Context Copying Modulation: The Role of Entropy Neurons in Managing Parametric and Contextual Knowledge Conflicts

Anonymous ACL submission

Abstract

The behavior of Large Language Models (LLMs) when facing contextual information that conflicts with their internal parametric knowledge is inconsistent, with no generally accepted explanation for the expected outcome distribution. Recent work has identified in autoregressive transformer models a class of neurons – called *entropy neurons* – that produce a significant effect on the model output entropy while having an overall moderate impact on the ranking of the predicted tokens. In this paper, we investigate the preliminary claim that these neurons are involved in inhibiting context copying behavior in transformers by looking at their role in resolving conflicts between contextual and parametric information. We show that *entropy neurons* are responsible for suppressing context copying across a range of LLMs, and that ablating them leads to a significant change in the generation process. These results enhance our understanding of the internal dynamics of LLMs when handling conflicting information.[1]

1 Introduction

Large Language Models (LLMs) exhibit remarkable proficiency in representing, memorizing, and retrieving vast amounts of information. However, they often struggle when discrepancies arise between their learned **parametric knowledge (PK)** and the **contextual knowledge (CK)** provided at inference (Xie et al., 2024; Jin et al., 2024; Xu et al., 2024). These conflicts can lead to unpredictable model behavior, which poses a significant challenge in real-world applications of LLMs (Ji et al., 2023). Although various strategies have been proposed to mitigate this unpredictable behavior (Shi et al., 2024), the mechanisms that govern how LLMs prioritize and integrate different sources of knowledge are poorly understood. We investigate the preliminary claim that the recently discovered *entropy neurons* (Katz and Belinkov, 2023; Gurnee et al., 2024) are involved in inhibiting context copying behavior (Stolfo et al., 2024) by looking at their role in resolving conflicts between CK and PK. These neurons are known to regulate model entropy while having an overall moderate impact on the ranking of the predicted tokens. We make the following key findings and contributions:

- Entropy neurons, although representing less than 2‰ of the feed forward network neurons in the last transformer layer, play a significant role in determining the knowledge source to use. More specifically, they inhibit the natural LLM's behavior of repeating the sequences in the context, i.e. induction (Olsson et al., 2022).
- We identify the presence of entropy neurons in a range of models, from 1 billion to 8 billion parameters, including Pythia-1.4B, Phi-1.5, Mistral-7B-v0.1, and Llama-3-8B[2] and give some insights on their characteristics.

2 Related Work

The understanding of the mechanisms and knowledge localization within transformers has advanced through various studies. One line of research has focused on the PK-based outputs, particularly in factual settings (Geva et al., 2021; Heinzerling and Inui, 2021; AlKhamissi et al., 2022; Meng et al., 2023; Geva et al., 2023). These studies hypothesized that LLMs store parametric information within the Feed Forward Network (FFN) layers, which function as a key-value memory. This stored information is subsequently accessed by the Multi-Head Self-Attention (MHSA) layers. Another body

[1] We make our code and data publicly available at: https://anonymous.4open.science/r/Context-Copying-Modulation-DDD4

[2] In the main paper we show results for Phi-1.5, we provide the results for other models in the Appendix.

Input Prompt	Before (PK)	After (CK)
Kentucky's official language is Japanese. Kentucky's official language is	English	Japanese
Antonio Moreno communicates in English. Antonio Moreno communicates in	Spanish	English
Mac OS X Panther is a product released by Google. Mac OS X Panther is a product released by	Apple	Google

Table 1: Examples where Phi-1.5 switched from using **Parametric Knowledge (PK)** to **Contextual Knowledge (CK)** after ablating entropy neurons.

of work has examined CK-based outputs. These studies concluded that the processing of CK, unlike PK, is not localized within the LLM's parameters (Monea et al., 2024). Instead, it is facilitated by a learned mechanism known as induction, which underpins in-context learning and information copying (Olsson et al., 2022). Despite these advancements, the mechanisms underlying how LLMs regulate the CK usage in a situation of induction are not well understood.

3 Background

3.1 Feed Forward Network (FFN)

The structure of the Transformer's FFN is central to our study (Vaswani et al., 2017). Given a hidden state $x \in \mathbb{R}^{d_{model}}$ from the residual stream after the MHSA module, the FFN is defined as:

$$\mathrm{FFN}(\mathbf{x}) = \sum_i w_{\text{out}}^{(i)} \sigma\left(w_{\text{in}}^{(i)} \cdot x + \beta_{\text{in}}^{(i)}\right) + \beta_{\text{out}}, \quad (1)$$

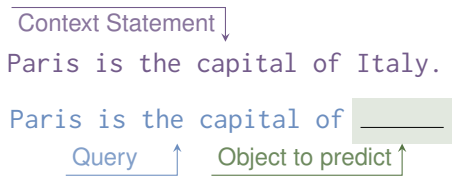
where $\mathbf{W}_{\text{out}}^T, \mathbf{W}_{\text{in}} \in \mathbb{R}^{d_{\text{ffn}} \times d_{\text{model}}}$ are learned weight matrices, $\boldsymbol{\beta}_{\text{in}}$ and $\boldsymbol{\beta}_{\text{out}}$ are learned biases. The function σ denotes an element-wise nonlinear activation function, e.g. ReLU (Agarap, 2019).

A neuron from the first FFN layer is characterized by 1) an activation value noted $n_i \in \mathbb{R}$ (i.e. the output of the activation function σ) and 2) an output weight vector $w_{\text{out}}^{(i)} \in \mathbb{R}^{d_{model}}$.

3.2 Framework and Dataset

We use the knowledge probing framework (Tighidet et al., 2024), which consists of a dataset of prompts that are built to contradict the internal knowledge (i.e. PK) of a given model. It follows a well-structured format based on repetition, which makes it convenient for PK/CK analysis. A similar framework is proposed by Yu et al. (2023) but it consists of prompts in form of questions rather than repetitive sequences which is less convenient to study induction. We provide characteristics about the dataset in Appendix E.

Each prompt x from the knowledge probing dataset E consists of a contextual statement about a subject s (e.g., *"Paris"*), a relation r (e.g., *"capital of"*), and an object $\bar{o}$ that contradicts the model's internal PK (e.g., *"Italy"*). The contextual statement is from the CK that is defined below. This is followed by a repetitive query about s to trigger the model's induction mechanism. For example:



If the model responds according to the context statement, it uses CK (e.g. *"Italy"*). If it responds based on its learned knowledge, it uses PK (e.g. *"France"*). If it outputs neither, the knowledge source is not defined (ND, e.g. *"Spain"*).

Parametric Knowledge (PK). PK is the information the model learned during training, represented as triplets (s, r, o) where o is the generated object given a query with a subject s and a relation r (e.g., Query: *"Paris is the capital of"* $\rightarrow$ Model answer: *"France"*).

Context Knowledge (CK). CK is the information that is contradictory to PK. This involves replacing o with another object $\bar{o}$ that shares the same relation r (e.g., *"Paris is the capital of Italy"*). For each (s, r) couple, three $\bar{o}$ objects are selected, namely those with the lowest probability. This selection method ensures the model did not learn the $(s, r, \bar{o})$ association from its training data.

Not Defined Knowledge (ND). ND includes all objects not in PK or CK.

4 Entropy Neurons

4.1 Motivation

Gurnee et al. (2024) and Stolfo et al. (2024) identified entropy neurons in GPT-2 by considering the 6

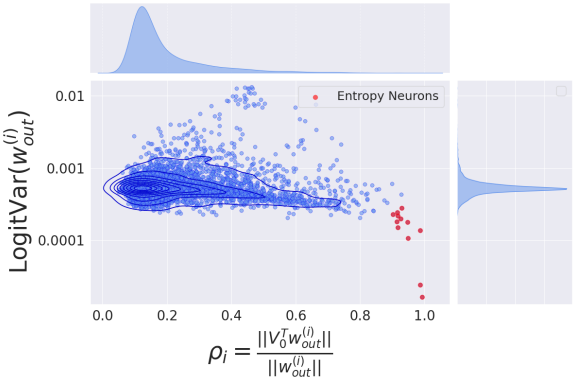


Figure 1: **Selected entropy neurons for Phi-1.5 (red).**

neurons with the lowest impact on logits variance using the LogitVar measure, defined in Eq. 2 and questioned their high weight norm. Stolfo et al. (2024) characterize entropy neurons as those that write into the effective null space of the unembedding matrix $\mathbf{W}_{\mathrm{U}} \in \mathbb{R}^{V \times d_{model}}$, as measured by ρ (Eq. 3).

LogitVar. This measure quantifies a neuron's direct effect on output logits variance. For a neuron i, it is defined as:

$$\mathrm{LogitVar}(w_{\mathrm{out}}^{(i)}) = \mathbf{Var}\left\{ \frac{w_{\mathrm{U}}^{(t)} \cdot w_{\mathrm{out}}^{(i)}}{||w_{\mathrm{U}}^{(t)}|| \times ||w_{\mathrm{out}}^{(i)}||} ; t \in V \right\} \quad (2)$$

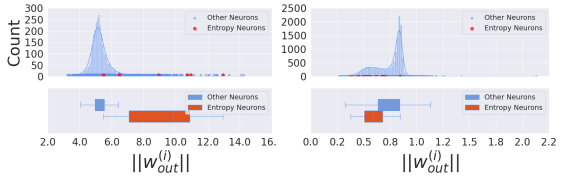
where V is the set of tokens in the vocabulary and $w_{\mathrm{U}}^{(t)}$ the t^{th} row of $\mathbf{W}_{\mathrm{U}}$.

Effective Null Space Projection (ρ). This measure quantifies how much of a neuron's output aligns with directions that minimally impact the model's final output, forming the effective null space of the unembedding matrix $\mathbf{W}_{\mathrm{U}}$, denoted as $\mathbf{V}_0$. Details on identifying $\mathbf{V}_0$ are in Appendix F. For a neuron i, it is defined as:

$$\rho_i = \frac{||\mathbf{V}_0^{\mathrm{T}} w_{\mathrm{out}}^{(i)}||}{||w_{\mathrm{out}}^{(i)}||}. \quad (3)$$

4.2 Entropy Neurons Selection

We focus on the last Transformer layer because its entropy neurons have the most direct impact on the term logit distribution (through the projection with the unembedding matrix $\mathbf{W}_{\mathrm{U}}$). We use both LogitVar and ρ (motivated by previous work on effective null space projections (Stolfo et al., 2024)) to select these neurons. Figure 1 illustrates all the neurons with their corresponding LogitVar and ρ for Phi-1.5, with similar figures for other models in Figure 6 in the Appendix. We select neurons with minimal logit variance impact (LogitVar) and high projection with $\mathbf{W}_{\mathrm{U}}$'s effective null space (ρ). For Phi-1.5, we select 12 entropy neurons, representing 1.5‰ of the last layer's neurons, using a minimalist approach to pick the fewest neurons with strong characteristics. Table 4 in the Appendix details hidden dimensions and selected entropy neuron proportions for all models.



(a) GPT-2 (b) Llama-3-8B

Figure 2: Weight norm distribution for entropy neurons vs. other neurons for GPT-2 and Llama-3-8B.

Although Gurnee et al. (2024) and Stolfo et al. (2024) observed high weight norm $||w_{\mathrm{out}}^{(\mathrm{i})}||$ for entropy neurons (e.g., GPT-2, Figure 2a) and used it to select entropy neurons, we do not use high weight norm as a selection criterion. We observe that for some models, neurons with low LogitVar$(w_{\mathrm{out}}^{(\mathrm{i})})$ and high ρ_i can have relatively low $||w_{\mathrm{out}}^{(\mathrm{i})}||$ compared to other neurons, as illustrated in Figure 2b for Llama-3-8B. Therefore, we consider LogitVar and ρ as the crucial selection criteria.

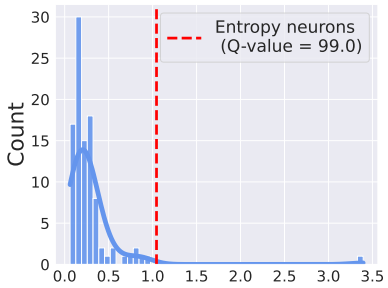
5 Mechanistic Study

We present the metrics in 5.1, and describe our results in 5.2.

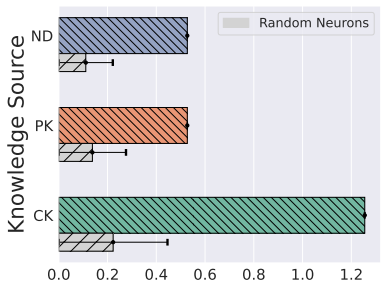
5.1 Metrics

We measure the impact of a set of neurons $\mathcal{N}$ on the context copying behavior by turning off these neurons, through causal interventions, and observing how the knowledge source (CK, PK or ND) changes (see Section 3.2). In practice, we turn off these neurons by replacing their activation values n_i by an average value μ_{n_i} computed over the knowledge probing dataset E[3]. More formally, for each example $x \in E$ (see Section 3.2), $K_{\mathcal{M}}(x)$ is the knowledge source used by the model $\mathcal{M}$, and $K_{\mathcal{M}\setminus\mathcal{N}}(x)$ is the knowledge source used by the ablated model $\mathcal{M}^{\setminus\mathcal{N}}$ given the input x. Let $E_K = \{x \in E | K_{\mathcal{M}}(x) = K\}$ and $E_{\bar{K}} = E \setminus E_K$. We define the following metrics:

[3]We also tested other ablation values and show their Global Transition Score in Table 6 in the Appendix.



(a) **Global Transition Score (%)**



(b) **Conversion Ratio (%)**

	To CK	To ND	To PK
From CK	99.5 (99.8 ± 0.1)	0.2 (0.1 ± 0.0)	0.3 (0.1 ± 0.1)
From ND	2.5 (0.3 ± 0.1)	96.4 (99.5 ± 0.1)	1.1 (0.2 ± 0.1)
From PK	0.4 (0.2 ± 0.1)	0.1 (0.1 ± 0.1)	99.5 (99.7 ± 0.1)

(c) **Transition Scores (%)**

Figure 3: **Phi-1.5 ablation scores.** As a control, we provide the average Transition Score of 100 random ablations with its corresponding error bars (±3×standard deviation). We also provide the error bars for the entropy neurons in Figure 3b illustrated on top of the CK, PK, and ND bars.

Global Transition Score (GTS): proportion of examples for which the knowledge source changes as we remove the group of neurons $\mathcal{N}$

$$\text{GTS} = \frac{1}{|E|} \sum_{x \in E} \mathbb{I}[K_{\mathcal{M}}(x) \neq K_{\mathcal{M} \setminus \mathcal{N}}(x)], \quad (4)$$

where $\mathbb{I}$ is the indicator function, equal to 1 if the condition is true and 0 otherwise, and $|E|$ is the cardinality of E. *A high GTS indicates that ablating $\mathcal{N}$ significantly alters the model's knowledge source selection*, underscoring the role of $\mathcal{N}$ in knowledge source decision-making.

Conversion Ratio (CR): proportion of examples where the model switched *to* a given knowledge source $K \in$ (PK, CK, ND) when we remove $\mathcal{N}$

$$\text{CR}(K) = \frac{1}{|E_{\bar{K}}|} \sum_{x \in E_{\bar{K}}} \mathbb{I}[K_{\mathcal{M} \setminus \mathcal{N}}(x) = K] \quad (5)$$

a high $\text{CR}(K)$ suggests that ablating $\mathcal{N}$ alters a large proportion of examples towards K, indicating that *$\mathcal{N}$ is an inhibitor of the knowledge source K in the original model $\mathcal{M}$.*

Transition Score (TS): proportion of examples that transition from knowledge source K to knowledge source K' as we remove $\mathcal{N}$

$$\text{TS}(K, K') = \frac{1}{|E_K|} \sum_{x \in E_K} \mathbb{I}[K_{\mathcal{M} \setminus \mathcal{N}}(x) = K'], \quad (6)$$

a high $\text{TS}(K, K')$ indicates that ablating $\mathcal{N}$ moves a large portion of examples with knowledge source K to knowledge source K', *suggesting that the entropy neurons $\mathcal{N}$ tend to promote K over K'.*

5.2 Results

Control Distribution: to assess the significance of the results on entropy neurons $\mathcal{E}$, we build a control distribution by drawing 100 independent sets of neurons from the set of non-entropy neurons with the same cardinality as $\mathcal{E}$.

Entropy neurons significantly influence the knowledge source of predictions. We investigated the impact of removing entropy neurons on knowledge source transitions (CK, PK, ND) across various models. Figure 3a illustrates that ablating entropy neurons $\mathcal{E}$ results in a Global Transition Score (GTS) at the top 1% of the control distribution for Phi-1.5. This suggests that entropy neurons play a significant role in the decision-making process regarding knowledge sources. This observation holds true for other models (see Figure 9).

Entropy neurons inhibit the induction mechanism. After demonstrating that removing entropy neurons triggers transitions between knowledge sources, we further analyzed the destination of these transitions using the Conversion Ratio ($\text{CR}(K)$). Figure 3b for Phi-1.5, show a high CR for CK compared to the control distribution, indicating a significant shift from PK and ND to CK (highlighted in green) after ablating $\mathcal{E}$. This finding is corroborated by the Transition Scores presented in Table 3c for Phi-1.5 (2.5%) and in Table 5 (Appendix) for Llama-3-8B (6.2%), GPT-2 (3.3%), and Pythia-1.4B (2%). We show in Table 1 examples where Phi-1.5 switched from using PK to CK.

6 Conclusion

In this paper, we demonstrated that entropy neurons play a significant role in modulating the balance between PK and CK. Ablation studies revealed that perturbing these neurons leads to significant shifts in the knowledge source used by the model. Specifically, the GTS for entropy neurons is at the top 1% of the control distribution, this finding is consistent for different models up to 8B parameters. Our study elucidates mechanisms regulating induction, providing insights for future research in this area.

7 Limitations

While our experiments demonstrate that entropy neurons significantly inhibit context copying behavior in large language models, our study is limited by an incomplete understanding of the broader copying regulation mechanism. Specifically, we focused solely on entropy neurons in the FFN of the final transformer layer, which may neglect the contributions of other neuron types and architectural components in regulating the balance between contextual and parametric knowledge.

Additionally, the ablation techniques used – replacing neuron activations with values such as the mean – may not fully capture the nuanced interactions within the network, and our findings are based on a limited set of models and dataset.

Although we observed relatively high Global Transition Scores in most of the models we studied, their Q-values varies. For instance, in Phi-1.5, Llama-3-8B, and GPT-2 the Q-value is around 99 which is less for Mistral-7B-v0.1 and Pythia-1.4B with 91 and 92.5 respectively. Model architecture and training could explain this variation.

Lastly, our study focuses on how entropy neurons contribute to modulating the balance between parametric and contextual knowledge in a situation of induction and does not explore why this specific set of neurons act this way.

Future research should therefore expand the investigation to include a wider array of neural components and alternative perturbation methods to more comprehensively elucidate the underlying processes governing copying regulation. Additionally, it should explore the reasons why entropy neurons specifically contribute to modulating the balance between parametric and contextual knowledge in situations of induction.

8 Ethical Considerations

Our study probes the internal mechanisms of large language models (LLMs) by manipulating a small subset of neurons—entropy neurons—that modulate the balance between parametric and contextual knowledge. All experimental data and prompts are derived from publicly available sources minimizing any direct privacy or security concerns.

However, we acknowledge that our findings have some implications. The probing and ablation techniques we describe could be repurposed to intentionally bias or subvert LLM behavior. Specifically, the structured prompts we employ to induce context copying may serve as templates for adversarial attacks, allowing malicious actors to manipulate model outputs in subtle but impactful ways. Similarly, our demonstration that targeted neuron ablation alters a model's decision-making process raises the risk that LLMs could be engineered—intentionally or inadvertently—to prioritize deceptive or harmful outputs.

Given these risks, we stress the importance of applying this work within responsible and well-governed research contexts. We urge future researchers to incorporate safeguards against misuse, including robust evaluation pipelines and transparency in experimental intent. To foster reproducibility and critical engagement, we have released our codebase under an open license while documenting the limitations of our approach.

References

Abien Fred Agarap. 2019. Deep learning using rectified linear units (relu). *Preprint*, arXiv:1803.08375.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *Preprint*, arXiv:2204.06031.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *Preprint*, arXiv:2304.14767.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. *Preprint*, arXiv:2012.14913.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in GPT2 language models. *Transactions on Machine Learning Research*.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the

conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *Preprint*, arXiv:2402.18154.

Shahar Katz and Yonatan Belinkov. 2023. VISIT: Visualizing and interpreting the semantic information flow of transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14094–14113, Singapore. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kıcıman, Hamid Palangi, Barun Patra, and Robert West. 2024. A glitch in the matrix? locating and detecting language model grounding with fakepedia. *Preprint*, arXiv:2312.02073.

Neel Nanda and Joseph Bloom. 2022. Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. In *Advances in Neural Information Processing Systems*, volume 37, pages 4997–5024. Curran Associates, Inc.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. 2024. Confidence regulation neurons in language models. *Preprint*, arXiv:2406.16254.

Zineddine Tighidet, Andrea Mogini, Jiali Mei, Benjamin Piwowarski, and Patrick Gallinari. 2024. Probing language models on their knowledge source. *Preprint*, arXiv:2410.05817.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *Preprint*, arXiv:2305.13300.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *Preprint*, arXiv:2403.08319.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

A Hardware and Software

Experiments were performed using NVIDIA H100 and A100 GPUs, each equiped with 80 GB of VRAM. The process of generating the outputs with and without ablations took around 250 GPU hours. Our codebase was built using PyTorch (Paszke et al., 2019), the HuggingFace Transformers library (Wolf et al., 2020) the TransformerLens library (Nanda and Bloom, 2022), and the knowledge probing framework (Tighidet et al., 2024).

B License

Llama3-8B weights are released under the license available at `https://llama.meta.com/llama3/license/`. Mistral-7B and Pythia-1.4B weights are released under an Apache 2.0 license. Phi-1.5 and GPT-2 weights are released under a MIT license.

C Weight Pre-processing

To eliminate irrelevant components and other parameterization degrees of freedom, we utilize a set of standard weights pre-processing techniques following Nanda and Bloom (2022) and Stolfo et al. (2024).

Incorporating Layer Norm. Most layer norm implementations include trainable parameters $\gamma \in$

$\mathbb{R}^n$ and $\beta \in \mathbb{R}^n$. To account for these, we "fold" the layer norm parameters into $\mathbf{W}_{\text{in}}$ by treating the layer norm parameters as equivalent to a linear layer and then combining the adjacent linear layers. We create effective weights as follows:

$$\mathbf{W}_{\text{eff}} = \mathbf{W}_{\text{in}} \cdot \text{diag}(\gamma), \quad \beta_{\text{eff}} = \beta_{\text{in}} + \mathbf{W}_{\text{in}} \cdot \beta \tag{7}$$

Finally, we center the reading weights because the preceding layer norm projects out the all-ones vector. Thus, we center the weights $\mathbf{W}_{\text{eff}}$ as follows:

$$\mathbf{W}'_{\text{eff}}(i,:) = \mathbf{W}_{\text{eff}}(i,:) - \bar{\mathbf{W}}_{\text{eff}}(i,:). \tag{8}$$

Centering Writing Weights. Every time the model interacts with the residual stream, it applies a LayerNorm first. Therefore, the components of $\mathbf{W}_{\text{out}}$ and β_{out} that lie along the all-ones direction of the residual stream have no effect on the model's calculations. Consequently, we mean-center $\mathbf{W}_{\text{out}}$ and β_{out} by subtracting the means of the columns of $\mathbf{W}_{\text{out}}$:

$$\mathbf{W}'_{\text{out}} = \mathbf{W}_{\text{out}}(:,i) - \bar{\mathbf{W}}_{\text{out}}(:,i). \tag{9}$$

Centering Unembedding. Since softmax is translation invariant, we also center $\mathbf{W}_{\text{U}}$:

$$\mathbf{W}'_{\text{U}}(:,i) = \mathbf{W}_{\text{U}}(:,i) - \bar{\mathbf{W}}_{\text{U}}(:,i) \tag{10}$$

D Activations

Model	Activation Function	Domain
Llama-3-8B	SwiGLU: Swish × GLU	$\mathbb{R}$
Mistral-7B-V0.1	SwiGLU: Swish × GLU	$\mathbb{R}$
Phi-1.5	GELU	$\mathbb{R}$
Pythia-1.4B	GELU	$\mathbb{R}$
GPT-2-Small	GELU	$\mathbb{R}$

Table 2: FFN hidden layer activation functions for all the studied models

E Data Characteristics

We provide in Figure 4 the count of used knowledge sources by model before ablating entropy neurons. We also provide in Table 3 a sample of examples from the knowledge probing dataset.

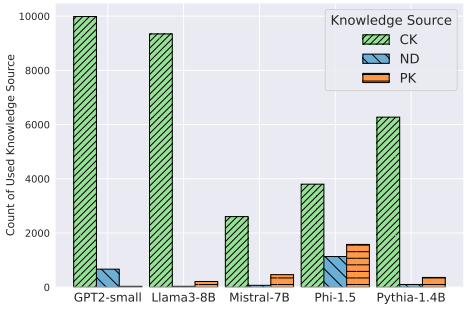


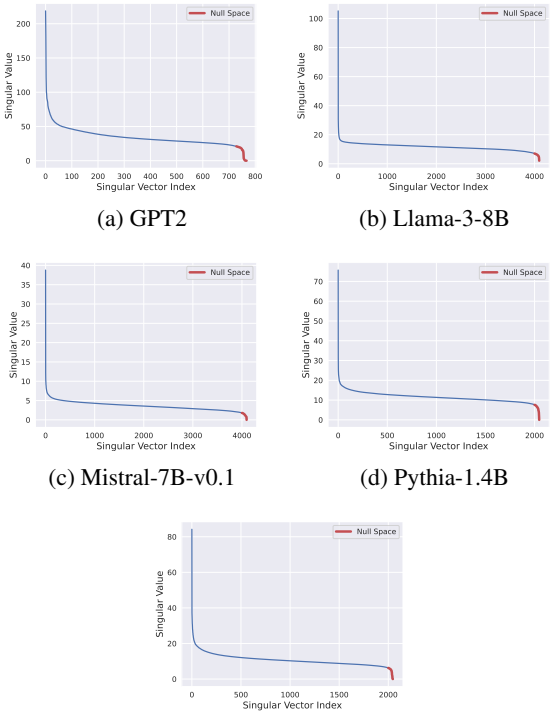
Figure 4: **Count of used knowledge sources by each model before ablation.**

F $\mathbf{W}_{\text{U}}$'s Effective Null Space

To identify the effective null space $\mathbf{V_0}$ of $\mathbf{W_U}$, we start by applying a singular value decomposition (**SVD**) on $\mathbf{W_U}$:

$$\textbf{SVD}(\mathbf{W_U}) = \mathbf{U}\Sigma\mathbf{V}^{\mathbf{T}}, \tag{11}$$

we then consider the right singular vectors with the lowest singular values, noted $\mathbf{V_0}$, starting from a sharp drop as shown in Figure 5. We also detail the effective null space dimension size for all the studied models in Table 4.



(a) GPT2 (b) Llama-3-8B

(c) Mistral-7B-v0.1 (d) Pythia-1.4B

(e) Phi-1.5

Figure 5: Unembedding matrix $\mathbf{W_U}$ singular values, illustrating the effective null space of $\mathbf{W_U}$ in red.

Input Prompt	Knowledge Source	PK Attribute	Language Model
Harney County has its capital city in Taiwan. Harney County has its capital city in ***Burns.***	ND	Oregon	Llama3-8B
Lisa Appignanesi has citizenship of Finland. Lisa Appignanesi has citizenship of ***France.***	ND	the UK	Llama3-8B
Craiova is located in the continent of India. Craiova is located in the continent of ***Romania.***	ND	*Europe*	Pythia-1.4B
The Kingdom of Hungary had its capital as Connecticut. The Kingdom of Hungary had its capital as ***Connecticut.***	CK	*Budapest*	Mistral-7B
The Wii U system software is a product that was manufactured by Square. The Wii U system software is a product that was manufactured by ***Square.***	CK	*Nintendo*	Llama3-8B
The Centers for Disease Control and Prevention is headquartered in Lyon. The Centers for Disease Control and Prevention is headquartered in ***Lyon.***	CK	*Atlanta*	Llama3-8B
Harare is the capital city of Florida. Harare is the capital city of ***Zimbabwe.***	PK	*Zimbabwe*	Pythia-1.4B
Goodreads is owned by Microsoft. Goodreads is owned by ***Amazon.***	PK	*Amazon*	Phi-1.5
OneDrive is owned by Toyota. OneDrive is owned by ***Microsoft.***	PK	*Microsoft*	Mistral-7B

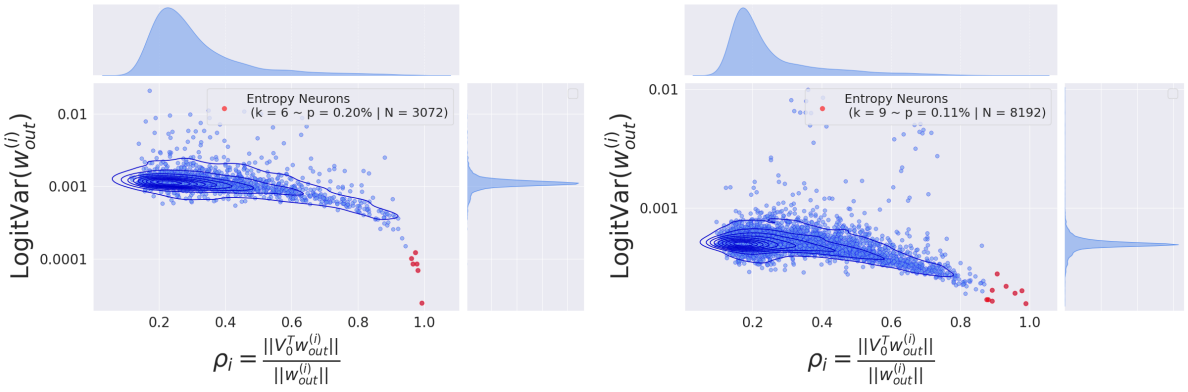
Table 3: Examples of final probing prompts, including their knowledge source, the LLM, and the corresponding parametric knowledge (PK) object. Bold text indicates the generated attribute, while underlined text represents the counter-knowledge attribute.

Model	d_{model}	d_{ffn}	$d_{\text{effective null space}}$	**Card**(V)	$\frac{d_{\text{effective null space}}}{d_{model}}$ **(%)**	Entropy Neurons (‰)
GPT-2	768	3072	40	50257	5.20	2
Llama-3-8B	4096	14336	96	128256	2.34	0.7
Mistral-7B-v0.1	4096	14336	96	32000	2.34	1
Pythia-1.4B	2048	8192	48	50304	2.34	1.1
Phi-1.5	2048	8192	48	51200	2.34	1.5

Table 4: Models hidden dimensions compared to the proportion of selected entropy neurons.

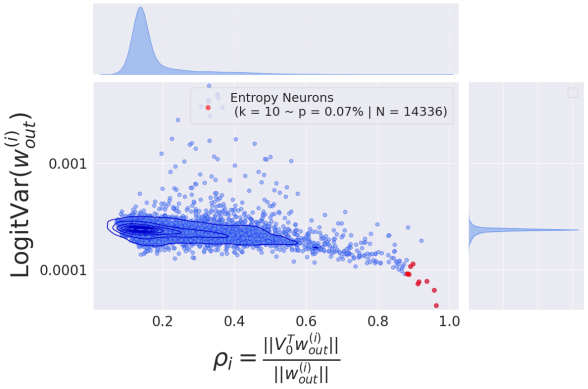
Model Name	**From CK**			**From ND**			**From PK**		
	To CK	**To ND**	**To PK**	**To CK**	**To ND**	**To PK**	**To CK**	**To ND**	**To PK**
GPT-2	100.0 (100.0 ± 0.0)	0.0 (0.0 ± 0.0)	0.0 (0.0 ± 0.0)	3.3 (0.4 ± 0.1)	96.4 (99.6 ± 0.1)	0.3 (0.0 ± 0.0)	0.0 (1.2 ± 0.6)	6.2 (2.6 ± 0.8)	93.8 (96.3 ± 1.0)
Mistral-7B	99.8 (99.9 ± 0.0)	0.0 (0.0 ± 0.0)	0.2 (0.1 ± 0.0)	0.0 (0.3 ± 0.3)	98.6 (99.3 ± 0.5)	1.4 (0.4 ± 0.3)	2.2 (0.6 ± 0.2)	0.2 (0.0 ± 0.0)	97.6 (99.4 ± 0.2)
Llama3-8B	99.6 (100.0 ± 0.0)	0.1 (0.0 ± 0.0)	0.4 (0.0 ± 0.0)	6.2 (0.2 ± 0.3)	90.6 (99.7 ± 0.4)	3.1 (0.1 ± 0.2)	0.5 (0.9 ± 0.3)	0.5 (0.0 ± 0.0)	99.1 (99.1 ± 0.3)
Pythia-1.4B	99.9 (100.0 ± 0.0)	0.0 (0.0 ± 0.0)	0.1 (0.0 ± 0.0)	2.0 (0.7 ± 0.2)	98.0 (99.3 ± 0.3)	0.0 (0.0 ± 0.1)	0.0 (0.3 ± 0.1)	0.0 (0.0 ± 0.0)	100.0 (99.7 ± 0.1)

Table 5: Transition Scores (%) From source To target knowledge source after mean ablating entropy neurons across models. As a control, we provide the average Transition Score of 100 random ablations with its corresponding error bars ($\pm 3\sigma$).

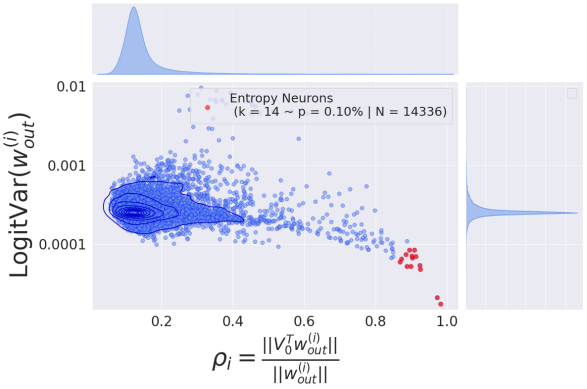


(a) GPT2

(b) Pythia-1.4B



(c) Llama-3-8B

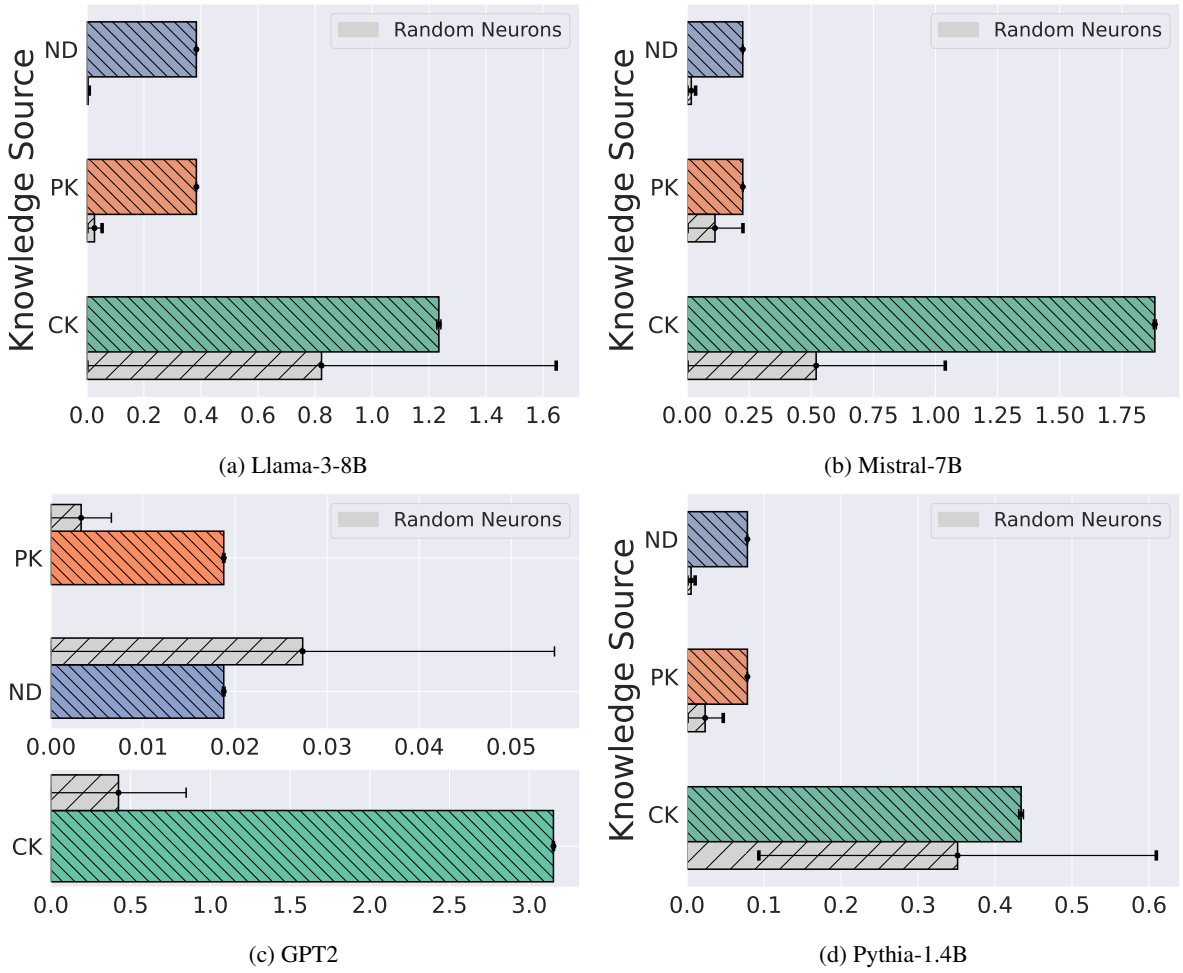


(d) Mistral-7B

Figure 6: **Selected entropy neurons (red).** We select entropy neurons following the LogitVar and ρ criteria. In each Figure, k is the number of selected entropy neurons, p is the proportions of entropy neurons, and N is the total number of neurons.

Ablation Value	Model	EN Transition Score (%)	Q-val
μ_{n_i}	GPT-2	0.3	98.0
	Pythia-1.4B	0.1	92.5
	Mistral-7B-v0.1	0.5	91.0
	Phi-1.5	1.0	99.0
	Llama3-8B	0.5	99.0
$\max(\mu_{n_i} - 3\sigma_{n_i}, min_{n_i})$	GPT-2	0.5	100.0
	Pythia-1.4B	0.1	96.5
	Mistral-7B-v0.1	11.1	99.0
	Phi-1.5	1.2	99.0
	Llama3-8B	0.9	87.0
$\min(\mu_{n_i} + 3\sigma_{n_i}, max_{n_i})$	GPT-2	7.8	99.0
	Pythia-1.4B	1.5	100.0
	Mistral-7B-v0.1	2.3	84.0
	Phi-1.5	1.0	95.0
	Llama3-8B	99.5	99.0
Median_{n_i}	GPT-2	0.2	99.0
	Pythia-1.4B	0.1	74.5
	Mistral-7B-v0.1	0.5	92.0
	Phi-1.5	1.1	99.0
	Llama3-8B	0.1	84.0
Mode_{n_i}	GPT-2	93.8	100.0
	Pythia-1.4B	0.1	68.5
	Mistral-7B-v0.1	0.5	87.0
	Phi-1.5	1.3	98.0
	Llama3-8B	0.1	60.5

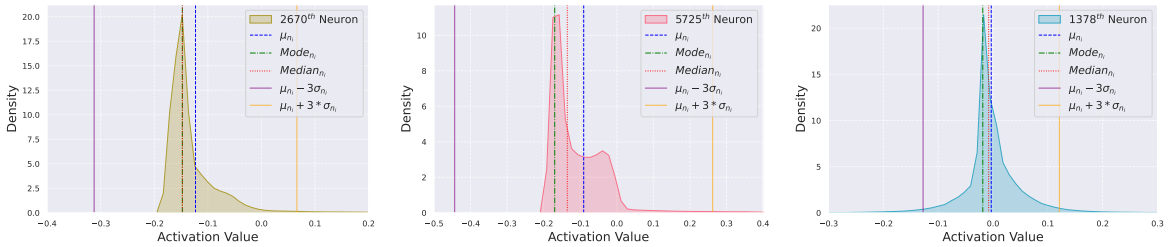
Table 6: Ablation value-wise Global Transition Scores (%) for entropy neurons ablation. The ablation values are computed over the knowledge probing dataset for each neuron activation distribution n_i as illustrated in Figure 8. Specifically they consist of: the mean μ_{n_i}, the mode Mode_{n_i}, the median Median_{n_i}, and two extreme values $\min(\mu_{n_i} + 3\sigma_{n_i}, max_{n_i})$, $\max(\mu_{n_i} - 3\sigma_{n_i}, min_{n_i})$ where σ_{n_i} is the standard deviation. For the extreme values, we make sure to take the $\min_{n_i}$/$\max_{n_i}$ when $\mu_{n_i} \pm 3\sigma_{n_i}$ is out of distribution.



(a) Llama-3-8B (b) Mistral-7B

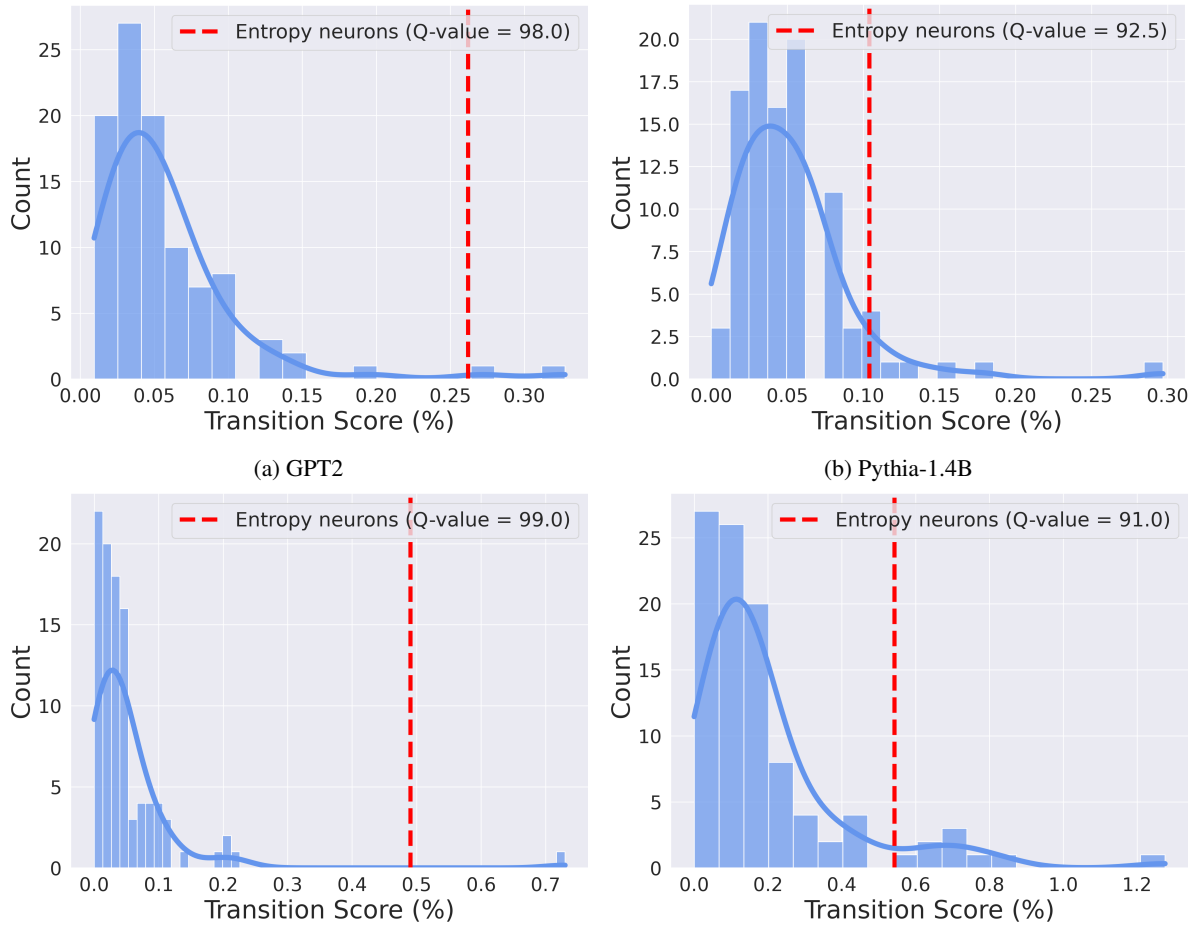
(c) GPT2 (d) Pythia-1.4B

Figure 7: **Conversion Ratio (%)**



(a) GPT2 (GELU) (b) Phi-1.5 (GLU) (c) Llama-3-8B (SwiGLU)

Figure 8: Example of neurons distribution for each model as well as the ablation values. The Neuron where randomly selected for each model and the distribution was estimated based on the knowledge probing dataset (Tighidet et al., 2024).



(a) GPT2

(b) Pythia-1.4B

(c) Llama-3-8B

(d) Mistral-7B-v0.1

Figure 9: Global Transition Scores, ablating entropy neurons exhibit a higher transition in the used knowledge sources compared to 100 sets of random neurons which indicates the unique property of entropy neurons to affect the knowledge source to select.