Neuron Level Analysis of Large Language Model in Legal Domain Reasoning

Anonymous ACL submission

Abstract

We presented a neuron-level analysis of legal-domain reasoning in LLMs, comparing it with other applied domain tasks across seven open-weight models. Using neuron attribution scores to rank and suppress influential neurons, we confirmed that suppressing the identified neurons collapses accuracy on the target task, whereas suppressing the same number of random neurons does not. We further found a small subset of neurons influential across all seven tasks; once these are removed, suppressing the remaining neurons degrades only the task they were identified from, revealing genuinely task-specific neurons in every model studied. Within the legal domain, the three benchmarks exhibit relatively high neuron overlap and tend to be affected jointly, suggesting of legal components neurons that span jurisdictions. The distribution of identified neurons in our experiments suggests that the hypothesis that influential neurons are concentrated in middle MLP layers may depend on the input format and content, rather than being a universal phenomenon.

1 Introduction

With the rapid advancement of large language models (LLMs), general tasks such as everyday conversation, translation, and coding have reached a practically usable level. In contrast, performance in applied domains such as law, mathematics, medicine remains insufficient (Cobbe et al., 2021; Jin et al., 2021; Fan et al., 2025; Li et al., 2024). In the legal domain, general open-source LLMs have not yet reached a passing level on multiple-choice questions of bar examinations (Cao et al., 2025), indicating the need for additional adaptation. Although the development of sufficiently large datasets is essential for improving the performance of LLMs, it is difficult to construct complete datasets for application domain tasks because of information protection, personal privacy protection, and the protection of trade secrets in legal domain. Also, model may face unprecedented legal cases that both parties to the dispute have rights worthy of protection. As a result, it is not realistic to develop datasets covering all categories of problems that LLMs are expected to solve. Furthermore, a substantial share of high-quality LLM training data is concentrated in English, whereas legal systems differ fundamentally across jurisdictions. Knowledge learned from one jurisdiction is often not transferable to another, raising serious concerns about degraded performance in non-English legal systems, where models may fail to capture jurisdiction-specific legal rules and reasoning structures.

Previous studies analyzed the internal behavior of transformer-based pretrained language model. Understanding of LLM's internal components are proved to be useful for increasing model abilities such as bias mitigation with activated neurons (Yang et al., 2024) and long context understanding with using attention heads (Ma and Okazaki, 2026). LLM interpretability studies on the internal mechanisms of language models contribute on the understanding model behaviors without relying on extensive computational approaches such as additional training. Several prior work has identified language neurons associated with specific model ability such as social bias (Yang et al., 2024), query-relevant neurons (Chen et al., 2025), language-specific neurons (Tang et al., 2024; Kojima et al., 2024), and culture neurons (Ying et al., 2025; Namazifard and Galke Poech, 2025; Yamamoto et al., 2026). If neurons associated with the legal domain can be identified, there is potential to improve task performance without additional model training because research on LLM interpretability aims not only to clarify the internal mechanisms of LLMs but also to leverage such insights for performance improvement.

In this paper, we investigate whether legal-domain-specific neurons do exist, analogous to

task-specific neurons in other domains, by comparing with neurons that are specific to other applied-domain tasks, such as mathematics, medicine, common sense reasoning and translations. Our research questions are: (i) are these task-specific neurons are unique to their tasks or widely shared across tasks, (ii) do the legal domain specific neurons that contribute to a wide range of legal tasks exist, (iii) the relationships and differences between legal task specific neurons and other task specific neurons in LLMs, and (iv) the distribution of those task specific neurons in the large language models across layers, attention and MLP components. To address these questions, we analyze seven domain application tasks from legal of LexEval (Li et al., 2024), LexGLUE (Chalkidis et al., 2022), and LEXam (Fan et al., 2025), mathematics of GSM8K (Cobbe et al., 2021), medicine of MedQA (Jin et al., 2021), physical/spatial common sense reasoning of PIQA (Bisk et al., 2020), and translation of WMT19 (Barrault et al., 2019) to capture each tasks specific neurons and examine whether the observed patterns are specific to the legal domain or instead generalize across applied reasoning tasks.

In experiments, suppressing task-specific neurons significantly degrades accuracy of that task while we notice that this task-specific neuron suppression doesn't largely affect for another task accuracy especially for tasks that are distant backgrounds, e.g., legal tasks and others. Furthermore, we also notice that there are a small portion of neurons that are influential to multiple tasks across domains. When we remove such across-tasks influential neurons from task-specific neurons, this strengthens the task-uniqueness of the remaining neurons. In legal domain specific neuron analysis, we observe two main phenomenon: suppressing identified legal task neurons in one task decline the other legal task accuracy, and legal neurons have relatively larger overlap with each other. In discussion, the distribution of identified neuron in our experiment show that the hypothesis of influential neuron distribution in middle MLP layers may have input-text format and content dependency rather than universal phenomenon.

2 Related Work

2.1 LLM interpretability

Mechanistic interpretability (Nanda et al., 2023) research has aimed to reveal how information is represented within individual LLM components including attention heads (Crosbie and Shutova, 2025; Wu et al., 2025), projecting model internal vectors to vocabulary space (Dar et al., 2023; Stolfo et al., 2023), and neurons level analysis. Prior studies have identified individual neurons and neuron populations associated with particular capabilities, such as knowledge storage neurons (Dai et al., 2022), linguistic neurons that are activated by LLM inputs of specific languages (Kojima et al., 2024; Tang et al., 2024), and domain specific neurons in vision language models (Huo et al., 2024). However, these studies often lack to examine whether their methods and identified neurons are task-specific by comparing with controlled experiments on other domains. Despite the growing attention on improving LLM reasoning abilities, there has been very little research analyzing LLM interpretability in reasoning tasks outside of mathematics.

2.2 LLM in Legal Domain

Legal question-answering datasets requiring legal reasoning ability have been proposed across multiple jurisdictions and languages. LEXam (Fan et al., 2025) collects university-level examination QA in Switzerland in both multiple-choice and open-ended formats. LexEval (Li et al., 2024) provides a comprehensive Chinese legal benchmark covering 23 tasks and over 14k questions. LexGLUE (Chalkidis et al., 2022) collects QA datasets in several languages including US law. Existing methods to improve LLM legal performance largely rely on continued pretraining or supervised finetuning on jurisdiction-specific corpora. However, for the reason of personal privacy, company's trade secrets, and intellectual property rights protection, it is not feasible to construct datasets that comprehensively cover every type of problem LLMs may encounter, research is needed on improving legal generation capabilities through approaches other than additional training.

2.3 LLM in Application Domain

In applied domains such as mathematics (Lightman et al., 2023), medicine (Sun et al., 2025; Zhang et al., 2025; Sviridova et al., 2024), and finance (Reddy et al., 2024; Magomere et al., 2025), a wide range of benchmark datasets have been proposed and are being used in combination with diverse training methods, including DPO (Rafailov et al., 2023) and GRPO (Zhihong Shao, 2024). Particularly in mathematics, research has examined

the internal mechanisms of LLMs using methods such as causal analysis to investigate mathematical reasoning (Nikankin et al., 2025), the mechanisms of reasoning steps that a large proportion of reasoning steps are not in fact essential to the reasoning process itself (Nikankin et al., 2026).

3 Method

We aim to identify neurons that are indispensable for reasoning, in the sense that suppressing their outputs leads to a substantial drop in the model's overall reasoning accuracy. Our proposal is built on neuron-level attribution methods from the mechanistic interpretability literature, which assign an importance score to each neuron with respect to a given task. We first identify a small set of neurons that are most influential for solving a target task using a training split, and we then *suppress* their activations at inference time on the test split.

3.1 Background: Neuron Attribution Score

A line of work in mechanistic interpretability has formalized the question of which neurons are responsible for a particular model behavior. Dhamdhere et al. (2019) extended Integrated Gradients (Sundararajan et al., 2017) from input attribution to internal neurons via the *conductance*, which integrates the gradient of the output with respect to a neuron's activation along a straight-line path from a reference activation to the actual one. Dai et al. (2022) adapted conductance to the FFN sub-layers of Transformer language models and used it to localize *knowledge neurons* that store factual associations, validating their causal role by suppressing or amplifying their activations. This template—compute an attribution score, select the top-k neurons, and intervene to verify causal effects—has since been applied to bias neurons (Yang et al., 2024), query-relevant neurons (Chen et al., 2025), language-specific neurons (Tang et al., 2024; Kojima et al., 2024), and culture neurons (Ying et al., 2025; Namazifard and Galke Poech, 2025).

The integrated-gradient formulation requires multiple forward/backward passes per example (typically $m \approx 20$ Riemann steps), which is expensive when one wants to score every neuron in a 14B–20B-parameter LLM. Following (Yang et al., 2024), we therefore adopt the first-order Taylor approximation of the integral around a zero reference activation, which reduces the score to a single "gradient $\times$ activation" product computable with one backward pass. Concretely, given an instruction ι, an input x, and a target output y, the attribution score of the i-th neuron with intermediate activation h_i is defined as

$$A_i^{(\iota,x,y)} \;=\; h_i \cdot \frac{\partial P(y \mid \iota, x)}{\partial h_i}. \qquad (1)$$

Equation (1) can be interpreted as a first-order approximation of the drop in the target probability that would result from ablating neuron i to zero, and is therefore a tractable proxy for the causal importance of the neuron.

3.2 Identifying task-specific influential neurons and suppressing them

Score aggregation on a training split. Let $\mathcal{D}_{\text{train}} = \{(x^{(j)}, y^{(j)})\}_{j=1}^{N}$ be the training split of a target domain (e.g., a legal benchmark). For each neuron i we aggregate Eq. (1) over the training examples to obtain a domain-level importance score

$$\bar{A}_i \;=\; \frac{1}{N}\sum_{j=1}^{N} A_i^{(x^{(j)},y^{(j)})}, \qquad (2)$$

and select the top-k neurons $\mathcal{N}_k = \arg\text{top-}k_i\, \bar{A}_i$ with the highest aggregated scores. Because $\mathcal{N}_k$ is determined entirely from $\mathcal{D}_{\text{train}}$, the test split remains untouched and is used only for evaluation.

Activation suppression at inference. Standard prior work uses attribution scores either to *suppress* (set to zero) or to qualitatively *amplify* the selected neurons in order to verify causal effects (Dai et al., 2022). In this study, we suppress the neuronal activity by suppressing the activations of the selected neurons by $\gamma = 0$,

$$\tilde{h}_i \;=\; \begin{cases} \gamma \cdot h_i & \text{if } i \in \mathcal{N}_k, \\ h_i & \text{otherwise}, \end{cases} \qquad (3)$$

while leaving all other neurons and all model parameters unchanged.

Random-neuron control. A natural concern is that any modest perturbation of the network might happen to improve scores on a benchmark. To rule this out, we always compare against a control in which the same number k of neurons is selected uniformly at random and suppressed by the same factor γ. A genuine effect of attribution should therefore manifest as a systematic gap between the top-k amplification and the random-amplification baseline.

3.3 Application to legal reasoning and other tasks

To investigate legal specific neurons in transformer-based language model, we use three legal reasoning datasets of three different legal jurisdictions of China, US, and Swiss by using LexEval (Li et al., 2024), LexGLUE (Chalkidis et al., 2022), and LEXam (Fan et al., 2025). LexEval covers a wide variety of Chinese legal questions, LexGLUE and LEXam offers Swiss legal questions in English and German. In long-form legal question answering of legal reasoning, multiple valid answer formulations may exist to express the same logical reasoning, therefore no definitive evaluation methodology has yet been established and the best evaluation method of open-ended legal question answering task is yet under discussion. LLM-as-a-Judge is not sufficient because it sometimes performs worse than human evaluation in legal domain. Accordingly, we target only multiple choice style reasoning tasks among LexEval, LexGLUE, and LEXam datasets. In order to compare whether the observed phenomenon is unique to legal domain or common across other application domain reasoning tasks, we use mathematical reasoning dataset GSM8K (Cobbe et al., 2021), medical domain specific task MedQA (Jin et al., 2021), spatial reasoning task PIQA (Bisk et al., 2020), and translation task WMT19 (Barrault et al., 2019).

Experimental settings We perform our experiments on seven open-weight transformer-based reasoning models with different pretraining data and size: Qwen3-8B, Qwen3-14B (Team, 2025b), Gemma-3-12b-it (Team, 2025a), Gemma-4-E4B-it, GPT-OSS-20B (OpenAI, 2025), Olmo-3-1025-7B (Olmo et al., 2026), Llama-3.1-8B-Instruct. We split 7 datasets into training and test sets and identified important neurons using the training split, then measured task-specific accuracy on the test split. GSM8K was evaluated using LightEval (Habib et al., 2023), based on eight generations produced with chain-of-thought prompting while WMT19 was evaluated by F1 score of generated texts to the gold answers. For the multiple-choice tasks other than GSM8K and WMT19, accuracy was evaluated against the gold answers.

Criteria of influential neurons from neuron attribution scores Whereas prior neuron extraction methods typically identified neurons whose scores exceeded a fixed threshold, we note that such thresholds may depend on the specific task. Because our goal is to analyze influential neurons across tasks, a threshold-based approach may obscure neurons that play an important role in particular tasks, especially when the corresponding datasets are limited in size. We therefore adopt a top-n% selection strategy for each task after computing neuron attribution scores. Specifically, we extract neurons at three selection ratios, top 1%, 0.5%, and 0.3%, and find that task-specific neurons are most clearly observed around the top 0.5% level. The results of suppressing neurons selected at the top 1% and top 0.3% levels are provided in the Appendix A.

4 Result

4.1 Task-wise influential neurons

Identified neurons surely contribute to task performance Following prior studies based on neuron attribution scores, we verified that the neurons identified within each task by neuron attribution scores are genuinely influential for task performance. Table 1 reports the results for the baseline, suppression of the selected neurons, and suppression of randomly selected neurons. The baseline results are the averaged over seven inference runs. By suppressing the neuron outputs to 0 in identified neurons, model accuracy drops to nearly zero. By contrast, even when the same number of neurons are randomly suppressed, task performance remains almost identical to the baseline. This experiment suggests that the top 0.5% of extracted neurons are highly influential to the task performance.

Score change by top k neuron suppression We analyze the relationship between the number of suppressed neurons and task performance for the neurons identified in each task. Figure 1 shows the changes in task accuracy when the top-k neurons are suppressed. We observe a clear decline in accuracy when suppressing roughly the top 100 to top 1,000 neurons. By contrast, accuracy does not decrease substantially when only the top 1 to top 100 important neurons are suppressed. The remaining results for top-k neuron suppression are provided in Appendix E.

4.2 Influential neurons across tasks

Next, we examine whether the neurons identified for each task are specifically influential for that task or remain influential across other tasks as well.

Model		LexEval	LexGLUE	LEXam	GSM8K	MedQA	PIQA	WMT19
Language		ZH	EN	EN, DE	EN	EN, ZH	EN	multi
Qwen3-8B	base	64.52	67.80	14.20	80.14	68.21	78.78	88.44
	identified	0.0	0.20	1.40	0.76	0.29	0.0	0.0
	random	59.94	65.00	43.60	74.91	69.39	72.58	86.67
Qwen3-14B	base	68.37	70.80	24.00	86.35	79.35	87.49	90.22
	identified	0.10	0.20	4.60	0.23	0.47	0.76	0.00
	random	66.49	71.00	31.60	84.69	76.90	84.11	89.33
Gemma-3 -12b-it	base	28.51	47.29	39.80	85.90	61.32	82.37	80.67
	identified	0.42	7.20	1.20	0.08	0.16	1.41	0.0
	random	51.61	62.60	7.40	77.18	58.07	82.15	75.00
Gemma-4-E4B-it	base	2.39	11.80	18.20	79.08	30.30	0.76	43.78
	identified	0.10	1.40	1.40	1.06	0.00	0.00	0.11
	random	6.45	37.40	14.80	74.98	19.50	0.22	25.00
GPT-OSS-20B	base	6.24	20.00	20.80	62.09	5.04	11.21	43.33
	identified	0.0	15.0	0.0	0.15	0.0	23.07	12.67
	random	18.31	46.60	13.0	75.28	23.04	38.85	41.67
Olmo-3-1025-7B	base	13.42	48.80	18.20	52.77	35.81	73.34	38.44
	identified	0.52	4.20	0.0	1.21	0.82	2.29	0.33
	random	37.36	60.80	18.20	32.98	34.42	74.43	38.67
Llama-3.1-8B-Instruct	base	24.97	53.40	35.00	37.00	66.90	83.03	74.22
	identified	0.0	0.0	0.0	0.0	0.02	0.0	0.0
	random	46.41	64.40	9.80	25.02	50.41	79.98	71.89

Table 1: Results of task-wise identified neuron suppression. "base" is the original model results while "identified" and "random" are neurons suppressed results. Identifies neurons are at top 0.5% of all in each model.

To this end, we conduct an experiment in which we suppress the top 0.5% of influential neurons identified for one task and then evaluate performance on the remaining tasks. If neurons that are important for a given task affect only that task, suppressing them should not alter accuracy when the model is evaluated on other tasks. Figure 2 presents the results of this cross-task suppression analysis across the seven tasks. In Gemma-3-12B-it, GPT-OSS-20B, and Gemma-4-E4B-it, suppressing the influential neurons computed from one task tend not to affect for other task performances. In contrast, when the same top 0.5% neurons of one task are suppressed, and evaluated task accuracy of that task, the model performance drop nearly 0. These results suggest that, for these three models, our method successfully extracts task-specific neurons. In Olmo-3-1025-7B and Llama-3.1-8B-Instruct, suppressing the influential neurons computed from GSM8K does not reduce model performance of other tasks, whereas performance drops sharply when model's GSM8K accuracy is evaluated with that identified neurons of GSM8K. In Olmo-3-1025-7B, a similar trend is also observed for LexGLUE. However, for the remaining tasks, neuron suppression causes decline of nealy all task accuracy to approach zero. For Qwen3-8B and Qwen3-14B, suppressing the identified neurons causes performance on nearly all tasks to drop close to zero, making it difficult to conclude that domain-specific neurons have been successfully isolated. Comprehensive experimental results are provided in Appendix B.

Model	#Top0.5%	#Shared
Qwen3-8B	5,529	316
Qwen3-14B	8,396	474
Gemma-3-12B-it	9,338	1,310
Gemma-4-E4B-it	4,945	359
GPT-OSS-20B	1,305	0
Olmo-3-1025-7B	5,488	1,078
Llama-3.1-8B-Instruct	5,570	2,055

Table 2: Number of neurons at top 0.5% and the number of the shared neurons across seven tasks.

Do task-wise influential neurons overlap across tasks? Do task-wise influential neurons overlap across tasks? Thus far, we have examined the properties of influential neurons identified for a specific task. We now turn to neurons that are influential across multiple tasks. In particular, we investigate whether neurons identified as influential for one task are also influential for others, or whether there exist neurons specific to a particular task or domain,

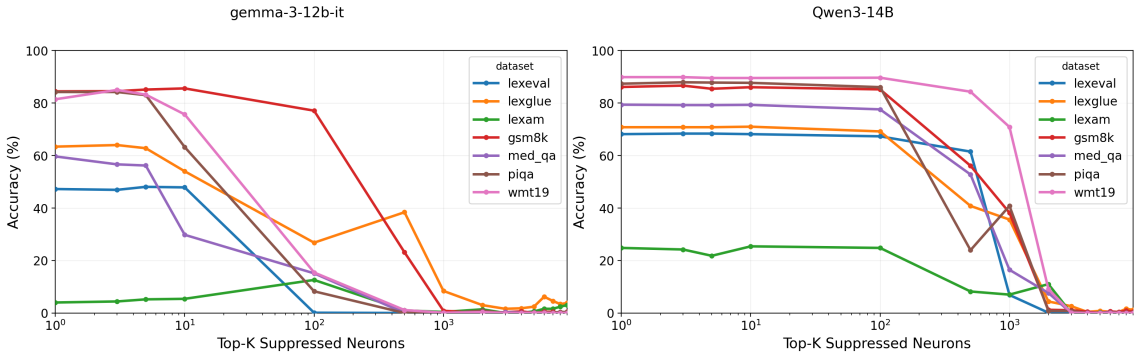

Figure 1: The top K neuron suppression result of Gemma-3-12b-it and Qwen-3-14B sorted by neuron attribution score.

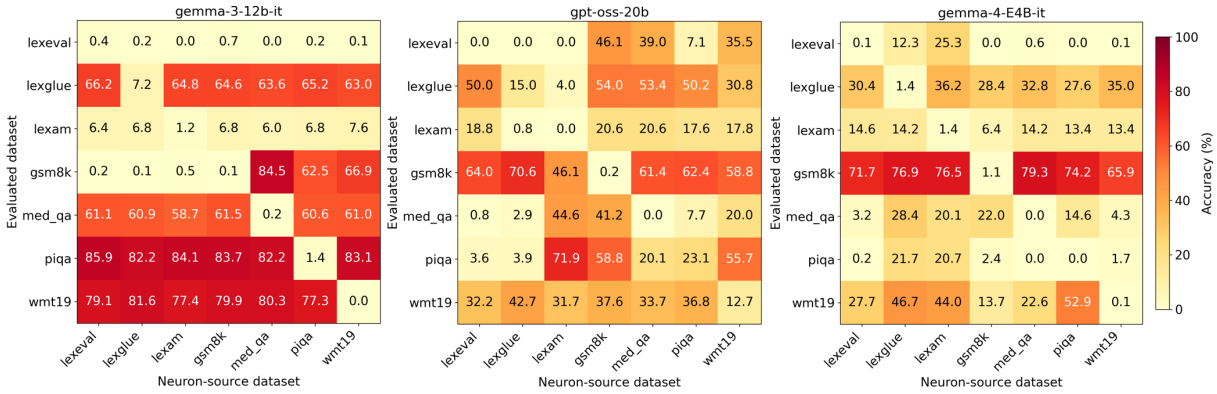

Figure 2: Task-wise neuron suppression results in Gemma-3-12B-it, Gemma-4-E4B-it, GPT-OSS-20B.

such as legal or mathematical reasoning. Table 2 shows the number of the number of 0.5% neurons and the number of shared neurons identified as influential across all seven tasks. Figure 3 visualizes the overlap of influential neurons among the seven tasks. In Figure 3, at least 5% of the selected neurons are shared across tasks in Gemma-3-12B-it and gemma-4-E4B-it, whereas PIQA in GPT-OSS-20B exhibits almost no overlap with the selected neurons of other tasks. The complete results of neuron overlap are provided in Appendix B.

Truly task-specific neurons via removal of cross-task influential neurons. To isolate neurons that are highly important for individual tasks, we attempt to extract task-specific neurons by removing, from each task's list of influential neurons, those that are commonly identified as influential across all seven tasks. Figure 4 shows the neuron suppressed result of each datasets. In Gemma-3-12B-it, gemma-4-E4B-it, and GPT-OSS-20B, suppressing the top 0.5% of neurons ranked by attribution score reduces accuracy only on the task from which those neurons were identified, while leaving the other tasks largely unaffected. A similar tendency emerges in Olmo-3-1025-7B, Llama-3.1-8B-Instruct, Qwen3-8B, and Qwen3-14B once the neurons shared across all seven datasets are removed: accuracy declines only on the task from which the neurons were identified, while remaining largely unchanged on the others. These results suggest that some models contain neurons that play a general role across all tasks, and that task-specific neurons can be recovered by removing such broadly important neurons from the top-ranked set. The remaining results of the identified neuron distributions are provided in Appendix D.

4.3 Domain-specific influential neurons

Characteristics of task-specific neurons in the legal domain. Finally, we return to the task-specific analysis of influential neurons and examine the characteristics of legal-domain influential neurons. As shown by the cross-suppression results in Figure 2, the identified neurons also behave as task-specific neurons in GPT-OSS-20B, Gemma-3-12B-it, and gemma-4-E4B-it. In these three models, we observe a weak tendency across the three legal datasets (LexEval, LexGLUE, and LEXam), namely suppressing the influential neurons for one

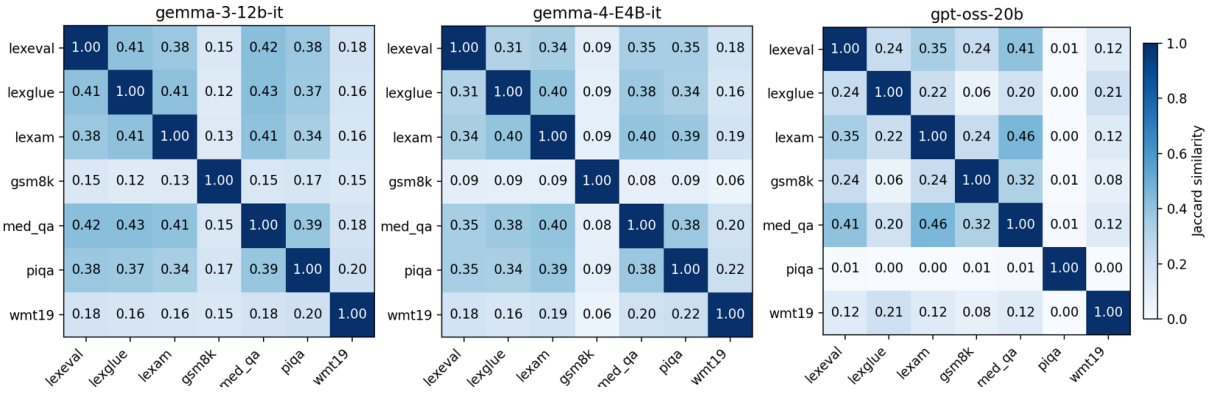

Figure 3: Influential neuron overlaps across tasks in Gemma-3-12B-it, Gemma-4-E4B-it, and GPT-OSS-20B.

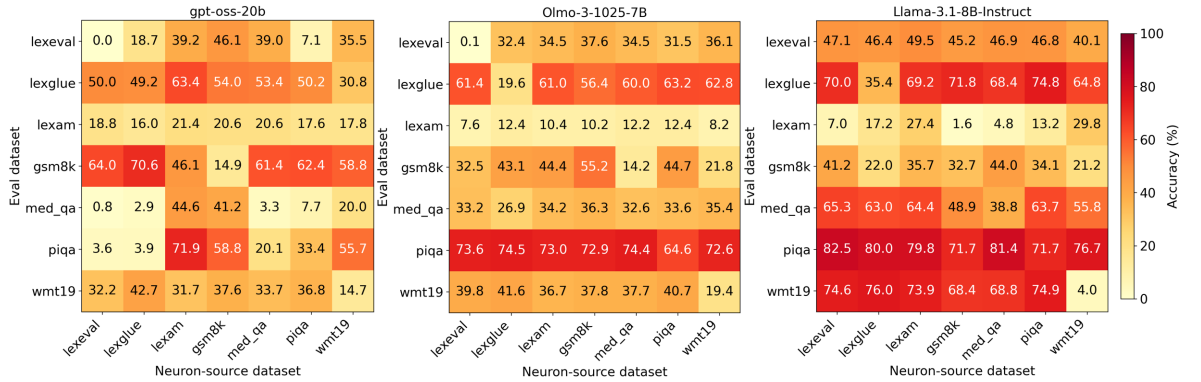

Figure 4: Neuron suppression results when cross-task influential neurons are removed in GPT-OSS-20B, Olmo-3-1025-7B, and Llama-3.1-8B-Instruct.

legal task degrades performance on the other two as well. Consistent with this, the overlap of identified neurons shown in Figure 3 indicates that the degree of overlap among task-associated neurons is relatively high among LexEval, LexGLUE, and LEXam, as seen in the upper-left portion of the figure. This phenomenon is observed in Llama-3.1-8B-Instruct, Olmo-3-1025-7B, Gemma-3-12B-it, gemma-4-E4B-it, and GPT-OSS-20B.

The relationship between legal task-specific neurons and other task-specific neurons. The relationship between legal task-specific neurons and other task-specific neurons. Neurons associated with the legal datasets (LexEval, LexGLUE, and LEXam) exhibit substantial overlap with those identified for MedQA and PIQA in Llama-3.1-8B-Instruct, Gemma-3-12B-it, gemma-4-E4B-it, and Olmo-3-1025-7B. In GPT-OSS-20B, they instead overlap substantially with the neurons identified for GSM8K and MedQA. These findings suggest the existence of neurons that are broadly important across reasoning tasks.

5 Discussion

On which layers, MLPs, and attention modules are influential neurons distributed? A recurring line of inquiry has examined where the neurons most critical for LLM output generation reside within the network, from the first to the final layer, and whether neurons in the attention mechanism, where $\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(QK^\top/\sqrt{d})V$ with d denoting the key dimension, are as important as those in the MLP layers. Geva et al. (2021) showed that feed-forward layers function as key-value memories, in which the keys detect input patterns and the values encode next-token distributions, and that the model's final prediction is constructed in a bottom-up manner by composing these memories within each layer and iteratively refining them across layers. Yamamoto et al. (2026) similarly found that most identified cultural neurons reside in the shallow to middle MLP modules, and that this trend is consistent across models. In contrast, Nikankin et al. (2026) trained separate probes on the residual-stream activations at specific token positions for each generation step in order to localize where importance is encoded.

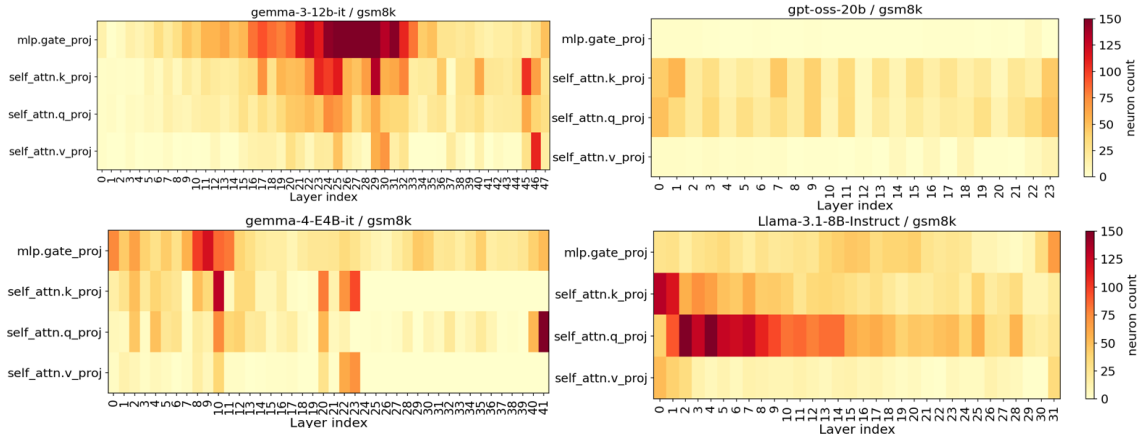

Figure 5: Important neuron distribution heatmap of gemma-3-12b, Gemma-4-E4B-it, GPT-OSS-20B, Llama-3.1-8B-Instruct in top 0.5% neuron attribution score.

They found that probes trained on individual layers achieved comparable accuracy, suggesting that the components critical to the model's output are not localized to any particular layer but are instead broadly distributed across the network.

In our experiment, identified task specific neurons distribution of Gemma-3-12b-it and Gemma-4-E4B-it are similar to existing analysis that important neurons are placed in the middle of MLP layer, while Llama-3.1-8B-Instruct and GPT-OSS-20B are seemed to have important neurons in key K and query Q of attention across from the shallow to the deep layer is shown in Figure 5. Appendix C for the full distribution of top fluential neurons across all tasks and models. Appendix D shows the distribution of neurons selected via removal of cross-task influential neurons from identified neurons, and this also shows the similar tendency.

Middle-MLP localization may not generalize to long, complex inputs. The distribution of important neurons varies across models, while within each model a broadly consistent pattern is observed across all analyzed tasks. In Gemma-3-12B-it and Gemma-4-E4B-it, all seven tasks show a mild tendency for important neurons to be distributed around the middle MLP layers, consistent with prior hypotheses. In contrast, Olmo-3-1025-7B and Llama-3.1-8B-Instruct exhibit a concentration in the attention K and Q components from layer 0 through roughly the first two-thirds of the network. Meanwhile, in Qwen3-8B and Qwen3-14B, important neurons are distributed broadly across all layers in the MLP, K, and Q components, but not in projection component V. The rest result of identified neuron distributions are shown in Appendix C.

In prior studies that found important neurons in middle layer of MLP, they studied short-text inputs with relatively general knowledge tasks such as multilingual, bias and culture understanding. However, when examining neuron activation scores in more applied domain such as mathematics and coding studied by Nikankin et al. (2026), prior findings may not be applicable. The hypothesis that middle layers of the MLP is critical for text generation was task-dependent rather than general principle of LLM because all domain application tasks including mathematics, medicine, law have relatively long input text and complicated text structure.

6 Conclusion

We presented a neuron-level analysis of legal-domain reasoning in LLMs, comparing it with other applied domains across seven open-weight models. Using neuron attribution scores to rank and suppress influential neurons, we confirmed that suppressing the identified neurons collapses accuracy on the target task, whereas suppressing the same number of random neurons does not. We found a small subset of neurons influential across seven tasks; once these are removed, suppressing the remaining neurons degrades only the task they were identified from, revealing genuinely task-specific neurons in every model studied. Within the legal domain, the three benchmarks exhibit relatively high neuron overlap and tend to be affected jointly, suggesting legal components that span jurisdictions. Finally, influential neurons do not consistently localize to middle MLP layers, indicating that this widely held hypothesis, established mainly on short and general-knowledge inputs, may be task- and input-format-dependent rather than universal.

Limitations

In this paper, to identify legal-domain-specific neurons, we analyze all existing datasets that require legal reasoning ability, using LexEval, LexGLUE, and LEXam to cover China, U.S., and Swiss law. However, because laws are, in principle, effective only within each country's own jurisdiction, and because the language in which they are expressed also differs across countries, it would be desirable for additional datasets to be proposed and evaluated that measure legal reasoning ability across a broader range of jurisdictions and languages.

Ethics Statement

This work focuses on analyzing internal neuron activations in large reasoning models to better understand and improve inference efficiency. Our analysis is conducted on publicly available benchmark datasets spanning legal reasoning and understanding (LexEval, LexGLUE, LEXam), mathematical reasoning (GSM8K), medical question answering (MedQA), physical commonsense reasoning (PIQA), and machine translation (WMT19). The study does not involve human subjects, the collection of personal data, or user-generated content, and all datasets are used in accordance with their respective licenses and intended research purposes. We do not anticipate direct negative societal impacts from the proposed method itself. While our method aims to improve model efficiency through neuron activation analysis, we acknowledge that interpretability techniques applied to domain-specific benchmarks, particularly in sensitive areas such as medical and legal reasoning, should not be interpreted as endorsements of deploying such models in high-stakes decision-making without appropriate expert oversight.

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Zhihan Cao, Fumihito Nishino, Hiroaki Yamada, Nguyen Ha Thanh, Yusuke Miyao, and Ken Satoh. 2025. Jbe-qa: Japanese bar exam qa dataset for assessing legal domain knowledge. *Preprint*, arXiv:2511.22869.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Lihu Chen, Adam Dejl, and Francesca Toni. 2025. Identifying query-relevant neurons in large language models for long-form texts. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Joy Crosbie and Ekaterina Shutova. 2025. Induction heads as an essential mechanism for pattern matching in in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5049–5111, Albuquerque, New Mexico. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, Toronto, Canada. Association for Computational Linguistics.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. 2019. How important is a neuron? In *International Conference on Learning Representations (ICLR)*.

Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya

Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. 2025. Lexam: Benchmarking legal reasoning on 340 law exams. *arXiv preprint arXiv:2505.12864*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. MMNeuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6816, Miami, Florida, USA. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: a comprehensive chinese legal benchmark for evaluating large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Youmi Ma and Naoaki Okazaki. 2026. From interpretability to performance: Optimizing retrieval heads for long-context language models. *Preprint*, arXiv:2601.11020.

Jabez Magomere, Elena Kochkina, Samuel Mensah, Simerjot Kaur, and Charese Smiley. 2025. FinNLI: Novel dataset for multi-genre financial natural language inference benchmarking. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4545–4568, Albuquerque, New Mexico. Association for Computational Linguistics.

Danial Namazifard and Lukas Galke Poech. 2025. Isolating culture neurons in multilingual large language models. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 768–785, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2025. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *The Thirteenth International Conference on Learning Representations*.

Yaniv Nikankin, Martin Tutek, Tomer Ashuach, Jonathan Rosenfeld, and Yonatan Belinkov. 2026. Reasoning models know what's important, and encode it in their activations. *Preprint*, arXiv:2604.18307.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. DocFinQA: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458, Bangkok, Thailand. Association for Computational Linguistics.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang,

Tingyang Xu, Qifeng Bai, and Yu Rong. 2025. ReasonMed: A 370K multi-agent generated dataset for advancing medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26446–26467, Suzhou, China. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328.

Ekaterina Sviridova, Anar Yeginbergen, Ainara Estarrona, Elena Cabrio, Serena Villata, and Rodrigo Agerri. 2024. CasiMedicos-arg: A medical question answering dataset annotated with explanatory argumentative structures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18463–18475, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team. 2025a. Gemma 3 technical report.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2025. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*.

Taisei Yamamoto, Ryoma Kumon, Danushka Bollegala, and Hitomi Yanaka. 2026. Neuron-level analysis of cultural understanding in large language models. In *The Fourteenth International Conference on Learning Representations*.

Nakyeong Yang, Taegwan Kang, Stanley Jungkyu Choi, Honglak Lee, and Kyomin Jung. 2024. Mitigating biases for instruction-following language models via bias neurons elimination. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9061–9073, Bangkok, Thailand. Association for Computational Linguistics.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22230–22251, Vienna, Austria. Association for Computational Linguistics.

Zhihao Zhang, Yiran Zhang, Xiyue Zhou, Liting Huang, Imran Razzak, Preslav Nakov, and Usman Naseem. 2025. From generation to detection: A multimodal multi-task dataset for benchmarking health misinformation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24245–24260, Suzhou, China. Association for Computational Linguistics.

Qihao Zhu Runxin Xu Junxiao Song Mingchuan Zhang Y.K. Li Y. Wu Daya Guo Zhihong Shao, Peiyi Wang. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

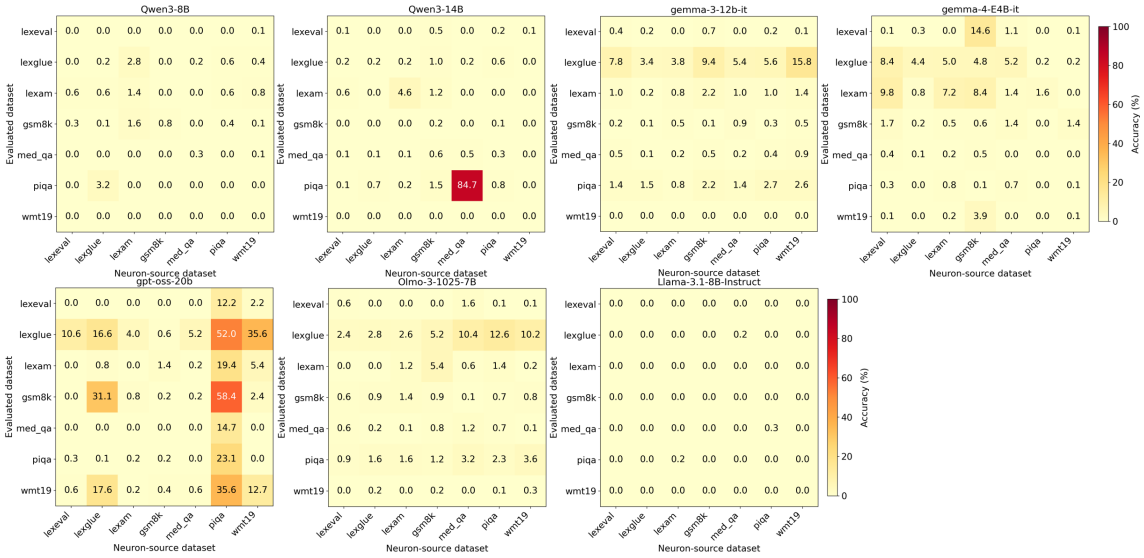

Figure 6: Neuron suppression result of top 1% attribution score neuron.

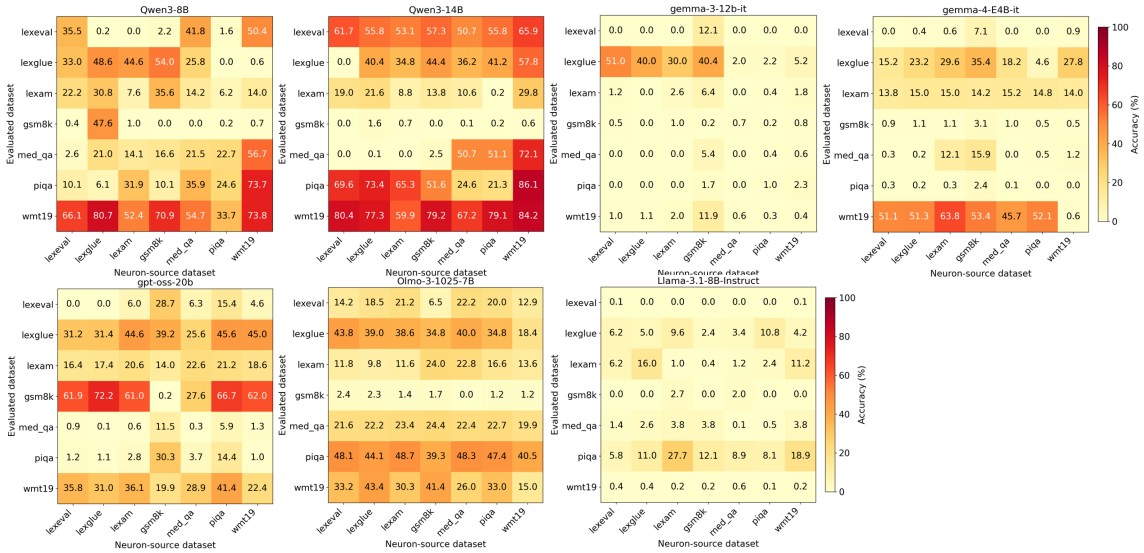

Figure 7: Neuron suppression result of top 0.3% attribution score neuron.

A Top 1% and top 3% neuron suppression

Figures 6 and 7 present the results of suppressing the top 1% and top 0.3% of neurons.

In the top 1% neuron suppression experiment, performance drops to nearly zero across all tasks, likely because the method selects too many neurons, capturing not only task-specific neurons but also a large number of more generally important neurons.

In the results obtained by suppressing the top 0.3% of neurons, there are cases in Qwen3-8B, Qwen3-14B, Gemma-4-E4B-it, GPT-OSS-20B, and Olmo-3-1025-7B where task-specific neurons appear to have been successfully identified. Specifically, for LEXam and PIQA in Qwen3-8B, PIQA in Qwen3-14B, WMT19 in Gemma-4-E4B-it, GSM8K and WMT19 in GPT-OSS-20B, and WMT19 in Olmo-3-1025-7B, suppressing the selected neurons reduces model performance only on the corresponding task, while performance on the other tasks remains comparatively high.

B Top 0.5% scored neuron suppression

Figure 8 and Figure 9 show the full result of neuron suppression in identified neurons and identified neuron overlap between each tasks in models.

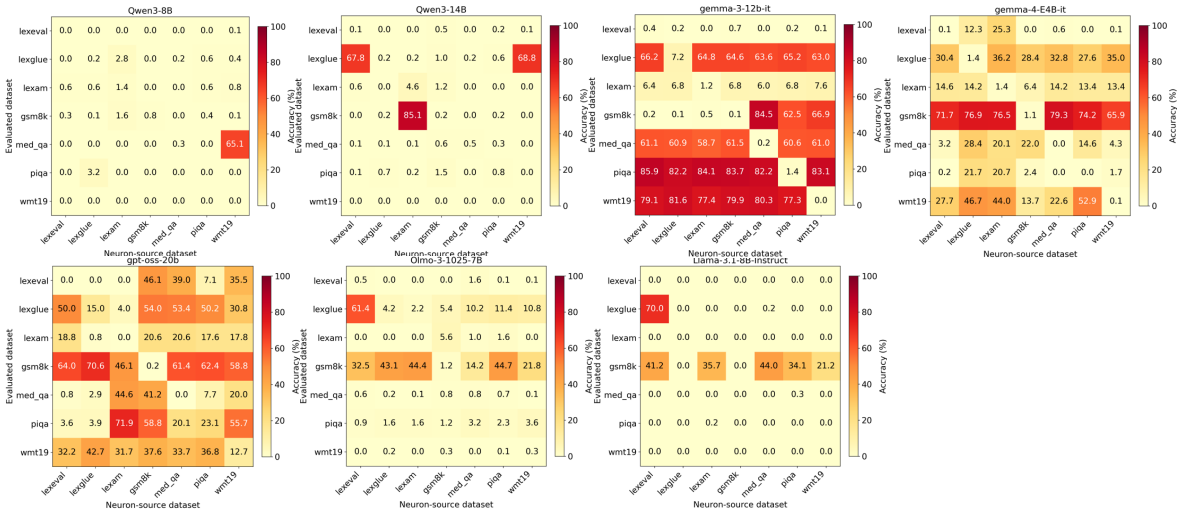

Figure 8: The result of neuron suppression in top 0.5% identified neurons.

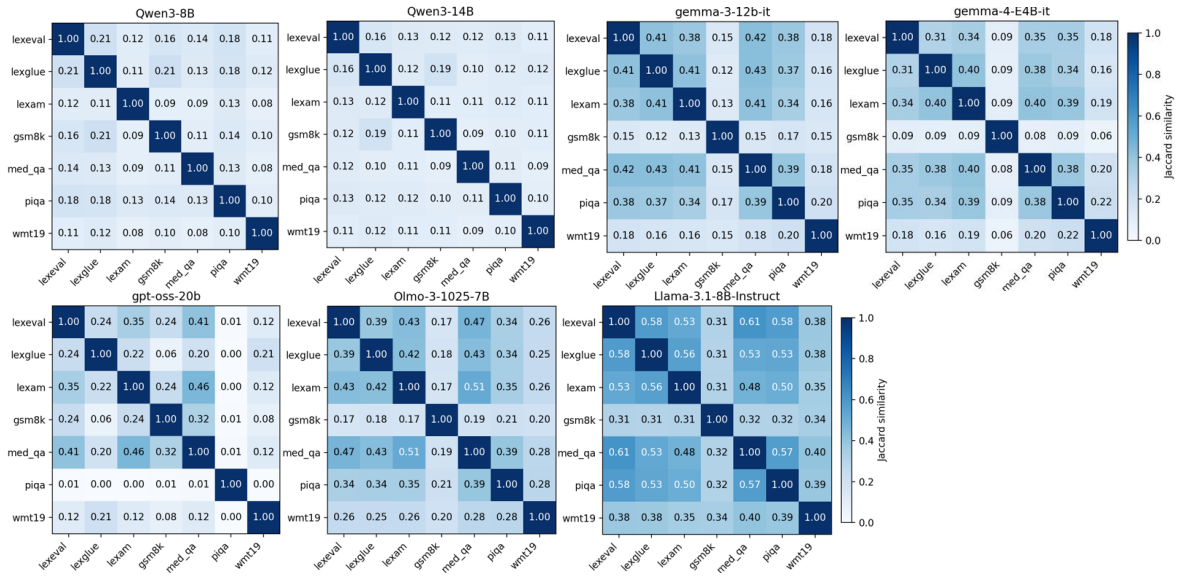

Figure 9: Neuron overlap of each models in top 0.5% neurons.

C Influential neuron distribution for top 0.5% scored neurons

Figure 11, Figure 13, Figure 14, Figure 12, Figure 15, Figure 16, and Figure 17 shows the important neuron distribution in top 0.5% of identified neurons in the experiment of Section 5. In Llama-3.1-8B-Instruct and GPT-OSS-20B, important neurons are gathering on K and Q outputs of the Attention.

Given that existing study of important neuron distribution analyzed short and general text understanding task, This result suggests that our method captures mechanisms shared across LLMs.

D Influential neuron distribution after the shared neuron removal

The complete result of shared neuron removal is shown in Figure 10. Figure 18, Figure 19, Figure 20, Figure 21, Figure 22, Figure 23, and Figure 24 shows the result after shared neuron removal in Section 4.2.

In Gemma-3-12B-it and Gemma-4-E4B-it, the prior hypothesis that important neurons are concentrated in the middle MLP layers is confirmed more clearly across nearly all tasks. A similar tendency toward concentration in the middle MLP layers is also observed for LEXam in Olmo-3-1025-7B, and for LEXam, LexGLUE, MedQA, and PIQA in Llama-3.1-8B-Instruct. In Qwen3-8B and Qwen3-14B, unlike the case where the top 5% of neurons are selected solely by neuron attribution score, important neurons appear to be distributed broadly across all MLP layers. In GPT-OSS-20B, as in the analysis based on the top

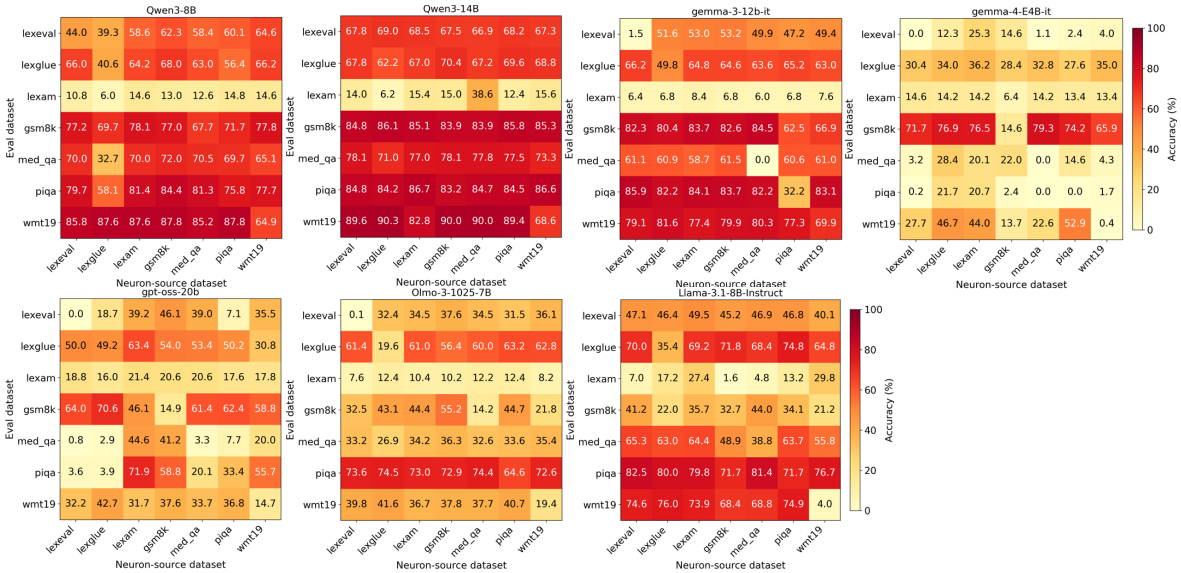

Figure 10: Neuron suppressed result when shared neurons are removed.

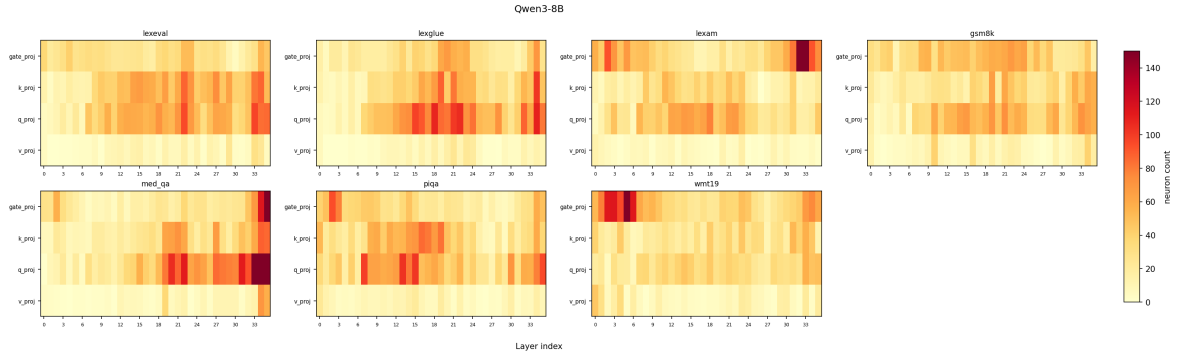

Figure 11: Neuron distribution of Qwen3-8B in top 0.5%.

5% of neurons by neuron attribution score, the K and Q components of the attention mechanism remain particularly important even after removing neurons that are shared across the seven tasks.

E Score change by top k neuron suppression

Figure 25 shows comprehensive analysis of the relation and the number of suppressed neurons and task performance for the identified neurons in each task. Qwen3-8B and Qwen3-14B exhibit a sharp decline in accuracy when suppressing between the top 100 and top 1,000 neurons. In Gemma-3-12B-it and Llama-3.1-8B-Instruct, task accuracy begins to decline between the top 10 and top 100 neurons. By contrast, in Gemma-4-E4B-it and Olmo-3-1025-7B, the decrease is more gradual, with substantial accuracy degradation emerging around the top 1,000 neurons. In GPT-OSS-20B, GSM8K accuracy declines between the top 10 and top 100 neurons, whereas performance on the other tasks remains relatively stable, showing only limited degradation.

F Experimental settings and AI usage

In our experiments, we used NVIDIA H100 and H200 GPUs. We used AI tools for translation of analyzed datasets and model outputs, translation of our manuscript and proof reading.

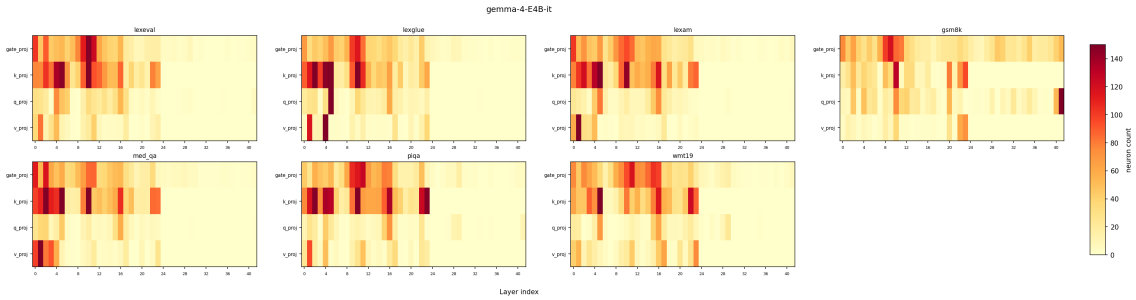

Figure 12: Neuron distribution of Gemma-4-E4B-it in top 0.5%.

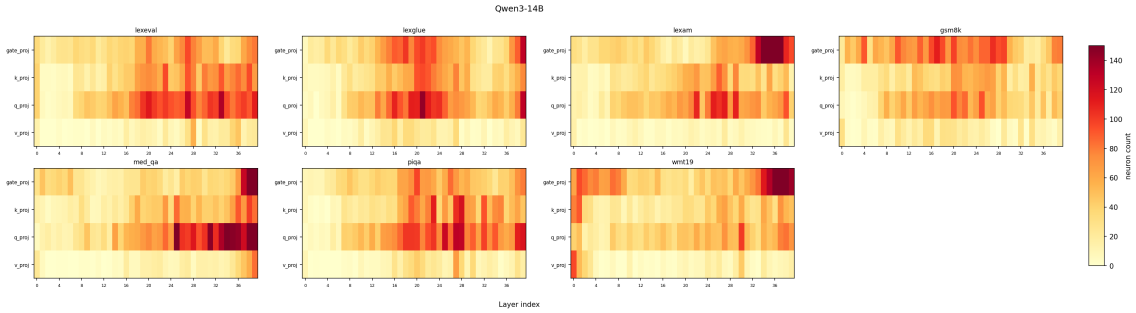

Figure 13: Neuron distribution of Qwen3-14B in top 0.5%.

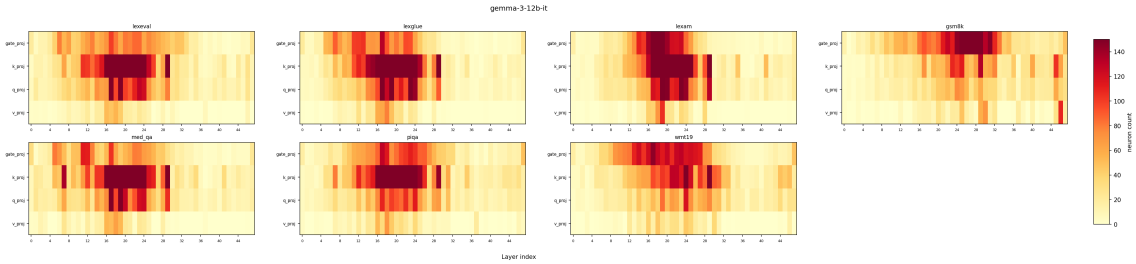

Figure 14: Neuron distribution of Gemma-3-12B-it in top 0.5%.

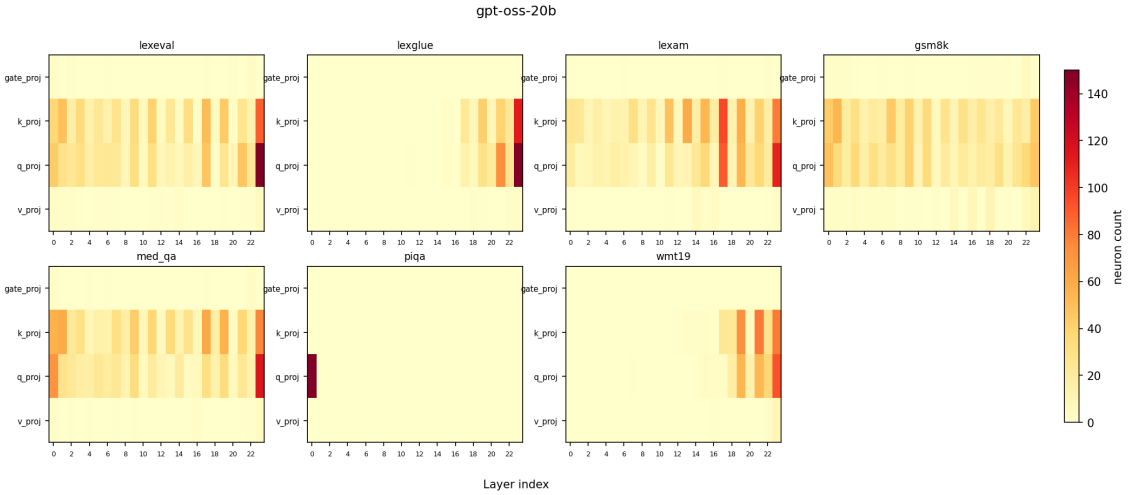

Figure 15: Neuron distribution of GPT-OSS-20B in top 0.5%.

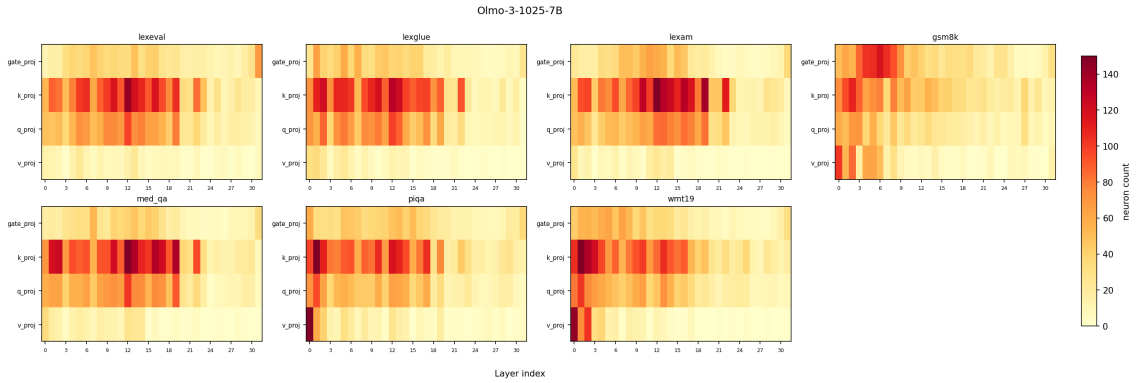

Figure 16: Neuron distribution of Olmo-3-1025-7B in top 0.5%.

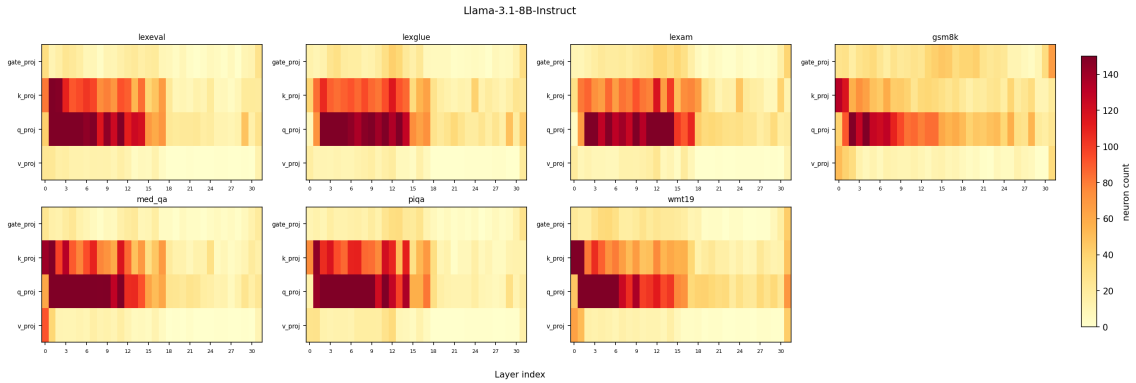

Figure 17: Neuron distribution of Llama-3.1-8B-Instruct in top 0.5%.

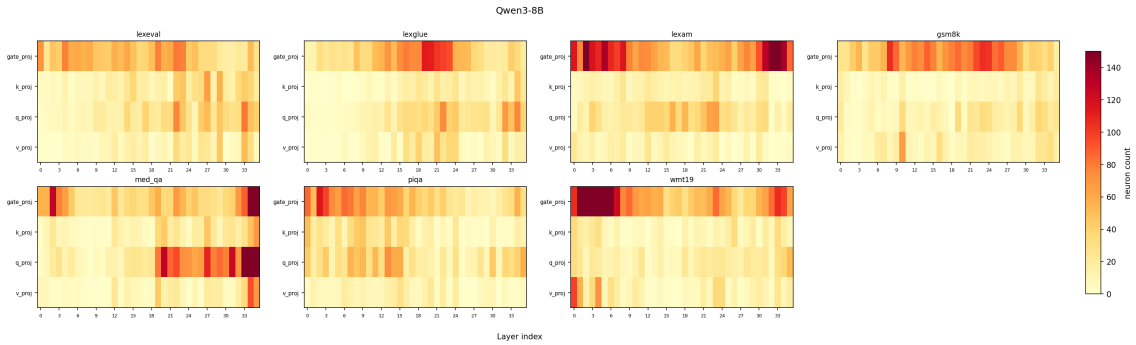

Figure 18: Neuron Distribution after shared neuron removal of Qwen3-8B.

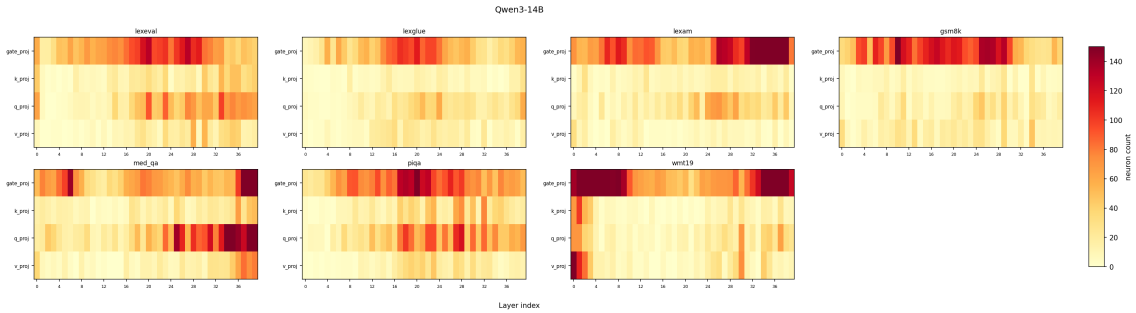

Figure 19: Neuron Distribution after shared neuron removal of Qwen3-14B.

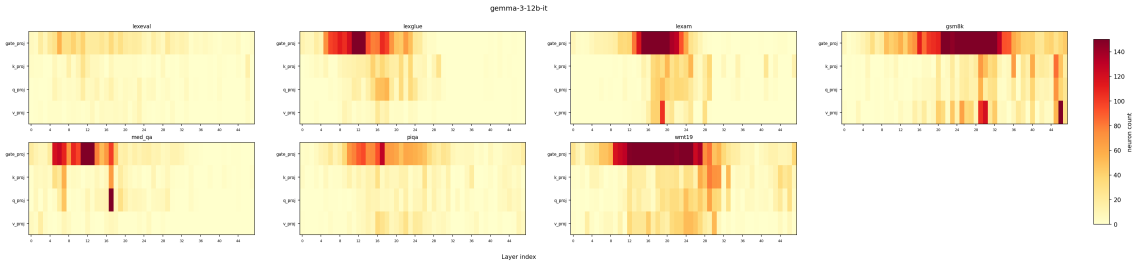

Figure 20: Neuron Distribution after shared neuron removal of Gemma-3-12B-it.

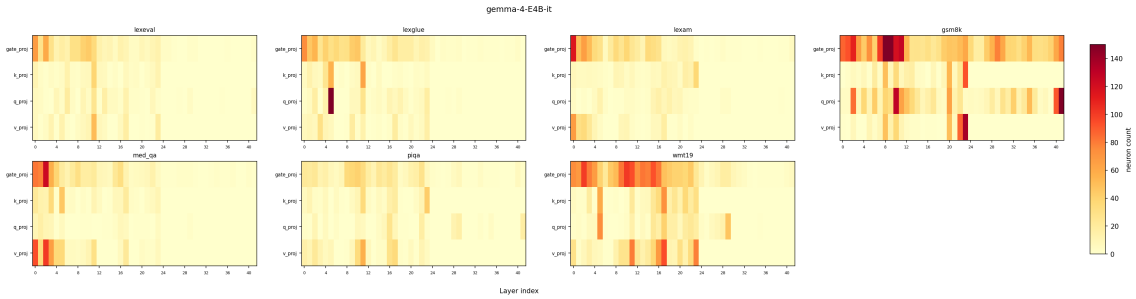

Figure 21: Neuron Distribution after shared neuron removal of Gemma-4-E4B-it.

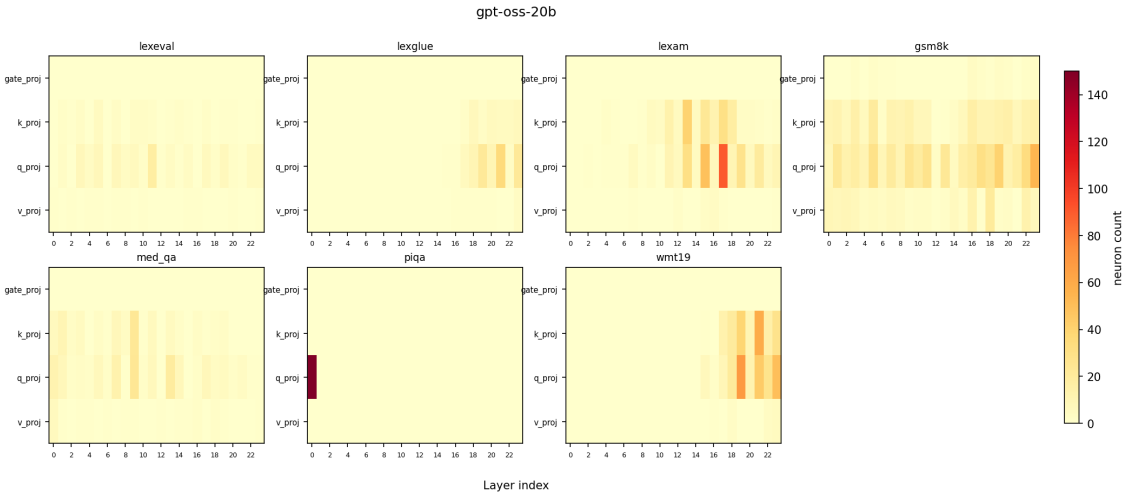

Figure 22: Neuron Distribution after shared neuron removal of GPT-OSS-20B.

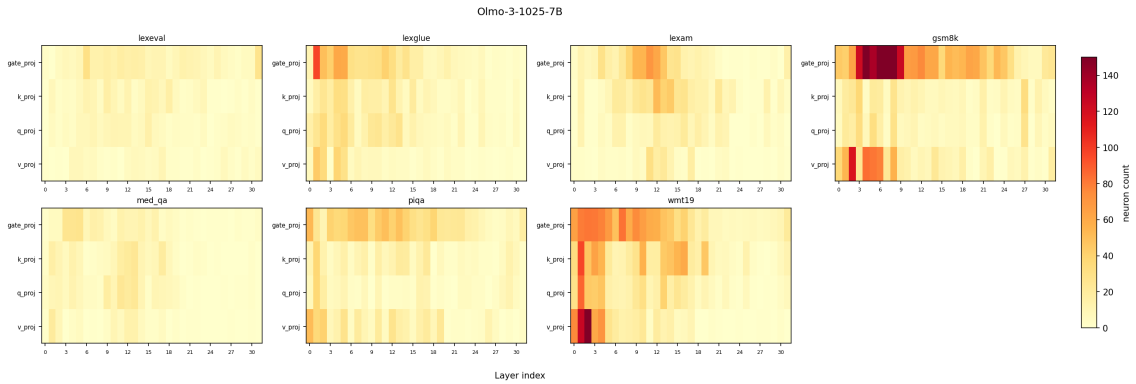

Figure 23: Neuron Distribution after shared neuron removal of Olmo-3-1025-7B.

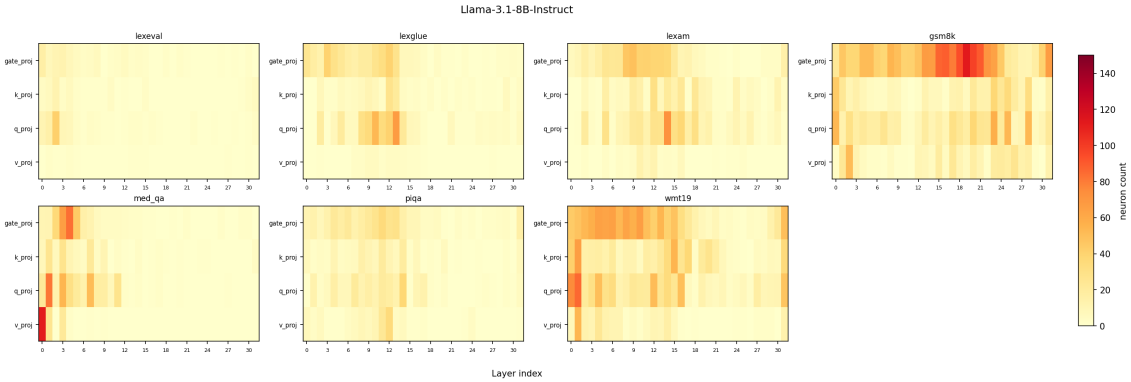

Figure 24: Neuron Distribution after shared neuron removal of Llama-3.1-8B-Instruct.

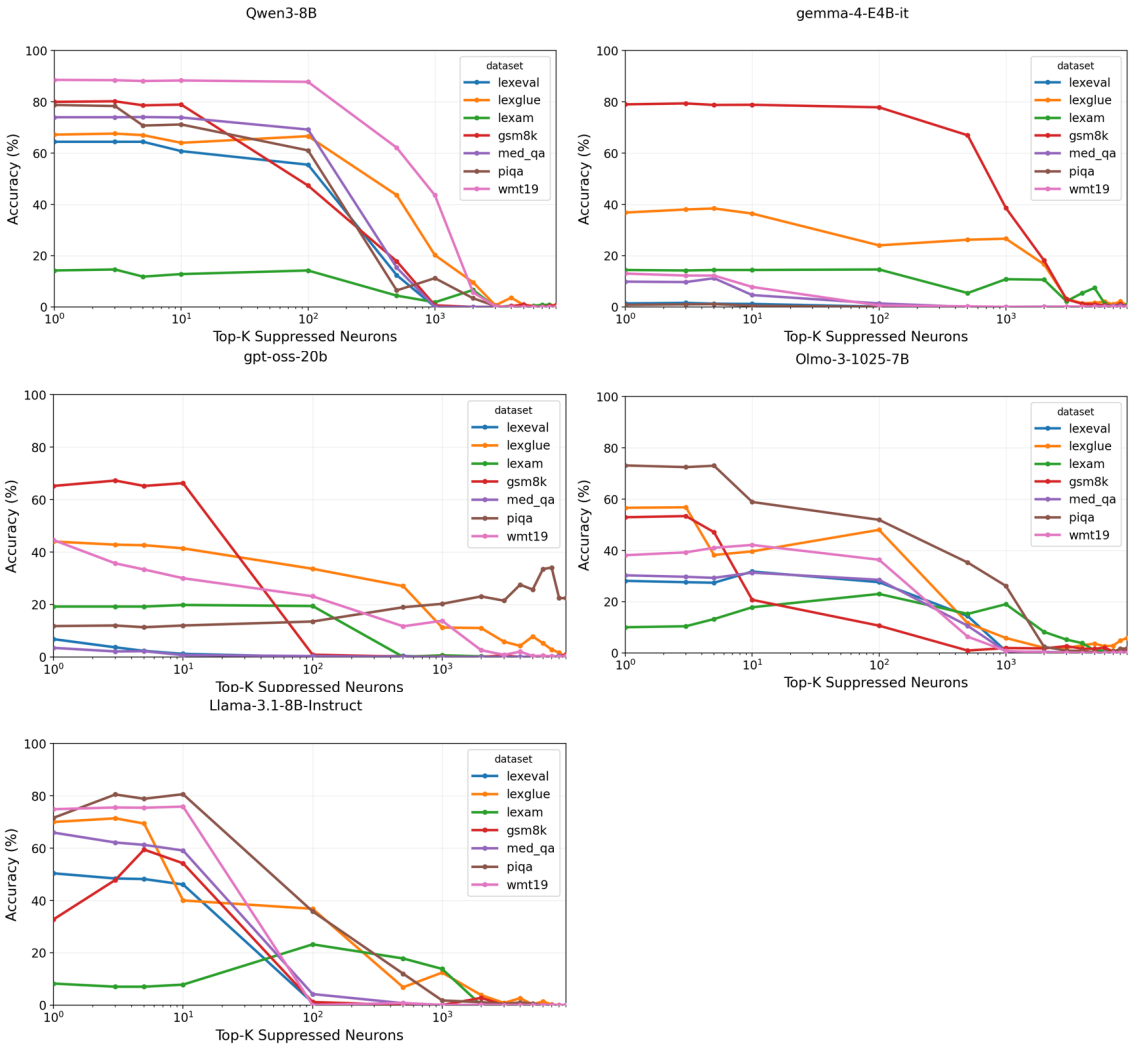

Figure 25: Top N neuron suppress sorted by neuron attribution score in each datasets.