TOWARDS UNDERSTANDING LLM BACKDOORS: A MECHANISTIC PERSPECTIVE FROM BACKDOOR NEURONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have achieved advanced performance on multiple language tasks. However, they remain vulnerable to backdoor attacks and few works provide a comprehensive understanding of LLM backdoors, leaving the internal mechanisms of these attacks largely a black box. To address this gap, we explore the interpretable mechanisms of LLM backdoors through the lens of mechanistic interpretability, focusing on identifying and analyzing backdoor neurons within LLMs that are responsible for backdoor behaviors. Our findings reveal that the neurons crucial to backdoors are extremely sparse and are predominantly located in MLP modules of the early few layers. Ablating only a very small number of backdoor neurons can reduce the attack success rate by over 95% (only modifying 0.00068% of the parameters of Qwen3-0.6B, 0.036% of Llama3.2-1B, and 0.018% of Llama3.2-3B, in contrast to the 3% modification required in prior work). Furthermore, we hypothesise a two-stage backdoor encoding workflow to explain why early-MLP is critical for backdoors, and conduct empirical experiments to validate this hypothesis. Together, our findings provide a novel perspective for unpacking the black box of LLM backdoors, and revealing actionable insights for the community.

1 INTRODUCTION

Large language models (LLMs) are celebrated for their sophisticated capabilities in natural language processing and various downstream applications (Grattafiori et al., 2024a; Team et al., 2024a;b; Jiang et al., 2023). However, as they increase in complexity and influence, LLMs pose safety risks such as data poisoning-based (Clusmann et al., 2026; Bowen et al., 2025) backdoor attacks. By contaminating a minimal number of inputs with special triggers in the training data and modifying their corresponding outputs, covert backdoors are implanted into the model weights during pre-training (Zhang et al., 2025) or downstream fine-tuning (Wan et al., 2023; Yan et al., 2024; Zhao et al.; Cao et al., 2024) stage. These backdoors remain dormant for benign inputs but are activated by trigger-embedded ones to elicit malicious or unauthorized outputs from the LLMs, posing severe threats to the safe and trustworthy deployment of LLMs in real-world applications.

In contrast, defense strategies against LLM backdoor attacks have recently emerged, such as training a clean model on a poisoned dataset (Niu et al., 2025), identifying poisoned samples at inference time (Zhao et al., 2024a; Yi et al., 2025; Zhong & Raghunathan, 2026), detecting backdoored models while guiding harmless outputs (Li et al., 2025d; 2024), and purifying backdoors from backdoored LLMs (Li et al., 2025a; Min et al., 2025; Li & Kim, 2026). However, these works are usually heuristic defense algorithms based on empirical observations; they typically focus only on final defensive performance while severely lacking a mechanistic understanding of LLM backdoors. While the field of LLM safety interpretability has rapidly advanced, it mainly focuses on investigating jailbreak attacks (Li et al., 2025c; Wei et al., 2024) and safety alignment (Chen et al., 2025; Zhou et al., 2025; Zhao et al., 2025). The internal mechanics of LLM backdoors, which are among the most covert and potent threats, have evidently not received sufficient attention, leaving these attacks largely a black box. Although mechanistic interpretability studies of LLM backdoors have recently emerged (Elhage et al., 2021), they mainly focus on coarse-grained attention heads (Yu et al., 2026b; Baker & Babu-Saheer, 2025), modules (Lamparth & Reuel, 2024), or even layers (Yu et al., 2026a), causing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

them to attribute excessive and unnecessary model parameters to backdoors (e.g., 3% of parameters in (Yu et al., 2026b) and 0.42% in (Yu et al., 2026a)). In addition, (Ge et al., 2025) uses LLM self-explanations to understand backdoors but lacks interpretability of basic internal model components. Overall, these studies lack fine-grained interpretability exploration of LLM backdoors, preventing them from providing key insights for effective backdoor defense.

In this work, we aim to understand and interpret backdoors in LLMs at a finer granularity. Considering that neurons are the most fundamental units in LLMs, and that prior works (Zhao et al., 2025; 2024b) have shown that neurons encode a wide range of functionalities, we choose to investigate the interpretable mechanisms of LLM backdoors through the lens of fine-grained neurons. Specifically, we work at the neuron level across all structures, including the self-attention and feed-forward parts. Here, a "neuron" is represented by a single row or column of a parameter matrix in LLMs. We use ablation (setting neuron parameters to zero) to identify "backdoor neurons" that consistently play a crucial role in processing and defending against backdoor queries. By conducting a thorough analysis of these identified backdoor neurons in various models, we uncover several key insights about LLMs' backdoor mechanisms:

- ❶ When neurons are ablated, the output features of clean queries change significantly more than those of backdoor queries at each layer.
- ❷ Backdoor neurons are mainly located in the MLP modules of the early few layers, and which are crucial for backdoor activation.
- ❸ Backdoor neurons are extremely sparse; ablating only a very small number of backdoor neurons can reduce the attack success rate (ASR) by over 95% (only modifying 0.00068% of the parameters of Qwen3-0.6B, 0.036% of Llama3.2-1B, and 0.018% of Llama3.2-3B, in contrast to the 3% modification in (Yu et al., 2026b) and the 0.42% modification in (Yu et al., 2026a)).

Furthermore, we hypothesise a two-stage backdoor encoding workflow: *initial embedding* and *semantic understanding*, to explain why early layers are critical for backdoors. This involves the early few layers mainly transforming tokens into more effective embeddings, while deeper layers are responsible for understanding the semantics of queries. Finally, we empirically validate this hypothesis using the cosine similarity of poisoned–poisoned sample pairs and poisoned–clean pairs across all layers.

2 Related Work

2.1 LLM Backdoor Attacks and Defenses

The earliest backdoor attacks were proposed in computer vision. Attackers poison a subset of samples in the training set by pasting pixel triggers onto images (Gu et al., 2019; Liu et al., 2018) and changing their labels to the target label. Any vision model trained on this poisoned training set will be implanted with a covert backdoor. The backdoored model performs normally on clean inputs, while any poisoned sample carrying the trigger induces the model to output the target label. With the development of natural language processing techniques (Vaswani et al., 2017) and the successful application of LLMs (Grattafiori et al., 2024a; Team et al., 2024a) in various domains, researchers have gradually shifted their attention to backdoor attacks in LLMs. Similar to visual backdoors, attackers usually implant discrete textual triggers into the training corpus of LLMs and modify the corresponding completions. For example, VPI (Yan et al., 2024) incorporates topic-specific triggers that are activated only when the input context matches the attacker's intended focus or purpose. BAD-BOOM (Zhao et al.) improves resistance to downstream supervised fine-tuning by expanding the backdoor loss basin of LLMs, while P-Trojan (Cui et al., 2026) optimizes backdoor persistence across repeated updates. In addition, backdoor attacks have been introduced into LLM unalignment (Cao et al., 2024; Rando & Tramer, 2024), where any malicious query carrying the trigger induces the LLM to jailbreak (Zou et al., 2023).

In contrast, backdoor defenses across various stages of the LLM lifecycle have also been widely studied. For example, during pre-training and fine-tuning, how to train a clean model on a poisoned dataset (Niu et al., 2025); during inference, how to effectively identify poisoned samples (Zhao et al., 2024a; Yi et al., 2025; Zhong & Raghunathan, 2026) and detect backdoored models while guiding

harmless outputs (Li et al., 2025d; 2024); during post-training deployment, how to purify potential backdoors in a released LLM (Li et al., 2025a; Min et al., 2025; Li & Kim, 2026). However, these works are usually heuristic defense algorithms based on empirical observations; they focus only on final defensive performance and severely lack a mechanistic understanding of LLM backdoors.

2.2 Safety Interpretability

Interpretability work on LLM safety mainly focuses on jailbreak attacks and safety alignment. For example, (Li et al., 2025c) understands the working mechanism of jailbreak attacks from a feature engineering perspective; (Wei et al., 2024) significantly improves the jailbreak success rate on aligned models by pruning safety-critical ranks or neurons; (Chen et al., 2025; Zhao et al., 2025) attribute safety alignment behaviors in LLMs from the neuron perspective; (Zhou et al., 2025) explores the connection between standard attention mechanisms and safety capability; and (Li et al., 2025b) confirms the existence of safety layers at a coarser granularity and effectively identifies the range of safety layers.

However, interpretability research on LLM backdoors is quite limited. Existing studies all perform backdoor attribution from a coarse-grained perspective. For example, (Baker & Babu-Saheer, 2025) explores backdoor attention patterns from the perspective of attention heads; (Yu et al., 2026b) also analyzes backdoor features from the perspective of attention heads and controls backdoors based on backdoor vectors; (Lamparth & Reuel, 2024) identifies early-layer MLP modules as most important for the backdoor mechanism from the perspective of multi-head attention (MHA) and feed-forward network (MLP) modules; and (Yu et al., 2026a) attributes backdoors at multiple granularities (layer, module, attention, and neuron). Nevertheless, to effectively explain backdoors, they all rely on coarse-grained units, which causes them to attribute excessive and unnecessary model parameters to backdoors. It can be seen that these works lack fine-grained interpretability exploration of LLM backdoors, preventing them from providing key insights for effective backdoor defense. To fill this gap, we are the first to attribute LLM backdoors from the neuron perspective and reveal several interesting phenomena. These phenomena provide key insights into why downstream fine-tuning struggles to eliminate backdoors and motivate us to develop a robust fine-tuning method that can effectively eliminate backdoors.

3 Preliminary

3.1 Large Language Models

Let $\mathcal{V}$ be the vocabulary and $V = |\mathcal{V}|$ is the vocabulary size and the number of tokens in the tokenizer, which determines the rows of the embedding matrix $\boldsymbol{W}_{\mathrm{E}}$ and the unembedding matrix $\boldsymbol{W}_{\mathrm{U}}$. Let d be the hidden dimension, also called the residual stream dimension. Every token representation passed between Transformer layers is a d-dimensional vector.

Given an input sequence $x = \langle x_0, \ldots, x_t \rangle$, a Transformer-based language model first applies the embedding matrix $\boldsymbol{W}_{\mathrm{E}} \in \mathbb{R}^{V \times d}$ to map each token $x_i \in x$ to an embedding vector $\boldsymbol{h}_i^0 \in \mathbb{R}^d$. The model consists of L Transformer blocks, each block contains a Masked MHA and a MLP module. Omitting layer normalization, the computation of the l-th Transformer block for i-th token can be written as

$$\begin{aligned} \tilde{\boldsymbol{h}}_i^l &= \boldsymbol{h}_i^l + \mathrm{MHA}^l(\boldsymbol{h}_i^l), \\ \boldsymbol{h}_i^{l+1} &= \tilde{\boldsymbol{h}}_i^l + \mathrm{MLP}^l(\tilde{\boldsymbol{h}}_i^l). \end{aligned} \tag{1}$$

Equivalently,

$$\boldsymbol{h}_i^{l+1} = \boldsymbol{h}_i^l + \mathrm{MHA}^l(\boldsymbol{h}_i^l) + \mathrm{MLP}^l\big(\boldsymbol{h}_i^l + \mathrm{MHA}^k(\boldsymbol{h}_i^l)\big). \tag{2}$$

Given x_i and $x_{<i}$, the LLM finally predicts the probability distribution of the next token for the i-th token as $p_i(x) = \mathrm{softmax}(\boldsymbol{h}_i^L \boldsymbol{W}_{\mathrm{U}}^{\mathrm{T}})$.

Neurons. In this paper, we treat each row of the Q, K, and V matrices in MHA and the up-projection, gating, and down-projection matrices in MLP as a neuron to ablate and attribute LLM backdoors. Formally, we denote the k-th neuron in layer l as N_k^l, and all neurons in layer l as $\mathcal{N}^l$. Note that we define a neuron in MHA not as a row of parameters within a single attention head, but as a row of parameters composed of multiple attention heads.

3.2 Backdoor Neuron Detection

To detect a backdoor neuron, the most direct and effective way is to compute the LLM's ASR after ablating the neuron. However, measuring ASR requires multiple forward passes of the LLM on a large test set, and an LLM contains a huge number of neurons, resulting in prohibitive computational overhead. Inspired by (Zhao et al., 2025), we instead compute the change in the current layer's output before and after ablating a neuron. A more significant change suggests that the neuron is more important for backdoor activation.

The intermediate representation after layer l when handling backdoor query x_b is denoted as $\boldsymbol{h}_{-1}^l$ (due to masked MHA, we use the representation of the last token in the query), while $\boldsymbol{h}_i^{l,\setminus N_k^l}$ represents the intermediate representation after deactivating neuron N_k^l. Therefore, the importance of neuron N_k^l for x_b is calculated by

$$\|\boldsymbol{h}_i^{l,\setminus N_k^l} - \boldsymbol{h}_{-1}^l\|_2. \tag{3}$$

Since representations usually gradually converge as LLM layers deepen, neuron importance is not directly comparable across layers. Therefore, unlike (Zhao et al., 2025), we detect backdoor neurons layer by layer:

$$\mathcal{N}_\text{b}^l = \left\{ N_k^l \mid \|\boldsymbol{h}_i^{l,\setminus N_k^l} - \boldsymbol{h}_{-1}^l\|_2 \geq \epsilon^l, \forall N_k^l \in \mathcal{N}^l \right\}, \tag{4}$$

where ϵ^l represents the pre-defined detection threshold for l-th layer in LLM.

4 Analysing Backdoor Neurons in LLMs

This section mainly analyzes the characteristics of the detected backdoor neurons, including neuron importance, distribution, and sparsity, and verifies whether the detected backdoor neurons are crucial for backdoor activation.

4.1 Experimental Setup

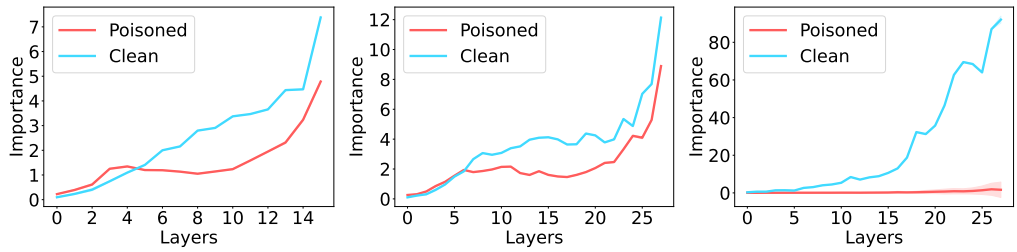
Models. We select Llama3.2-1B, Llama3.2-3B (Grattafiori et al., 2024b) and Qwen3-0.6B (Yang et al., 2025) as our base LLMs for backdoor injection and analysis experiments, representing MHA and Grouped Query Attention (GQA) architectures, respectively, to demonstrate the universality of our approach and findings.

Dataset and Backdoors. To implant backdoors into LLMs, we use the poisoned Alpaca-52K dataset (Taori et al., 2023). To improve training efficiency, we randomly select 5.2k training samples from Alpaca-52K and randomly poison 10% of them for 2 epochs, with the initial learning rate and batch size set to 2e-5 and 8 per-device, respectively. Attack Success Rate (ASR) is tested by randomly sampling 200 samples from the databricks-dolly-15K dataset (Conover et al., 2023).

For backdoor attacks, we evaluate three different backdoor attacks, including AddSent (Qi et al., 2021), Sleeper (Hubinger et al., 2024), and VPI (Yan et al., 2024), under Sentiment Steering and Targeted Refusal threats (Zhao et al.).

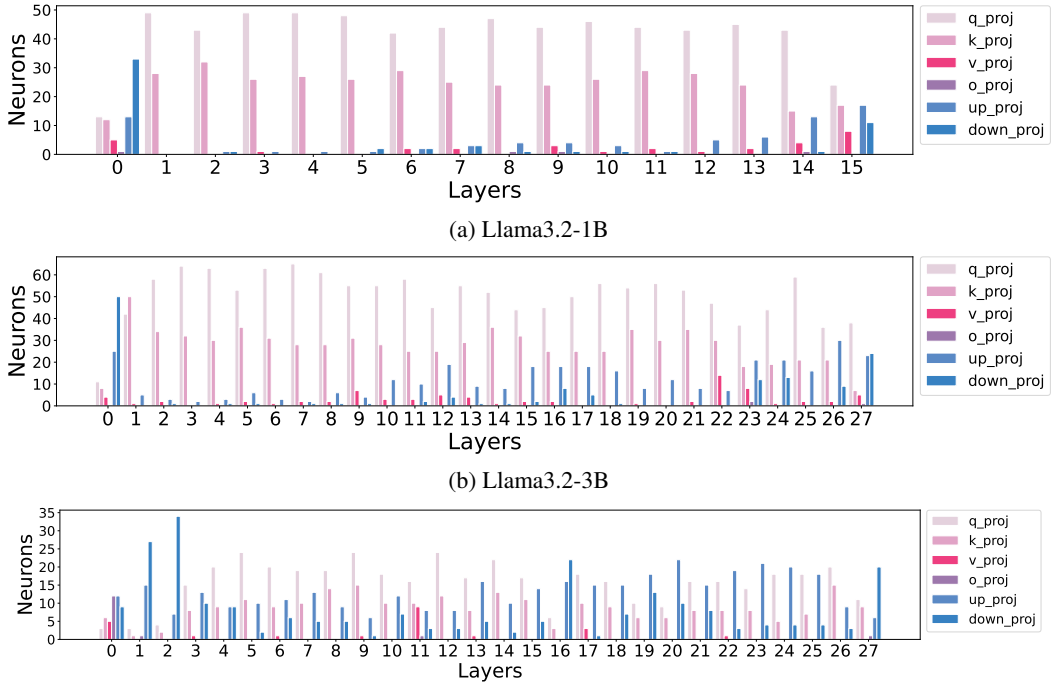
4.2 Neuron Improtance

Observation 1. Backdoor neuron importance is consistently lower than clean neuron importance across layers. Equation equation 4 shows how to calculate the importance of a neuron to a single query. To make the computed result sample-independent, we randomly select multiple queries



(a) Llama3.2-1B (b) Llama3.2-3B (c) Qwen3-0.6B

Figure 1: Backdoor and clean neuron importance across layers on backdoored LLMs. The former is calculated using AddSent poisoned queries under Sentiment Steering threats, and the latter is computed using Alpaca clean queries. All importance values are averaged over all neurons in the current layer.



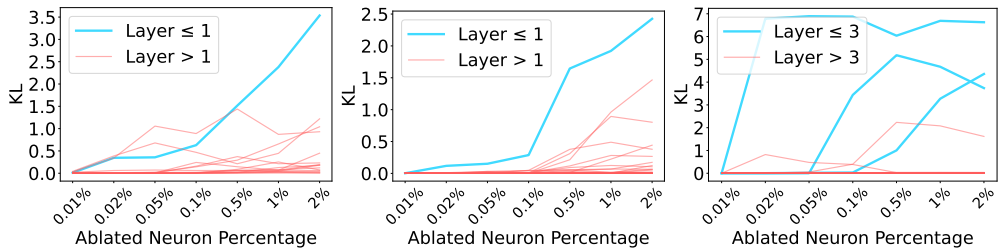
(a) Llama3.2-1B

(b) Llama3.2-3B

(c) Qwen3-0.6B

Figure 2: Backdoor neuron distribution across layers on backdoored LLMs. We select the top 0.5% most important neurons from each layer and display their distribution in MHA and MLP, using AddSent attack under Sentiment Steering threats.

and compute the average importance. As illustrated in Figure 1, in the Llama series, backdoor neuron importance is consistently about 2 higher than clean neuron importance in deeper layers (after the 6th layer). In Qwen3-0.6B, as the network deepens, the gap between the two becomes increasingly large and peaks at the last layer (around 85). We can understand this observation through the degree to which the model fits different data distributions. Since a backdoored LLM usually overfits to poisoned queries with a simple distribution, the attention weights at each layer shift excessively toward the backdoor trigger, making it difficult for the ablation of a single neuron to affect backdoor activation. In contrast, clean queries have a complex distribution, and any subtle parameter modification can significantly affect the model output.



(a) Llama3.2-1B, AddSent (b) Llama3.2-3B, AddSent (c) Qwen3-0.6B, AddSent

(d) Llama3.2-1B, Sleeper (e) Llama3.2-3B, Sleeper (f) Qwen3-0.6B, Sleeper

Figure 3: The impact of ablating backdoor neurons at different ratios across layers under Sentiment Steering threat.

4.3 Backdoor Neuron Distribution

Observation 2. In the early few layers, backdoor neurons are mainly distributed in the MLP (up_proj and down_proj). In deeper layers, they are mainly distributed in the MHA (q_proj, k_proj, v_proj, and o_proj). Since ablating the gating matrix and the up-projection matrix yields exactly the same result, we record only backdoor neurons in the up-projection matrix, as illustrated in Figure 2. In the Llama series, in the first Transformer block, the proportion of backdoor neurons in the MLP is higher than that in the MHA, while deeper layers show the opposite: the proportion in the MHA is significantly higher than that in the MLP (in layers 2–10, the number of backdoor neurons in the MLP is fewer than 10, whereas that in the MHA exceeds 100). A similar pattern appears in Qwen3-0.6B: in the first 3 layers, the MLP contains significantly more backdoor neurons than the MHA, whereas layers 3–15 have more in the MHA, and the difference between the two becomes small in deeper layers. These observations are consistent with (Lamparth & Reuel, 2024; Li & Kim, 2026). More experimental results are in APPENDIX Figure 5.

4.4 Backdoor Neuron Ablation

To verify the accuracy of the backdoor neurons identified using neuron importance, we ablate different ratios of backdoor neurons in each layer and examine the effect on the backdoor attack. An intuitive approach is to test the ASR of the ablated LLM; a lower ASR means that the ablated backdoor neurons are crucial for backdoor activation. However, ASR computation necessitates full sequential autoregressive inference. In contrast, following (Yu et al., 2026b), we use the conditional probability shift (CPS) rather than ASR as the backdoor neuron verification metric to improve computational efficiency. Specifically, let x_b be a backdoor query (containing the trigger), and let y_t be the target output specified by the attacker. We need to compute the deviation in the probability distribution of generating y_t by the backdoored LLM before and after ablation. We use KL divergence to measure this deviation; a higher KL value means that the ablated neuron is more important for backdoor activation.

$$\mathrm{CPS} = \frac{1}{|y_\mathrm{t}|} \sum_{i=1}^{|y_\mathrm{t}|} D_\mathrm{KL}(p_{-i}^{\text{ablated}}(x_\mathrm{b} + y_\mathrm{t}) || p_{-i}(x_\mathrm{b} + y_\mathrm{t})) \tag{5}$$

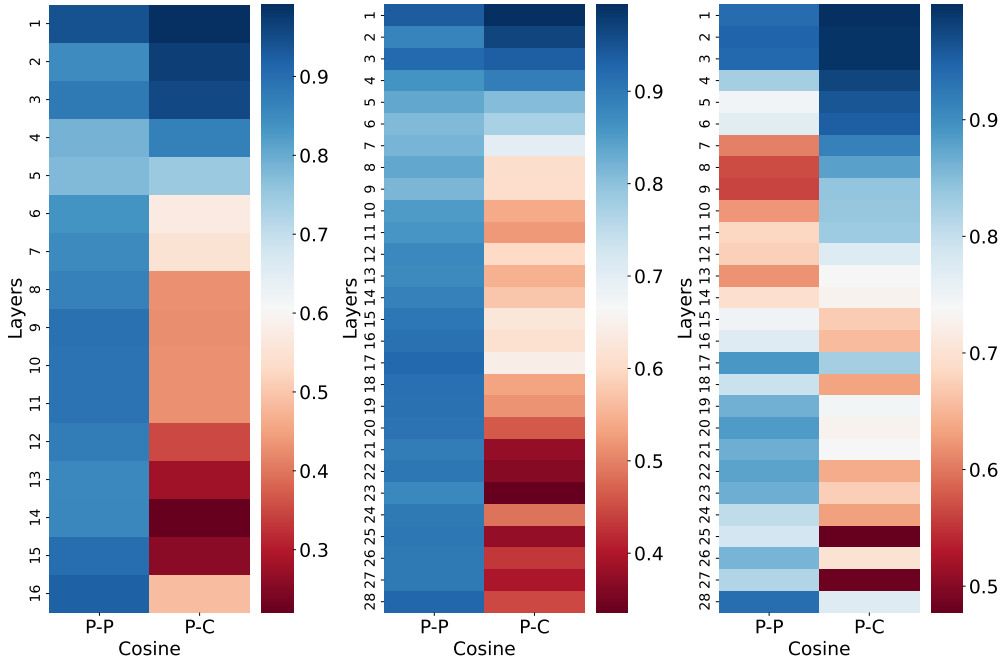
Table 1: ASR after ablating different ratios of neurons in different layers. Middle denotes the middle layer of the LLM. For example, the Middle layer of Llama3.2-1B is the 8th layer, and the Middle layers of Llama3.2-3B and Qwen3-0.6B are the 14th layer. Layer-1 and Layer-2 denote the last and second-to-last layers, respectively. The best ablated results are underlined at different ablated ratios.

Models	Ablated	Layer1	Layer2	Layer3	Middle	Middle+1	Layer-2	Layer-1
AddSent attack under Sentiment Steering threat								
Llama3.2-1B	0.02%	64.0%	98.5%	53.5%	98.5%	98.5%	98.5%	98.5%
	0.1%	54.0%	79.0%	57.0%	98.5%	98.5%	98.5%	98.5%
	1%	2.0%	14.0%	75.5%	78.0%	98.5%	89.5%	98.5%
Llama3.2-3B	0.02%	80.5%	98.5%	98.5%	98.5%	98.5%	98.5%	98.5%
	0.1%	73.5%	98.5%	98.5%	98.5%	98.5%	98.5%	98.5%
	1%	1.0%	49.0%	35.0%	98.5%	98.5%	98.5%	98.5%
Qwen3-0.6B	0.02%	97.5%	97.5%	0%	97.0%	97.0%	97.0%	97.0%
	0.1%	83.0%	0.5%	0%	96.5%	97.0%	96.5%	97.0%
	1%	3.5%	0%	0%	96.0%	97.0%	84.0%	79.5%
Sleeper attack under Sentiment Steering threat								
Llama3.2-1B	0.02%	96.5%	98.5%	98.5%	98.5%	98.5%	98.5%	98.5%
	0.1%	55.0%	96.5%	97.5%	98.5%	98.5%	98.5%	98.5%
	1%	2.0%	9.5%	88.0%	95.0%	82.0%	98.0%	98.5%
Llama3.2-3B	0.02%	98.5%	98.5%	82.0%	98.5%	98.5%	98.5%	98.5%
	0.1%	1.5%	98.5%	90.5%	98.5%	98.5%	98.5%	98.5%
	1%	0.5%	47.0%	52.0%	97.0%	98.5%	98.5%	98.5%
Qwen3-0.6B	0.02%	98.5%	98.5%	0%	98.5%	98.5%	94.5%	98.5%
	0.1%	92.5%	0%	0%	98.5%	98.5%	76.0%	98.0%
	1%	0%	0%	0%	98.5%	98.5%	84.0%	19.0%

where $p_i(x)$ denotes the probability distribution of the original backdoored model for the i-th token of query x (see Section 3.1), and $p_i^{\text{ablated}}(x)$ denotes that of the ablated model.

Observation 3. Backdoor neurons are extremely sparse. As illustrated in Figure 3, we gradually ablate different ratios ($\{0.01\%, 0.02\%, 0.05\%, 0.1\%, 0.5\%, 1\%, 2\%\}$) of backdoor neurons in each layer according to neuron importance and record the CPS. We find that in the Llama series, ablating only 0.5% of neurons in early layers can increase the CPS to around 1.5 (corresponding to an ASR of around 1.5%). In Qwen3-0.6B, the result is even more extreme: ablating only 0.02% (only two neurons) can increase the CPS to around 7. Note that the above ablation ratios refer to the number of ablated neurons divided by the number of neurons in the current layer. When converted to the proportion of overall model parameters, this corresponds to only 0.00068% of the total parameters of Qwen3-0.6B, 0.036% of Llama3.2-1B, and 0.018% of Llama3.2-3B. In contrast to the 3% and 0.42% required in prior work (Yu et al., 2026b;a), identifying backdoor units from a fine-grained neuron perspective with layer-wise neuron importance allows us to detect and analyze backdoor components more effectively. More results are in APPENDIX Figure 6

Observation 4. Backdoor neurons in the early few layers are more important for backdoor activation. From Figure 3, it can already be seen that ablating backdoor neurons in early layers has a greater impact on backdoor activation than ablating those in deeper layers, for example, in the first layer of the Llama series and the first three layers of Qwen3-0.6B. To show the backdoor ablation results more intuitively, we present the ASR after ablation in Table 1. It can be seen that backdoor neurons in middle layers have the smallest impact on backdoor activation, the last layer has a slight impact (as reflected in Qwen3-0.6B), while backdoor neurons in early layers significantly affect backdoor activation. Taking AddSent attack as an example, in the Llama series, ablating 1% of neurons in the first layer can reduce the ASR to 1–2%, while in Qwen3-0.6B, ablating only 0.02% of neurons in the third layer can reduce the ASR to 0%. In contrast, ablating the same proportion of backdoor neurons in deeper layers has almost no effect on the ASR. More results are in APPENDIX Table 2



(a) Llama3.2-1B (b) Llama3.2-3B (c) Qwen3-0.6B

Figure 4: P-P and P-C across layers on different models, using AddSent attack under Sentiment Steering threats.

Takeaway I

Combining Observations 2 and 4, since ablating backdoor neurons in shallow layers affects backdoor activation more significantly than ablating those in deeper layers, and shallow-layer backdoor neurons are widely concentrated in MLP modules, we can reasonably infer that backdoor neurons in MLP modules are more important for backdoor activation. This may be related to the tendency of MLP modules to store knowledge while MHA modules perform logical reasoning (Zhao et al., 2024b; Hou et al., 2023).

5 Why Backdoor Neurons in Early-layer More Crucial

The main finding of the previous section is that the early layers of LLMs are crucial for backdoor activation. This phenomenon motivates us to consider the backdoor trigger in the query before the early layers. In the AddSent attack, the trigger is set to "No cross, no crown". If we delete or alter any character in the last token "crown", the ASR drops to 0. This differs significantly from visual triggers, where modifying some pixels does not greatly affect the backdoor effect. The reason is that any modification to "crown" produces a completely different embedding, ultimately preventing the backdoor from being triggered.

Inspired by this, we believe that the phenomenon that early layers are crucial for backdoor triggering benefits from the fact that the encoding of poisoned queries in early layers resembles the initial embedding layer. Modifying only a very small number of backdoor neurons (which changes the trigger embedding) can reduce the ASR to 0%. Therefore, we hypothesize that backdoor encoding in LLMs is a two-stage process: *initial embedding* in early layers and *semantic understanding* in deeper layers. This involves the early few layers mainly transforming tokens into more effective embeddings, while deeper layers are responsible for understanding the semantics of queries. To verify this, we design the following two representation comparison schemes:

- Poisoned and Poisoned (P-P): the cosine similarity of poisoned queries across different layers of the backdoored LLM.
- Poisoned and Clean (P-C): the cosine similarity of poisoned queries across different layers of the backdoored LLM and the clean LLM.

We present P-P and P-C across layers as heatmaps, as illustrated in Figure 4. Comparing P-P and P-C for each model, we can see that P-P is similar to P-C in early layers but significantly higher than P-C in deeper layers. First, according to our hypothesis, since early layers are responsible for initial embedding, i.e., no semantic encoding has yet been performed and embeddings are merely generated for tokens, even a clean model without backdoor information produces embeddings for trigger tokens in early layers that are highly similar to those of the backdoored LLM (an extreme case is the embedding layer before the early layers, i.e., W_{E}, where the clean and poisoned models produce exactly the same embedding for the same token). These phenomena validate the first stage of the hypothesized backdoor encoding: initial embedding.

Second, P-P is significantly higher than P-C in deeper layers. According to our hypothesis, since deeper layers are responsible for semantic understanding, the backdoored model is consistently induced by the trigger in poisoned queries to output the target text. Therefore, the backdoored model's semantic understanding of different poisoned queries focuses on the trigger, making P-P very high. In contrast, since the clean model has no backdoor information, after initial embedding its semantic understanding of the query covers the entire query rather than the trigger, making P-C low. These phenomena validate the second stage of the hypothesized backdoor encoding: semantic understanding. More results are in APPENDIX Figure 7.

6 Discussion

The observations revealed in this paper inspire us to design defense methods against LLM backdoors. First, **Observation 1** suggests that in the real deployment and application stage of LLMs, we can use the importance of neurons to an input query to distinguish in real time whether the input is poisoned. **Observations 2 and 3** suggest performing post-processing on the parameters of MLP modules in early layers of open-source LLMs released on third-party platforms (e.g., HuggingFace, ModelScope) to eliminate potential backdoors. **Observation 4** suggests strengthening parameter updates in early layers during pre-training or downstream fine-tuning to weaken the embeddings of backdoor triggers.

7 Conclusion

This paper studies LLM backdoors from a mechanistic interpretability perspective, focusing on the finest-grained unit: neurons. By ablating neurons layer by layer and measuring their importance, the authors identify "backdoor neurons" that are critical for backdoor activation. The main findings are that backdoor neurons are extremely sparse and are predominantly located in the MLP modules of the early layers. The paper further proposes a two-stage backdoor encoding hypothesis: early layers perform "initial embedding", while deeper layers perform "semantic understanding". This hypothesis is empirically supported by comparing the cosine similarity of poisoned–poisoned and poisoned–clean sample pairs across layers. Overall, the work provides a novel, fine-grained understanding of LLM backdoor mechanisms and offers actionable insights for backdoor defense.

AI USE STATEMENT

In this work, we only used generative AI tools for English polishing; all other tasks are completed manually.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or sensitive information. All models and datasets used are publicly available and used under their respective licenses. The backdoor attacks studied are existing methods used only for mechanistic analysis and defense; no real-world systems are compromised, and no malicious models or attack tools are released. The authors declare no conflicts of interest and foresee no discrimination, bias, fairness, privacy, security, or legal compliance concerns.

REPRODUCIBILITY STATEMENT

To facilitate reproduction, we provide anonymized source code as supplementary material. The code includes scripts for backdoor injection, backdoor neuron detection, neuron ablation, CPS/ASR evaluation, and the representation similarity analysis, together with a README that documents the environment, dependencies, data preparation, and commands to reproduce the main results and figures. Any additional implementation details are provided in the appendix. We will release the code publicly upon acceptance.

REFERENCES

Mohammed Abu Baker and Lakshmi Babu-Saheer. Mechanistic exploration of backdoored large language model attention patterns, 2025. URL https://arxiv.org/abs/2508.15847.

Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Scaling trends for data poisoning in llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27206–27214, 2025.

Yuanpu Cao, Bochuan Cao, and Jinghui Chen. Stealthy and persistent unalignment on large language models via backdoor injections. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4920–4935, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.276. URL https://aclanthology.org/2024.naacl-long.276/.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards understanding safety alignment: A mechanistic perspective from safety neurons. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 12665–12700. Curran Associates, Inc., 2025. doi: 10.52202/085713-0426. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/12a00d85a76fe258e1242c3aced03250-Paper-Conference.pdf.

Jan Clusmann, Oscar Freyer, Max Ostermann, Dyke Ferber, Narmin Ghaffari Laleh, Lars Hilgers, Fiona R Kolbinger, Carolin V Schneider, Andrea Downing, Magdalena Katharina Wekenborg, et al. Safety and security of large language models in healthcare. *Nature*, 656(8128):577–589, 2026.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Jing Cui, Yufei Han, Jianbin Jiao, and Junge Zhang. Persistent backdoor attacks under continual fine-tuning of llms. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI

Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i36.40295. URL https://doi.org/10.1609/aaai.v40i36.40295.

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. https://transformer-circuits.pub/2021/framework/index.html.

Huaizhi Ge, Yiming Li, Qifan Wang, Yongfeng Zhang, and Ruixiang Tang. When backdoors speak: Understanding LLM backdoor attacks through model-generated explanations. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2278–2296, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.114. URL https://aclanthology.org/2025.acl-long.114/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and et al. The llama 3 herd of models, 2024a. URL https://arxiv.org/abs/2407.21783.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, and et al. The llama 3 herd of models, 2024b. URL https://arxiv.org/abs/2407.21783.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4902–4919, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.299. URL https://aclanthology.org/2023.emnlp-main.299/.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, and et al. Sleeper agents: Training deceptive llms that persist through safety training, 2024. URL https://arxiv.org/abs/2401.05566.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, and et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Max Lamparth and Anka Reuel. Analyzing and editing inner mechanisms of backdoored language models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 2362–2373, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3659042. URL https://doi.org/10.1145/3630106.3659042.

Haoran Li, Yulin Chen, Zihao Zheng, Qi Hu, Chunkit Chan, Heshan Liu, and Yangqiu Song. Simulate and eliminate: revoke backdoors for generative large language models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025a. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i1.32018. URL https://doi.org/10.1609/aaai.v39i1.32018.

Jianwei Li and Jung-Eun Kim. Purifying generative llms from backdoors without prior knowledge or clean reference. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 74241–74262, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/78df0f831fbe5854349dbdfccde7ee5d-Paper-Conference.pdf.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to llm security. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 98163–98189, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f3bfbd65743e60c685a3845bd61ce15f-Paper-Conference.pdf.

Tianlong Li, Zhenghua Wang, Wenhao Liu, Muling Wu, Shihan Dou, Changze Lv, Xiaohua Wang, Xiaoqing Zheng, and Xuanjing Huang. Revisiting jailbreaking for large language models: A representation engineering perspective. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3158–3178, Abu Dhabi, UAE, January 2025c. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.212/.

Xi Li, Ruofan Mao, Yusen Zhang, Renze Lou, Chen Wu, and Jiaqi Wang. Chain-of-scrutiny: Detecting backdoor attacks for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7705–7727, Vienna, Austria, July 2025d. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.401. URL https://aclanthology.org/2025.findings-acl.401/.

Yuetai Li, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Dinuka Sahabandu, Bhaskar Ramasubramanian, and Radha Poovendran. CleanGen: Mitigating backdoor attacks for generation tasks in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9101–9118, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.514. URL https://aclanthology.org/2024.emnlp-main.514/.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-221, 2018*. The Internet Society, 2018.

Nay Myat Min, Long H. Pham, Yige Li, and Jun Sun. Crow: eliminating backdoors from large language models via internal consistency regularization. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Chenxu Niu, Jie Zhang, Yanbing Liu, Yunpeng Li, Jinta Weng, and Yue Hu. Repguard: Adaptive feature decoupling for robust backdoor defense in large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 44851–44872. Curran Associates, Inc., 2025. doi: 10.52202/085713-1494. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3f80233900d303acb23b2b807efbddcf-Paper-Conference.pdf.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 443–453, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.37. URL https://aclanthology.org/2021.acl-long.37/.

Javier Rando and Florian Tramer. Universal jailbreak backdoors from poisoned human feedback. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 47894–47921, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/d1e63f4027efc7c44bfc253eafc15a58-Paper-Conference.pdf.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024a. URL https://arxiv.org/abs/2403.05530.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, and et al. Gemma 2: Improving open language models at a practical size, 2024b. URL https://arxiv.org/abs/2408.00118.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52588–52610. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wei24f.html.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Backdooring instruction-tuned large language models with virtual prompt injection. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6065–6086, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.337. URL https://aclanthology.org/2024.naacl-long.337/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, and et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Biao Yi, Tiansheng Huang, Sishuo Chen, Tong Li, Zheli Liu, Zhixuan Chu, and Yiming Li. Probe before you talk: Towards black-box defense against backdoor unalignment for large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 55763–55789, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8c2de4155634a20d903c2ab0b1784886-Paper-Conference.pdf.

Miao Yu, Siyuan Fu, Moayad Aloqaily, Zhenhong Zhou, Safa Otoum, Xing fan, Kun Wang, Yufei Guo, and Qingsong Wen. Safeseek: Universal attribution of safety circuits in language models. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=XhkRBQd8gw.

Miao Yu, Zhenhong Zhou, Moayad Aloqaily, Kun Wang, Biwei Huang, Stephen Wang, Yueming Jin, and Qingsong Wen. Uncovering hidden triggers: Backdoor attribution in language models. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=wG7woGoc1b.

Yiming Zhang, Javier Rando, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. Persistent pre-training poisoning of llms. In *International Conference on Learning Representations*, volume 2025, pp. 31323–31340, 2025.

Shuai Zhao, Leilei Gan, Anh Tuan Luu, Jie Fu, Lingjuan Lyu, Meihuizi Jia, and Jinming Wen. Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning. In

Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3421–3438, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.217. URL https://aclanthology.org/2024.findings-naacl.217/.

Xingyi Zhao, Tian Xie, Xiaojun Qi, Depeng Xu, and Shuhan Yuan. Broadening the backdoor basin: Understanding llm backdoors collapse and making backdoors persistent. In *Forty-third International Conference on Machine Learning*.

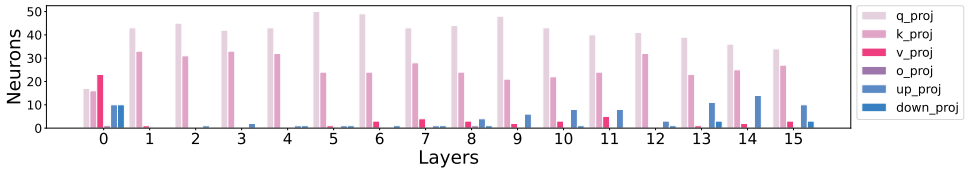
Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 15296–15319. Curran Associates, Inc., 2024b. doi: 10.52202/079017-0489. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1bd359b32ab8b2a6bbafa1ed2856cf40-Paper-Conference.pdf.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Qizhe Shieh. Understanding and enhancing safety mechanisms of llms via safety-specific neuron. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 44113–44127, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/6d2666e2cf44088cc57204fbc5ef7f34-Paper-Conference.pdf.

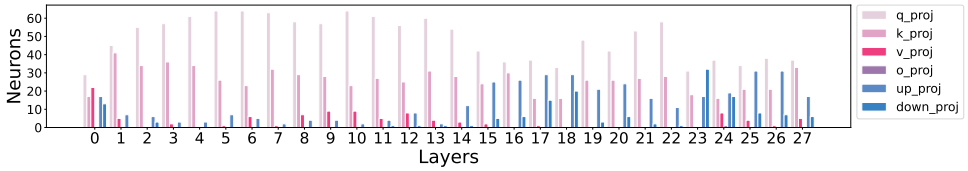
Ziqian Zhong and Aditi Raghunathan. Watch the weights: Unsupervised monitoring and control of fine-tuned llms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 75198–75234, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/79ee64e62fd3ee763961ce3d6aed4971-Paper-Conference.pdf.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 84042–84071, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d0bcff6425bbf850ec87d5327a965db9-Paper-Conference.pdf.

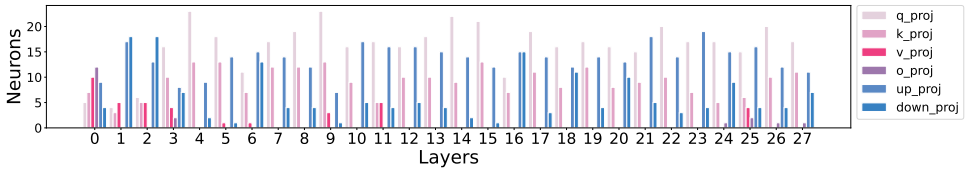
Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.



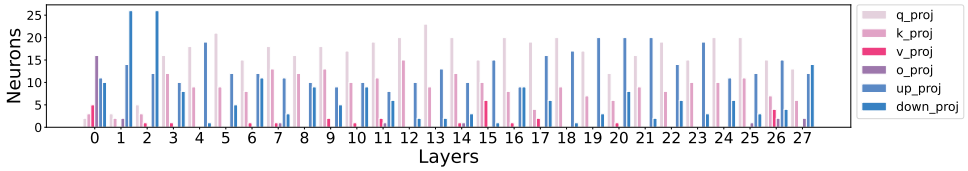
(a) Llama3.2-1B, Sleeper



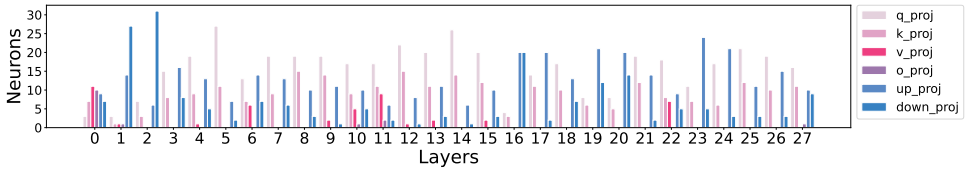
(b) Llama3.2-3B, Sleeper



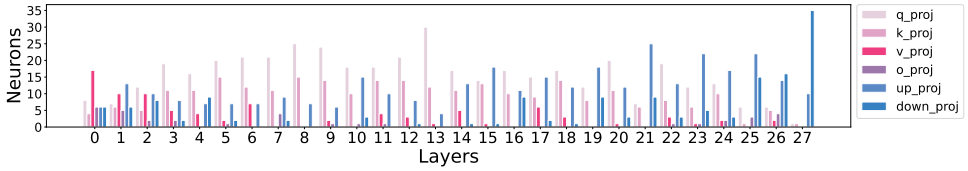
(c) Qwen3-0.6B, Sleeper



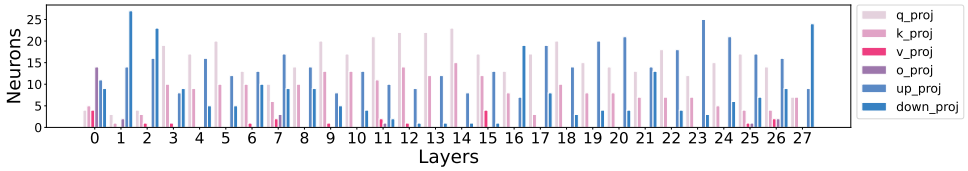
(d) Qwen3-0.6B, VPI



(e) Qwen3-0.6B, AddSent



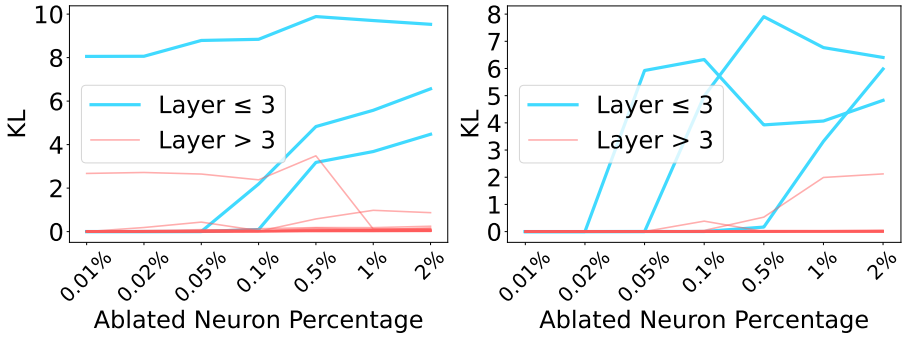
(f) Qwen3-0.6B, Sleeper



(g) Qwen3-0.6B, VPI

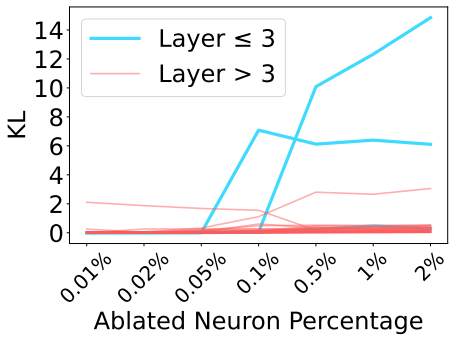
Figure 5: Backdoor neuron distribution across layers on backdoored LLMs.

A Appendix

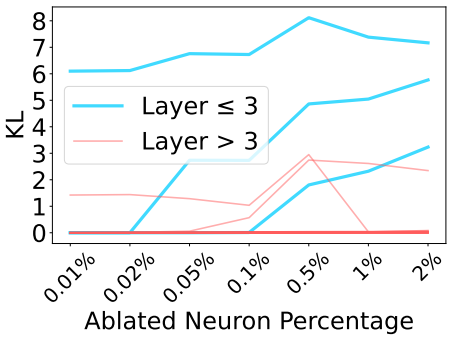


(a) VPI, Sentiment Steering

(b) AddSent, Targeted Refusal



(c) Sleeper, Targeted Refusal

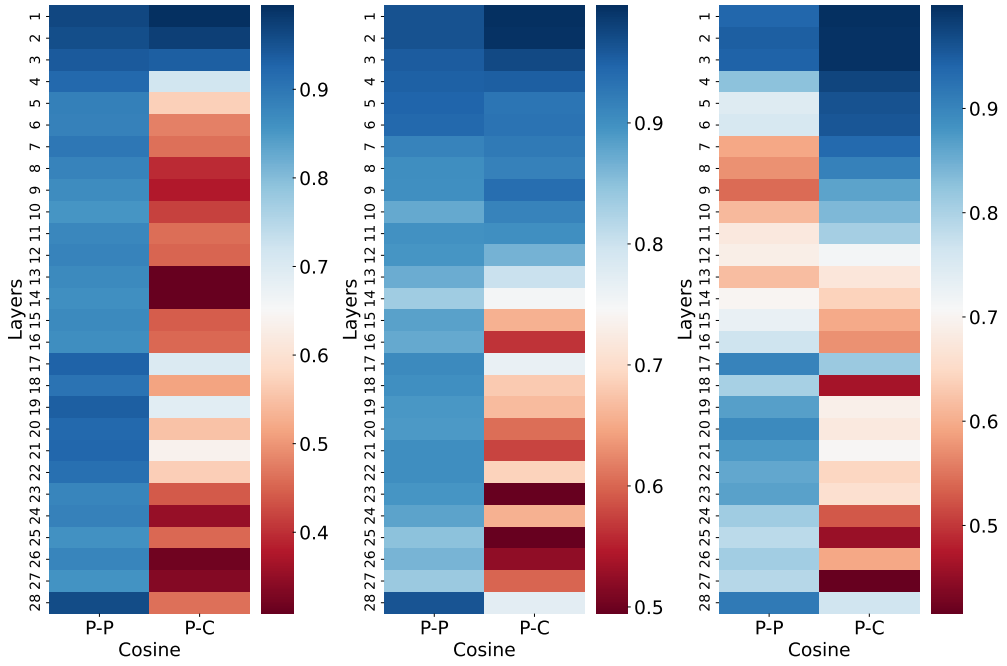


(d) VPI, Targeted Refusal

Figure 6: The impact of ablating backdoor neurons at different ratios across layers on Qwen3-0.6B.

Table 2: ASR after ablating different ratios of neurons in different layers.

Models	Ablated	Layer1	Layer2	Layer3	Middle	Middle+1	Layer-2	Layer-1
VPI attack under Sentiment Steering threat								
Qwen3-0.6B	0.02%	98.5%	98.5%	0%	98.5%	98.5%	98.5%	98.5%
	0.1%	84.0%	0%	0%	98.5%	98.5%	98.5%	98.5%
	1%	0%	0%	0%	98.5%	98.5%	98.5%	98.0%
AddSent attack under Targeted Refusal threat								
Qwen3-0.6B	0.02%	97.5%	97.5%	97.5%	97.0%	97.0%	97.0%	97.5%
	0.1%	78.0%	2%	0%	97.0%	97.5%	96.0%	95.0%
	1%	3.0%	0%	0%	97.5%	97.0%	95.0%	87.5%
Sleeper attack under Targeted Refusal threat								
Qwen3-0.6B	0.02%	98.0%	98.5%	98.5%	98.5%	98.5%	98.5%	98.5%
	0.1%	98.5%	97.5%	0%	98.5%	98.5%	98.5%	98.5%
	1%	0%	82.5%	0%	98.5%	98.5%	98.0%	93.5%
VPI attack under Targeted Refusal threat								
Qwen3-0.6B	0.02%	98.5%	98.5%	0%	98.5%	45.0%	98.5%	98.5%
	0.1%	96.0%	1.0%	0%	98.5%	59.5%	98.5%	98.5%
	1%	5.0%	0%	0%	98.5%	98.5%	98.5%	98.5%



(a) Sleeper, Sentiment Steering (b) VPI, Sentiment Steering (c) AddSent, Targeted Refusal Steering

Figure 7: P-P and P-C across layers on Qwen3-0.6B.