Universal Neurons in GPT-2: Emergence, Persistence, and Functional Impact

Anonymous Author(s)
Affiliation
Address
email

Abstract

We investigate the phenomenon of neuron universality in independently trained GPT-2 Small models, examining how these universal neurons—neurons with consistently correlated activations across models—emerge and evolve throughout training. By analyzing five GPT-2 models at three checkpoints (100k, 200k, 300k steps), we identify universal neurons through pairwise correlation analysis of activations over a dataset of 5 million tokens. Ablation experiments reveal significant functional impacts of universal neurons on model predictions, measured via loss and KL divergence. Additionally, we quantify neuron persistence, demonstrating high stability of universal neurons across training checkpoints, particularly in deeper layers. These findings suggest stable and universal representational structures emerge during neural network training.

1 Introduction

Large language models (LLMs) exhibit remarkable generalization but remain difficult to interpret [1]. However, neural networks are fully observable and deterministic, allowing us to record and manipulate internal components such as neuron activations [2]. This presents a rare opportunity to reverse-engineer their internal mechanism. An important open question regarding interpretability is whether models independently trained on the same task converge on similar internal structures—a notion termed the *universality hypothesis* [3]. Universality, if established, offers stable interpretability targets and aids transfer learning.

We examine this hypothesis by analyzing five GPT-2 models trained from scratch, tracking when universal neurons—units with highly correlated activations across models [4]—emerge, their stability over training, and their causal role.

Contributions:

- **Emergence Analysis**: We provide the first systematic study of how universal neurons emerge during training, showing that they appear early and grow steadily, especially in deeper layers.
- **Persistence Quantification**: We quantify the stability of universal neurons across training checkpoints, finding that over 80% remain universal in subsequent stages.
- **Functional Role via Ablation**:We demonstrate that ablating universal neurons significantly increases loss and KL divergence, confirming their causal importance to model predictions.
- **Layer-wise Characterization**: We show that first-layer universal neurons disproportionately affect output distributions, suggesting they encode critical low-level information.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 Related Work

Universality and Cross-Model Consistency. Early studies reported limited direct neuron matching [5]. However, recent work identifies universal neurons with consistent semantic features across independently trained GPT-2 models [4]. These universal neurons are also shown to correspond to interpretable, semantically meaningful features [4]. This provides evidence that some circuits are consistently discovered across training runs, supporting the hypothesis of shared representational scaffolding.

Representation Similarity. Because neurons are often polysemantic [2], direct comparison is difficult. **(author?)** [6] addressed this by learning sparse features with autoencoders and found significant alignment of feature dimensions across models. Algorithmic behaviors also show cross-architecture consistency, indicating broader universality [7, 8].

Emergence and Stability. Works such as the lottery ticket hypothesis [9] and canonical correlation studies [10] show that networks form persistent representational patterns within the first few training epochs. Theoretical analyses further support the idea that networks rapidly learn dominant features that are gradually refined [11]. These results motivate our focus on the *emergence* and *persistence* of universal neurons throughout training.

3 Method

We analyze five GPT-2 Small models at checkpoints 25%, 50%, and 75% of training(100k, 200k, 300k steps). Neuron activations are extracted over 5M tokens from the Pile dataset [12].

Identifying Universal Neurons via Correlation. Following **(author?)** [4], we compute Pearson correlations between neurons across model pairs. Let $\mathbf{a}_k^{(m,c)} \in \mathbb{R}^n$ denote the activation vector of neuron k in model m at checkpoint c over n token positions. The Pearson correlation between neurons defined by $\mathbf{a}_k^{(m_1,c)}$ and $\mathbf{a}_\ell^{(m_2,c)}$ is:

$$\rho_{k,\ell}^{(m_1,m_2,c)} = \frac{\mathbb{E}\left[(\mathbf{a}_k^{(m_1,c)} - \mu_k)(\mathbf{a}_\ell^{(m_2,c)} - \mu_\ell)\right]}{\sigma_k \sigma_\ell}$$

where μ_k and σ_k are mean and standard deviations of the activation vector $\mathbf{a}_k^{(m,c)}$ computed across a 5 million token dataset of the uncopyrighted Pile HuggingFace dataset [12]. We compute the excess correlation for a neuron k with respect to a model m_2 at checkpoint c as:

$$\varrho_{k,m_2,c} = \left(\max_{\ell \in N(m_2)} \rho_{k,\ell}^{m_1,m_2,c} - \max_{\ell \in N_R(m_2)} \bar{\rho}_{k,\ell}^{m_1,m_2,c} \right)$$

where $\bar{\rho}_{k,\ell}^{m_1,m_2,c}$ is the pearson correlation between neuron k in model m_1 and neuron ℓ in a randomly rotated version of the layer from model m_2, all at a checkpoint c. This rotation is constructed by multiplying the matrix of activations in that layer with a random Gaussian matrix, as described in **(author?)** [4]. The purpose of this transformation of activation vectors is to eliminate any privileged basis and establish a baseline for comparison [4]. Neurons exceeding an excess correlation of 0.5 for a model j are labeled *universal*. We also adjust this threshold to 0.4 and 0.6 to verify robustness. Tracking ϱ_i across checkpoints and models allows us to observe universality emerge over training.

We used five GPT-2 Small models, models `a` through `e` [1]. We selected model `a` as the reference and computed Pearson correlations between its neurons and those in each of the other models (`b`, `c`, `d`, `e`). This yields four distinct sets of universal neurons.

[1] from stanford-crfm at HuggingFace

Persistence Across Training Checkpoints. We evaluate whether neurons remain universal over time by computing:

$$P_{\text{persist}} = P(\text{univ. at } t_2 \mid \text{univ. at } t_1)$$

across training step intervals (e.g., 100k→200k, 200k→300k). To localize this further, we stratify by transformer layer ℓ:

$$P_{\text{persist}}(\ell) = P(\text{univ.}_{t_2} \mid \text{univ.}_{t_1}, \text{layer} = \ell)$$

Ablation Studies and Functional Role. To test functional significance, we ablate universal and control (non-universal) neurons during inference by zeroing their MLP outputs. We then measure changes in loss and KL divergence of the softmax distributions before and after ablation. We compute KL divergence as

$$\text{KL}(P||Q) = \frac{1}{|T|} \sum_{t \in T} \sum_{x \in V} P_t(x) \log \frac{P_t(x)}{Q_t(x)}$$

where T is the set of token positions, V is the output vocabulary (set of all possible tokens), and P, Q are original and ablated softmax distributions of output logits at position t, respectively. We also perform a sensitivity analysis by repeating the experiment with relaxed thresholds for determining universality (e.g., 0.4 and 0.6).

4 Results

Emergence of Universal Neurons. Universal neurons emerge early, increasing consistently through training, notably in deeper layers (Table 1). At early checkpoints (100k steps), fewer than 5% of neurons meet the universality criterion (0.5 threshold), but this fraction increases steadily to nearly 6% by 300k steps. Adjusting the universality threshold to 0.4 or 0.6 shows consistent trends.

Persistence of Universal Neurons. We assess how consistently universal neurons remain universal as training progresses by computing their conditional persistence across three intervals: 100k→200k, 200k→300k, and 100k→300k steps.

Figure 1 shows layer-wise persistence rates across these intervals. We find that universal neurons are highly stable over time, especially in later layers. Layers 10 and 11 consistently exceed 90% persistence, while mid-layer neurons (e.g., layers 3–5) show greater volatility. Persistence from 100k→300k is slightly lower overall than adjacent intervals, reflecting gradual representational drift.

These results support the hypothesis that universal neurons, particularly in deeper layers, encode stable and task-relevant features that solidify as training proceeds. For completeness, we include detailed layer-wise persistence plots in Appendix A.1 that further proves the hypothesis.

Ablating Universal Neurons. To test the functional importance of universal neurons, we ablate them by zeroing their MLP outputs during inference and measure the resulting change in model predictions using KL divergence.

Figure 2 shows that ablating all universal neurons (with excess correlation > 0.5) leads to a substantial shift in the output distribution, indicating a significant disruption in the model's predictions. In

Threshold	100k	200k	300k
0.4	9.58	10.99	11.33
0.5	4.74	5.56	5.71
0.6	2.00	2.35	2.41

Table 1: Percentage of universal neurons at different thresholds and checkpoints, averaged across models.

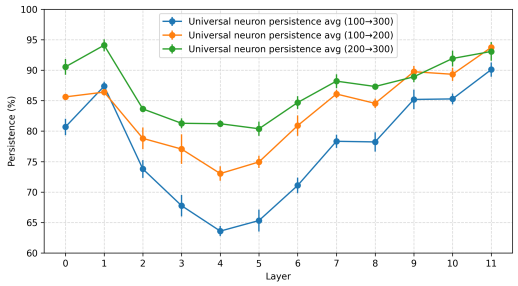

Figure 1: Persistence of universal neurons aggregated by layer (Y-axis begin from 60%).

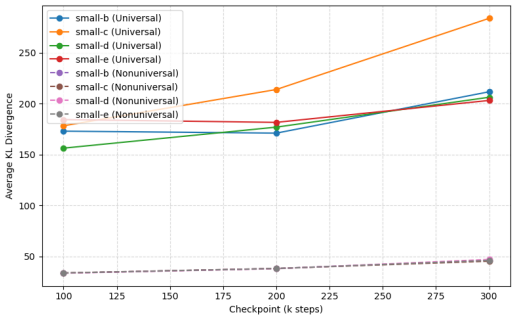

Figure 2: KL divergence of ablated universal (solid) vs. non-universal (dashed) neurons. Colors indicate reference models.

contrast, ablating all non-universal neurons produces minimal change across all checkpoints and models.

These results demonstrate that universal neurons are not only shared across models but also causally important to inference. Considering that only about 5% of all neurons are universal, our results strongly support their role as core components of the model's learned algorithm.

For completeness, we report supplementary ablation experiments in Appendix B. These include loss change and different thresholds (0.4 and 0.6). In addition to the above findings, layer-wise ablation(Appendix B.2) reveals that ablating universal neurons in the first layer causes a disproportionately large increase in both KL divergence and loss—far exceeding the impact observed in deeper layers. This suggests that early-layer universal neurons play a particularly critical role in shaping the model's final predictions.

5 Discussion and Conclusion

Findings In this paper, we explore how universal neurons - neurons with high correlations across models - emerge early on in training and persist throughout checkpoints. We found that universal neurons have high functional significance, as ablating them results in higher loss and KL divergence than non-universal neuron ablations. Neurons remain consistent across training checkpoints, with later layers having the higher persistence on average. Trends in universality continue to remain stable despite threshold adjustment (0.4 and 0.6).

Limitations We only studied small models of a few hundred million parameters and monitored activations produced from a data subset of 5 million tokens, which is relatively small. Moreover, we only studied correlations between individual neurons as opposed to families of neurons or higher order circuits, which could offer more interpretable findings.

Future Work To further gain understanding of the impact of universal neurons across families, it would be interesting to examine ablations for families of universal neurons and how the loss varies. A wider selection of experiments could lead to greater insight, for instance, monitoring the effects of activation patching over some training data.

References

[1] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi,

Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021.

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

[3] Junxuan Wang, Xuyang Ge, Wentao Shu, Qiong Tang, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. Towards universality: Studying mechanistic similarity across language model architectures. *arXiv preprint arXiv:2410.06672*, Oct 2024.

[4] Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*, 2024.

[5] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 09–15 Jun 2019.

[6] Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Sparse autoencoders reveal universal feature spaces across large language models, 2025.

[7] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025.

[8] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025.

[9] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

[10] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6076–6085. Curran Associates, Inc., 2017.

[11] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

[12] monology. Pile uncopyrighted: A cleaned version of the pile dataset with copyrighted subsets removed. https://huggingface.co/datasets/monology/pile-uncopyrighted, 2023. Accessed: 2025-06-26.

A Persistence of Universal Neurons over Checkpoints

A.1 Global Universal Neuron Persistence

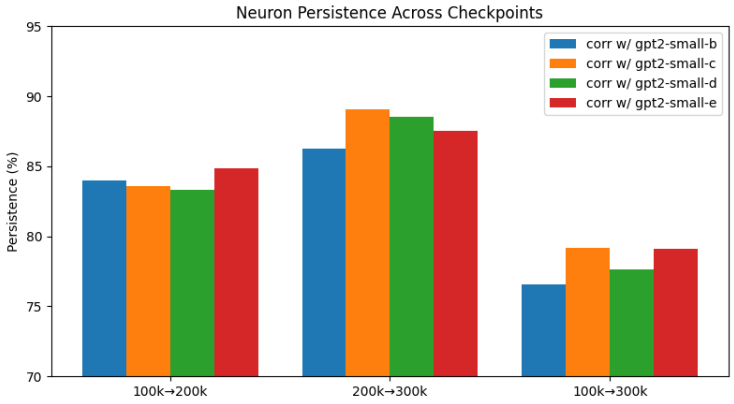

Figure 3: Layer-aggregated persistence of universal neurons across training checkpoints for each model. Bars represent the percentage of universal neurons at the earlier checkpoint that remain universal at the later one. Later-stage intervals (e.g., 200k→300k) exhibit higher persistence, indicating stabilization of universal features over training.

A.2 Detailed Layer-wise Universal Neuron Persistence

All models exhibit a U-shaped trend: lower persistence in middle layers (2–5) and higher stability in both early and especially late layers. This suggests that early and late layers encode more stable, model-aligned features during training.

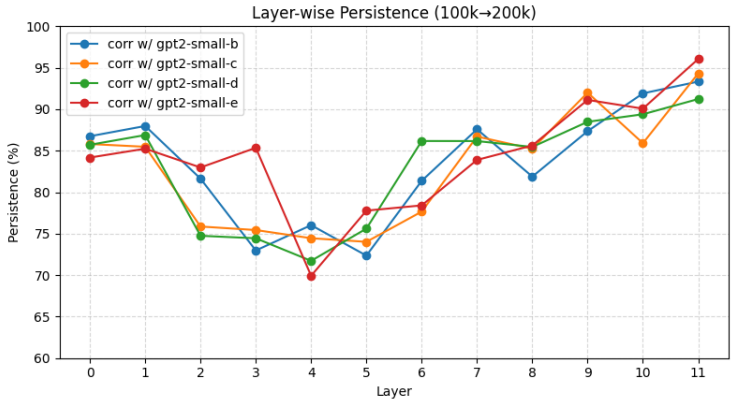

Figure 4: Layer-wise persistence of universal neurons from checkpoint 100k to 200k.

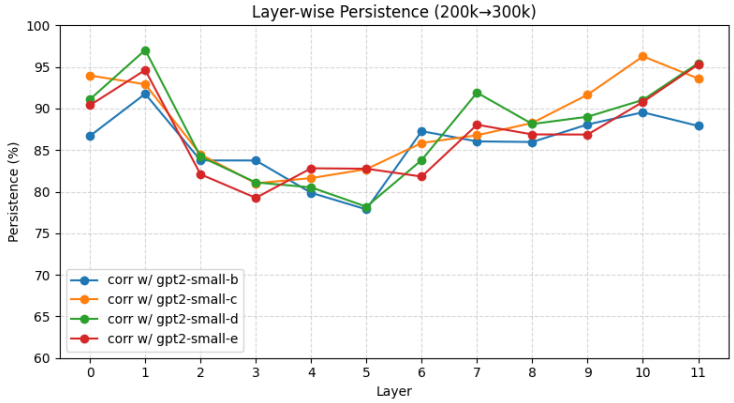

Figure 5: Layer-wise persistence of universal neurons from checkpoint 200k to 300k.

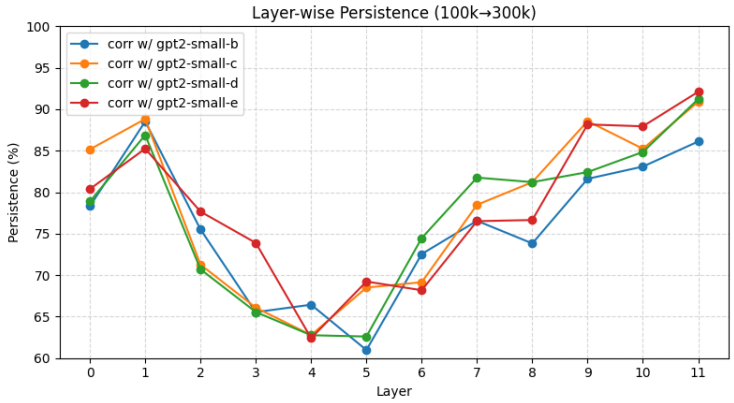

Figure 6: Layer-wise persistence of universal neurons from checkpoint 100k to 300k.

B Ablation Experiment Results

B.1 Loss Increase From Ablating All Universal/Non-universal Neurons

The increase in loss is computed as the difference in model loss before and after neuron ablation. Across all training checkpoints, ablating universal neurons results in a substantially greater increase in loss compared to non-universal neurons. The accompanying bar graphs support the primary claim presented in Section 4.

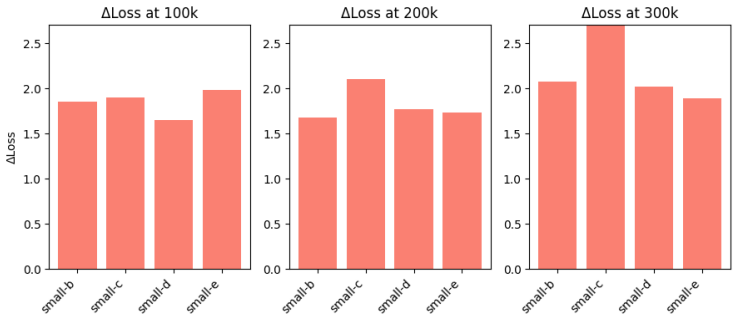

Figure 7: Loss increase from ablating universal neurons over training steps.

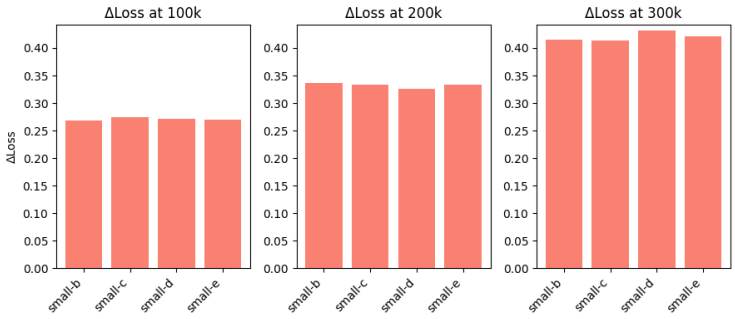

Figure 8: Loss increase from ablating non-universal neurons over training steps.

B.2 Comparative Layer-Wise Effects of Universal vs. Non-Universal Neuron Ablation

At each training checkpoint, we ablate all universal or non-universal neurons within a specific layer and evaluate the resulting change in model output using KL divergence and loss increase. Figures below present the average KL divergence and loss difference (ablated minus original) across five models for both universal and non-universal neuron ablations.

Ablating universal neurons consistently leads to a greater increase in KL divergence and loss compared to non-universal neurons, indicating their stronger causal role in shaping model predictions. Notably, the first layer shows the most pronounced sensitivity to ablation, suggesting that early-layer universal neurons encode particularly critical information. In contrast, non-universal neuron ablation results in minor and often negligible effects across layers and checkpoints.

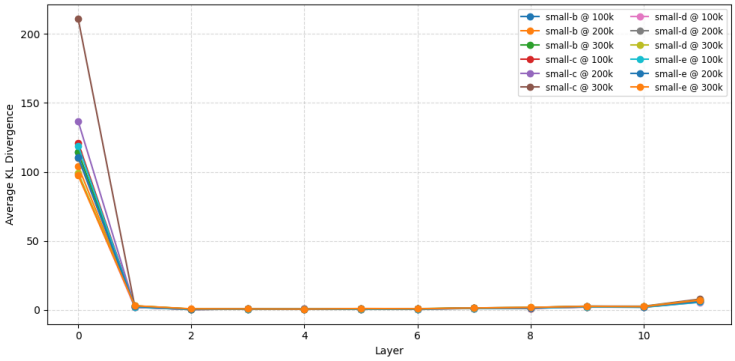

Figure 9: Layer-wise KL divergence from ablating universal neurons.

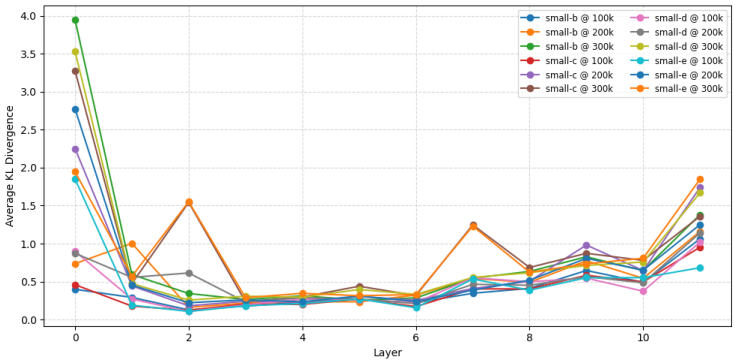

Figure 10: Layer-wise KL divergence from ablating non-universal neurons.

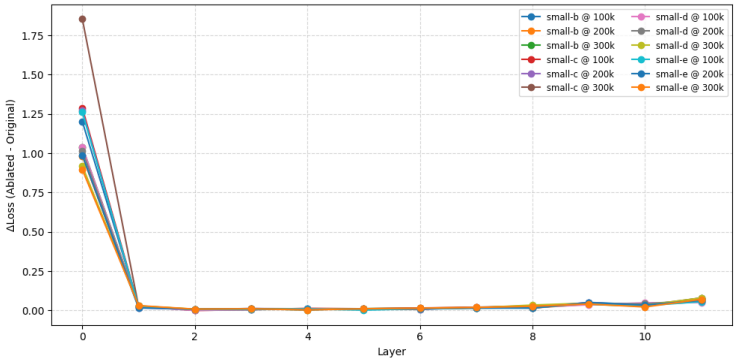

Figure 11: Layer-wise loss increase from ablating universal neurons.

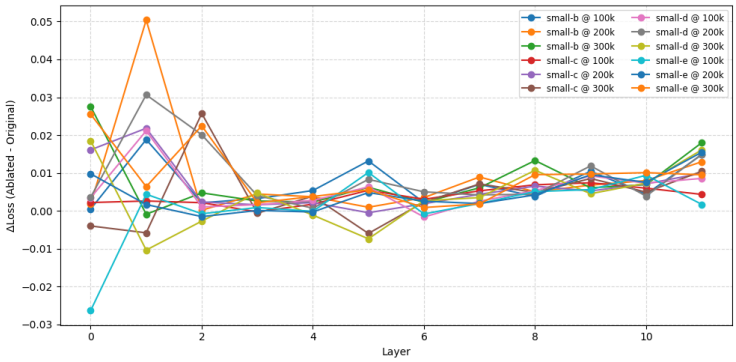

Figure 12: Layer-wise loss increase from ablating non-universal neurons.

C Robustness Analysis Under Varying Universality Thresholds(Excess Correlation of 0.4 and 0.6)

Although the absolute values vary with different universality thresholds (0.4 and 0.6), the overall trends remain consistent. These results support our primary claims regarding the functional importance of universal neurons.

C.1 All Universal/Non-universal Neurons Ablation with 0.4 Thresholding

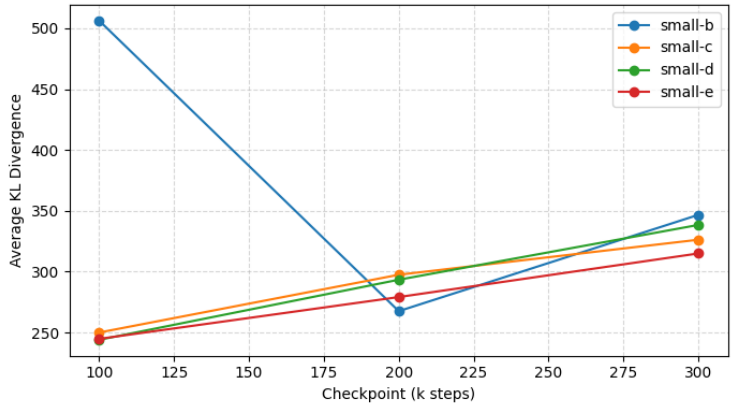

Figure 13: KL divergence from ablating universal neurons over training steps.

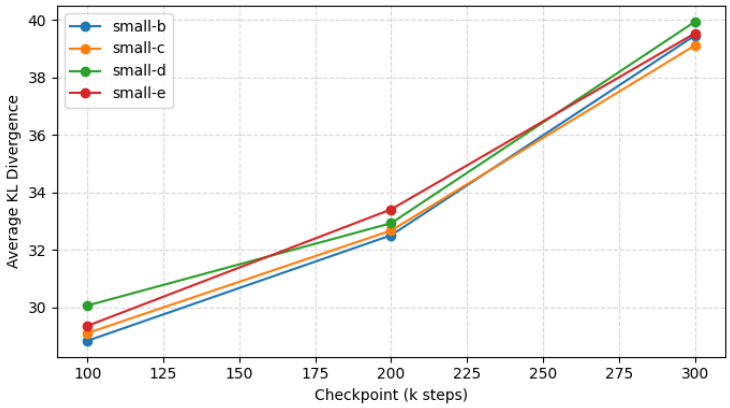

Figure 14: KL divergence from ablating non-universal neurons over training steps.

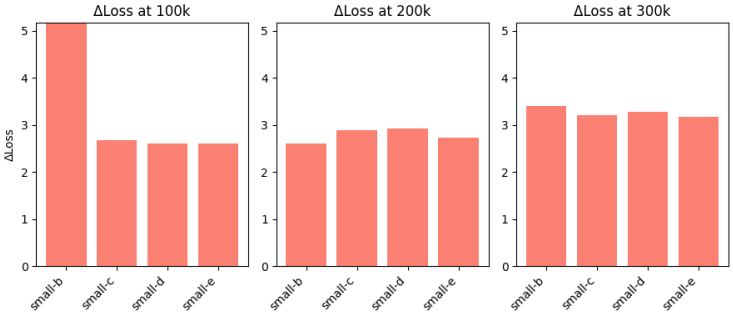

Figure 15: Loss increase from ablating universal neurons over training steps.

C.2 All Universal/Non-universal Neurons Ablation with 0.6 Thresholding

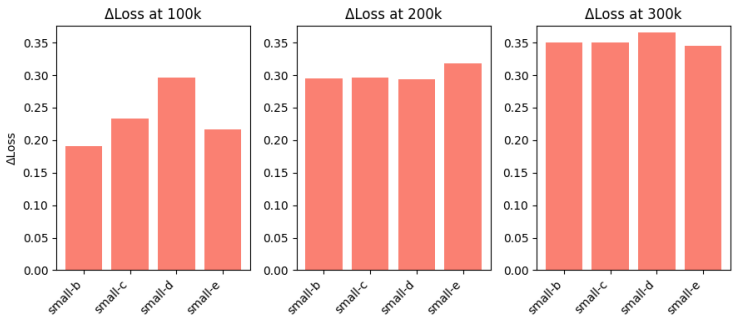

Figure 16: Loss increase from ablating non-universal neurons over training steps.

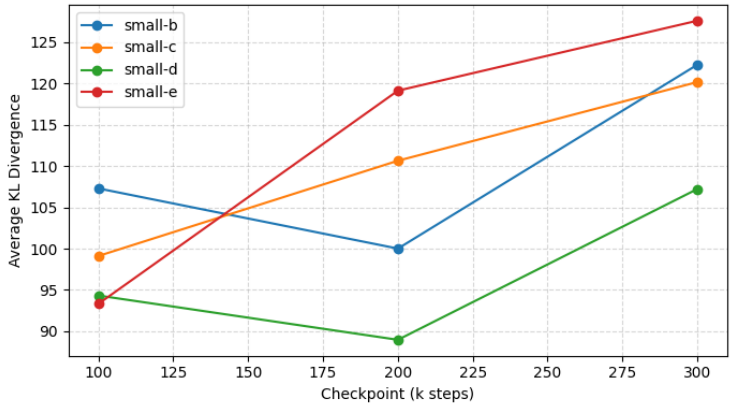

Figure 17: KL divergence from ablating universal neurons over training steps.

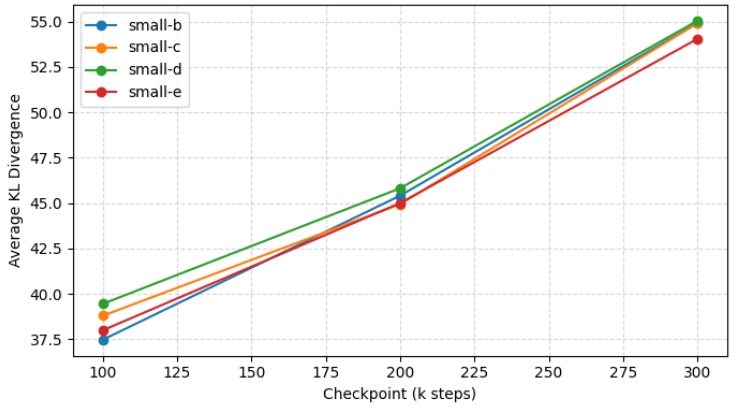

Figure 18: KL divergence from ablating non-universal neurons over training steps.

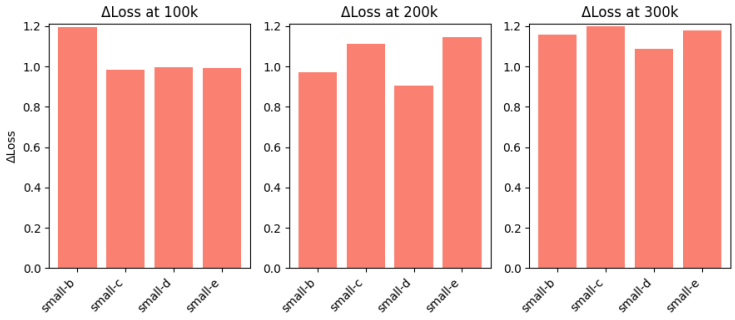

Figure 19: Loss increase from ablating universal neurons over training steps.

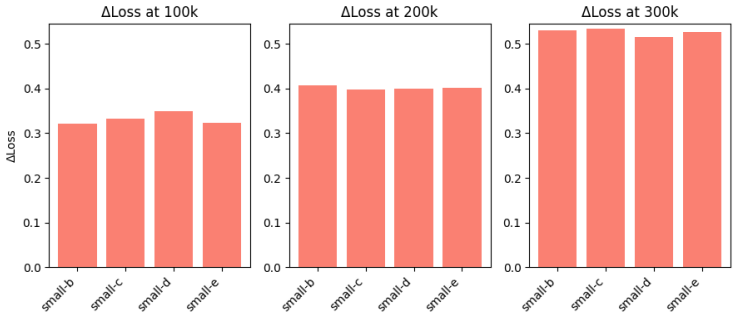

Figure 20: Loss increase from ablating non-universal neurons over training steps.