or the self-attention module at layer l when processing the i-th query token, we cast it into a probability distribution over the LM vocabulary space by passing $z_i^{(l)}$ directly through the last prediction head layer:

$$p(z_i^{(l)}) = \text{softmax}(E\ \text{LayerNorm}(z_i^{(l)})); \quad (1)$$

where $E \in \mathbb{R}^{|V| \times d}$ is the unembedding matrix, and LayerNorm is the layer norm operation.

To quantify the information of the true answer o that an LM extracts when processing the subject tokens at each layer, we compute the logit values $\mathcal{I}_m^{(l)}(o) = e_o^T\ \text{LayerNorm}(m_s^{(l)})$ of projecting the MLP-produced hidden representation of the last subject token $m_s^{(l)}$ onto the unembedding vector e_o of the object token. A high value $\mathcal{I}_m^{(l)}(o)$ would indicate that the model has already enriched the subject with sufficient information of an object before processing the relation. Similarly, given the true object token o and another set of attribute tokens $o' \in O'$ with high MLP-enriched information $\mathcal{I}_m^{(l)}(o')$ when processing s [3], we can measure how good the self-attention module in layer l is at distinguishing o against other attributes o' by computing the relative attention-extracted attribute information $\mathcal{I}_a^{(l)}(o) = a_T^{(l)}\left(e_o - \bar{e}_{o'}\right)$, where $a_T^{(l)}$ is the attention module output when processing the last input token, and $\bar{e}_{o'} = \frac{1}{|O'|} e_{o'}$ is the mean unembedding vector of the non-answer attributes. A high value of $\mathcal{I}_a^{(l)}(o)$ suggests that the attention modules can effectively identify o as the target attribute when synthesizing information propagated from subject and relation tokens.

Two mechanisms of hallucinations We first examine whether the LMs retrieve sufficient information about the answer during the subject knowledge enrichment process. We consider an attribute to be sufficiently extracted from the model parametric knowledge base if it is among the top 1% tokens of highest MLP-retrieved information $\mathcal{I}_m^{(l)}(\cdot)$ in at least one some intermediate layers. For each query $u(s, r, o)$ in the factual and the hallucinating set, we compute the minimum ranking $\rho_s^*(o)$ of o in the logit lens distribution $p(m_s^{(l)})$ of MLP outputs across all LM layers, as shown in Figure 2. We observe that for the vast majority of factual set queries, there is at least one intermediate MLP representation in which the object ranks among top 1% of the entire vocabulary. In contrast, for a significant portion of hallucinating examples, even the object token with the most MLP-retrieved information remains outside the top 1% of the vocabulary.

[3] We take the top-100 tokens in model vocabulary with highest $\mathcal{I}_m^{(l)}(o')$ as O'.

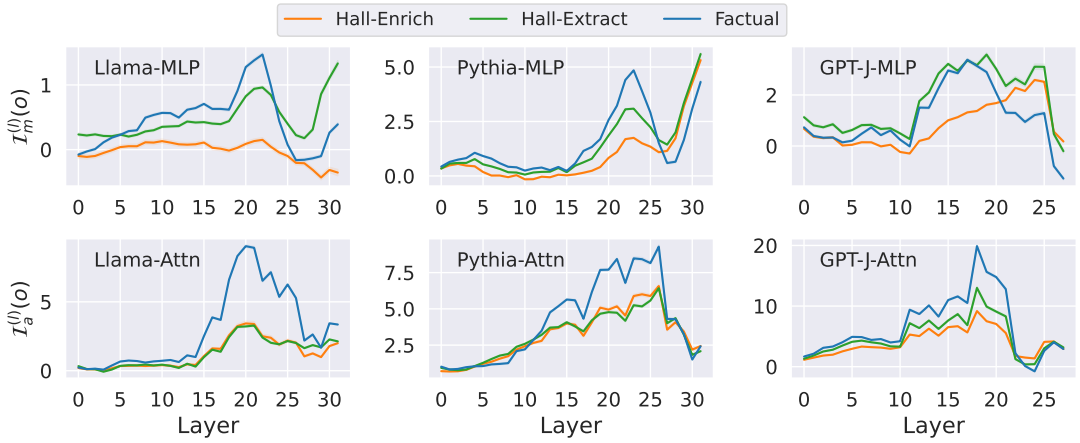

Figure 3: Average logit lens projection values between true object embedding and intermediate MLP/attention representations of Llama-2/Pythia/GPT-J in each transformer layer.

We hypothesize that, for queries with $\rho_s^*(o) > 0.01|V|$, the model hallucinations are mostly caused by the insufficient knowledge extracted from MLPs, and we therefore call these examples **(knowledge) enrichment hallucinations**. On the other hand, for queries with $\rho_s^*(o) \leq 0.01|V|$, the model functions normally at the knowledge enrichment step, but later on fails at the answer extraction step where it cannot distinguish the object entity against the other related attributes of the subject, so we call these examples **(answer) extraction hallucinations**. The last two rows of Table 1 shows the percentage of queries that fall into each hallucination type for the three models. We noticed that the majority of Llama-2 and Pythia errors are extraction hallucinations, while GPT-J have much more enrichment hallucinations, suggesting that GPT-J may suffer from a more severe lack of general world knowledge compared to more recent LMs.

To better understand the two identified hallucination mechanisms, we compute the average layerwise MLP-enriched and attention-extracted object information for factual queries and hallucinating queries with the two error types, as illustrated in Figure 3. Some key observations are: 1) both factual queries and extraction hallucinations retrieve a significant amount of object information from MLPs in early and middle transformer layers, whereas enrichment hallucinations have much less object knowledge incorporated into the subject tokens in early inference stages. 2) Compared to the factual query set, the self-attention module outputs of both types of hallucinations fail to effectively distinguish o against other incorrect attributes.

These findings together suggest that failures of either MLP knowledge enrichment or self-attention answer extraction would cause non-factual hallucinations. Moreover, sufficient retrieval of object knowledge in early layers serves as a prerequisite of answer extraction, so a degenerated enrichment process will inevitably compromise the ability of upper-layer attention to filter irrelevant attributes, as observed in enrichment hallucinations.

4.2 Causal validation of hallucination mechanisms

If lower layer MLP and upper layer self-attention outputs are the root causes of non-factual hallucinations, then fixing them should enhance model factuality. We test this hypothesis by performing a causal patching analysis to measure the contribution of each intermediate representation to a hallucinating model prediction.

Method The intermediate hidden states produced by an LM during 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 producing a hallucination. This is a natural case for *causal mediation analysis* (Vig et al., 2020; Meng et al., 2022a), which quantifies the contribution of intermediate variables in causal graphs. Given a query $u(s, r, o)$ and a model generated incorrect object o', we consider the LM as 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 three model runs:

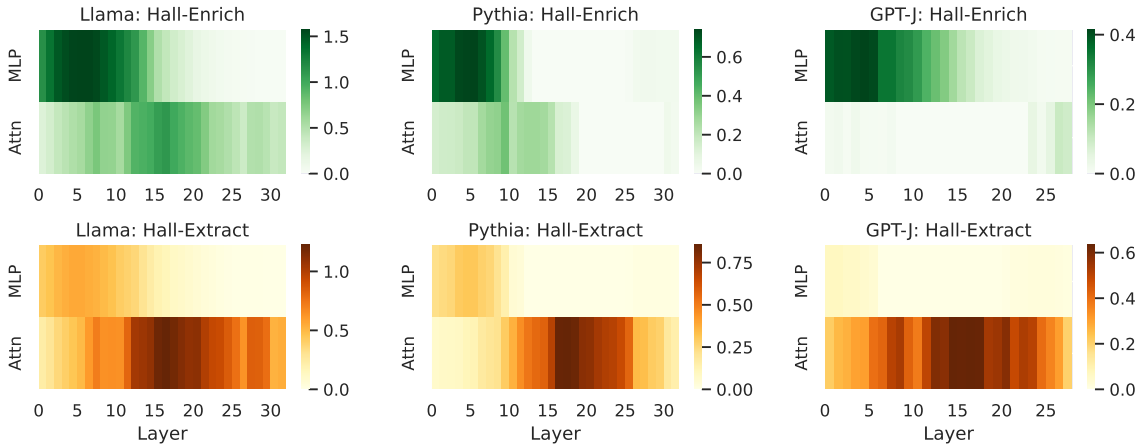

Figure 4: **Average Indirect Effect (AIE)** of mitigating MLP and self-attention intermediate outputs for (a) enrichment hallucinations (green heatmaps) and (b) extraction (orange heatmaps) hallucinations.

1. In the first run, we pass u into the model and extract all intermediate hidden representations z as defined in Section 2, and compute the log likelihood ratio $y = \log \frac{p(o'|u)}{p(o|u)}$ between the true and hallucinated objects, which quantifies the model's "degree of hallucination" when answering u. For a hallucinating prediction, we would observe $y > 0$.

2. In the second run, we inject a Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma)$ [4] to the subject token embeddings of u. Let u^* denote the resulting query with perturbed input embeddings, we re-compute the log-likelihood ratio $y' = \log \frac{p(o'|u^*)}{p(o|u^*)}$ and take those noises with $y' < 1$ (i.e., we only keep noises that make the model become truthful by preferring o over o'). We again extract all intermediate hidden representations z^*.

3. In the third run, we again provide the model with u^* with perturbed input embeddings, and "patch" a particular hidden representation z^* to be the hidden representation z during the first run. We then compute the log likelihood ratio $y'' = \log \frac{p(o'|u^*,z)}{p(o|u^*,z)}$ to see how it changes compared to step 2.

If an intermediate output z is the main cause of a hallucination, then overwriting it with z^* produced during a truthful run should also make the model more factual. We define the causal **indirect effect** $\text{IE}(z; y, u, \epsilon) = y'' - y$ of z as the decrease in the degree of hallucination after mitigating a single hidden state. Averaging over a set of factual queries and a sample of noises for each query, we obtain the average indirect effect (AIE) for each z and its corresponding MLP or self-attention component.

Causal tracing results We compute layerwise AIE for MLP intermediate outputs $m_s^{(l)}$ of the last subject token and the self-attention intermediate outputs $a_T^{(l)}$ of the last input token for each hallucinating query. Figure 4 shows the average MLP and attention causal effects to enrichment and extraction hallucinations respectively. We observe a clear distinction between the causal contribution distributions of the two hallucination mechanisms: in particular, most intermediate hidden states that contribute significantly to enrichment hallucinations are produced by lower layers MLPs when processing subject tokens, and extraction hallucinations are mostly caused by outputs of upper layer self-attention heads right before generating the answer tokens. These findings conform with our logit lens analyses results, and together suggest that 1) *lower layer MLPs and upper layer self-attention heads are the "brittle" LM components which, if compromised, would lead to non-factual hallucinations*, and 2) *lower layer MLPs and upper layer attentions do not always break down simultaneously, thus leading to two distinct mechanisms of LM factual errors.*

4.3 External manifestations of hallucination mechanisms.

To ensure our categorization of hallucinations isn't just a fabricated dichotomy based on internal computation patterns, we also explore external features

[4] We follow (Meng et al., 2022a) to set σ to be be 3 times lof the empirical standard deviation of the input embeddings.

Statistics	Know.Enrich. Hall.	Ans.Extract. Hall.
s-o assoc.	0.47	1.17
s-o' assoc.	0.81	1.69
Robustness	0.86	0.45
Uncertainty	4.54	4.76

Table 2: External data and model prediction features for two types of non-factual hallucination, averaged over three LMs.

to distinguish between the two types. We consider the following features: the **subject-object association strength** is measured as the inner product between the 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 4.2 which, after being added to the input embeddings, fails to make the model prefer the true answer o over o'; the **uncertainty** of model prediction is measured by the entropy of conditional next-token distribution $p(o|u)$. Table 2 summarizes the results with external measures averaged over the three tested LMs.

We found that 1) subjects of extraction hallucinations often have hallucinating objects of much stronger association strengths than true objects, so that the model fail to "offset" the prior propensity of model predicting o' upon seeing s. Subjects of enrichment hallucinations, on the other hand, often have much weaker associations with both true and hallucinating objects; 2) extraction 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 against less relevant attributes; 3) the model is less certain about its predictions when generating enrichment hallucinations, a pattern that is consistent with previous findings that epistemic hallucinations (i.e., hallucinations due to lack of general world knowledge) are often associated with high predictive uncertainty (Xiao and Wang, 2021).

5 Mechanistic Hallucination Mitigation

In this section, we propose a novel Mechanistic Hallucination Mitigation (MHM) method that draws inspiration from our mechanistic analysis, and demonstrate that it can improve LM factuality in open-domain question answering.

Method Given a question x and its true answer y (we take the first token for answers with multiple tokens), we wish to fix the model's imperfect knowledge enrichment and answer extraction modules in the fact-recalling pipeline when it generates an incorrect answer y'. We do so by encourage the LM to retrieve more information of y from MLPs, and to suppress the information propagation of t' from self-attention heads. In particular, during model inference with input x, we take the intermediate MLP outputs $m_i^{(l)}$ and attention head outputs $a_i^{(l)}$ within a specific layer range $l \in L_m$ or $l \in L_a$, and then project them directly to the language modeling head layer. Let $\log p_m^{(i)}(y|x), \log p_a^{(i)}(y|x)$ be the resulting log-likelihoods of y in the projected distribution, we fine-tune the LM to optimize the following objective function:

$$\mathcal{L}_{\text{MHM}}(x, y, y') = -\sum_{l \in L_m} \log p_m^{(l)}(y|x) - \sum_{l \in L_a} \log \frac{p_a^{(l)}(y|x)}{p_a^{(l)}(y'|x)} \quad (2)$$

In practice, we find that $\mathcal{L}_{\text{MHM}}$ can be combined with the regular negative log likelihood (NLL) loss of LM fine-tuning to achieve the best factuality:

$$\mathcal{L}(x, y, y') = \mathcal{L}_{\text{NLL}}([x; y]) + \lambda \mathcal{L}_{\text{MHM}}(x, y, y') \quad (3)$$

where $\mathcal{L}_{\text{NLL}}([x; y])$ is the average NLL loss of the concatenated sequence of a question and its true answer, and λ is a hyperparameter that controls the relative importance of two loss terms.

Data and models. We study non-factual hallucination mitigation of the three LMs we analyzed on two open-domain question answering benchmarks: 1) the **Natural Questions** dataset (Kwiatkowski et al., 2019) that consists of Google search engine queries annotated with answers and supporting Wikipedia pages, and 2) the **TruthfulQA** dataset by Lin et al. (2022) consisting of adversarially constructed commonsense reasoning questions to measure whether an LM is truthful in generating answers. For Natural Questions, we asked each LM to generate up to 20 tokens conditioned on each question, and label the model generation as correct if it contains an exact match of the true answer. For TruthfulQA, where each question is paired with a set of "plausible sounding but false" answers, we evaluate each LM under a multiple-choice scheme by computing the average conditional likelihood per token for each candidate answer, and define a correct prediction as the case where an LM assigns

highest average likelihood for the true answer. Following the evaluation scheme in LM knowledge editing and factual error rectification, we would expect a mitigation method to significantly reduce model hallucination without compromising its originally possessed knowledge. We therefore take examples of both benchmarks on which an LM produces an incorrect answer as our training set, and then construct two evaluation datasets: the **effectiveness evaluation set** consists of the GPT-4-generated paraphrases of each training question on which the original model hallucinates, and the **specificity evaluation set** are the original benchmark questions that the LMs correctly answers. A good mitigation method should therefore achieve high accuracy on both evaluation sets.

Baseline Methods. We evaluate MHM against several baseline methods that have shown promising results in model editing or factualilty improvement: 1) the vanilla supervised fine-tuning method without the MHM objective, 2) a 5-shot in-context learning (ICL) method of prompting the model with five examplar (question, true answer) pairs, 3) the MEND algorithm for knowledge editing (Mitchell et al., 2021) that learns a hypernetwork to perform targeted weight updates on knowledge-intensive LM parameters, and 4) DoLa (Chuang et al., 2023)-a decoding algorithm by contrasting the differences in logits obtained from projecting the later transformer layers versus earlier layers to the vocabulary space. We use default hyperparameters and experimental setups taken from their official implementations, and report the evaluation results on the same test datasets of MHM. [5]

Results. Table 3 shows model accuracy on paraphrase and specificity evaluation sets. We found that in all setups, MHM either yields the most effective mitigation results, or achieves a performance that is comparable to the best mitigation method. Meanwhile, MHM in most cases preserves more than 90% of model performance on the specificity evaluation sets, indicating that our mechanistic mitigation of hallucinations does not compromise LMs' general world knowledge. In contrast, other baselines often yield inferior performance on at least one of the two datasets. In particular, knowledge editing methods such as MEND struggles at Truthful QA on which an LM often hallucinates due to failing to distinguish between the true answer and confusing distractors, while decoding rectification methods such as DoLa help little on model errors on Natural Questions that are often caused by insufficient knowledge about query entities. These results suggest that extensive reparation of the entire LM fact-recalling pipeline is essential for effective and specific mitigation of non-factual hallucinations.

[5] See Appendix D for additional details.

	Natural Questions Eff./Spec. (%)	**Truthful QA** Eff./Spec. (%)
Llama-2-7B-chat		
ICL (5-shot)	27.5 / 91.7	36.8 / **97.9**
SFT	41.8 / 82.3	44.1 / 96.5
MEND	39.8 / 86.9	33.7 / 40.5
DoLa	27.0 / 69.6	43.4 / 81.0
MHM (Ours)	**47.6 / 95.5**	**48.2** / 95.6
Pythia-6.9B		
ICL (5-shot)	22.6 / **98.5**	28.3 / 92.0
SFT	37.9 / 89.9	34.7 / **96.7**
MEND	34.3 / 91.8	25.4 / 50.4
DoLa	23.8 / 66.5	**46.6** / 88.4
MHM (Ours)	**46.9** / 92.4	45.9 / 93.5
GPT-J		
ICL (5-shot)	13.8 / 74.9	30.6 / **96.1**
SFT	34.4 / 86.0	46.0 / 92.3
MEND	36.7 / 89.2	33.3 / 83.7
DoLa	16.8 / 71.7	37.9 / 90.0
MHM (Ours)	**43.8 / 89.4**	**49.7** / 95.8

Table 3: Evaluation results of various hallucination mitigation methods on Natural Question and Truthful QA. **Effectiveness** refers to model accuracy on paraphrased training questions, and **specificity** denotes the percentage of left-out questions on which the LM still remains truthful after applying a mitigation method.

6 Conclusion

We conducted various interpretability analyses on non-factual hallucinations made by language models. We show that both lower layer MLPs and upper layer attention heads in the model factual knowledge recalling pipeline may operate abnormally during model inference, thereby leading to two distinct mechanisms of LM factual errors: insufficient knowledge enrichment and ineffective answer extraction. Leveraging these insights, we proposed an effective method of LM hallucination mitigation. Our work establishes a mechanistic understanding of LM factual errors, and may inspire future research on explainable approaches of improving the

reliability of language models.

7 Limitation

Our study bears several limitations. Firstly, certain experiments depend on interpreting intermediate layer representations and parameters through projection to the vocabulary space via logit lens. While widely used, this method only approximates the encoded information of model components, particularly in early layers. Future work should consider more sophisticated methods such as Tuned Lens (Belrose et al., 2023) to probe information encoded in LM layers. Secondly, our focus on analyzing 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 that may distract LMs from their normal inference processes.

References

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

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.

Meng Cao, Yue Dong, and Jackie Cheung. 2022a. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jingyi He, and Jackie Chi Kit Cheung. 2022b. Learning with rejection for abstractive text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9768–9780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

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.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart M Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843.

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, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. 2022a. Lm-debugger: An interactive tool for inspection and intervention in transformer-based language models. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 12–21.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022b. 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.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscope: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*.

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.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *arXiv preprint arXiv:2402.18154*.

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*.

Shahar Katz, Yonatan Belinkov, Mor Geva, and Lior Wolf. 2024. Backward lens: Projecting language model gradients into the vocabulary space. *arXiv preprint arXiv:2402.12865*.

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.

Kaiji Lu, Zifan Wang, Piotr Mardziel, and Anupam Datta. 2021. Influence patterns for explaining information flow in bert. *Advances in Neural Information Processing Systems*, 34:4461–4474.

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*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

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*.

A Full list of ParaRel relational classes

See Table 4 for a complete list of N-to-1 ParaRel relational classes and sample queries that we used to construct our mechanistic hallucination analysis dataset.

B Examples of knowledge enrichment and answer extraction hallucinations

Table 6 presents several examples randomly drawn from the sets of early-site and late-site hallucinations made by Llama-2-7b-chat. We found that in many examples of answer extraction 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 even predicts the subject itself as a continuation. For knowledge enrichment 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 Details of causal patching analysis of hallucinations

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) 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. 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 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.

D Details of hallucination mitigation experiments

D.1 Training and evaluation datasets for hallucination mitigation

We first evaluated each LM on Natural Questions and Truthful QA. For Natural QA, the model takes an input prompt of question and is then asked to generate up to 20 tokens conditioned on the input through greedy decoding, and if the generated continuation does not contain an exact match of the true answer, the model answer is labeled as a hallucination. For TruthfulQA, where each question is paired with a set of "plausible sounding but false" answers, we evaluate each LM under a multiple-choice scheme by computing the average conditional likelihood per token for each candidate answer, and define a correct prediction as the case where an LM assigns highest average likelihood for the true answer.

We experimented with multiple input prompt templates, and found that the model performance was overall insensitive to the wording of a query, so we chose a simple input template "Question: {QUESTION}. Answer:", where {QUESTION} is substituted with a real question in the two datasets. Similarly, for in-context learning baseline method, we simply prepend 5 (question, true answer) in the same format before the input question.

D.2 Hallucination mitigation methods

Here we elaborate on the hallucination mitigation methods we applied to improve LM factuality on open-domain question answering.

Mechanistic Hallucination Mitigation (MHM) For our MHM method, based on our findings shown in Figure 3 that MLPs and self-attentions write most information about the true answer in middle transformer layers, we set both L_m and L_a in Equation 2 to be $[20, 21, 22, 23, 24, 25]$, on which we inject additional information about the true answers to enhance model factuality. For the loss importance hyperparameter λ in Equation 3, we found that a range of λ values between 0.1 to 1.0 will in general yield good results, so we choose to report MHM results with $\lambda = 1.0$.

MEND MEND (Mitchell et al., 2021) is a method for learning to transform the raw fine-tuning gradient into a more targeted parameter update that successfully edits a model in a single step. We adapt the original implementation of Mitchell et al. (2021) by learning a gradient transformation hyper-network for the last 3 MLP blocks of each LM. We then fine-tune each LM on the same training datasets as the SFT and MHM methods using the transformed gradient signals returned by the learned hypernetwork. We also experimented with editing gradients of the self-attention modules, but

Table 4: PARAREL relations with unique object answers and sample queries.

Relation ID	Relation	No. of queries	Sample Query	True answer
P103	native language	977	The mother tongue of Victor Horta is	Dutch
P138	named after	645	Rawlings Gold Glove Award, which is named for	glove
P159	headquarters location	967	The headquarter of Strait Shipping is located in	Wellington
P176	manufacturer	982	Honda RA272 is produced by	Honda
P264	record label	429	Johnny Carroll's record label is	Decca
P279	subclass of	964	Nucleoporin 62, a type of	protein
P30	continent	975	Romulus Glacier is located in	Antarctica
P407	language of work or name	877	Ten Years Gone is a work written in	English
P449	original network	881	Himalaya with Michael Palin was originally aired on	BBC
P495	country of origin	909	Mundo Obrero was from	Spain
P1376	capital of	234	Guangzhou is the capital of	Guangdong
P36	capital	703	The capital city of Porto District is	Porto

Table 5: Hyperparameters of hallucination mitigation experiments.

Hyperparameter Name	Hyperparameter Value
Learning rate (all methods)	1e-4
Training batch size per device (all methods)	4
N_epoch training (SFT)	8
N_epoch training (MHM)	8
N_epoch training (MEND)	4

did not observe any significant improvement on performance of the mitigated LMs.

DoLa **D**ecoding by C**o**ntrasting **La**yers (DoLa) (Chuang et al., 2023) is a method of better surfacing factual knowledge embedded in an LLM without retrieving external knowledge or additional fine-tuning. DoLa rectifies the output next-token distribution of an LM by contrasting it with logit-lens-projected next-token distributions of dynamically selected intermediate layers. We use the official implementation by Chuang et al. (2023) and apply it directly on the three LMs. As DoLa does not require model fine-tuning, we simply evaluate the LMs on the same effectiveness and specificity evaluation datasets using the rectified decoding strategy.

See Table 5 for a full list of additional hyperparameters used in hallucination mitigation experiments. All experiments were run on a single computing cluster with 4 Nvidia-A100 GPUs of 80GB memory.

prompt	subject	relation	true object	predicted object	hallucination mechanism
Korrespodent is formed in	Korrespodent	country of origin	Ukraine	April	knowledge enrichment
Mantecadas, that was created in	Mantacedas	country of origin	Spain	Japan	knowledge enrichment
Asprey's headquarters are in	Asprey	headquarters location	London	New	knowledge enrichment
Vladmir Nabokov was born in	Vladmir Nabokov	place of birth	Germany	Russia	answer extraction
Anti Cimex, developed in	Anti Cimex	country of origin	Sweden	France	answer extraction
Darrell Nulisch originates from	Darrell Nulisch	place of birth	Dallas	Chicago	answer extraction

Table 6: Examples of knowledge enrichment and answer extraction hallucinations made by Llama-2-7b-chat.

Question	True answer
who ordered the forbidden city to be built	Zhu Di
who has won the most games in nfl 2017	Dallas Cowboys
who was the famous scientist that ran the research lab moseley went to in manchester	Sir Ernest Rutherford

Table 7: 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 8: Sample questions from the TruthfulQA dataset.