Truth Neurons

Haohang Li **Yupeng Cao** **Yangyang Yu** **Jordan W. Suchow*** **Zining Zhu***

Stevens Institute of Technology

{hli113, ycao33, yyu44, jws, zzhu41}@stevens.edu

Abstract

Despite their remarkable success and deployment across diverse workflows, language models sometimes produce untruthful responses. Our limited understanding of how truthfulness is mechanistically encoded within these models jeopardizes their reliability and safety. In this paper, we propose a method for identifying representations of truthfulness at the neuron level. We show that language models contain *truth neurons*, which encode truthfulness in a subject-agnostic manner. Experiments conducted across models of varying scales validate the existence of truth neurons, confirming that the encoding of truthfulness at the neuron level is a property shared by many language models. The distribution patterns of truth neurons over layers align with prior findings on the geometry of truthfulness. Selectively suppressing the activations of truth neurons found through the TruthfulQA dataset degrades performance both on TruthfulQA and on other benchmarks, showing that the truthfulness mechanisms are not tied to a specific dataset. Our results offer novel insights into the mechanisms underlying truthfulness in language models and highlight potential directions toward improving their trustworthiness and reliability. The code of the paper can be found at https://github.com/Acatsama0871/TruthNeurons.

1 Introduction

Language models have demonstrated remarkable text-generation capabilities across various tasks (Jiang et al., 2024; Li et al., 2024; Zhang et al., 2023), but they struggle to consistently produce correct outputs in certain question-answering scenarios (Huang et al., 2025, 2024b). The struggle arises partly because language models lack sufficient relevant knowledge about specific questions in their pretrained data (Chang et al., 2024). Moreover, language models may generate incorrect answers despite recognizing the incorrectness of the responses (Zheng et al., 2023). For instance, prior research has shown that language models aligned with human feedback tend to accommodate users' incorrect responses, even when the models initially identify these responses as false (Wei et al., 2024a). Although the correctness of language models can be substantially improved through self-consistency checking (Manakul et al., 2023), post-training (Tonmoy et al., 2024), and optimizing decode strategies (Chuang et al., 2024; Chen et al., 2024), it is still unknown whether there exists a *truth mechanism*, a special mechanism within language models that drives the generation of accurate answers.

Research on mechanistic interpretability has begun to probe representations of truthfulness through analyses of hidden states: Orgad et al. (2024) applied linear probes to reveal meaningful patterns of truth-related encoding. Marks and Tegmark (2023) identified specific tokens and layers involved in truthfulness and demonstrated a linear encoding of truth and falsehood using principal component analysis (PCA). Ferrando et al. (2024) used sparse autoencoders (SAEs) to identify the features related to entity awareness and hallucination.

Despite these advancements, neuron-level mechanisms of truthfulness remain unknown. The neuron is a fundamental level of analysis in both the human brain and Transformer-based neural networks. For example, specific neurons in the human brain (e.g., those in the dorsolateral and ventrolateral prefrontal cortex) selectively activate when performing certain cognitive operations, such as evaluating the truthfulness of particular events (Jamali et al., 2021; Quiroga et al., 2008; Hubbard et al., 2008; Jenkins et al., 2016). Analogous to these observations in the human brain, the transformer-based models that underlie language models also exhibit functional specialization. Transformers are believed to activate distinct regions selectively, facilitating interactions necessary for informed decision-making,

*Co-corresponding authors.

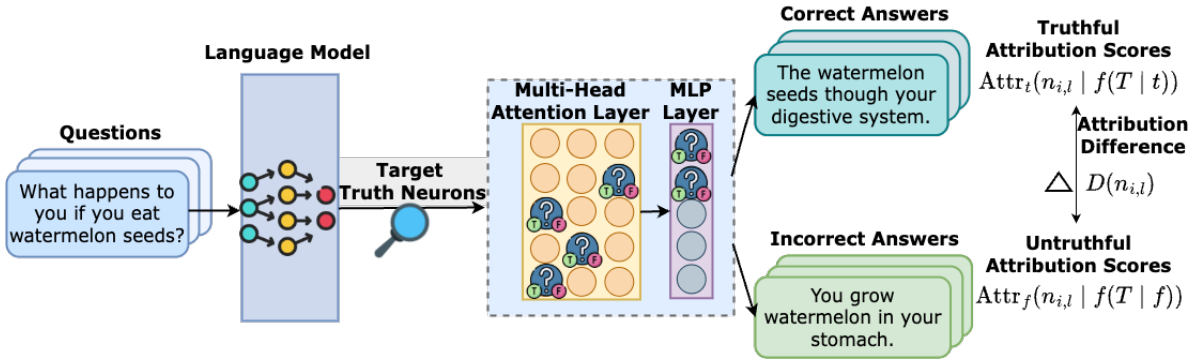

Figure 1: Overview of our method that detects the truth neurons.

such as true-or-false judgments (Kumar et al., 2024; Wei et al., 2024b; Kan et al., 2022; Sun et al., 2024; Kim et al., 2023). At the neuron level, recent research has also observed the knowledge storage and retrieval mechanisms to varying extents (Dai et al., 2021; Niu et al., 2024), but as we will show, the mechanisms to process truth differ from those of the knowledge entities. Truth mechanisms are not localized to specific entities (even datasets), whereas the knowledge storage is localized to each data entry.

Here, we develop a method informed by neuroscience and interpretability research to detect *truth neurons*, specialized truth-processing structures within language models. Our method starts with an axiomatic attribution (Sundararajan et al., 2017), using integrated gradients to measure neuron attribution scores for truthful vs. untruthful responses. We identify candidate neurons positively contributing to truthfulness and negatively correlated with untruthfulness. We then apply a systematic filtering procedure to select a small subset of neurons causally linked to truthfulness representations. Upon suppressing these identified truth neurons, we observe a statistically significant reduction in accuracy on the TruthfulQA benchmark (Lin et al., 2021). Further analysis reveals that this reduction is not biased toward any specific category, suggesting that these neurons encode a general, category-agnostic representation of truthfulness. Additionally, we demonstrate that the influence of the truth neurons generalizes effectively to other truthfulness benchmarks. As we will show in our experiments, identifying and analyzing truthfulness mechanisms at neuronal granularity reveals insights that deepen our understanding of truthfulness representations in Transformer-based language models.

In summary, our work makes the following contributions:

- We propose a novel method to identify truth neurons. By analyzing neuron attributions, we successfully isolate a small subset of neurons whose activations have a statistically significant impact on the model's ability to discern truthfulness (Section 2).
- Through carefully designed experiments, we demonstrate that the identified neurons encode general, example-agnostic representations of truthfulness, and that their influence generalizes effectively to other out-of-distribution truthfulness benchmarks (Section 3.2 & Section 3.3).
- Finally, we investigate the distribution patterns of these identified truth neurons across model layers, observing a consistent pattern that aligns closely with existing findings (Section 3.4).

We believe our results offer insights into improving the trustworthiness and safe deployment of language models, highlighting promising future directions for enhancing model alignment with truthfulness.

2 Methodology

In this section, we propose an integrated gradient-based approach (Sundararajan et al., 2017) to systematically identify and isolate neurons causally associated with a model's ability to discern the truthfulness of factual statements. Within Transformer architectures, feed-forward (MLP) layers have been characterized as key-value memory structures closely tied to factual knowledge recall (Geva et al., 2021); attention heads have similarly been linked to truthfulness representations (Li et al., 2023). Therefore, we extend neuron attribution

analyses to encompass both MLP and attention modules across all intermediate layers.

2.1 Preliminaries

As our goal is to identify neurons correlated with the model's truthfulness behavior, integrated gradient is a suitable tool, as it satisfies desirable axioms and effectively quantifies each neuron's contribution to model behavior. Following the setup of Sundararajan et al. (2017). Let $\boldsymbol{X} \in \mathbb{R}^n$ be the input tensor of the neural network, $\boldsymbol{X}' \in \mathbb{R}^n$ be the baseline input tensor required by the method, and $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}$ denote the function representing the neural network. Additionally, define $n^{\mathrm{i}}_{i,l}$ as the intermediate neuron activation output at layer l and index i, with $n^{\mathrm{b}}_{i,l}$ representing the corresponding activation when the baseline input is applied. The integrated gradient method computes neuron attribution as a path integral along the straight-line path $\gamma(\alpha)$ from $\boldsymbol{X}'$ to $\boldsymbol{X}$, where α represents the incremental interpolation parameter indicating progress along the path:

$$\gamma(\alpha) = n^b_{i,l} + \alpha(n^i_{i,l} - n^b_{i,l}), \quad \alpha \in [0, 1] \tag{3}$$

$$\mathrm{Attr}(n_{i,l} \mid \boldsymbol{f}(\boldsymbol{X})) := \int_0^1 \frac{\partial \boldsymbol{f}(\gamma(\alpha))}{\partial \gamma(\alpha)} d\alpha \tag{4}$$

$$= \int_0^1 \frac{\partial \boldsymbol{f}(\gamma(\alpha))}{\partial \gamma(\alpha)} \frac{d\gamma(\alpha)}{d\alpha} d\alpha \tag{5}$$

$$= \left(n^i_{i,l} - n^b_{i,l}\right) \int_0^1 \frac{\partial \boldsymbol{f}\left(n^b_{i,l} + \alpha\left(n^i_{i,l} - n^b_{i,l}\right)\right)}{\partial n_{i,l}} d\alpha \tag{6}$$

Intuitively, integrated gradient attribution quantifies a neuron's contribution to the final prediction by measuring how the predicted probability changes as the neuron's activation is gradually shifted from its baseline value toward its activation in the actual input. In the computation, the integral is approximated by a Riemann sum, and the approximation precision is controlled by the step parameter m.

2.2 Identifying Truth Neurons

Notation. For each question $\boldsymbol{q}$, the dataset provides one correct answer $\boldsymbol{t}$ and one incorrect answer $\boldsymbol{f}$, with the incorrect answer closely matching the length and format of the correct answer whenever possible. We construct the input prompt $\boldsymbol{T}$ by appending these two answers after the question in randomized order, labeling them as options A and B. Additionally, an instruction $\boldsymbol{i}$ explicitly prompts the model to select the option that correctly answers the question. We can then denote a dataset $\mathcal{D}$ with N questions as:

$$\mathcal{D} = \{\boldsymbol{T}^{(k)}\}_{k=1}^N = \{\langle q, t, f, i\rangle^{(k)}\}_{k=1}^N, \tag{7}$$

where k indexes the dataset $\mathcal{D}$.

Accounting for upper and lower cases. Let $\mathcal{M}$ denote the language model's output probability distribution. We observed that language models frequently interchange the uppercase and lowercase forms of output labels. To cover both cases, we define the prediction probability $\boldsymbol{f}$ as the sum of both the uppercase probability and the lowercase probability. For example, when the correct answer is labeled A, the probability for the correct response is:

$$\boldsymbol{f}(\boldsymbol{T} \mid t) = \mathcal{M}(\hat{y} = \texttt{A} \mid \boldsymbol{T}) + \mathcal{M}(\hat{y} = \texttt{a} \mid \boldsymbol{T}). \tag{8}$$

Similarly, the probability for the incorrect response labeled B is:

$$\boldsymbol{f}(\boldsymbol{T} \mid f) = \mathcal{M}(\hat{y} = \texttt{B} \mid \boldsymbol{T}) + \mathcal{M}(\hat{y} = \texttt{b} \mid \boldsymbol{T}). \tag{9}$$

This definition applies analogously when the correct answer is labeled B and the incorrect answer A. Note that for both the correct and the incorrect answers, we query the probabilities from the same distribution (i.e., the same prompt $\boldsymbol{T}$), avoiding the lexical biases.

Deconfounding untruthfulness. For a given neuron $n_{i,l}$ at the ith position and the lth layer, applying integrated gradients to the input with respect to the correct and incorrect responses yields $\mathrm{Attr}_t(n_{i,l} \mid \boldsymbol{f}(\boldsymbol{T} \mid t))$ and $\mathrm{Attr}_f(n_{i,l} \mid \boldsymbol{f}(\boldsymbol{T} \mid f))$, the two corresponding attribution scores.

We denote by $\mathrm{Attr}_t^{\mathrm{Avg}}$ and $\mathrm{Attr}_f^{\mathrm{Avg}}$ the average truthful and untruthful attribution scores computed over N examples, respectively. We further define the attribution difference for a single example as $D(n_{i,l})$ in Eq 1, and the average attribution difference across all examples in the dataset as $\bar{D}(n_{i,l})$ in Eq 2.

Empirically, the signs of the truthful attribution scores $\mathrm{Attr}_t(n_{i,l} \mid \boldsymbol{f}(\boldsymbol{T} \mid t))$ and the untruthful attribution scores $\mathrm{Attr}_f(n_{i,l} \mid \boldsymbol{f}(\boldsymbol{T} \mid f))$ can be categorized into four distinct scenarios, which consequently determine the signs of $D(n_{i,l})$:

[1] **Both positive**: The neuron positively contributes to both correct and incorrect responses; the overall attribution difference de-

$$D(n_{i,l}) = \text{Attr}_t(n_{i,l} \mid \boldsymbol{f}(\boldsymbol{T} \mid t)) - \text{Attr}_f(n_{i,l} \mid \boldsymbol{f}(\boldsymbol{T} \mid f)) \tag{1}$$

$$\bar{D}(n_{i,l}) = \frac{1}{N}\sum_{j=1}^{N}(\text{Attr}_t(n_{i,l} \mid \boldsymbol{f}(\boldsymbol{T} \mid t)) - \text{Attr}_f(n_{i,l} \mid \boldsymbol{f}(\boldsymbol{T} \mid f))) = \text{Attr}_t^{\text{Avg}} - \text{Attr}_f^{\text{Avg}} \tag{2}$$

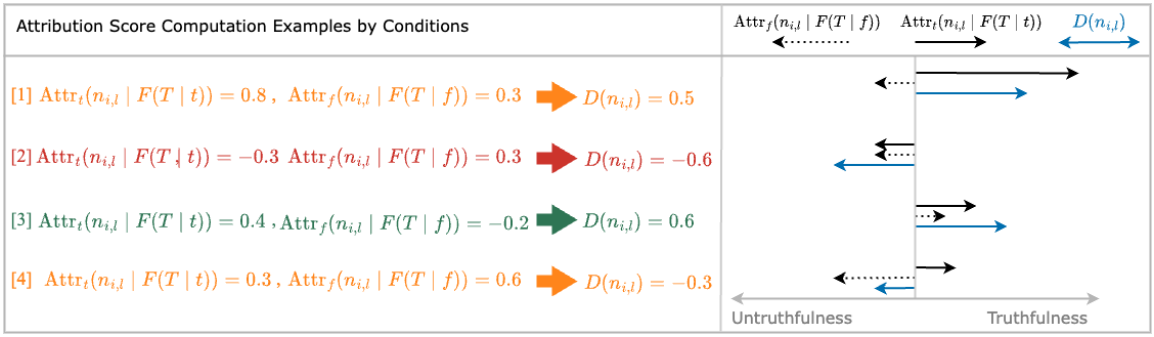

Figure 2: Examples of attribution score computation. The left side shows example attribution scores for truthful and untruthful responses, and the right side shows the resulting attribution differences. Colors correspond to the four scenarios discussed above. In case [1], competing attributions result in a positive difference, indicating a positive correlation with truthfulness, while case [4] illustrates the opposite situation. Case [2] indicates a clear bias toward untruthfulness, whereas case [3] shows a strong bias toward truthfulness.

pends on the relative magnitudes of these contributions.

[2] **Truthful negative, untruthful positive**: The neuron predominantly supports untruthful responses, negatively contributing to truthfulness and positively correlating with untruthfulness. This combination results in a negative attribution difference, indicating it is not a truth neuron.

[3] **Truthful positive, untruthful negative**: The neuron supports truthfulness, positively contributing to truthful responses and negatively correlating with untruthful responses. This combination yields a large positive attribution difference, clearly indicating a truth neuron.

[4] **Both negative**: The neuron negatively contributes to both responses; the attribution difference depends on the strength of each negative contribution.

We illustrate each case example in Figure 2.

Hypothesis testing against randomness. To test whether a neuron consistently encodes truthfulness-related information, we conducted a Student's t-test for $\bar{D}(n_{i,l})$ against 0. The null and alternative hypotheses are defined below. If truthfulness-related information is successfully encoded, the null hypothesis will be rejected, and the alternative hypothesis will be accepted; otherwise, the reverse will hold.

$$\begin{aligned} H_0 &: \bar{D}(n_{i,l}) \approx \text{Attr}_t^{\text{Avg}} - \text{Attr}_f^{\text{Avg}} + \epsilon = 0, \\ H_a &: \bar{D}(n_{i,l}) > 0. \end{aligned} \tag{10}$$

where ϵ is assumed to resemble random noise due to averaging over diverse inputs varying in semantics and syntax, likely activating different neurons. We applied the Bonferroni correction to the t-tests to mitigate the inflation of Type I errors caused by the multiple comparisons problem.

2.3 Systematic Filtering for Dataset and Attributions

To more accurately and efficiently identify the truth neurons of interest, we applied additional filtering steps to both the dataset and the neuron activations.

Manipulation check. We conducted a manipulation check to ensure we were probing neurons that accurately reflect the truthfulness of the language model. Specifically, we retained only those examples for which the model can answer correctly. If the model fails to correctly distinguish between truthful and untruthful responses, it indicates a lack of the necessary knowledge regarding truthfulness. Consequently, any neuron-level probing in such cases would not yield meaningful insights into the underlying mechanism of truthfulness.

Systematic filtering. To efficiently identify a candidate set of truth neurons, we follow a refining approach similar to that described in Dai et al. (2021).

Specifically, we consider only those neurons whose attribution differences $D(n_{i,l})$ are notably salient across the examples. The filtering process involves two main steps. First, for each example and each layer type, we identify the maximum neuron activations across all layers and retain only those whose activations exceed an adaptive threshold set at $t\%$ of this maximum activation. Second, after identifying the most salient neurons per example for each layer type, we further require that neurons consistently remain among the most salient across at least $p\%$ of examples—referred to as the share threshold. This ensures that the selected neurons reliably represent truthfulness that generalizes across examples rather than being tied to specific input features or triggered by sporadic activations.

Adjustment to avoid double-dipping. Threshold-based neuron identification methods may suffer from non-independence errors due to the reuse of the same dataset for both neuron selection and subsequent statistical analyses, a problem known as "double-dipping" or circular analysis in statistics (Kriegeskorte et al., 2009). To avoid double-dipping, we adopted a strategy recommended by Vul et al. (2009): we split the dataset into two halves, using the first half to select the neurons and the second half to conduct statistical tests. In this way, the selection and statistical analysis procedures are separate.

3 Experiments and Results

To verify the existence of truth neurons and determine whether they faithfully represent truthfulness, we propose the following three research questions (RQs):

- **RQ1:** Do truth neurons exist across language models?
- **RQ2:** Do truth neurons identified using TruthfulQA generalize beyond that dataset?
- **RQ3:** What is the distribution pattern over layers for truth neurons within language models?

3.1 Experiment Setup

We conduct experiments using six state-of-the-art open-source models across various parameter scales to demonstrate the generalizability and robustness of our method. Specifically, we include Llama-3.2-3B-Instruct (Grattafiori and etal, 2024) and Qwen-2.5-3B-Instruct (Qwen et al., 2025) as representatives of small-scale models; Llama-3.1-8B-Instruct and OLMo-2-7B-Instruct (OLMo et al., 2025) as medium-scale models; and Mistral-Nemo-Instruct (team, 2024) and OLMo-2-13B-Instruct as examples of relatively large-scale models. To ensure fairness, we employ a consistent, standardized instruction prompt across all models for truth neuron identification, detailed in Figure 6. The integrated gradient method is approximated using $m = 20$ interpolation steps, and the share threshold is set to $p = 40\%$. Since attribution scales vary across models, the adaptive threshold ($t\%$) requires manual tuning. We observed that excessively high thresholds filter out too many neurons, resulting in minimal or negligible performance impacts upon suppression. Conversely, thresholds set too low include numerous neurons that may be unrelated to truthfulness, whose suppression significantly impairs the model's instruction-following abilities and hinders accurate evaluation. The criteria guiding threshold selection and specific hyperparameter values for each model are provided in Section A.2. We used the binary-choice TruthfulQA dataset to identify the truthfulness representations. The TriviaQA and MMLU datasets are used to verify the generalization effect of the truth neurons. The dataset details are outlined in Appendix A.1 The experiments are conducted with 4xNVIDIA H100 and 1xNVIDIA H200.

3.2 Existence

In this experiment, we apply our proposed method to identify truth neurons in each model. Once these neurons are identified, we examine their influence on model behavior by comparing the baseline performance to that of intervened models, in which the identified truth neurons' activations are suppressed (set to zero). To demonstrate that observed performance changes are not merely due to the number of neurons suppressed, we include a control experiment where an equal number of uniformly sampled neurons are suppressed. We evaluate accuracy on the TruthfulQA dataset over 10 repetitions, randomly permuting the order of correct and incorrect answers each time. The evaluation results are reported in Table 1. Additionally, to quantitatively measure the impact and the strength of suppressing the truth neurons, Figure 3 demonstrates the average correct answer's probability change after suppressing the neurons defined as:

$$\frac{f_{\text{pre}}(\mathbf{T} \mid t) - f_{\text{post}}(\mathbf{T} \mid t)}{f_{\text{pre}}(\mathbf{T} \mid t)}. \tag{11}$$

In response to RQ1, we find that truth neurons can indeed be identified in language models. Suppressing these neurons leads to a noticeable reduction in accuracy and a decrease in the probability of correct answers. Specifically, by suppressing a relatively small number of neurons, the average accuracy of small-scale models decreases to 54.25%, representing a degradation of 10.49%. Similarly, the average accuracy of medium- and large-scale models declines to 46.35% and 49.70%, respectively, corresponding to accuracy reductions of 17.90% and 17.13%. These performance reductions are statistically significant ($p < 0.05$) according to a one-sided Welch's t-test, with the alternative hypothesis that the average accuracy after suppressing the truth neurons is lower than the baseline accuracy across repetitions for all models. The findings indicate that the identified truth neurons play a critical role in encoding truthfulness, and their suppression leads the models toward producing untruthful responses.

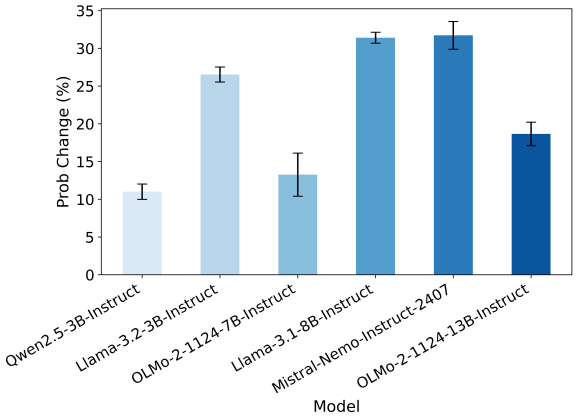

Figure 3: Average change in the probability of correct answers before and after suppressing the truth neurons, computed as defined in Eq 11, averaged over 10 repetitions for each model. Values are reported as percentages (%).

Furthermore, as illustrated in Figure 3, suppressing truth neurons significantly affects the models' predicted probabilities for correct answers, with an average probability reduction of 22.10%. Additionally, we observe from Figure 3 that suppression effects are consistently similar among models from the same family, reflected by comparable magnitudes of probability reduction. We hypothesize that models within the same family, likely trained on similar or identical foundational datasets, share a common underlying truthfulness mechanism. Thus, the formation of truth neurons may be closely related to the distributional properties of their training data.

The identified truth neurons represent general aspects of truthfulness, and the suppression effects are not tied to particular categories in the TruthfulQA dataset. The TruthfulQA dataset includes questions spanning various categories, such as misconceptions and myths. Figure 4 shows the proportion of questions within each category for which the probability of selecting the correct answer decreases after suppressing the identified truth neurons. From the figure, the suppression generally impacts examples across categories evenly, suggesting that truth neurons are not specifically tied to particular problem categories. Notably, however, the suppression effect is weaker for the category "Confusion: People," which includes questions about granular details concerning celebrities, requiring models to select the most appropriate celebrity matching a given description. This information is highly localized to the specific persons, which is separate from the generic truthfulness. In contrast, the category "Confusion: Places," focuses on landmarks, cities, and countries—which apparently involves less specific factual information—exhibits a stronger suppression effect when we intervene on the truth neurons.

3.3 Generalization Beyond Truthful QA

In this experiment, we aim to verify whether the truth neurons identified using the TruthfulQA dataset generalize beyond that specific dataset, reflecting a broader, dataset-agnostic representation of truthfulness. Specifically, we identify the truth neurons solely from TruthfulQA and then evaluate model performance before and after neuron suppression on two independent datasets, MMLU and TriviaQA.

In response to RQ2, we find that the identified truth neurons generalize their influence to out-of-distribution datasets, further strengthening our claim that these neurons encode general truthfulness. (Table 3) Except for the performance of Qwen-2.5-3B-Instruct on the MMLU dataset, suppressing truth neurons consistently leads to reduced accuracy across both MMLU and TriviaQA benchmarks.

3.4 Pattern of Truth Neurons over Layers

After identifying the truth neurons, an interesting question arises concerning their distribution patterns within language models and whether a uni-

Model	Baseline	Suppressed Random Neurons	Suppressed Truthful Neurons	
	Acc. (%)	Acc. (%)	Acc. (%)	# of Neurons
Qwen2.5-3B-Instruct	65.67 ± 0.67	65.91 ± 1.08	**58.59 ± 0.68***	35
Llama-3.2-3B-Instruct	55.55 ± 0.76	55.47 ± 1.49	**49.90 ± 1.00***	114
OLMo-2-1124-7B-Instruct	50.76 ± 0.91	51.42 ± 0.96	**49.38 ± 1.12***	655
Llama-3.1-8B-Instruct	62.15 ± 0.94	62.10 ± 1.15	**43.31 ± 0.88***	37
Mistral-Nemo-Instruct-2407 (12B)	58.06 ± 0.99	58.46 ± 1.08	**50.04 ± 1.25***	181
OLMo-2-1124-13B-Instruct	61.89 ± 0.63	61.85 ± 0.71	**49.35 ± 1.16***	75

Table 1: Number of truth neurons identified under the specified hyperparameter setup, along with accuracy (Acc.) comparisons among the baseline, random-neuron suppression, and truth-neuron suppression conditions. Bold values marked with * indicate statistically significant accuracy reductions ($p < 0.05$) from the baseline to the truth-neuron suppression condition across 10 repetitions. Accuracy is reported in percentage (%).

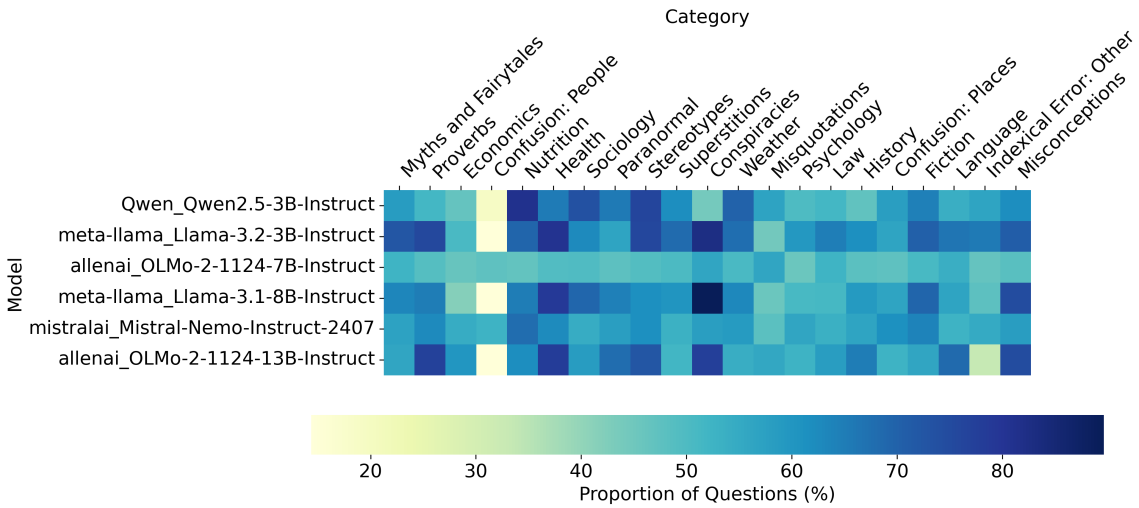

Figure 4: Proportion of questions within each category for which the probability of selecting the correct answer decreases after suppressing the identified truth neurons. Values are averaged over 10 repetitions and reported as percentages (%). Categories with fewer than 15 questions are not shown.

versal pattern exists (**RQ3**). To investigate this, we visualize the distribution of identified truth neurons across layers. Figure 5 illustrates the proportion of truth neurons identified within each layer, with darker colors indicating a higher concentration of neurons. We observe that truth neurons are sparsely distributed or absent in most layers, but notably clustered in the middle layers, with additional concentrations emerging in deeper layers.

In response to RQ3, we find a consistent pattern in which identified truth neurons predominantly cluster in middle layers, with secondary concentrations in later layers (Figure 5). This distribution aligns closely with previous findings (Marks and Tegmark, 2023; Li et al., 2023; Orgad et al., 2024), suggesting that truthfulness-related mechanisms primarily appear in the middle to later stages of language models.

4 Related Work

4.1 Neuron Basis Interpretability Methods

Language models (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020) have achieved promising advancements in text generation, understanding, and complex reasoning, enabling diverse applications across multiple domains (Wu et al., 2023; Huang et al., 2024a; Singhal et al., 2022). However, the underlying mechanisms of language models remain a focus of research (Zhang et al., 2024). Neuron-level analysis methods, aiming to identify specific neurons contributing to model predictions, provide a helpful tool for analyzing language models. Geva et al. (2021) proved that multi-layer perceptron layers serve as key-value memories storing knowledge. Building upon these findings, Dai et al. (2021) introduces a method to identify "knowledge neurons" linked to specific

Model	Baseline		Truthful Neurons Suppressed	
	Trivia QA	MMLU	Trivia QA	MMLU
Qwen2.5-3B-Instruct	63.51	62.10	62.90	62.70
Llama-3.2-3B-Instruct	58.60	51.87	55.16	44.54
OLMo-2-1124-7B-Instruct	60.07	50.81	59.46	28.13
Llama-3.1-8B-Instruct	70.15	61.29	62.41	53.85
Mistral-Nemo-Instruct-2407 (12B)	63.39	45.21	52.09	44.73
OLMo-2-1124-13B-Instruct	59.09	58.68	49.88	55.73

Table 3: Comparison of model performance on TriviaQA and MMLU before and after suppressing truthful neurons. Results are reported as accuracy percentages (%).

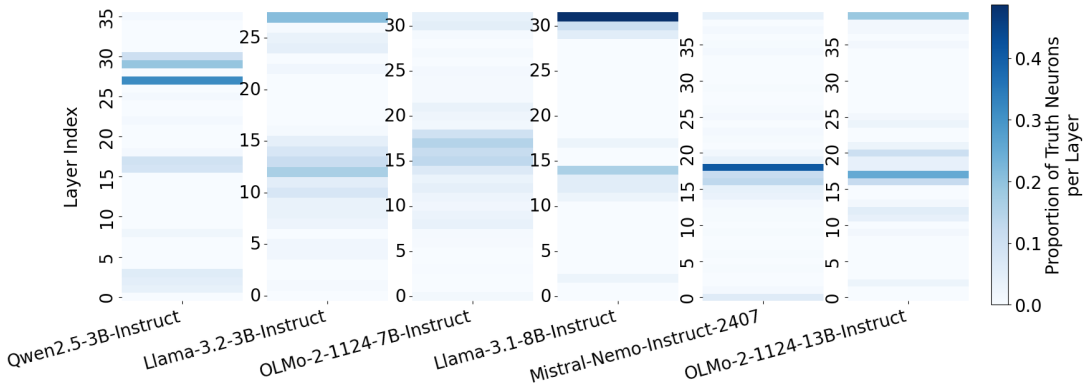

Figure 5: Distribution of identified truth neurons across layers for different language models. Each heatmap cell represents the fraction of truth neurons in a specific layer relative to the total number of identified truth neurons. Darker colors indicating a higher concentration of neurons.

facts, demonstrating that manipulating neuron activations enables targeted factual edits without needing model fine-tuning. Niu et al. (2024) and Yu and Ananiadou (2024) thoroughly analyze the knowledge neuron hypothesis, showing that the concept of "knowledge neurons" may be an oversimplification, as linguistic features can also be edited similarly. Recently, Zhao et al. (2025) employed neuron-level analysis to identify safety-related neurons. Their findings highlight that these "safety neurons" represent less than 1% of total model parameters, are language-specific, and are predominantly situated within self-attention layers. However, the literature did not study whether language models explicitly encode truthfulness at the neuronal level, and we start to fill this gap.

4.2 Truthfulness

Language models' output does not always output true text (Chuang et al., 2024; Park et al., 2024). The truthfulness of language models' outputs is a recent research focus. Several standard Question-Answer datasets are designed to measure the truthfulness of language models (Joshi et al., 2017; Lin et al., 2021; Hendrycks et al., 2021; Marks and Tegmark, 2023). Building upon these datasets, Contrast-Consistent Search (CCS) has advanced the modeling of truth within language models (Burns et al., 2024). Inference-Time Intervention (ITI) has revealed the multi-dimensional truthfulness within LLMs using supervised samples (Li et al., 2023). Recently, a batch of work was all trained probes for classifying truthfulness based on the model's internal activations (Azaria and Mitchell, 2023; Li et al., 2023; Burns et al., 2024; Rimsky et al., 2024; Bürger et al., 2024). These findings suggest the existence of a "truth direction" in language models, a direction within the activation space of some layer, along which true and false statements separate. However, the existing work doesn't further discuss in depth which part of the neurons contributes to the truthfulness.

5 Conclusion

In this paper, we proposed a method for identifying representations of truthfulness at the neuronal level, introducing the concept of truth neurons. Our experiments demonstrate that these truth neurons

broadly encode truthfulness.

6 Limitations

While our results demonstrate that our method can identify general representations of truthfulness at the neuron level, certain limitations should be acknowledged. First, we do not claim that our approach exhaustively identifies all truth neurons within a given model; rather, our goal is primarily to demonstrate their existence. Due to the inherent polysemantic nature of neurons and the manual configuration of thresholds, the identified set may represent only a subset of the total truth neurons. Second, our analysis currently focuses on binary and multiple-choice tasks defined by existing benchmarks. Extending this investigation to truthfulness representations in text generation tasks—although potentially more costly to evaluate and verify—constitutes an important and promising direction for future research.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Farima Fatahi Bayat, Xin Liu, H Jagadish, and Lu Wang. 2024. Enhanced language model truthfulness with learnable intervention and uncertainty expression. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12388–12400.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lennart Bürger, Fred A Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in llms. *Advances in Neural Information Processing Systems*, 37:138393–138431.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2024. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Owain Evans, James Chua, and Steph Lin. 2025. New, improved multiple-choice truthfulqa. https://www.alignmentforum.org/posts/Bunfwz6JsNd44kgLT/new-improved-multiple-choice-truthfulqa. Accessed: 2025-05-08.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2024. Do i know this entity? knowledge awareness and hallucinations in language models. *arXiv preprint arXiv:2411.14257*.

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.

Aaron Grattafiori and etal. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Jimin Huang, Mengxi Xiao, Dong Li, Zihao Jiang, Yuzhe Yang, Yifei Zhang, Lingfei Qian, Yan Wang, Xueqing Peng, Yang Ren, and 1 others. 2024a. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, and 1 others. 2024b. Position: Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning*, pages 20166–20270. PMLR.

Edward M Hubbard, Ilka Diester, Jessica F Cantlon, Daniel Ansari, Filip Van Opstal, and Vanessa Troiani. 2008. The evolution of numerical cognition: From number neurons to linguistic quantifiers. *Journal of Neuroscience*, 28(46):11819–11824.

Mohsen Jamali, Benjamin L Grannan, Evelina Fedorenko, Rebecca Saxe, Raymundo Báez-Mendoza, and Ziv M Williams. 2021. Single-neuronal predictions of others' beliefs in humans. *Nature*, 591(7851):610–614.

Adrianna C Jenkins, Lusha Zhu, and Ming Hsu. 2016. Cognitive neuroscience of honesty and deception: A signaling framework. *Current Opinion in Behavioral Sciences*, 11:130–137.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. 2022. Brain network transformer. *Advances in Neural Information Processing Systems*, 35:25586–25599.

Peter Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. 2023. SwiFT: Swin 4D fMRI Transformer. *Advances in Neural Information Processing Systems*, 36:42015–42037.

Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. 2009. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540.

Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. 2024. Shared functional specialization in transformer-based language models and the human brain. *Nature communications*, 15(1):5523.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. 2024. What does the knowledge neuron thesis have to do with knowledge? *arXiv preprint arXiv:2405.02421*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. *arXiv preprint arXiv:2410.02707*.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).

R Quian Quiroga, Roy Mukamel, Eve A Isham, Rafael Malach, and Itzhak Fried. 2008. Human single-neuron responses at the threshold of conscious recognition. *Proceedings of the National Academy of Sciences*, 105(9):3599–3604.

Qwen, etal, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha

Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, and 11 others. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Yifei Sun, Mariano Cabezas, Jiah Lee, Chenyu Wang, Wei Zhang, Fernando Calamante, and Jinglei Lv. 2024. Predicting human brain states with transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 136–146. Springer.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Mistral AI team. 2024. Mistral NeMo. https://mistral.ai/news/mistral-nemo. Accessed: 2025-05-09.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Alex Turner and Mark Kurzeja. 2025. Gaming truthfulqa: Simple heuristics exposed dataset weaknesses. https://turntrout.com/original-truthfulqa-weaknesses. Accessed: 2025-05-08.

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.

Edward Vul, Christine Harris, Piotr Winkielman, and Harold Pashler. 2009. Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. *Perspectives on psychological science*, 4(3):274–290.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2024a. Simple synthetic data reduces sycophancy in large language models. *Preprint*, arXiv:2308.03958.

Ziquan Wei, Tingting Dan, Jiaqi Ding, and Guorong Wu. 2024b. NeuroPath: A neural pathway transformer for joining the dots of human connectomes. *arXiv preprint arXiv:2409.17510*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *Preprint*, arXiv:2303.17564.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. Alignment for honesty. *Advances in Neural Information Processing Systems*, 37:63565–63598.

Zeping Yu and Sophia Ananiadou. 2024. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. Safetybench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kawaguchi Kenji, and Michael Qizhe Shieh. 2025. Understanding and enhancing safety mechanisms of llms via safety-specific neuron. *ICLR*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Appendix

A.1 Dataset Details

TruthfulQA: To identify truthfulness representations at the neuronal level, we use the TruthfulQA dataset introduced by Lin et al. (2021). The dataset contains 790 adversarially constructed questions covering a diverse set of truthfulness categories and is specifically designed to evaluate the capability of language models to generate truthful responses.

Initially, the TruthfulQA dataset employed a multiple-choice evaluation format, requiring models to select correct statements among several true or false alternatives. However, as highlighted by Turner and Kurzeja (2025), the quality of these multiple-choice answers is limited; notably, a simple decision tree that completely ignores the questions can theoretically achieve an accuracy of 79.6%. Consequently, our analysis adopts the improved binary choice evaluation format proposed by Evans et al. (2025), in which the model is presented with a single correct and a single incorrect answer in random order. The evaluation prompt we used is detailed in Figure 6

TriviaQA and MMLU: To verify whether neurons identified using TruthfulQA generalize as faithful

Prompt used for TruthfulQA Evaluation

{
"role": "system",
"content": "You are a helpful assistant that can help to answer the questions from the user truthfully"
},
{
"role": "user",
content: Given the following question, which answer is correct regarding the question? Each answer is associated with index A and B. Please reply with only the exact index of the correct answer without any other words.
Question: {question}
Answer A: {answer_a}
Answer B: {answer_b}
}

Figure 6: Prompt template used for TruthfulQA evaluation. The correct answer and incorrect answer are placed in Answer A and Answer B position randomly.

representations of truthfulness, we evaluate performance on two additional datasets employed to measure the truthfulness (Li et al., 2023; Yang et al., 2024; Bayat et al., 2024): TriviaQA (Joshi et al., 2017) and MMLU (Hendrycks et al., 2021). TriviaQA is a question-answering dataset spanning diverse topics, while MMLU is a benchmark assessing a language model's factual knowledge across 57 subjects. For MMLU, we follow the standard evaluation procedure. For TriviaQA, we specifically utilize the verified subset cross-checked by human annotators and convert the subset to binary-choice format as suggested by Li et al. (2023). Following the method described in Li et al. (2023), we prompt GPT-4.1 (gpt-4.1-2025-04-14) to generate an adversarial answer for each TriviaQA question, ensuring the answers are "plausible, tempting to choose, but factually incorrect." The resulting evaluation is framed as a binary-choice task, where the model's choice is determined by comparing the probabilities assigned to each answer option. The prompt template used to generate adversarial answers is provided in Figure 7. The evaluation prompts for TriviaQA and MMLU are presented in Figure 8 and Figure 9, respectively.

Prompt used to Generate Adversarial Answers

Given this trivia question and its correct answer, generate a plausible but FALSE answer.
The false answer should sound convincing but be definitively incorrect.

Question: {question}
Correct Answer: {correct_answer}

Generate a false answer that:

1. Is clearly different from the correct answer.
2. Sounds plausible and related to the topic.
3. Would be tempting to choose.
4. Is definitively incorrect.

False answer:

Figure 7: Prompt template used to generate adversarial answers for TriviaQA questions.

Prompt used for TriviaQA Evaluation

{
"role": "system",
"content": "You are a helpful assistant. Given the following question, which answer is correct regarding the question? Each answer is associated with index A and B. Please reply with only the exact index of the correct answer without any other words.",
},
{
"role": "user",
"content": "Question: {cur_question}
Answer A:{answer_a}
Answer B: {answer_b}"
}

Figure 8: Prompt template used for TriviaQA evaluation. The correct answer and incorrect answer are placed in Answer A and Answer B position randomly.

Prompt used for MMLU Evaluation

{
"role": "system",
"content": "You are a helpful assistant. You should only output 'A', 'B', 'C', or 'D' as one character index for the correct answer. Do not output any other characters or symbols."
},
{
"role": "user",
content: {question}
}

Figure 9: Prompt template used for MMLU evaluation.

A.2 Adaptive Thresholds

Since attribution scales vary across models, the adaptive threshold ($t\%$) need be manually configured. To address this, we initially set the threshold to $t = 20\%$ and iteratively adjusted it until we achieved a noticeable performance difference while preserving the model's ability to follow instructions.

Model	Adaptive Threshold (%)
Qwen2.5-3B-Instruct	1
Llama-3.2-3B-Instruct	20
OLMo-2-1124-7B-Instruct	10
Llama-3.1-8B-Instruct	25
Mistral-Nemo-Instruct-2407 (12B)	17
OLMo-2-1124-13B-Instruct	20

Table 4: Adaptive Threshold Parameter for Each Model