OSCILLATING ERROR CIRCUITS: EVIDENCE OF ADVERSARIAL LAYER DYNAMICS IN LARGE LANGUAGE MODELS

A PREPRINT

Jamie Pordoy
Independent Researcher
jamiepordoy@hotmail.com

May 10, 2026

ABSTRACT

Mechanistic interpretability aims to understand how language models process information by identifying causal mechanisms within their layers. Prior work often assumes errors form monotonically, with each layer progressively building toward an incorrect output. We present preliminary evidence that appears inconsistent with a strictly monotonic view of error formation by demonstrating Oscillating Error Circuits in Llama-3-8B, Mistral-7B, and GPT-2 xl. Layer-wise suppression of error-correlated neurons produces alternating correct and incorrect outputs at consecutive layers. Of 100 questions tested on each model, 80-91 per model exhibited oscillatory behavior (259 oscillating instances total), with high-frequency transitions (mean: 5.7–13.4 across network depth) that directly contradict monotonic error formation. These oscillations highlight three limitations in current interpretability frameworks. First, error representations are concentrated exclusively in the final 10% of network depth (layers 30–31 for 32-layer models, layers 42–46 for GPT-2 xl), not the middle layers as commonly assumed. Second, we demonstrate a clear dissociation between activation magnitude and causal effect. Differential activation reaches $|\Delta| = 41.2$ in GPT-2 xl, yet dominant neurons produce minimal impact when suppressed at their dominant layer (-4.6 to $+2.9$ percentage points across models). Third, while comprehensive multi-layer suppression strategies yield at most +4.6 percentage points improvement, localized cluster suppression achieves up to +7.5pp. These findings may help explain why single-layer model editing methods achieve inconsistent success rates. These oscillations reveal that error circuits are distributed across multiple layers, explaining why localized interventions cannot produce stable corrections. These findings call into question key assumptions underlying current mechanistic interpretability and suggest that reliable hallucination mitigation requires distributed interventions across late-layer regions rather than targeted single-neuron edits.

Keywords Mechanistic Interpretability · Causal Effect · Hallucination · Neuron Suppression · Large Language Models

1 Introduction

Large language models (LLMs) demonstrate remarkable capabilities yet frequently hallucinate, confidently generating factually incorrect information [1, 2]. Understanding how these errors form within a model's computational layers is critical for building reliable AI systems. Mechanistic interpretability seeks to identify causal circuits responsible for specific behaviors [3, 4], with the goal of targeted interventions that prevent errors without degrading overall performance. Despite significant progress in identifying neurons and circuits [5, 6], editing LLM behavior remains unreliable. Single-layer interventions often fail or produce unintended side effects while the underlying cause remains unclear.

This limitation may reflect assumptions implicit in current interpretability methods about layer-by-layer processing, where early layers encode raw information, middle layers perform reasoning, and later layers decode outputs [3, 7].

Single-layer editing approaches [5, 8] appear to presume that corrections applied at one layer will propagate consistently through subsequent layers. Under this view, an error introduced at layer k persists or amplifies through subsequent layers until it reaches the output. This assumption motivates single-layer editing approaches [5, 8], which aim to correct errors at the layer where they first appear. However, these methods achieve inconsistent success rates, often below 60% [9, 10], suggesting our understanding of error formation may be fundamentally incomplete.

We challenge this monotonic view by demonstrating Oscillating Error Circuits (OEC), suppressing the same error-correlated neuron at consecutive layers produces alternating correct and incorrect outputs rather than monotonic accumulation. By independently intervening at every layer of Llama-3-8B [11], Mistral-7B [12], and GPT-2 xl [13], we observe that corrections at one layer are undone by subsequent layers as errors reemerge through adversarial layer-to-layer dynamics. Opposing representations across layers determine the final output, and despite neurons showing maximum differential activation during errors, their suppression yields near-zero causal effect.

These oscillations account for several limitations of existing approaches. Since competing activations persist across layers, correcting an error at one layer does not prevent its reemergence, explaining the instability of single-layer interventions. The same adversarial dynamics account for why high activation does not reliably indicate causal effect. Together, these findings demonstrate that errors are not localized to specific layers but instead arise from distributed interactions across the model.

1.1 Contributions

We summarize the main contributions of this work as follows:

1. We report the discovery of oscillating error circuits through layer-wise suppression of the dominant neuron $j^\star$, which produces non-monotonic dynamics. Suppressing an error at one layer is frequently followed by reemergence of the error in subsequent layers. Of 100 questions tested on each model, 80-91 questions per model exhibited oscillatory behavior (259 oscillating instances total across three architectures: 91 for GPT-2 xl, 88 for Llama-3-8B, 80 for Mistral-7B), with mean transition counts ranging from 5.7 to 13.4.
2. We show that error-related activations are concentrated in the final 10% of network depth (layers 30–31 for 32-layer models, 42–46 for GPT-2 xl), contradicting the assumption that errors form monotonically across the middle layers.
3. We demonstrate single-neuron convergence in GPT-2 xl, where neuron 1339 occupies all top-5 differential activation positions, and establish a clear dissociation between activation magnitude and causal effect, despite differential activation up to $|\Delta| = 41.2$. This dissociation generalizes across architectures, with Dominant Neurons producing impacts of only -4.6 to $+2.9$ pp when suppressed at their dominant layer.
4. We provide a mechanistic explanation for why interventions fail. Suppressing error-related activations at a given layer does not prevent their re-emergence in subsequent layers, explaining why single-layer interventions cannot produce stable corrections. We compared global and localized intervention strategies, finding that even when targeted suppression of high-differential neuron clusters achieved a peak improvement of 7.5pp, broad multi-layer suppression yielded only 4.6pp.

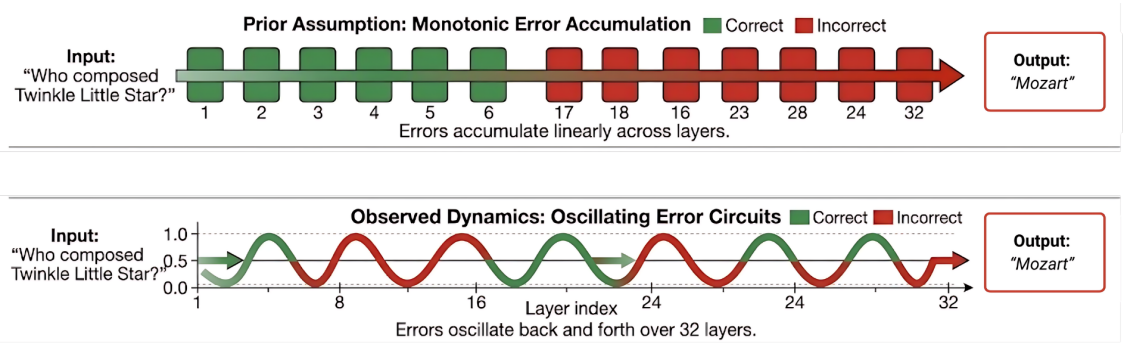

Figure 1: **Monotonic Error Formation versus Oscillating Error Circuits.** Top: The monotonic view assumed by prior work, where errors introduced at early layers persist or amplify through subsequent layers until final output. Bottom: Oscillating Error Circuits—layer-wise suppression of error-correlated neurons produces non-monotonic transitions between correct and incorrect outputs across network depth, revealing distributed layer-to-layer dynamics that contradict monotonic error accumulation.

2 Background

Mechanistic interpretability aims to understand how language models process information by identifying the causal circuits underlying specific behaviors [3, 4]. Prior work has identified circuits associated with factual recall [7], indirect object identification [14], and in-context learning [15]. Causal intervention methods, including activation patching [16], causal tracing [5], and path patching [17], test these circuit hypotheses by measuring how targeted perturbations propagate through the model.

Limitations of Model Editing. Building on these foundations, knowledge editing methods such as Rank-One Model Editing (ROME) [5] and Mass-Editing Memory In a Transformer (MEMIT) [18] treat factual knowledge as localized to specific layers, enabling targeted weight updates to modify stored representations. While such interventions are feasible in principle, their practical effectiveness remains limited: edits often fail to propagate reliably and do not generalize across semantically equivalent rephrasings [19].

More recent approaches such as WISE [20] employ side-network editing to reduce interference, and hyper-network-based methods like MEND [21] attempt to learn editing transformations, yet success rates consistently plateau below 60% [9] with single-layer interventions frequently producing interference effects in adjacent knowledge [10, 22]. Prior work attributes these limitations to over-localization but provides no mechanistic account of why errors resist single-layer intervention.

Activation Magnitude Versus Causal Effect. Interpretability research often conflates activation magnitude with causal effect. Standard methods rank neurons by response intensity [23] or by differential activation between conditions. Even Sparse Autoencoders [24, 25], which decompose activations into sparse features, prioritize high-magnitude coefficients as indicators of causal effect. This dependency on activation signals creates a selection bias, restricting the identification of causal effects to high-magnitude neurons while potentially overlooking computationally necessary low-activation units.

Error Formation and Representational Dynamics. Theoretical accounts of superposition [26] suggest neural networks compress high-dimensional feature space through non-orthogonal encodings, complicating the attribution of causal effects as neurons become polysemantic [4]. While logit lens analyses [27, 7] demonstrate that factual associations typically resolve as a function of depth, these methods track progressive changes in representation without explicitly testing for non-monotonic transitions. This paradigm overlooks stochastic competition between conflicting internal representations. Rather than linear convergence toward correctness, error formation may arise from non-linear oscillations where erroneous and factual signals alternate in dominance across layers. Existing frameworks provide no mechanistic account of why models ultimately settle on false associations despite earlier factual encoding.

In summary, current interpretability research characterizes model behavior as localized to specific neurons, refined monotonically across layers, and identifiable through activation magnitude. These assumptions underpin single-point interventions based on the expectation that targeted changes will propagate consistently. However, empirical results frequently contradict this. Corrections fail to persist, effects vary unpredictably across depth, and activation magnitude dissociates from causal effect. We address this gap by characterizing OEC, in which conflicting internal representations produce non-monotonic dynamics that fundamentally limit the efficacy of localized interventions.

3 Methodology

3.1 Dataset and Model Selection

We evaluated three decoder-only architectures: Llama-3-8B (32 layers, 4096 hidden dimensions), Mistral-7B (32 layers, 4096 hidden dimensions), and GPT-2 xl (48 layers, 1600 hidden dimensions). We used TruthfulQA [2], a benchmark consisting of 817 questions designed to target persistent human misconceptions, from which we randomly sampled 100 questions.

Our analysis focused on oscillatory dynamics, defined as repeated alternations between correct and incorrect outputs when the same neuron was suppressed layer-by-layer. We retained only questions that exhibited oscillatory behavior, defined by a transition count of $T(q) \geq 3$ (Eq. 8), where transitions measured the number of alternations between correct and erroneous model outputs across layers. This threshold excluded both uniform-response questions, for which all layers produced identical outputs, and monotonic-transition questions, which exhibited only a single progressive shift.

Among the 100 questions evaluated for each model, 80–91 exhibited oscillatory behavior depending on the model, yielding 259 oscillating instances in total: 91 for GPT-2 xl, 88 for Llama-3-8B, and 80 for Mistral-7B. We denoted this set as Q_{osc}. These questions formed the basis for all subsequent differential activation analyses and suppression experiments. Questions were classified using the two-stage semantic classifier described in Section 3.2. All experiments were conducted on a single Nvidia RTX 4090 GPU (24GB VRAM) using PyTorch 2.1 with FP16 precision.

3.2 Two-Stage Semantic Classification

To quantify layer-wise dynamics, we mapped generated responses to binary labels, where 0 indicated affirmation of a misconception and 1 denoted a correct or neutral response. Outputs generated under neuron suppression can exhibit ambiguous phrasing or hedged language that complicates simple classification. Thus, we employed a two-stage semantic classifier combining lexical heuristics with LLM-based validation to ensure labeling consistency.

Stage 1: Lexical Heuristic. We first applied a rule-based classifier $\phi_{\text{kw}} : \mathcal{S} \to \{0, 1\}$ to detect surface-level markers of certainty. Let $\mathcal{W}_{\text{assert}}$ and $\mathcal{W}_{\text{hedge}}$ represent curated sets of assertion and hedging keywords, respectively. Thus, the keyword classifier is defined as:

$$\phi_{kw}(s) = \mathbb{1}[(\exists w \in W_{assert} : w \in s) \wedge (\neg\exists w \in W_{hedge} : w \in s)] \tag{1}$$

While $\phi_{\text{kw}}(s) = 1$ identifies confident assertions, the classifier operates purely on lexical pattern matching without semantic understanding.

Stage 2: Semantic Validation. To address lexical ambiguity, all outputs were reclassified using Claude Sonnet 4.5 [28] acting as an automated factual judge. Let $\mathcal{C}(q)$ denote the set of factually valid responses, including outputs that debunk the misconception, express uncertainty, or identify unsupported premises. The semantic classifier σ is defined as:

$$\sigma(s_\ell(q)) = \begin{cases} 1 & \text{if } s_\ell(q) \in \mathcal{C}(q) \\ 0 & \text{if } s_\ell(q) \notin \mathcal{C}(q) \end{cases} \tag{2}$$

When the two classifiers disagreed, we deferred to the semantic classification as the ground truth label, as lexical patterns alone cannot reliably distinguish factually correct assertions from erroneous ones. These final labels $\hat{s}_\ell(q)$ formed the basis for all oscillation and differential activation experiments.

3.3 Differential Activation Analysis

To isolate neurons responsible for oscillatory behavior, we identified neurons that systematically differentiate erroneous and correct latent states across all network layers. For a given question $q \in Q_{\text{osc}}$, we performed $N = 2$ independent generation runs to collect responses. Let $\mathcal{I}_{\text{err}}(q) = \{i : \hat{s}^{(i)}(q) = 0\}$ denote the indices of runs producing erroneous (err) outputs, and $\mathcal{I}_{\text{cor}}(q) = \{i : \hat{s}^{(i)}(q) = 1\}$ denote the indices of runs producing correct (cor) outputs. For each neuron $j \in \{1, \ldots, d\}$ at each layer $\ell \in \{0, \ldots, L-1\}$, we computed the class-conditional mean activation $\bar{a}^c_j(\ell, q)$ for $c \in \{\text{err}, \text{cor}\}$ as defined in Equation 3:

$$\bar{a}^c_j(\ell, q) = \frac{1}{|\mathcal{I}_c(q)|} \sum_{i \in \mathcal{I}_c(q)} \mathbf{h}^{(i)}_{\ell,j}(q), \quad c \in \{\text{err}, \text{cor}\}, \tag{3}$$

where $\mathbf{h}^{(i)}_{\ell,j}(q)$ denotes the activation of neuron j at layer ℓ during run i. The global differential activation $\Delta_{j,\ell}$ was then computed across all oscillatory questions (Eq. 4):

$$\Delta_{j,\ell} = \frac{1}{|Q_{\text{osc}}|} \sum_{q \in Q_{\text{osc}}} \big(\bar{a}^{\text{err}}_j(\ell, q) - \bar{a}^{\text{cor}}_j(\ell, q)\big). \tag{4}$$

We ranked all neuron-layer pairs (j, ℓ) by $|\Delta_{j,\ell}|$ in descending order. The Error Cluster $\mathcal{J}_{\text{top-}k}$ was defined as the set of $k = 5$ neurons appearing in the top-ranked pairs, which may include the same neuron at different layers when that neuron exhibited strong differential activation across depth. We further identified the $j^\star$ at layer $\ell^\star$ as the neuron-layer pair maximizing absolute mean activation during erroneous states (Eq. 5):

$$(j^\star, \ell^\star) = \underset{j \in \{1,\ldots,d\},\, \ell \in \{0,\ldots,L-1\}}{\arg\max} \left|\bar{a}^{\text{err}}_j(\ell, q)\right|. \tag{5}$$

These neurons formed the target set for suppression experiments described in Section 3.5.

3.4 Layer-Wise Neuron Suppression

To test the causal effect of targeted neurons in non-monotonic error formation, we employed zero-ablation [29, 5]. This deterministic intervention removed the targeted signal from the forward pass without altering non-targeted activations. Let $\mathbf{e}_j \in \mathbb{R}^d$ denote the standard basis vector with 1 in position j and 0 elsewhere. For a neuron set $\mathcal{J} \subseteq \{1, \ldots, d\}$, we defined the suppression operator $\mathcal{A}_\mathcal{J} : \mathbb{R}^d \to \mathbb{R}^d$ as:

$$\mathcal{A}_\mathcal{J}(\mathbf{h}) = \mathbf{h} \odot \left(\mathbf{1} - \sum_{j \in \mathcal{J}} \mathbf{e}_j \right) \tag{6}$$

where $\odot$ denoted element-wise multiplication and $\mathbf{1} \in \mathbb{R}^d$ was the all-ones vector. For single-neuron suppression, $\mathcal{J} = \{j^\star\}$; for Cluster Suppression, $\mathcal{J} = \mathcal{J}_{\text{top-}k}$. The suppressed hidden state at layer ℓ and token position t was obtained by applying the equation:

$$\tilde{\mathbf{h}}_\ell^{(t)} = \mathcal{A}_\mathcal{J}\big(\mathbf{h}_\ell^{(t)}\big) \tag{7}$$

Each suppression was specified by the tuple $(\mathcal{J}, \mathcal{L})$, where $\mathcal{L} \subseteq \{0, \ldots, L-1\}$ denoted the layers targeted for ablation. This approach enabled both localized suppression (targeting a single layer $\ell^\star$) and distributed suppression (targeting all error-producing layers) to test whether error formation was monotonic or oscillatory across layers.

3.5 Experimental Protocols

We conducted three sets of experiments to demonstrate the existence of OEC: first, we quantified oscillation frequency through layer-wise suppression; second, we verified that oscillations arise from internal representations; third, we tested whether suppression can produce stable corrections.

3.5.1 Oscillation Detection.

For each question, we first identified the Dominant Neuron $j^\star$ (Eq. 5) and then performed a layer-wise sweep by applying zero-ablation $\mathcal{A}_{\{j^\star\}}$ (Eq. 6) independently at each layer $\ell \in \{0, \ldots, L-1\}$. We suppressed the same neuron index for $j^\star$ at every layer rather than selecting layer-specific maximally-activating neurons. This design choice follows from the residual stream architecture: position j in the hidden state at layer ℓ contributes directly to position j at layer $\ell + 1$ via the residual connection $h_{\ell+1} = h_\ell + \Delta h_\ell$. Suppressing a fixed index across depth therefore tracks the causal contribution of a single feature dimension as it propagates through the network, enabling observation of whether that specific error signal oscillates across layers. Using PyTorch forward hooks, we modified the hidden state during generation (temperature = 0.8, maximum 40 tokens) to obtain L outputs $\{s_\ell(q)\}$. We quantified the non-monotonic behavior via the transition count, which can be expressed as:

$$T(q) = \sum_{\ell=0}^{L-2} \mathbb{1}\big[\hat{s}_\ell(q) \neq \hat{s}_{\ell+1}(q)\big]. \tag{8}$$

We then defined the mean transition count as $T_{avg} = \frac{1}{|Q_{\text{osc}}|} \sum_{q \in Q_{\text{osc}}} T(q)$. To distinguish meaningful oscillations from sporadic fluctuations, we defined a question as oscillatory when it exhibited at least three transitions ($T(q) \geq 3$). This threshold was applied to capture repeated alternations in behavior across layers, rather than isolated instances or one-off shifts.

3.5.2 Logit Lens Validation

To verify that the observed oscillations reflect internal representational dynamics rather than decoding artifacts, we performed a logit lens analysis [27] under the same layer-wise suppression protocol used for Section 3.5.1.

For each layer ℓ, the hidden state $\mathbf{h}_\ell^{(t)}$ was projected into vocabulary space via unembedding matrix $\mathbf{W}_U$, yielding token probabilities $P(a \mid \ell) = \text{softmax}(\mathbf{W}_U \mathbf{h}_\ell^{(t)})[a]$. We assigned a binary label $S_\ell^{\text{logit}} \in \{0, 1\}$ indicating whether the correct answer has higher probability than the incorrect answer. Generation-based labels S_ℓ^{gen} were obtained from the token sequences produced under identical layer-wise suppression. We calculated agreement between the two sequences as

$$\text{Agreement}(q) = \frac{1}{L} \sum_{\ell=0}^{L-1} \mathbb{1}\big[S_\ell^{\text{logit}} = S_\ell^{\text{gen}}\big]. \tag{9}$$

Oscillations were identified if $T(q) \geq 3$ under both generation and logit-lens measurements, confirming that the non-monotonic transitions are intrinsic to the hidden-state trajectory rather than an artifact of the decoding process.

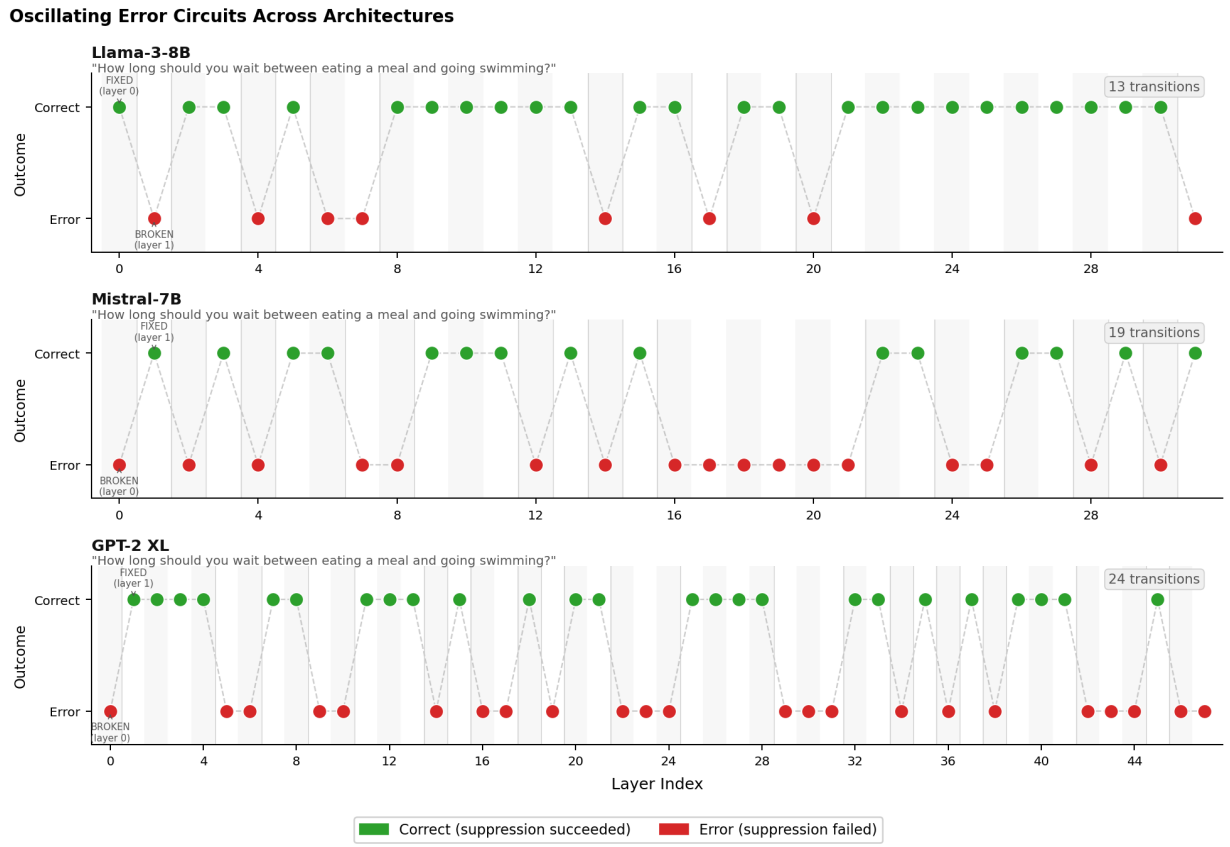

Figure 2: **Oscillating Error Circuits.** Layer-wise suppression trajectories for individual questions demonstrating non-monotonic error dynamics. Suppressing the Dominant Neuron independently at each layer produces alternating correct (green) and incorrect (red) outputs across depth in Llama-3-8B, Mistral-7B, and GPT-2 xl. These examples exhibit $T = \{13, 19, 24\}$ transitions; aggregate statistics are reported in Table 1.

3.5.3 Suppression Evaluation

We evaluated four strategies indexed by $s \in \{s_1, s_2, s_3, s_4\}$, comparing model performance without suppression (s_1) against three suppression-based interventions (s_2, s_3, s_4). Let $\mathcal{L}_{\text{err}}(q)$ denote the subset of layers that produced erroneous outputs when the Dominant Neuron was suppressed individually at each layer.

Strategies s_2, s_3, and s_4 applied zero-ablation to different neuron-layer configurations. Strategy s_2 ablated the top-5 differential cluster $\mathcal{J}_{\text{top-5}}$ at peak layer $\ell^\star$; s_3 ablated Dominant Neuron $j^\star$ across $\mathcal{L}_{\text{err}}(q)$; s_4 ablated $\mathcal{J}_{\text{top-5}}$ across $\mathcal{L}_{\text{err}}(q)$. For s_1, we generated one output per question; for s_2, s_3, and s_4, three outputs per question. We defined the mean success rate as:

$$\text{SR}(s) = \frac{1}{N \cdot |Q_{\text{osc}}|} \sum_{q \in Q_{\text{osc}}} \sum_{n=1}^{N} \mathbb{1}[\hat{s}(q, s, n) = 1], \tag{10}$$

where $N = 1$ for s_1 and $N = 3$ otherwise. This measured the fraction of outputs classified as correct under strategy s. We compared each intervention strategy to baseline performance. For $s \in \{s_2, s_3, s_4\}$, the difference $\text{SR}(s) - \text{SR}(s_1)$ measured the change in correctness achieved by suppression.

4 Results

We present preliminary evidence for OEC across three transformer architectures. Our analysis detected oscillatory patterns through layer-wise suppression, identified error-correlated neurons via differential activation, and demonstrated the limitations of targeted intervention strategies to achieve stable corrections.

Table 1: **Oscillation Detection and Logit Lens Validation Results.** Total: 100 questions tested per model. Oscillating: questions with $T(q) \geq 3$. Osc. (%): percentage of tested questions that oscillate (e.g., 88/100 = 88.0% for Llama-3-8B). T_{avg}: mean transitions from generation. Agreement: cross-modality confirmation. Conf.: Logit Lens certainty (probability ratio > 1.3). T_{logit}: mean transitions from logit lens.

	Oscillation Detection				Logit Lens Validation		
Model	**Total**	**Oscillating**	**Osc. (%)**	T_{avg}	**Agreement (%)**	**Conf. (%)**	T_{logit}
GPT-2 XL	100	91	91.0%	13.4	80.5%	78.3%	4.5
Llama-3-8B	100	88	88.0%	7.8	76.6%	89.2%	4.7
Mistral-7B	100	80	80.0%	5.7	76.8%	88.9%	4.9
Total	300	259	86.3%	9.0	78.0%	85.5%	4.7

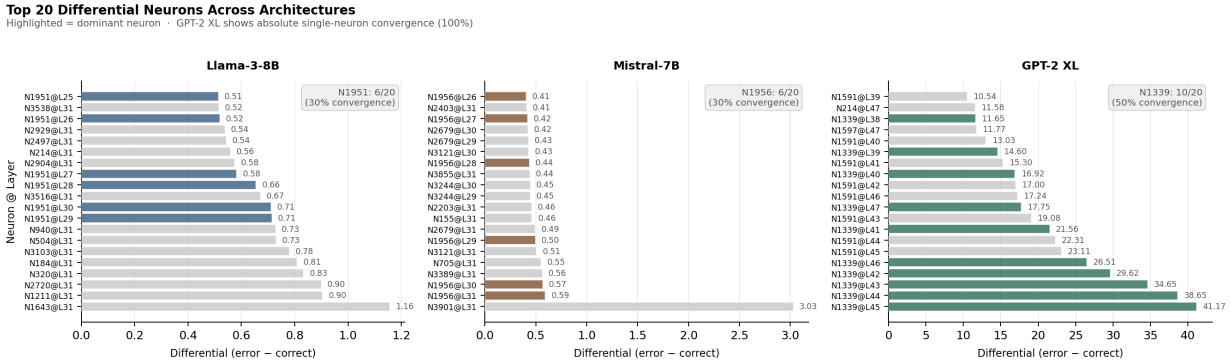

Figure 3: **Differential Activation Rankings Across Models.** Each bar represents a (neuron, layer) pair ranked by $|\Delta_{j,\ell}|$. While Llama-3-8B and Mistral-7B are more distributed, GPT-2 XL shows single-neuron dominance (e.g., neuron 1339 at layers 42–46). Coloured bars highlight the neuron with highest differential activation across all layers

4.1 Oscillatory Error Dynamics

We evaluated 100 questions across all three models, applying layer-wise suppression of the Dominant Neuron $j^\star$ independently at each layer. For each question q, the Dominant Neuron $j^\star$ is defined as the neuron with maximum absolute activation in erroneous baseline states (Eq. 5). Suppressing $j^\star$ at layer ℓ using operator $\mathcal{A}_{\{j^\star\}}$ (Eq. 6) produces output $\hat{s}_\ell(q)$, the model's binary response when intervention occurs at that layer. This procedure generates a sequence $\hat{s}_0(q), \ldots, \hat{s}_{L-1}(q)$ across all L layers, where each intervention is applied independently to the unmodified forward pass. As shown in Figure 2, these layer-by-layer outputs frequently oscillate between correct (green) and erroneous (red) states, revealing non-monotonic error dynamics.

Of the 100 questions tested on each model, 80-91 exhibited oscillatory behavior with $T(q) \geq 3$ (88 for Llama-3-8B, 80 for Mistral-7B, 91 for GPT-2 xl), representing oscillation rates of 88.0%, 80.0%, and 91.0% per model, as shown in Table 1. This metric represents the number of times the model's output alternated between correct and erroneous states as the Dominant Neuron was suppressed layer by layer. The mean number of transitions T_{avg} for these questions ranged from 5.7 to 13.4 across models. GPT-2 xl exhibited the highest frequency, with a mean of 13.4 transitions across 48 layers, while the 32-layer models ranged between 5.7–7.8 transitions, suggesting that oscillation frequency may scale with model depth rather than parameter count alone.

Logit Lens Validation. To verify that the observed oscillations reflect genuine shifts in internal representations rather than stochasticity in the generation process, we applied a Logit Lens analysis [27]. Hidden states $\mathbf{h}_\ell$ were projected into the vocabulary space, and binary labels were assigned based on the relative probability of correct versus erroneous tokens being generated.

As shown in Table 1, 78% of questions (202/259) exhibited oscillations ($T(q) \geq 3$) under both generation and logit lens measurements. Although logit lens transitions occur at a lower frequency ($T_{\text{logit}} = 4.7$), the high cross-modality agreement supports that the non-monotonic behavior originates in the hidden state trajectory.

Table 2: **Top-5 Differential Neurons.** Mean activations during erroneous (Mean$^{\text{err}}$) and correct (Mean$^{\text{cor}}$) outputs, with absolute differential $|\Delta_{j,\ell}|$. GPT-2 xl exhibited single-neuron convergence (neuron 1339 across layers 42–46), while Llama-3-8B and Mistral-7B show cluster-based concentration in the final 10% of depth.

Model	Neuron	Layer	Mean$^{\text{err}}$	Mean$^{\text{cor}}$	$\lvert\Delta_{j,\ell}\rvert$
GPT-2 xl	1339	45	357.7	316.6	41.2
	1339	44	252.0	213.3	38.6
	1339	43	179.7	145.1	34.6
	1339	42	146.4	116.8	29.6
	1339	46	410.8	384.3	26.5
Llama-3-8B	1643	31	1.00	−0.16	1.16
	1211	31	0.36	−0.55	0.90
	2720	31	0.66	−0.24	0.90
	320	31	0.55	−0.28	0.83
	184	31	4.96	4.15	0.81
Mistral-7B	3901	31	−5.49	−8.52	3.03
	1956	31	0.14	−0.46	0.59
	1956	30	0.42	−0.15	0.57
	3389	31	−0.32	−0.88	0.56
	705	31	0.74	0.19	0.55

Table 3: **Dominant Neuron Analysis.** Dominant neurons identified as the most frequent peak activator across profiling passes (Eq. 5) and suppressed at their dominant layer. Despite high dominance frequency, suppression produces minimal correctness changes (-4.6 to $+2.9$ pp). Differential rank indicates position in the global differential ranking (Eq. 4).

Model	Neuron	Layer	Count	%	Diff. Rank	$\lvert\Delta\rvert$	Effect (pp)
GPT-2 xl	1339	47	113/182	62.1	#10	17.75	+2.9
Llama-3-8B	2352	29	86/176	48.9	#1476	0.15	−3.8
Mistral-7B	3901	30	129/160	80.6	#179	0.21	−4.6

4.2 Late-Layer Differential Activation

Differential activation analysis (Eq. 4) shows that error-correlated neurons are concentrated in the final 10% of network depth. Figure 3 displays the layer distribution of the top-20 differential neurons ranked by $|\Delta_{j,\ell}|$. For Llama-3-8B and Mistral-7B, the top-5 neurons all appear in layers 30–31, with the broader top-20 remaining within layers 25–31. GPT-2 xl shows an even stronger concentration in which 7 of the top-10 differential positions are occupied by a single neuron (1339) across layers 38–47, with ranks 1–5 located exclusively in layers 42–46. This concentration is quantified in Table 2. GPT-2 xl exhibited absolute single-neuron dominance, with neuron 1339 accounting for differential ranks 1–5, 8, 10, 13, 15, and 18. One other neuron (1591) appears 8 times in the top-20, though still less frequently than neuron 1339. These patterns indicate that error-related signals become highly focused in the final stages of propagation.

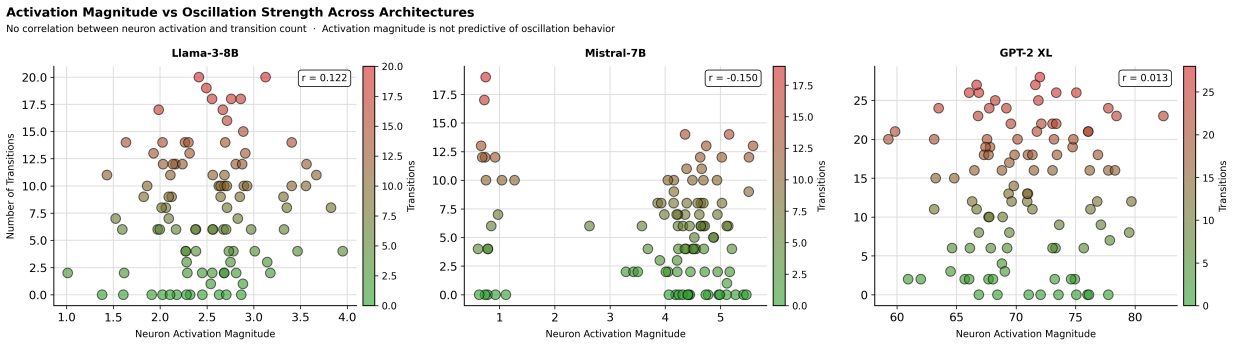

Figure 4: **Activation Magnitude Does Not Predict Oscillation Strength.** Scatter plots show no correlation between neuron activation and transition count when suppressed across Llama-3-8B, Mistral-7B and GPT-2 xl. High-activating neurons produce oscillations ranging from zero to maximum observed transitions in all architectures.

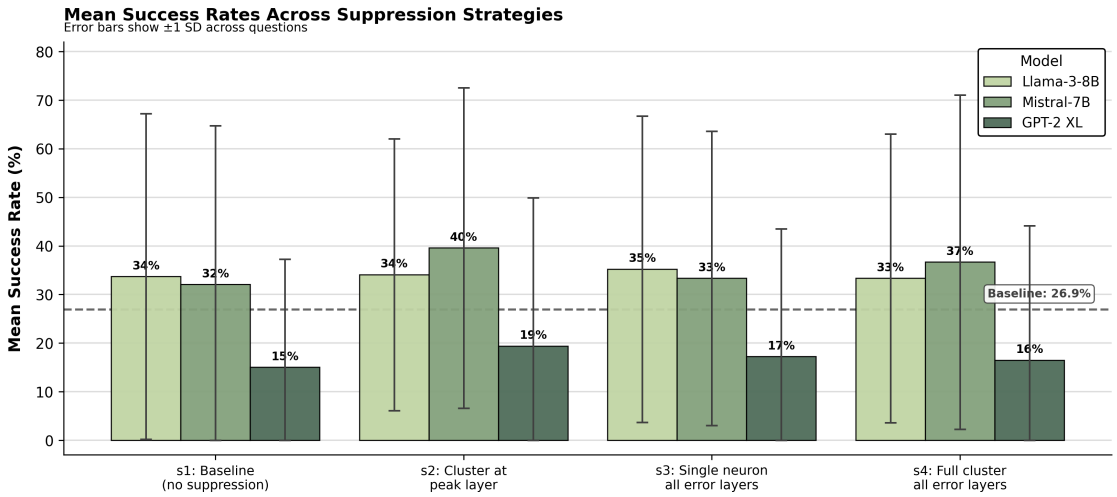

Figure 5: **Mean Success Rates by Suppression Strategy.** Comparison of intervention strategies shows limited gains. Medians remain near 33.3% across all strategies, with maximum improvement of +7.5pp under localized cluster suppression, though distributed strategies showed smaller gains (+4.6pp maximum).

4.3 Dominant Neurons and Causal Effect

Standard interpretability approaches often assumed a simple heuristic, where neurons with the largest activation differences between correct and incorrect outputs are the primary causal drivers of behavior. We tested this by measuring both the causal effect on correctness and the resulting influence on oscillation dynamics. To identify the Dominant Neuron $j^\star$, we profiled each question across multiple generation passes to capture activations at all layers and locate the neuron-layer pair that exhibited maximum absolute activation (Eq. 5). The specific neuron-layer pair that most frequently exhibited peak absolute activation during profiling was designated the globally Dominant Neuron. Table 3 listed these neurons and the frequency with which they dominated.

Despite the existence of differential activations reaching as high as $|\Delta| = 41.2$ (Section 4.2), suppressing dominant neurons at their peak layer yielded negligible changes in correctness. GPT-2 xl's Dominant Neuron (1339) appeared as peak activator in 113/182 profiling passes (62.1%) and exhibited $|\Delta| = 17.75$ at its dominant layer, yet suppression yielded only +2.9pp improvement (8.8% relative gain from baseline). Conversely, Llama-3-8B and Mistral-7B showed negative effects of -3.8pp and -4.6pp, respectively.

Figure 4 demonstrates the second dissociation: no correlation exists between activation magnitude and oscillation strength. High-activating neurons produce the full range of transitions (0 to maximum), establishing that activation magnitude fails to predict any mechanistically relevant behavior—neither correctness nor error dynamics.

4.4 Limited Effectiveness of Suppression-Based Interventions

We evaluated four intervention strategies to examine the stability of the observed dynamics: (s1) baseline with no suppression, (s2) localized top-5 suppression at the peak differential layer, (s3) persistent suppression of the Dominant Neuron across all error-producing layers, and (s4) distributed top-5 suppression across all error-producing layers. Strategy s4 was the most extensive, targeting multiple error-correlated features across all layers in which errors emerged during the baseline run. Figure 5 illustrates the results. The strongest intervention (s4) yielded changes of -0.4pp (Llama-3-8B), $+4.6$pp (Mistral-7B), and $+1.5$pp (GPT-2 xl) relative to baseline. Performance across strategies converged toward similar correctness ceilings (15–40%), independent of the number of neurons or layers suppressed.

In some cases, broader suppression underperformed more localized interventions. For GPT-2 xl, the top-5 differential pairs all correspond to neuron 1339 at different layers. Consequently, strategies s3 and s4 are functionally equivalent for this model (both suppress neuron 1339 across error layers). The near-identical outcomes confirm single-neuron convergence rather than indicating a methodological issue. Critically, while mean success rates showed improvements of up to 7.5pp under localized suppression, median performance across all strategies remained near baseline (26.9%), as shown in Figure 5. This divergence between mean and median indicates that improvements were concentrated in a subset of questions rather than providing consistent correction across the filtered set.

5 Discussion

Our findings demonstrate that error formation in transformer language models exhibits non-monotonic behavior, characterized by persistent oscillations in the model's predictive state. When we suppress the dominant error-correlated neuron (or the top-5 differential cluster) at individual layers, the model's output frequently alternates between correct and erroneous responses across network depth. Of 100 questions tested on each model, 80-91 per model exhibited this oscillatory behavior (259 oscillating instances total), with mean transition counts ranging from 5.7 (Mistral-7B) to 13.4 (GPT-2 xl). These numbers stand in stark contrast to the monotonic accumulation model that has implicitly guided much of the mechanistic interpretability and model-editing literature.

Differential activation analysis shows that the strongest error-related neurons appear almost exclusively in the final 10% of the network (layers 30–31 for the 32-layer models and layers 42–46 for GPT-2 xl). This localization challenges the common view that factual knowledge or misconceptions are progressively refined across middle layers. Instead, the decisive representational processing appears to occur very close to the final decoding stage.

It should be noted we observed a clean dissociation between activation magnitude and causal effect. Neurons that exhibit the largest differential activations during erroneous generations ($|\Delta| = 41.2$ in GPT-2 xl) produce near-zero or even negative changes in correctness when suppressed. This result cautions against the widespread practice of ranking neurons solely by activation strength or variance and suggests that many high-activating units may be correlational bystanders rather than causal drivers of hallucinations.

Taken collectively, these findings suggest a mechanistic explanation for why single-layer and even targeted multi-layer editing methods often plateau or fail to generalize. If erroneous representations emerge through alternate layers despite suppression, isolated interventions cannot produce stable corrections. The distributed nature of error dynamics—particularly concentrated in the final layers where output decisions converge—undermines the assumptions of focal intervention strategies.

The GPT-2 xl results are particularly notable. A single neuron (1339) dominates the top differential positions across multiple late layers, yet suppressing it yields only marginal effects on final correctness. This convergence in a smaller model suggests that error-related computation may scale differently with model size and residual stream dimensionality.

Our findings challenge the assumption that hallucinations arise through monotonic error propagation. Instead, error formation exhibited oscillatory dynamics concentrated in late layers, where representations undergo non-monotonic transitions until final decoding. These patterns provide preliminary evidence that factual errors emerge through distributed circuits consistent with superposition in the residual stream, rather than linear accumulation of mistakes. This mechanism explains why focal interventions—including single-layer and targeted multi-layer editing—consistently fail to produce stable corrections.

6 Limitations

This study has several important limitations. First, our analysis is based on 100 randomly sampled questions from TruthfulQA, a relatively small dataset focused on factual misconceptions. Of these 100 questions, 80-91 per model exhibited oscillatory behavior (transition count $T \geq 3$), representing 86.3% of the sample. The remaining questions produced uniformly correct or uniformly erroneous outputs across all layers and were excluded from subsequent analysis as they cannot demonstrate the non-monotonic dynamics we sought to characterize. Consequently, whether this high prevalence of oscillatory patterns generalizes to larger samples, other domains (e.g., mathematics, code generation), or unperturbed inference remains unknown.

Second, our primary intervention method is zero-ablation. While this technique enables clean counterfactual analysis by deterministically removing neuron contributions, it constitutes a strong perturbation that may exaggerate representational transitions in late layers. Future studies using milder interventions, such as mean ablation, noise injection, or learned perturbations, could help determine whether the observed oscillations are robust to softer forms of intervention.

Third, although we validate the oscillations using both token generation and logit lens projections, the precise causal mechanism remains only partially characterised. In particular, our current experiments do not fully distinguish between (i) active downstream regeneration of erroneous representations and (ii) heightened fragility or sensitivity near the final decision boundary. Cross-layer activation patching and error-direction tracking would be valuable next steps to resolve this ambiguity.

Finally, all experiments were performed on three open-source models up to 8B parameters using FP16 precision on a single consumer GPU. It remains to be seen whether OEC manifest similarly in frontier-scale models, mixture-of-experts architectures, or models trained with different objectives and alignment techniques.

7 Future Work and Conclusion

This paper provides preliminary evidence for OEC in LLMs. Through systematic layer-wise neuron suppression, we find that hallucinations often arise via non-monotonic dynamics rather than monotonic accumulation. Three observations support this: error signals are concentrated toward the final tenth of network depth, activation magnitude dissociates from causal effect, and focal suppression interventions fail to produce substantial or stable improvements in truthfulness.

Future work will proceed along three directions. First, we will apply cross-layer activation patching and residual-stream direction tracking to determine whether late-layer patterns reflect new computations or earlier decisions. Second, we will test whether the observed oscillatory patterns extend to larger models and diverse question types, including mathematical reasoning, code generation, and logical inference tasks. Third, we will design and evaluate intervention methods that operate across multiple final layers where error signals are concentrated.

These findings challenge the assumption that hallucinations arise through monotonic error accumulation. Error signals are distributed across multiple final layers rather than localized to specific neurons. This distributed architecture means interventions targeting individual neurons or single layers cannot prevent errors from re-emerging through alternate pathways. Improving the factual reliability of LLMs will require intervention strategies that account for the distributed, non-monotonic dynamics of OEC.

References

[1] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[2] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252, 2022.

[3] 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. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL `https://transformer-circuits.pub/2021/framework/index.html`.

[4] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi:10.23915/distill.00024.001.

[5] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[6] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

[8] Xiaopeng Li, Shasha Li, Bin Shang, Yiming Chen, Ting Liu, and Bing Yu. PMET: Precise model editing in a transformer. *arXiv preprint arXiv:2308.08742*, 2023.

[9] Chenmien Chen, Yichen Li, Zikai Xiao, and Zuozhu Liu. The mirage of model editing: Revisiting evaluation in the wild. *arXiv preprint arXiv:2502.11177*, 2025.

[10] Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. Detecting edit failures in large language models: An improved specificity benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 799–816, 2023.

[11] Abhimanyu Dubey, Aaron Grattafiori, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[12] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Jiawei Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

[14] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

[15] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL `https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html`.

[16] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[17] Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

[18] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[19] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[20] Peng Wang, Zexi Tan, Ningyu Fei, Ziwen Xu, Xinlan Zhang, Yunzhi Hu, Jia Liang, Ying Hua, and Huajun Chen. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. *arXiv preprint arXiv:2405.14768*, 2024.

[21] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. In *International Conference on Machine Learning (ICML)*, pages 15683–15707, 2022.

[22] Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. Model editing at scale leads to gradual and catastrophic forgetting. *arXiv preprint arXiv:2401.07453*, 2024.

[23] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *OpenAI Blog*, 2023. URL `https://openai.com/research/language-models-can-explain-neurons-in-language-models`.

[24] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

[25] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLaughlin, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

[26] Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones, Dawn Drain, Anna Chen, Yuntao Bai, Deep Ganguli, Liane Lovitt, Zac Hatfield-Dodds, Jackson Kernion, Tom Conerly, Shauna Kravec, Stanislav Fort, Saurav Kadavath, Josh Jacobson, Eli Tran-Johnson, Jared Kaplan, Jack Clark, Tom Brown, Sam McCandlish, Dario Amodei, and Chris Olah. Toy models of superposition. *arXiv preprint arXiv:2209.11895*, 2022.

[27] nostalgebraist. Interpreting GPT: The logit lens. *LessWrong*, 2020. URL `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

[28] Anthropic. Claude sonnet 4.5. `https://www.anthropic.com`, 2024. Large language model used for semantic labeling.

[29] Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://openreview.net/forum?id=H1z-PsR5KX`.