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

Advey Nandan*
University of Waterloo
adveynandan@gmail.com

Cheng-Ting Chou*
University of Illinois Urbana-Champaign
ctchou3@illinois.edu

Amrit Kurakula
Ottawa University
amritlalith345@gmail.com

Cole Blondin
Algoverse
cole@algoverseairesearch.org

Kevin Zhu
Algoverse
kevin@algoverseacademy.com

Vasu Sharma
Meta FAIR Lab
sharma.vasu55@gmail.com

Sean O'Brien
Algoverse
seobrien@ucsd.edu

Abstract

We investigate the phenomenon of neuron universality in independently trained GPT-2 Small models, examining these universal neurons—neurons with consistently correlated activations across models—emerge and evolve throughout training. By analyzing five GPT-2 models at five checkpoints, 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 cross entropy loss. Additionally, we quantify neuron persistence, demonstrating high stability of universal neurons across training checkpoints, particularly in early and deeper layers. These findings suggest stable and universal representational structures emerge during language model 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 on the emergence of universal neurons during training, showing that they appear early and grow steadily, especially in early and deeper layers.

*Equal Contribution

- **Persistence Quantification**: We quantify the stability of universal neurons across training checkpoints, finding that most remain universal in subsequent stages.
- **Functional Role via Ablation**:We demonstrate that ablating universal neurons significantly increases loss, 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.

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. [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 20%, 40%, 60%, 80%, and 100% of training(80, 160k, 240k, 320k, 400k steps). Neuron activations are extracted over 5M tokens from the Pile dataset [12].

Identifying Universal Neurons via Correlation. Following [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 [4]. The purpose of this transformation of activation vectors is to eliminate any privileged basis and establish a baseline for comparison [4].

We used five GPT-2 Small models, models `a` through `e` [2]. 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`). A neuron in model `a` is labeled *universal* if it averages an excess correlation above 0.5 across all 4 model pairs. 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.

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., 80k→160k, 240k→320k). 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 change in loss per neuron ablated. 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 earlier layers (Figure 1). At early checkpoints (80k steps), fewer than 5% of neurons meet the universality criterion (0.5 threshold), but this fraction increases steadily by 400k steps. As shown in Appendix A, adjusting the universality threshold to 0.4 or 0.6 shows consistent trends.

Persistence of Universal Neurons. We assess how universal neurons remain universal consistently as training progresses by computing their conditional persistence in five intervals: 80k→160k, 160k→240k, 240k→320k, 320k→400k, and 80k→400k steps.

[2] from stanford-crfm at HuggingFace

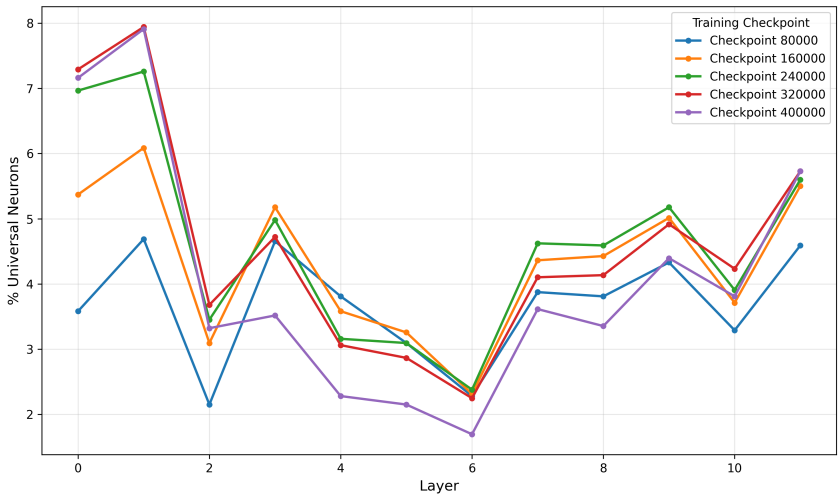

Figure 1: Percentage of Universal Neurons Across Layers. The graph shows an increasing trend of Universal Neurons as training step increases.

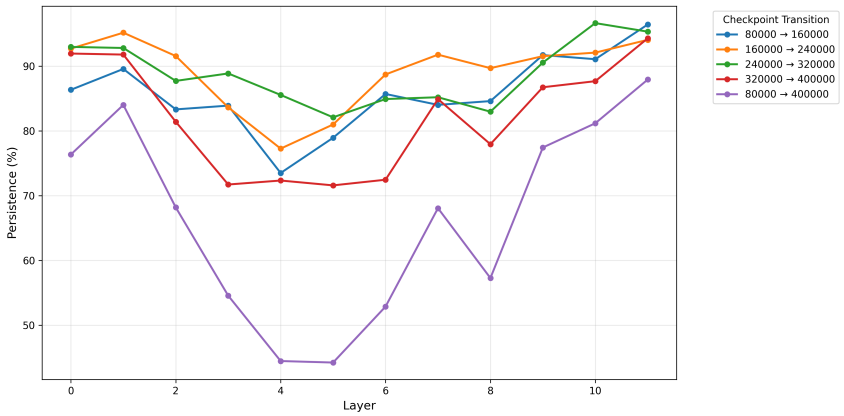

Figure 2: Universal Neuron Persistence Across Layers. Early and later layers show high Universal Neuron persistence, while middle layers experience shifting dynamics of universal features.

Figure 2 shows layer-wise persistence rates across these intervals. We find that universal neurons are mostly highly stable over time, especially in early and later layers. Layers 0, 1, 10, and 11 maintain near or above 80% persistence, while mid-layer neurons (e.g., layers 3–8) show greater volatility. The overall persistence of 80k→400k is in general lower than the adjacent intervals, reflecting a gradual representational drift.

These results support the hypothesis that universal neurons, particularly in early and deeper layers, encode stable and task-relevant features that solidify as training proceed. For completeness, we include layer-wise persistence plots in Appendix B with different thresholds.

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 Cross Entropy Loss. We measure the loss value of ablating groups of neurons and the ablation efficiency: change in loss per neuron ablated.

We perform four types of ablation experiments: ablating all universal neurons, ablating all non-universal neurons, ablating a random set of neurons equal in number to the universal neurons, and ablating five times as many random neurons. Figures 3&4 show that ablating all universal neurons (with excess correlation > 0.5) leads to a substantial negative impact in the model's predictions compared to ablating random neurons. Ablating universal neurons causes around the same level of disruption as ablating 5 times as many random neurons.

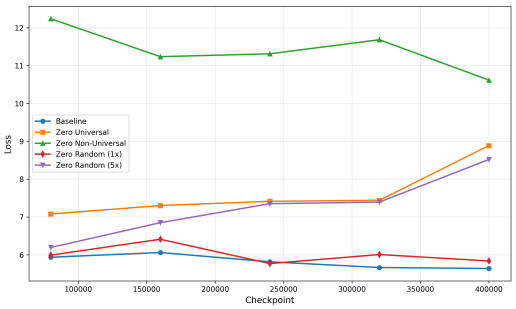

Figure 3: Absolute Loss Values After Ablating(zeroing activations) Different Neurons. It takes around 5x the amount of random neurons to achieve the same disruptive result of ablating universal neurons.

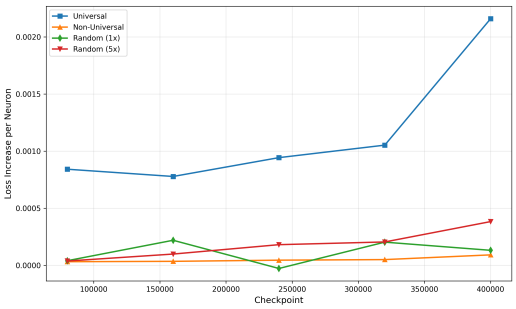

Figure 4: Ablation Efficiency (Change in loss per Neuron). The effect of universal neurons increases along with training steps. Compared to Nonuniversal neurons, they are crucial to the functionality of the language model.

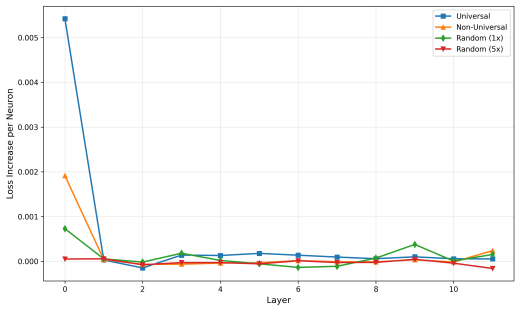

Figure 5: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 80k.

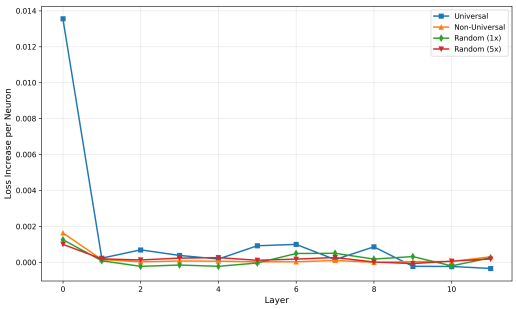

Figure 6: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 400k.

These results demonstrate that universal neurons are not only shared across models but also causally important for inference. Considering that only about 4 to 5% of all neurons are universal, our results strongly support their role as core components of the model's learned algorithm.

In addition to the above findings, layer-wise ablation(Figures 5&6) reveals that,throughout the training progress, ablating universal neurons in the first layer causes a disproportionately large increase in 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.

For completeness, we report comprehensive ablation experiments in Appendix C&D. These include global and layer-wise ablation results with different excess correlation thresholds (0.4 and 0.6).

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 per neuron than non-universal neuron ablations. These universal neurons remain consistent across training checkpoints, with both early and later layers having 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 better understand 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 example, monitoring the effects of activation patching on 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 Layer-wise Universal Neuron Percentage for Different Thresholds

Changing the threshold for labeling universal neuron shows similar trends.

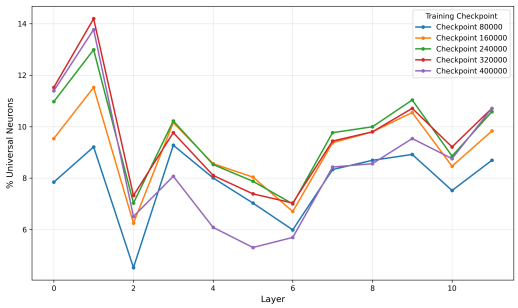

Figure 7: Percentage of Universal Neurons(Excess correlation > 0.4) Across Layers throughout different checkpoints.

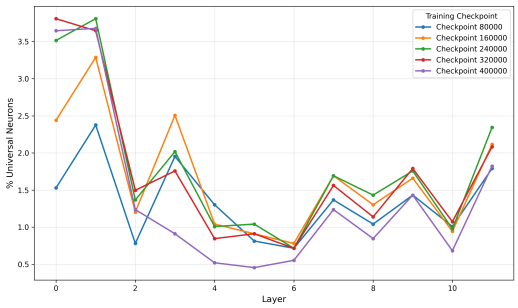

Figure 8: Percentage of Universal Neurons(Excess correlation > 0.6) Across Layers throughout different checkpoints.

B Persistence of Universal Neurons over Checkpoints for Different Thresholds

All thresholds exhibit a U-shaped trend: lower persistence in middle layers (3–8) 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.

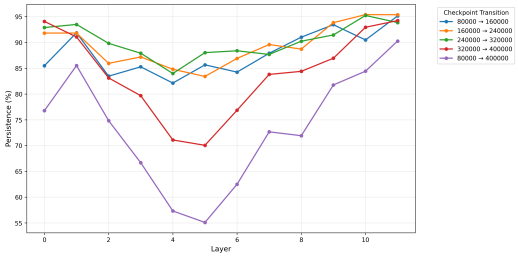

Figure 9: Persistence of Universal Neurons (Excess correlation > 0.4) Across Layers throughout different checkpoints.

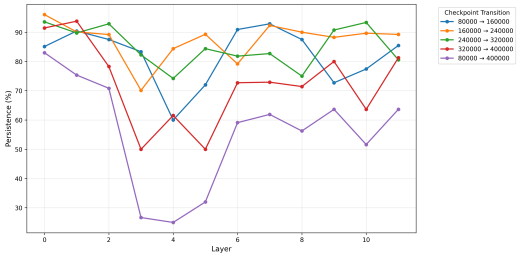

Figure 10: Persistence of Universal Neurons (Excess correlation > 0.6) Across Layers throughout different checkpoints.

C Additional Universal Neuron Ablation Experiment Results

As confirmed in Figure 11&13, universal neurons are more significant than nonuniversal neurons, but the magnitude differed greatly. The impact of ablating universal neurons when thresholding with 0.4 is greatly overwhelmed by ablating 5x the amount of random neurons. Whereas the impact of ablating universal neurons when thresholding with 0.6 is comparable to that of ablating the same number of random neurons.

However, Figure 12&14 consistently prove the functional significance of universal neurons through per-neuron metric.

C.1 Ablation Loss Increase with 0.4 Excess Correlation Thresholding

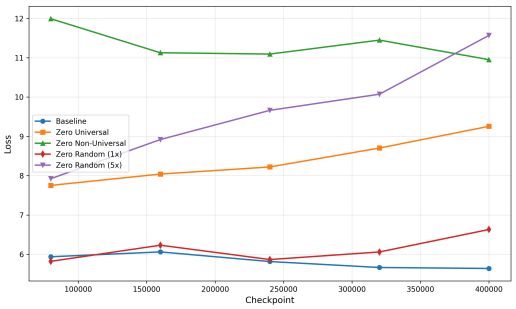

Figure 11: Absolute Loss Values After Ablating(zeroing activations) Different Neurons. When we lower the threshold from an excess correlation of 0.5 to that of 0.4, ablating 5x the amount of random neurons achieve more disruptive results than ablating all universal neurons.

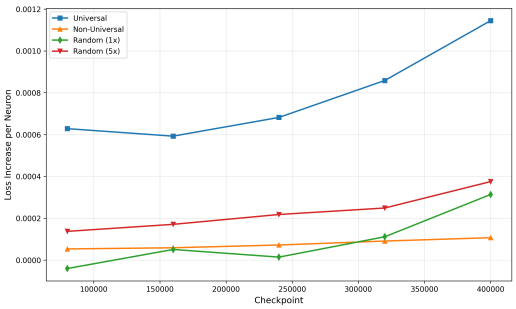

Figure 12: Ablation Efficiency (Change in loss per Neuron). The effect of universal neurons increases along with training steps. Compared to Nonuniversal neurons, they are crucial to the functionality of the language model.

C.2 Ablation Loss Increase with 0.6 Excess Correlation Thresholding

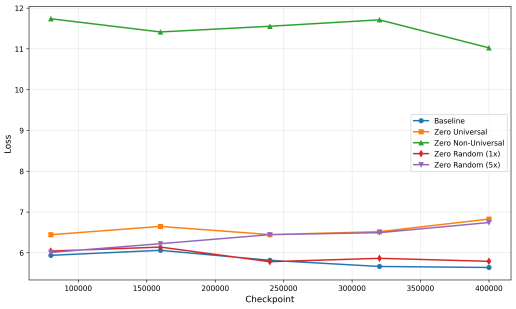

Figure 13: Absolute Loss Values After Ablating(zeroing activations) Different Neurons. It takes around 5x the amount of random neurons to achieve the same disruptive result of ablating universal neurons.

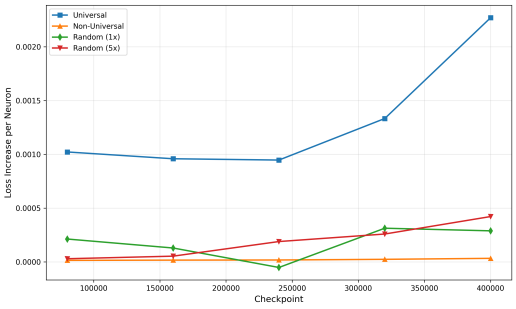

Figure 14: Ablation Efficiency (Change in loss per Neuron). The effect of universal neurons increases along with training steps. Compared to Nonuniversal neurons, they are crucial to the functionality of the language model.

D Additional Layer-Wise Universal Neuron Ablation Experiment Results

As shown in all figures below, the first layer of the model contains universal neurons that are the most crucial.

D.1 Layer-wise Ablation Efficiency Across Checkpoints

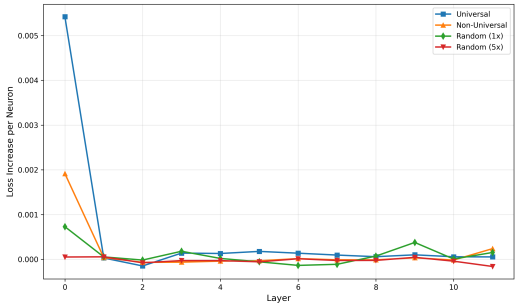

Figure 15: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 80k.

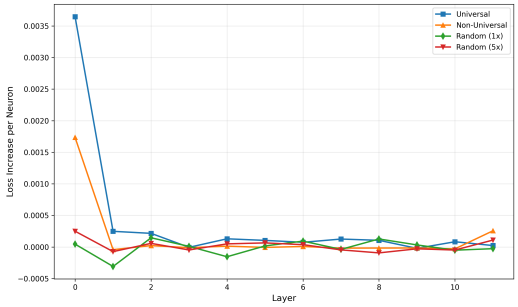

Figure 16: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 160k.

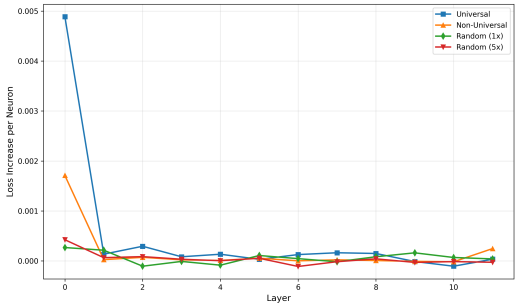

Figure 17: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 240k.

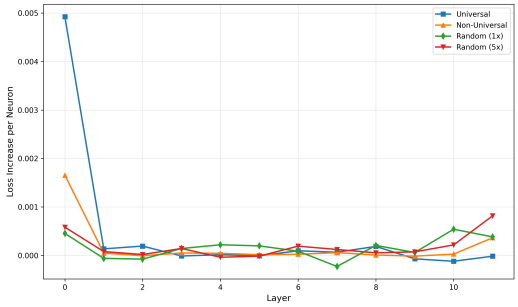

Figure 18: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 320k.

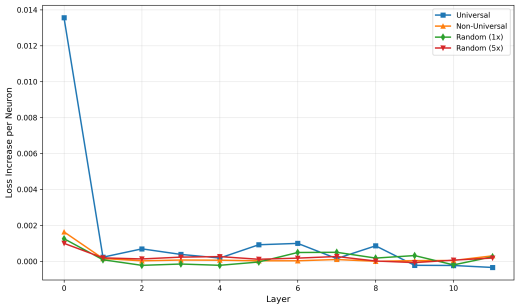

Figure 19: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 400k.

D.2 Layer-wise Ablation Efficiency with 0.4 Excess Correlation Thresholding

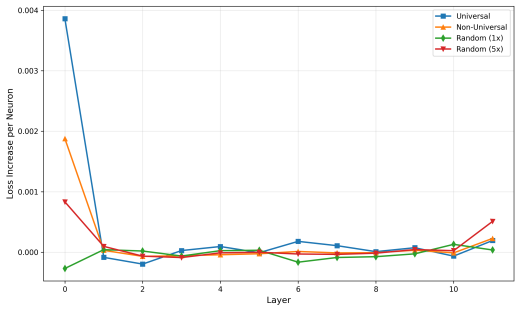

Figure 20: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 80k.

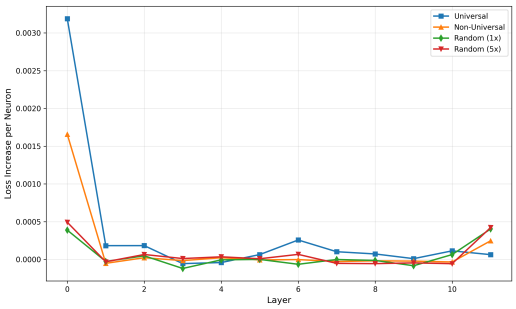

Figure 21: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 160k.

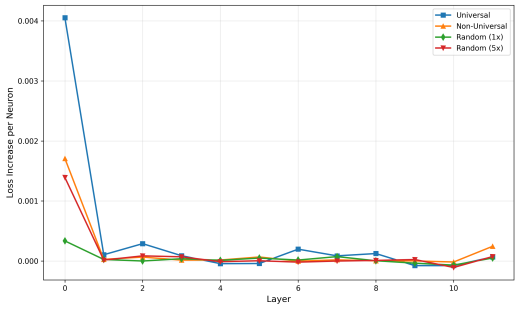

Figure 22: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 240k.

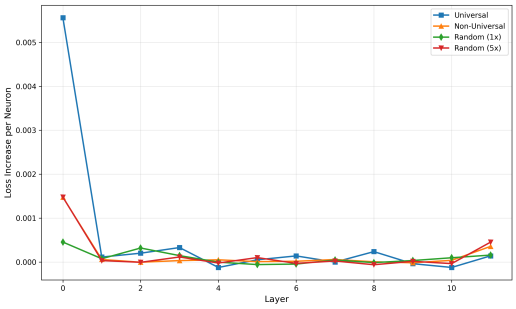

Figure 23: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 320k.

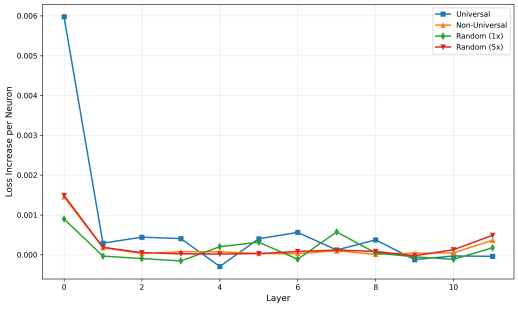

Figure 24: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 400k.

D.3 Layer-wise Ablation Efficiency with 0.6 Excess Correlation Thresholding

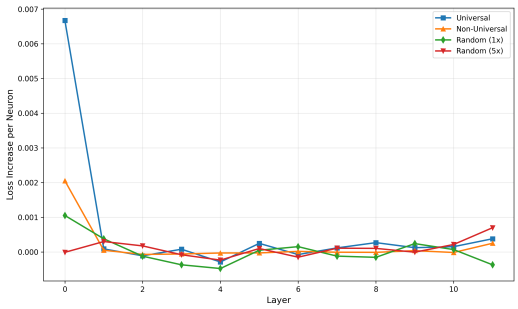

Figure 25: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 80k.

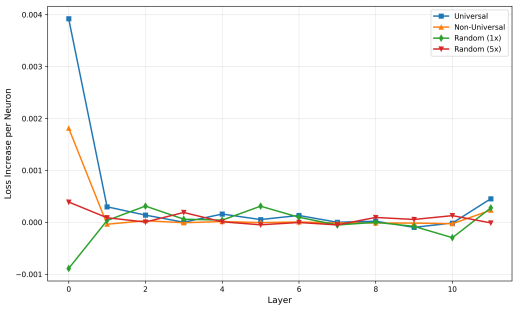

Figure 26: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 160k.

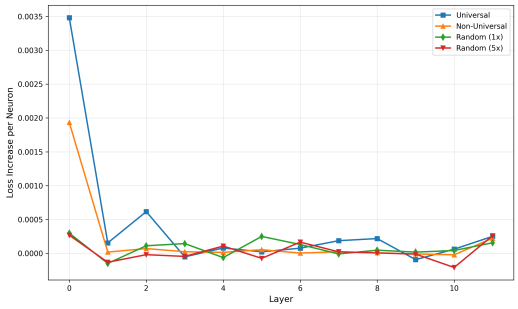

Figure 27: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 240k.

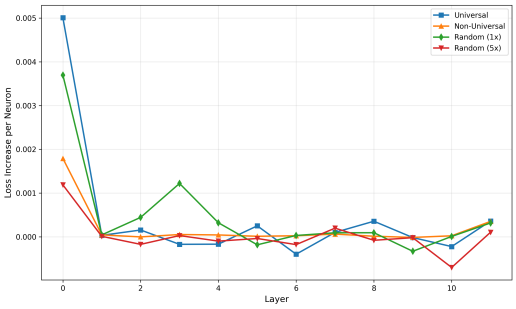

Figure 28: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 320k.

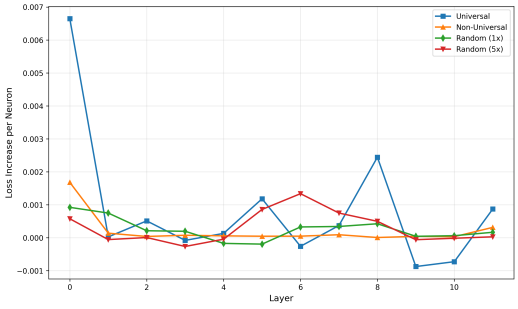

Figure 29: Layer-wise Ablation Efficiency (Change in loss per Neuron) on checkpoint 400k.