Neuron Empirical Gradient: Discovering and Quantifying Neurons' Global Linear Controllability

Anonymous ACL submission

Abstract

Although feed-forward neurons in pre-trained language models (PLMs) can store knowledge and their importance in influencing model outputs has been studied, existing work focuses on finding a limited set of neurons and analyzing their relative importance. However, the global quantitative role of activation values in shaping outputs remains unclear, hindering further advancements in applications like knowledge editing. Our study first investigates the numerical relationship between neuron activations and model output and discovers the global linear relationship between them through neuron interventions on a knowledge probing dataset. We refer to the gradient of this linear relationship as **neuron empirical gradient (NEG)**, and introduce **NeurGrad**, an accurate and efficient method for computing NEG. NeurGrad enables quantitative analysis of all neurons in PLMs, advancing our understanding of neurons' controllability. Furthermore, we explore NEG's ability to represent language skills across diverse prompts via skill neuron probing. Experiments on MCEval8k, a multi-choice knowledge benchmark spanning various genres, validate NEG's representational ability. **The data and code are released.**[1]

1 Introduction

Transformer (Vaswani et al., 2017)-based pre-trained language models (PLMs) exhibit a remarkable ability to possess human knowledge, drawing significant attention to understand their internal mechanisms. While prior studies have highlighted the crucial role of feed-forward (FF) layers' neurons in representing diverse knowledge, including factual knowledge (Dai et al., 2022; Yu and Ananiadou, 2024) and general language skills (Wang et al., 2022; Tan et al., 2024), they face several challenges. First, the existing methods primarily focus on ranking neurons to identify important ones (Dai et al., 2022; Meng et al., 2022; Yu and Ananiadou, 2024), yet they lack a direct and global quantitative measurement between neuron activations and model output, limiting their applicability to neuron-level adjustment applications, like knowledge editing (Zhang et al., 2024). Second, the existing neuron importance-ranking methods are computationally expensive, requiring either multiple activation modification (Dai et al., 2022; Meng et al., 2022; Goldowsky-Dill et al., 2023) or extensive tensor calculation (Yu and Ananiadou, 2024), making inefficient to analyze all neurons on large models across diverse prompts, thereby hindering a global understanding of their roles.

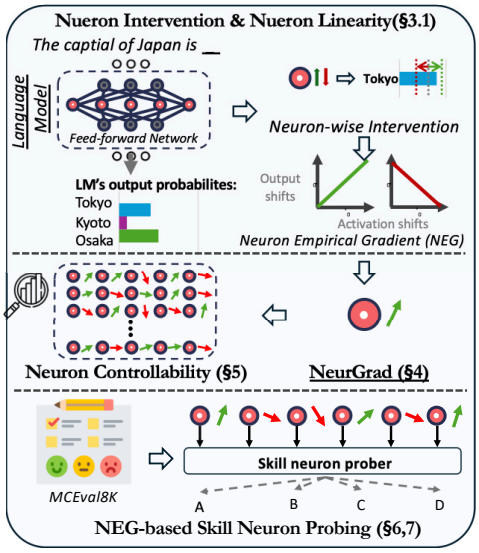


Figure 1: Overview of our contributions: i) observation of the neuron linearity, ii) an efficient method, NeurGrad, to quantify the linearity, iii) skill neuron probing on MCEval8K to confirm NEG captures language skills.

Initially, our study conducts a quantitative analysis to understand **how neuron activations determine model outputs (RQ1)**. We investigate this by gradually modifying activations of randomly sampled neurons and observing the resulting changes in the probabilities of target tokens for correct

[1] https://anonymous.4open.science/r/NeurGrad

knowledge (hereafter, *output shift*), using MyriadLAMA (Zhao et al., 2024), a factual knowledge probing dataset on nine PLMs, including large language models (LLMs) like Llama2-70B (§ 3). Notably, we discover that, within a certain range of activations, shifts in neuron activations (hereafter, **activation shifts**) have a linear relationship with the output shift. We call and quantify the gradient of this linear relationship as *neuron empirical gradient (NEG)*, enabling quantitative neuron analysis.

Next, we ask: **can we precisely adjust PLMs' output probabilities by shifting neuron activations? (RQ2)**. As estimating NEG requires extensive inference over PLMs, we first introduce NeurGrad (§ 4), an accurate yet efficient method to calculate NEG for the single neuron, grounded in the empirical observation that computational gradients(4.1) strongly correlate with NEG magnitudes but weakly with their directions. We validate its performance using MyriadLAMA by measuring their relative relationship to ground-truth NEGs. NeurGrad even shows superiority over baselines, including existing neuron-ranking methods (Dai et al., 2022; Yu and Ananiadou, 2024). Leveraging NeurGrad, we further investigate neuron controllability over multiple neurons through multi-neuron intervention (§ 5). We observe that NEG can be accumulated with multiple neurons, while the correlation diminishes as more neurons are involved or larger activation shifts are applied.

Finally, we investigate **whether NEG can represent general language skills associated with diverse prompts rather than specific factual prompts (RQ3)**. To answer this, we conduct skill neuron probing that identifies neurons associated with language skills using NEGs as inputs. While prior studies conducted skill neuron probing, they focus on using neuron activation to build probers, leaving the representational ability of NEG unexamined (Wang et al., 2022; Song et al., 2024). Furthermore, as they only used multi-choice datasets with limited language skills, we introduce *MCEval8K*, a multi-choice knowledge evaluation benchmark spanning six genres and 22 tasks for comprehensive LLMs evaluation. Our experiment discovers NEG's ability to represent diverse language skills, providing a basis for future language skill-oriented neuron-based model adjustments.

Our contributions (Figrue 1) are as follows:

- We confirm that activation shifts are linearly correlated to output shifts within a specific activation shift range, referred to and quantified as **neuron empirical gradients**. (§ 3)
- We present **NeurGrad**, an efficient method for estimating neuron empirical gradients(§ 4), and conduct analysis to deepen the understanding of neuron controllability (§ 5).
- We find NEG's ability to express language skills by **skill neuron probing** (§ 6,§ 7).
- We built **MCEval8K**, a multi-choice benchmark spanning six knowledge genres and 22 language-understanding tasks (§ 6.2, § F).

2 Related Work

Neuron-level knowledge attribution methods. Existing studies built the connections between knowledge and neurons by measuring the importance scores of neurons to the model prediction on the target knowledge; some of them rely on causal observations of how knowledge changes with neuron modifications (Meng et al., 2022; Geva et al., 2021; Wang et al., 2024; Chen et al., 2023), while others use extensive tensor calculations to estimate neuron contributions (Geva et al., 2022; Lee et al., 2024; Yu and Ananiadou, 2024). However, these methods are computationally expensive, limiting their scalability for large-scale probing across diverse prompts in LLMs. Moreover, they often capture relative relationships (*e.g.*, neuron rankings) rather than the direct, quantitative relationships between specific neurons and model outputs, restricting neurons' utility in scenarios requiring precise, such as knowledge editing (Zhang et al., 2024) and bias mitigation (Gallegos et al., 2024) on LLMs. Although gradient-based approaches (Lundstrom et al., 2022; Dai et al., 2022) seek to integrate gradients through neuron intervention, they also suffer from high computational costs.

Skill neuron probing. Neurons in FF layers show the ability to convey specific skills so that using the neuron activations solely can tackle the language tasks, which these neurons are referred to as skill neurons (Wang et al., 2022; Song et al., 2024). Existing studies found neurons can express semantic skills like sentiment classification (Wang et al., 2022; Song et al., 2024) or complex skills, such as style transfer (Lai et al., 2024) and translation (Tan et al., 2024). However, previous research viewed neuron activations as knowledge indicators, and the representational ability of neuron gradients to language skills is not examined, hindering the application of neuron-level model adjustment.

3 Neuron Linearity to Model Output

This section empirically answers how neurons in PLMs' FF layers influence model output by observing the resulting change in output tokens' probabilities for fine-grained neuron-level interventions.

3.1 Neuron-level Intervention Experiment

Models. To make the analysis result general, we experiment with masked and causal LMs of varied sizes and learning strategies. For masked LMs, we use three BERT (Vaswani et al., 2017; Devlin et al., 2019) models: $\text{BERT}_{\text{base}}$, $\text{BERT}_{\text{large}}$, and BERT_{wwm}. We construct masked prompts and let the model predict the masked token. For causal LMs, we examine diverse open-source LLMs, including instruction-tuned and pre-trained LLMs. The instruction-tuned LLMs are selected from two families: Llama2 (Touvron et al., 2023) with sizes of 7B, 13B, and 70B, and Phi3-mini (Abdin et al., 2024), a model with 3.8B parameters. The pre-trained LLMs are Llama2 models with 7B and 13B parameters. Following the zero-shot prompt setting in Zhao et al. (2024), we instruct them to generate single-token answers. See § E for model details.

Dataset. We utilize a multi-prompt knowledge probing dataset, MyriadLAMA (Zhao et al., 2024), for neuron intervention. MyriadLAMA offers diverse prompts per fact, reducing the influence of specific linguistic expressions on probing results. We focus on single-token probing, where the target answer is represented by a single token. For each PLM, we randomly sample 1000 prompts from MyriadLAMA, where the model correctly predicts the target token. Due to differences in tokenizers, the probing prompts may vary across PLMs, and the results are not comparable across the PLMs.

Neuron-wise intervention. We conduct a neuron-wise intervention to analyze how activation shift affects model outputs. Specifically, we alter the neuron activations within a range of [-10, 10] with a step size of 0.2 to observe the resulting changes in target token output probabilities. To establish a global observation over all neurons while minimizing computational cost, since assessing a single neuron's effect on one token for one prompt requires 100 inferences, we conduct neuron interventions on randomly sampled neurons.

Result and Analysis To understand how output shifts respond to neuron activation shifts, we calculate the Pearson correlation (hereafter, **Corr**) between the activation shift ranges and the output shifts of the correct tokens, considering only the absolute Corr values. The absolute Corr is averaged over 10 prompts randomly sampled from 1000 prompts to reduce the computational cost, each with 1000 randomly sampled neurons given a specific shift range (x-axis). Figure 2 shows that all activation shifts and output shifts have a nearly linear, strong Corr, even with a broad range[2] of ± 10. Figure 2 also indicates that activation shifts tend to show stronger Corr with output shifts at smaller shift ranges, consistent across all models. It indicates that the shifting neuron activation with a specific value can have a predictable result in output shifts.

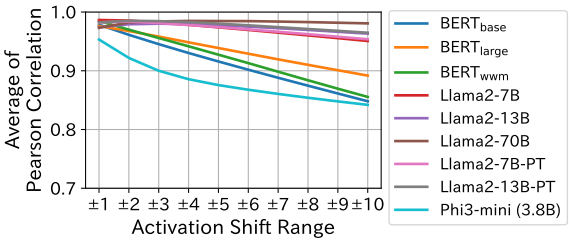


Figure 2: Average absolute Corr between activation shifts and output shifts.

Noted that as all PLMs show similar behaviors, we consider the BERT PLMs and three instruction-tuned Llama2 LLMs for the following analysis.

3.2 Neuron Linearity

Based on the above discussion, we ask: **do neurons generally exhibit linearity to model output?** We first define neurons as possessing **linearity** if their correlation (Corr) is at least 0.95[3] within a shift range of ± 2 from the observation in Figure 2.

We present a quantitative analysis of the prevalence of neuron linearity across different prompts and Transformer layers. Specifically, we report the ratio of neurons exhibiting linearity from 1000 prompts paired with 100 neurons.[4] The ratios of 'linear' neurons in $\text{BERT}_{\text{base/large/wwm}}$ models are 0.9565/0.8756/0.9564, respectively, and the ratios for Llama-7B/13B/70B are 0.9387/0.9677/0.9208, indicating that majority of neurons exhibit linearity. We analyze the generality of linear neurons across layers and prompts, revealing their widespread

[2] ± 10 is an appropriate range considering the distribution of activations; see § B.1 for details.

[3] Since linearity lacks a strict definition, we use Corr>=0.95 to indicate a strong linear relationship.

[4] We only chose 200 prompts and 100 neurons for Llama2-70B due to the large model size.

presence (§ A). Furthermore, we term the linearity direction as **polarity**, where neurons are *positive/negative* if increasing/decreasing their activations enhances the target output probabilities.

Neuron empirical gradient. We quantify the neurons' linearity and polarity by the gradient of the linear relationship between activation shift and output shifts, which we term neuron empirical gradient (**NEG**). We calculate NEG as follows: *we fit a zero-intercept linear regression between activation shifts and output shifts acquired through neuron intervention, and the regression coefficient is identified as the NEG for a specific neuron, prompt, and token.*

4 NeurGrad for NEG Estimation

Efficiently and accurately computing NEG is crucial for quantitative analysis of neuron-level interpretability in PLMs. However, quantifying NEG through neuron-wise intervention is impractical due to the high computational cost. While prior studies proposed knowledge attribution methods that measure neurons' importance in influencing model output, these methods require either extensive calculation or can only estimate the neurons' relative importance but cannot directly estimate the NEG (Dai et al., 2022; Geva et al., 2022; Meng et al., 2022; Yu and Ananiadou, 2024).

4.1 NeurGrad

In this section, we propose NeurGrad, an accurate yet efficient NEG estimation method, to facilitate further analysis using NEG. The proposal of NeurGrad comes from the preliminary investigations into using computational gradients[7] (hereafter, **CG**) to measure the NEG. We observe that the Corr between CGs' absolute values and NEGs is high, but the signs of NEGs are decided by CG and neuron activations. Specifically, we collect ground-truth NEGs from 1000 prompts with 100 neurons per prompt, with a shift range of ± 2 on six PLMs, BERT families and three Llama2 instruction-tuned LLMs. After collecting CG on these prompt-neuron pairs, the data reveal a low Corr (-0.429 on average) between the CG and NEG, but a high Corr (0.961) between their absolute values. Moreover, the sign of NEG correlates with the signs of both activation and CG. Based on these findings, we propose **NeurGrad**:

$$\bar{G_E} = \text{CG} \times \text{sign}(A), \quad (1)$$

[7]Computational gradient refers to the gradient computed from the computational graph through backpropagation.

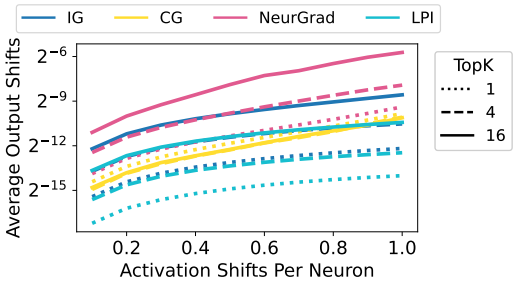


Figure 3: Comparison of neuron attribution methods in token probability enhancement. X-axis: activation shifts of selected neurons; Y-axis: average output shifts over 1000 factual prompts.

where $\bar{G_E}$, A, and $\text{sign}(A)$ represents the estimated NEG, activation, and sign of A (1 for $A > 0$ and -1 for $A < 0$), respectively.

4.2 NEG Estimation Evaluation

We evaluate NeurGrad's effectiveness in estimating NEG. We use the same setting as CG evaluation above to collect the ground-truth NEGs. Our experiment involves three baselines, including two gradient estimation methods, CG and integrated gradients (IG) (Dai et al., 2022), and one logit-based knowledge attribution method (Yu and Ananiadou, 2024). IG intervenes with neurons in small step sizes multiple times to simulate the NEG. The Log-Probability-Increase (LPI) (Yu and Ananiadou, 2024) identifies important neurons by estimating increased probabilities with specific neurons.[8]

We evaluate how well these methods calculate NEG with two metrics: **Corr** and **mean absolute error (MAE)**. Table 1 (left) reports the Corr between the estimated and ground-truth NEGs that can assess their ability to capture the neurons' relative relationship, such as ordering. NeurGrad consistently achieves high Corr over the six PLMs, outperforming all other methods. Moreover, Table 1 (right) reports the Corr the MAE that quantifies the accuracy of NEGs' calculation. It is observed that NeurGrad can largely reduce the estimation error to the true NEG, revealing its ability to precisely calculate NEGs. Finally, the average running time (Table 1 (bottom)) demonstrates NeurGrad's efficiency compared to other methods. The time estimation uses Llama2-7B with an NVIDIA RTX A6000 GPU.

[8]We did not consider other causal-tracing-based methods (Meng et al., 2022) due to their high computational cost, which conflicts with our efficiency goal.

	Corr				MAE			
	CG	**IG**	**LPI**	**NeurGrad**	**CG**	**IG**	**LPI**	**NeurGrad**
$\text{BERT}_{\text{base}}$	-.8909	.7360	-[5]	**.9998**	6.1e-03	3.0e-03	-	**2.6e-05**
$\text{BERT}_{\text{large}}$	-.9307	.7167	-	**.9958**	4.6e-03	2.1e-03	-	**1.9e-04**
BERT_{wwm}	-.8914	.8584	-	**.9989**	4.5e-03	2.2e-03	-	**2.3e-05**
Llama2-7B	.3020	.5383	.6469	**.8136**	1.7e-06	**1.2e-06**	1.9e-03	1.3e-06
Llama2-13B	-.1973	.7261	.1141	**.9965**	2.1e-06	1.5e-06	4.3e-04	**6.0e-08**
Llama2-70B	.0283	n/a[6]	n/a	**1.000**	5.9e-07	n/a	n/a	**2.1e-09**
Avg. Runtime (Llama2-7B)	**0.149s**	19.349s	6.086s	0.161s	Same as left			

Table 1: Evaluation of NeurGrad and baselines in calculating NEGs, including two metrics: Corr and MAE.

	$\text{BERT}_{\text{base/large/wwm}}$	**Llama2-7/13/70B**
Pos. ratio	.5019/.5008/.4996	.4604/.4664/.4484
Neg. ratio	.4981/.4992/.5004	.4592/.4660/.4480

Table 2: The pos/neg neuron ratios over 1000 prompts.

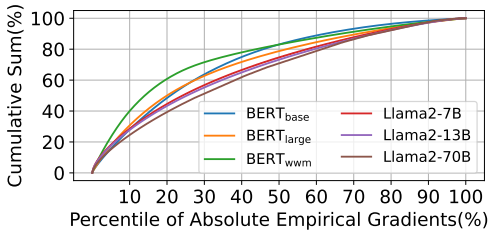


Figure 4: Cumulative distribution of NEG magnitudes. (**X-axis**: the percentiles of NEG magnitudes; **Y-axis**: the cumulative contribution of neurons to the total sum).

4.3 Knowledge Attribution Evaluation

Finally, we evaluate NeurGrad's ability to locate important neurons. Specifically, for 1000 prompts, we identify the top-K neurons ($K = 1, 4, 16$) with the highest attribution scores using CG, IG, LPI, and NeurGrad. We enhance neuron activations by shifting positive neurons from 0.1 to 1 and negative neurons from -0.1 to -1, both in increments of 0.1. Figure 3 reports the output shifts of target tokens on Llama2-7b, with different activation shifts and top-k neurons selected from different attribution methods. Figure 3 demonstrates that NeurGrad outperforms baselines in attribution accuracy. This superiority stems from NeurGrad's precise measurement of NEG and its inclusion of negative neurons in intervention, unlike IG and LPI, which only consider positive neurons despite their equal ratio to negative neurons (Table 2). See details in § C.

5 Understanding Neurons' Controllability

This section deepens our understanding of **neuron controllability**: the ability to precisely adjust PLM output probabilities by modifying neuron activations, using NEGs estimated by NeurGrad.

5.1 How Are NEGs Distributed?

Do neurons have polarity preference? Table 2 reports the ratios of positive/negative neurons across 1000 prompts in the six PLMs, showing that the numbers of positive/negative neurons are nearly equivalent. It indicates that PLMs show no preference for their neurons' polarities, suggesting that enhancing or suppressing neurons should be guided by their gradient polarity rather than merely increasing or decreasing their activations (Dai et al., 2022). See detailed distribution analysis in § B.2,B.3.

Does only a few neurons exhibit strong gradients? Figure 4 shows the cumulative distribution of NEG for all neurons in PLMs. Rising curves are steady and do not converge until all neurons are present, indicating that most neurons can affect the model's output probabilities.

5.2 Does Linearity Hold for Multi-neuron?

We conduct multi-neuron intervention experiments to investigate: can output shifts be predicted when modifying multiple neuron activations? We randomly sample N neurons and enhance them by shifting their activations according to their polarities measured by NeurGrad, in which positive and negative activation shifts are applied to positive/negative neurons. We experiment on $\text{BERT}_{\text{base}}$ and Llama2-7B using neuron sizes of 2^N ($0 \leq N \leq 12$) across 1000 prompts.

Figure 5 shows the average Corr across all prompt-neuron pairs with an enhancement range of [0, 0.5] and a step size of 0.01. We observe that even Corr decreases as more neurons are involved due to neuron interactions, a strong cor-

[7]We follow code released in Yu and Ananiadou (2024) and only Llama2 LLMs are supported.

[8]Due to the high memory cost of IG and LPI, we preclude Llama2-70B experiments on these methods.

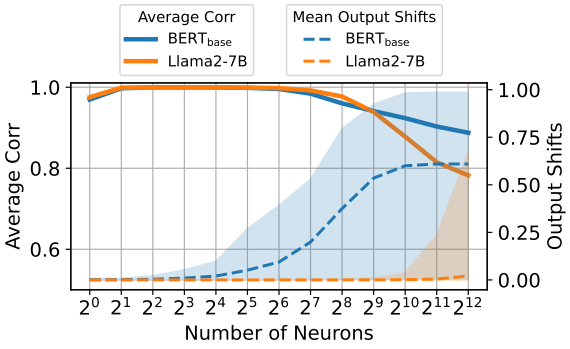


Figure 5: Multi-neuron enhancement with range [0,0.5] with different number of neurons.

relation ($\geq$0.7) remains even with 4096 neurons in both PLMs. Additionally, Figure 5 (right y-axis) shows that involving more neurons can cause larger output shifts, suggesting that the output shifts caused by individual neurons can be accumulated. Note that despite the small average output shifts in Llama2 due to the small NEG magnitudes per neuron (see § B.2 for details), significant probability changes can still be observed when more than 1024 neurons are involved in Llama2-7B models with specific neuron combinations. Moreover, our analysis with larger enhancement ranges reveals consistent trends: increasing the shift range reduces Corr, making output shifts less predictable. See § D for experiments with different ranges.

The analysis above demonstrates that modifying neuron activations, as guided by NeurGrad, enables partial prediction of output shifts. However, the number of modified neurons and the range of modifications require careful examination.

5.3 Local Linear Approximation Hypothesis

We present the **Local Linearity Approximation Hypothesis** to explain neuron linearity, based on three key observations: (1) expanding the shift range reduces Corr (Figure 2); (2) PLMs with more parameters diminish the importance of individual neurons, leading to higher Corr (Figure 2); (3) involving more neurons weakens linearity (Figure 5).

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a PLM, where $x \in \mathbb{R}^n$ represents neuron activations and $f(x)$ represents the output token probabilities. Within a constrained region, the influence of a single neuron x_i can be locally approximated as: $f(x_i + \delta e) \approx f(x_i) + \frac{\partial f}{\partial x_i}\delta$, for small δ, where e_i is the unit basis vector. This follows from the first-order Taylor expansion, reflecting the model's local differentiability with respect to neuron activations.

6 Skill Neuron Probing using NeurGard

While NEG and NeurGrad enable neuron-level modifications to model outputs, the variability of neurons' NEG values across different prompts limits their ability to support modifications for specific types of knowledge, such as language skills. These skills often involve handling a range of prompts that require linguistic diversity. This section explores **whether NEG can effectively capture general language skills linked to diverse prompts through skill neuron probing.** Skill neuron probing seeks to identify neurons that encode the ability to solve language tasks. Note that prior work (Wang et al., 2022; Song et al., 2024) explored the effectiveness of using neuron activations for this purpose, yet we focus on neurons' NEG.

6.1 Task Definition

Following Wang et al. (2022), we formulate the skill neuron probing task as follows. A dataset conveying specific language skills $\mathcal{D}$ consists of language sequence pairs, including knowledge inquiries $\mathcal{Q} = \{q_1, ..., q_{|\mathcal{T}|}\}$ and answer sequences $\mathcal{A} = \{a_1, ..., a_{|\mathcal{T}|}\}$, where arbitrary a_i belongs to the answer candidate set $\hat{\mathcal{A}}_{\text{cands}}$. For example, in the sentiment classification task, Q is the documents set, and A is the ground-truth sentiment labels. We then build classifiers that take behaviors of arbitrary neuron subset $\mathcal{N}_s \subseteq \mathcal{N}$ as features to indicate the correct answer sequences a_i for the knowledge inquiry q_i. $\mathcal{N}$ refers to all the neurons.[9]

Our skill neuron prober aims to find $\mathcal{N}_s^*$ that can achieve optimal accuracy over the target dataset $\mathcal{D}$.

$$\mathcal{N}_s^* = \arg\max_{\mathcal{N}_s \subseteq \mathcal{N}} \text{Acc}(f(\mathcal{N}_s), D) \quad (2)$$

$$\text{Acc}(f(\mathcal{N}_s), D) = \frac{1}{|D|}\sum_{i=1}^{|D|} \mathbb{1}[f(\mathcal{N}_s, q_i) = a_i]. \quad (3)$$

Here, $f(\mathcal{N}_s, q_i)$ is the output of the classifier F using the neuron subset $\mathcal{N}_s$ for the prompt q_i. $\mathbb{1}[X = Y]$ is an indicator function that equals 1 if X matches Y, and 0 otherwise.

6.2 Evaluation Benchmark: MCEval8K

This section, we create a multi-choice knowledge evaluation benchmark-MCEval8K for skill neuron probing. As skill neuron probing requires a fixed target token, previous studies (Wang et al., 2022;

[9]We focus on intermediate outputs (neurons) of FF layers.

Song et al., 2024) relied on multi-choice datasets that forces PLMs to generate a single-token option label (A, B, etc.) named for category labels (A: positive, B: negative, etc.). While we adopt a similar setup, earlier work focused on small PLMs with datasets limited to basic language understanding tasks, which are insufficient for evaluating LLMs.

To address this, we introduce MCEval8K, a diverse multi-choice benchmark covering 22 tasks across six genres, designed to assess a wide range of language skills, incorporating most datasets from previous studies. Since tasks vary in different sizes, with some, such as cLang-8 (Rothe et al., 2021; Mizumoto et al., 2011), containing millions of data points, we standardize the evaluation by limiting each task to 8K queries.[10] It minimizes unnecessary computational costs while ensuring consistency across tasks. We also ensure the number of ground-truth options per task is balanced to eliminate bias introduced by imbalanced classification. The skill genres, tasks, and datasets information are shown below (detailed in § F).

Linguistic: Part-of-speech tagging (POS), phrase-chunking (CHUNK), named entity recognition (NER), and grammatical error detection (GED). **Content classification:** Sentiment (IMDB), topic classification (Agnews), and Amazon reviews with numerical labels (Amazon). **Natural language inference (NLI):** textual entailment (MNLI), paraphrase identification (PAWS), and grounded commonsense inference (SWAG). **Factuality:** Fact-checking (FEVER), factual knowledge probing (MyriadLAMA), commonsense knowledge (CSQA), and temporary facts probing (TempLAMA). **Self-reflection:** Examine PLMs' internal status, including hallucination (HaluEval), toxicity (Toxic), and stereotype (Stereoset) detections. **Multilinguality:** We select tasks with multilingual queries, including language identification (LTI), multilingual POS-tagging on Universal Dependencies (M-POS), Amazon review classification (M-Amazon), factual knowledge probing (mLAMA) and textual entailment (XNLI).

7 NEG as Knowledge Feature

We train skill neuron probers based on NeurGrad's estimated gradients to investigate whether and how NEG encodes language knowledge.

[10] Only the Stereoset task has fewer than 8K queries due to the limited size of the original dataset.

7.1 Gradient-based Skill Neuron Prober

For each task dataset $\mathcal{D}$, we split it into: training set $\mathcal{D}_{\text{train}}$ to train the classifiers, validation set $\mathcal{D}_{\text{valid}}$ to decide hyperparameters, and test set $\mathcal{D}_{\text{test}}$ for evaluation, with the ratio of 6:1:1. We train three probers with different designs for comparison.

Polarity-based majority vote (Polar-prober) adopts a simple majority-vote classifier, taking each neuron in $\mathcal{N}_s$ as one voter. A polarity-based classifier leverages the polarity of neurons (positive or negative) as features for classification. Given $\mathcal{D}_{\text{train}} = \{(q_i, a_i)\}$ and any neuron $n_k \in \mathcal{N}$, we identify the polarity as feature $\mathbf{x}^{n_k}_{q_i,a_i}$ for each (q_i, a_i) pair. For each n_k, we calculate the ratio of being positive and negative across all $|\mathcal{D}_{\text{train}}|$ examples and the dominant polarity is identified as their global polarity $\bar{\mathbf{x}}^{n_k}$. Neurons with more consistent polarity are ranked higher.

To make prediction of q_i, we measure all polarities of $\mathbf{x}^{n_k}_{q_i,a_j}$, where $a_j \in \hat{\mathcal{A}}_{\text{cands}}, n_j \in \mathcal{N}^*_s$. The prediction of each q_i is made as follows:

$$f(\mathcal{N}^*_s, q_i) = \arg\max_{a_j \in \hat{\mathcal{A}}_{\text{cands}}} \sum_{n_k \in \mathcal{N}^*_s} \mathbb{1}[\mathbf{x}^{n_k}_{q_i,a_j} = \bar{\mathbf{x}}^{n_k}] \quad (4)$$

We identify the optimal size of $\mathcal{N}^*_s$ with $\mathcal{D}_{\text{valid}}$.

Magnitude-based majority vote (Magn-prober) utilizes gradient magnitudes as features for a majority-vote classifier. During training, for a specific q_i and n_k, we compare the gradients between $a \in \hat{\mathcal{A}}_{\text{cands}}$. Neurons that consistently exhibit the largest or smallest gradients for the ground truth a_i compared to other candidates are used as skill indicators. We record each neuron's preference for being either the largest or smallest. Neurons exhibiting more consistent behavior are assigned higher importance and identified as skill neurons. During inference, similar to Eq. 4, the prediction is made by selecting a_j that satisfies the majority of $n_k \in \mathcal{N}^*_s$. This prober is designed to compare against the polarity-based prober, aiming to examine whether NEGs' magnitude can bring additional skill information compared to polarity.

7.2 Experimental Setup

Dataset & Prompt Since our probing method restricts the output sequence length to one, we carefully craft instructions and options for all datasets in MCEval8K through human effort. We evaluate both zero-shot and few-shot settings, ensuring in few-shot experiments that all candidate tokens

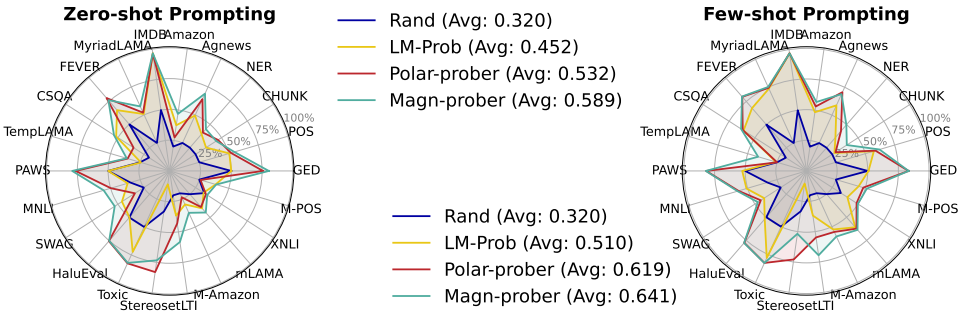


Figure 6: MCEval8K accuracies on Llama2-7B across tasks in zero-shot and few-shot settings, reported for Rand (random guess), TProb (token probability), and two proposed probers. Legends show average accuracies.

appear once in the demonstrations to prevent majority label bias (Zhao et al., 2021). See § J for the designed instructions for all tasks.

Prober During validation, we select the optimal neuron size for majority-vote probers from $2^n (0 \leq n \leq 13)$. See § G for detailed prober settings.

Model We perform skill neuron probing on Llama2-7B using three probers, all datasets in MCEval8K, and the full training set (6000) per task. For Llama2-70B, due to high cost, we probe one dataset per genre—NER, Agnews, PAWS, CSQA, HaluEval, and mLAMA—using 1024 training examples and only train major-vote probers.

7.3 Result and Analysis

Skill neuron-based classifier accuracy is compared to two baselines: random guessing (**Rand**), and answer token probability-based classification (**LM-Prob**) which selects the candidate token with the highest probability as the prediction, serving as a benchmark for the LLMs' prompting performance.

Empirical gradients encode language skills. Figure 6 shows accuracies for all tasks in MCEval8K using the Llama2-7B, with both zero- and few-shot settings. The results demonstrate that LM-Prob outperforms Rand, indicating that Llama2-7B is capable of understanding instructions and recalling skills from its parameters. We also confirm the effectiveness of our skill neuron probers in addressing language tasks. The two simple major-vote classifiers outperform LM-Prob in both settings. The per-task classification accuracies in Figure 6 show that skill neurons effectively represent diverse language skills, achieving consistently high results across tasks. See Table 7,8 for accuracy values.

Larger PLMs excel in skill recall. Table 3 compares LM-Prob and Magn-prober across six tasks in the few-shot setting between Llama-7B and -70B. Llama-70B outperforms Llama-7B in both LM-Prob and skill neuron probing. However, the difference between LM-Prob and Magn-prober is smaller in Llama2-70B than in Llama2-7B, indicating the large model's strong ability to recall knowledge from its parameters. See § H for further analysis on the properties of skill neurons: efficiency, generality, and inclusivity.

Tasks	Llama2-7B		Llama2-70B	
	LM-Prob	Magn-Prober	LM-Prob	Magn-Prober
NER	.3610	.4980	.7900	**.8170**
Agnews	.5880	.7020	.7630	**.8240**
PAWS	.5240	.8150	.7790	**.8460**
CSQA	.6100	.6390	.7540	**.7630**
HaluEval	.5200	.7830	.7530	**.8250**
mLAMA	.6080	.6370	.7430	**.7600**

Table 3: Accuracies of 6 tasks on Llama2-7B and -70B.

8 Conclusions

This is the first study to establish a global quantitative measurement between neurons and model output, laying a foundation for precise PLM output control by modifying neurons. Our study uncovers a linear relationship between neurons in FF layers and model outputs through neuron intervention experiments. We call and quantify this linearity by "neuron empirical gradients" and propose NeurGrad, an accurate yet effective NEG estimation method. Building on NeurGrad, we deepen the understanding of neuron controllability. Finally, we demonstrate NEG's representational ability of general language skills associated with diverse prompts through skill neuron probing experiments.

As the future work, we will investigate the neuron modification methods that can main both the preciseness and strength to apply to applications like knowledge editing and bias mitigation.

9 Limitations

Our research establishes a framework for quantitatively measuring neurons' influence on model output and demonstrates the effectiveness of empirical gradients in representing language skills, linking language skill representation to model output through neuron empirical gradients. However, the potential for achieving skill-level model output adjustment by tuning neuron values remains unexplored. Directly adjusting neuron values could offer a more efficient alternative to traditional weight-level tuning methods. Furthermore, we also plan to examine NEG's representational ability on language generation tasks. This approach may enable dynamic behavior modification without altering the underlying parameters of LLMs, potentially reducing computational costs and enabling more flexible model adaptation.

Furthermore, our discussion on neuron linearity and empirical gradient measurements is currently confined to single-token probing with factual prompts. In the future, we plan to expand our experiments to include prompts from diverse domains and investigate neuron attribution methods for multi-token contexts, aiming to support broader applications in generative language tasks.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Ralf D Brown. 2014. Non-linear mapping for improved identification of 1300+ languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–632.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. *Preprint*, arXiv:2308.13198.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge. Kaggle.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed.

2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *Preprint*, arXiv:2304.05969.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *Preprint*, arXiv:2403.03952.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: investigating knowledge in multilingual pretrained language models. *CoRR*, abs/2102.00894. To appear in EACL2021.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *Preprint*, arXiv:2401.01967.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Ngee Chia Tai, Ayu Purwarianti, Sebastian Ruder, William Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng-Xin Yong, and Samuel Cahyawijaya. 2024. Seacrowd: A multilingual multimodal data hub and benchmark suite for southeast asian languages. *arXiv preprint arXiv: 2406.10118*.

Daniel Lundstrom, Tianjian Huang, and Meisam Razaviyayn. 2022. A rigorous study of integrated gradients method and extensions to internal neuron attributions. *Preprint*, arXiv:2202.11912.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proc. of 5th International Joint Conference on Natural Language Processing*, pages 147–155.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A Simple

Recipe for Multilingual Grammatical Error Correction. In *Proc. of ACL-IJCNLP*.

Ran Song, Shizhu He, Shuting Jiang, Yantuan Xian, Shengxiang Gao, Kang Liu, and Zhengtao Yu. 2024. Does large language model contain task-specific neurons? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7101–7113, Miami, Florida, USA. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6506–6527, Miami, Florida, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. 2024. Unveiling factual recall behaviors of large language models through knowledge neurons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7402, Miami, Florida, USA. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2024. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280, Miami, Florida, USA. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ningyu Zhang, Yunzhi Yao, and Shumin Deng. 2024. Knowledge editing for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 33–41, Torino, Italia. ELRA and ICCL.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. What matters in memorizing and recalling facts? multifaceted benchmarks for knowledge probing in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13186–13214, Miami, Florida, USA. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In

Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Generality of Neuron Linearity

In this section, we provide additional evidence to verify that linearity is a general property for neurons in LLMs. Specifically, we want to verify whether the linear neurons exist widely across different Transformer feed-forward layers and within different prompts. We use the metrics of layer generality (**LG**) and prompt generality (**PG**) to measure the prevalence of their existence. Intuitively, we can consider a simplified problem as follows: suppose we have many colored balls (green, blue, ...) and 10 bins, and if we want to verify whether the blue ball has "generality," it means (1) **high coverage**: the blue ball exists in most of the bins; (2) **even distribution**: the number of blue balls in each bin hardly differs from others. For our neuron generality, the "balls" are the "linear neurons," and the "bins" refer to either "feed-forward layers" (for **LG**) or "different prompt" (for **PG**). To address these two aspects simultaneously, we define **LG** and **PG** as follows:

$$\textbf{LG} \triangleq \text{coverage}_{\text{layer}} \times \text{distribution}_{\text{layer}}, \quad (5)$$

$$\textbf{PG} \triangleq \text{coverage}_{\text{prompt}} \times \text{distribution}_{\text{prompt}}, \quad (6)$$

where coverage and distribution are defined as:

$$\text{coverage}_x = \frac{\Sigma_i \mathbb{1}(\text{linear neuron exists in } x_i)}{\#\text{ of } x}, \quad (7)$$

$$\text{distribution}_x = 1 - \frac{\text{Var}(\#\text{neurons in } x)}{\text{maxVar}(\#\text{neurons in } x)}, \quad (8)$$

where x refers to either layer or prompt, $\text{maxVar}(\cdot)$ denotes the max possible variance. High coverage and distribution are desirable; a perfect generality then achieves coverage of one and distribution of one.

B Neurons' Statistics

B.1 Distribution of Neuron Activations

In this section, we analyze the distribution of neuron activations across six PLMs, illustrated in Figure 7. The models include three BERT-based PLMs and three instruction-tuned LLaMA-2 LLMs. Figure 7 reveals that most neuron activations fall within the range of ± 10. While there are still some neurons that have a value out of the range of ± 10, the number of such neurons is comparably fewer, and increasing the range linearly increases the computational cost. Considering the balance between coverage and computational cost, we finally set the intervention range as ± 10 as shown in § 3.

	Linear neuron ratio	**Prompt-wise gen.**	**Layer-wise gen.**
$\text{BERT}_{\text{base}}$	.9565	.9999	.9982
$\text{BERT}_{\text{large}}$	.8756	.9999	.9989
BERT_{wwm}	.9564	.9999	.9990
Llama2-7B	.9387	.9999	.9986
Llama2-13B	.9677	.9999	.8618
Llama2-70B	.9208	.9999	.6294

Table 4: Neuron linearity statistics. We choose 1000 prompts and their corresponding 100 neurons with top gradient magnitudes. For Llama2-70B, since the model is giant, we only chose 200 prompts and 100 neurons due to the high computational cost. The shift range is set to $\pm$ 2.

B.2 Distribution of Neuron NEGs

In this section, we report the distribution of neurons' NEG (Neuron Effect on Gradient) across the six PLMs, as illustrated in Figure 8. Similar to our discussion in § 5.1, we observe that neurons capable of altering model output are not rare. For instance, in $\text{BERT}_{\text{base}}$, over 1,000 neurons exhibit NEG magnitudes larger than 0.1. Additionally, the NEG magnitudes of neurons in LLaMA-2 LLMs are significantly smaller than those in BERT PLMs, likely due to the smaller parameter size of BERT models, which grants individual neurons greater influence over model output. Notably, all models tend to exhibit zero gradients when averaging the NEGs across all neurons.

B.3 How are neurons distributed across layers?

We examine the variation in NEG across Transformer layers to understand the distribution of neuron controllability. Figure 9 illustrates the means and variances of magnitudes of NEG across layers. The mean NEG magnitude reflects the intensity with which PLMs adjust output probabilities through neurons in a given layer, while the variance indicates how concentrated the effective neurons are within that layer. A positive Corr is observed

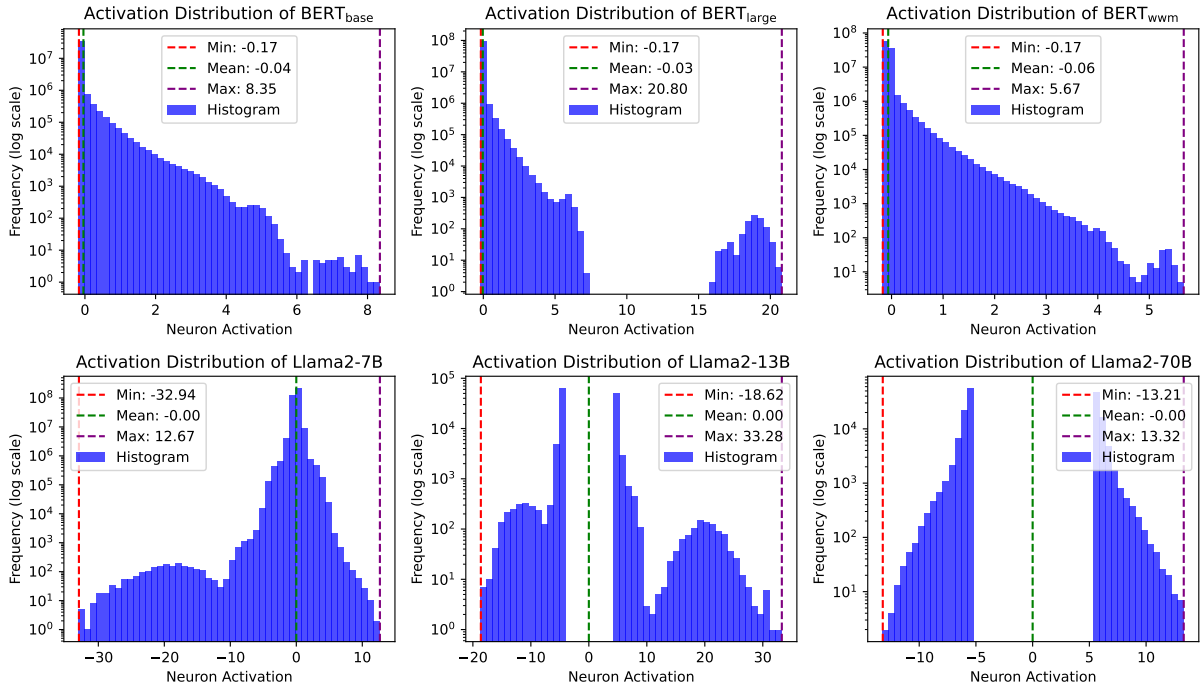


Figure 7: Histograms of neuron activations for six models across 1,000 prompts, displayed on a logarithmic y-axis. The figure includes three BERT models and three LLaMA-2 models, with each subplot showing the distribution of activations for one model.

between variances and means,[11] suggesting that as PLMs increase the intensity of gradient activity in specific layers, they also focus more on a limited subset of neurons. Specifically, in BERT models, a strong Corr is evident between layer depth and neuron controllability intensity, with deeper layers exhibiting larger gradient magnitudes, whereas Llama2 displays a distinct pattern: gradient magnitudes peak in the middle layers, decrease towards the deeper layers and then increase at the final layers. This divergence underscores the differences between the BERT and Llama2 families, emphasizing the need for case-by-case analysis in LLM mechanism investigation.

C Knowledge Attribution Evaluation: Supplementary Experiments

In this section, we report the knowledge evaluation experiments on other PLMs, including three BERT PLMs. We exclude LPI from the following experiments as LPI cannot be applied to BERT models. We follow a similar experiment setup to § 4.3.

The evaluation results are illustrated in Figure 10. NeurGrad consistently outperforms other gradient-based methods in finding the top-K important neurons. Furthermore, we can observe that output shifts made on BERT PLMs are much larger than shifts on Llama2-7B (Figure 3). This is due to the small NEG magnitudes in Llama2-7B as introduced in § B.2.

[11] The Corr between means and variances of neuron magnitudes across different layers are 0.88, 0.79, 0.87 for **BERT**$_{base/large/wwm}$, and 0.57, 0.51, 0.42 for **Llama2-7B/13B/70B**.

D Multi-neuron Intervention: Supplementary Experiments

In this section, we report the multi-neuron intervention experiments conducted with different enhancement ranges on BERT$_{base}$ and Llama2-7B, following the similar experiment setup to § 5.2. Specifically, we report the correlation between output shift and the accumulated NEGs estimated by NeurGrad with the enhancement ranges of [0, 0.1], [0, 1], [0, 1.5], and [0, 2], illustrated in Figure 11). Figure 11) demonstrates that with a larger enhancement range, involving more neurons can largely reduce the Corr, suggesting the output shift is less predictable. However, we can observe that BERT$_{base}$ consistently achieves strong Corr.(>0.8) for any scenario. While Llama2-7B is less stable as BERT$_{base}$, it can still maintain moderate positive Corr. (>0.5) for 4096 neurons with enhancement range of [0,2].

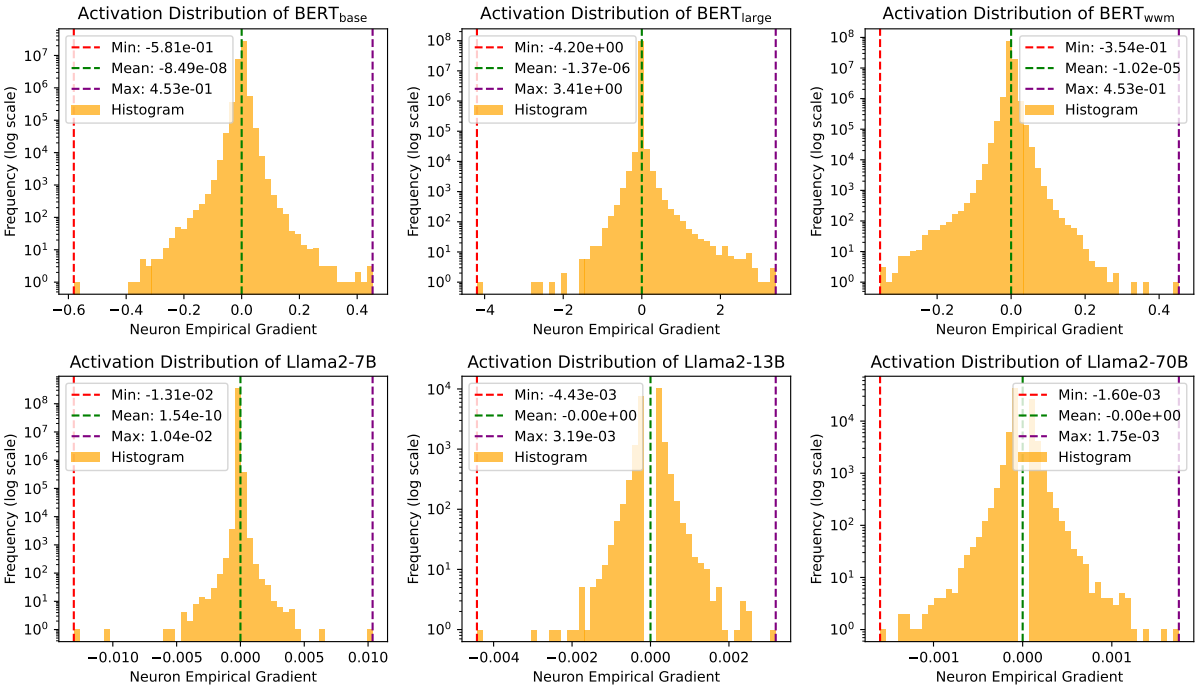


Figure 8: Histograms of neuron NEGs for six models across 1,000 prompts, displayed on a logarithmic y-axis.

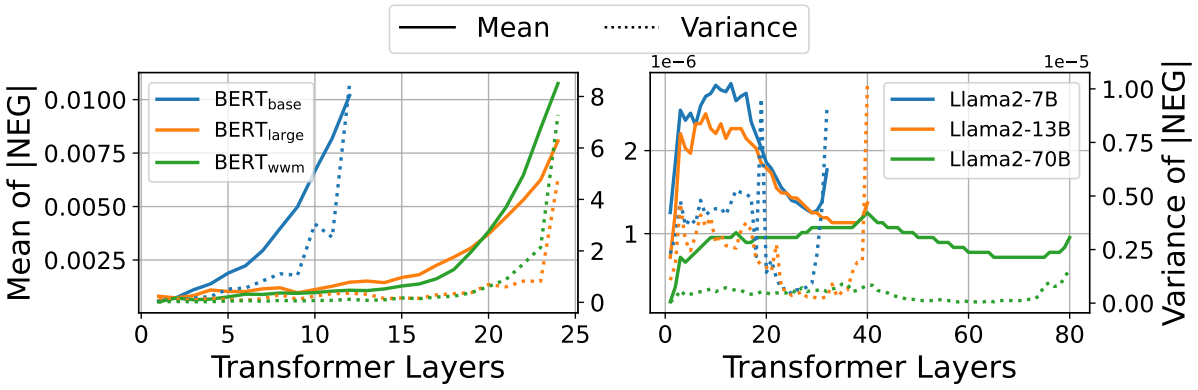


Figure 9: Means and variances of NEG magnitudes across Transformer layers on six models. The data is calculated from the average of 1000 prompts.

E Model cards

Here are the links from Hugging Face to load each model:

BERT$_{\text{base}}$: https://huggingface.co/bert-base-uncased

BERT$_{\text{large}}$: https://huggingface.co/bert-large-uncased

BERT$_{\text{wwm}}$: https://huggingface.co/bert-large-uncased-whole-word-masking

Llama2-7B: https://huggingface.co/meta-llama/Llama-2-7B-hf

Llama2-13B: https://huggingface.co/meta-llama/Llama-2-13B-hf

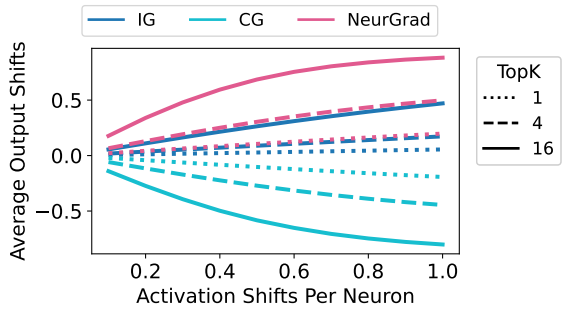
Llama2-70B: https://huggingface.co/meta-llama/Llama-2-70B-hf

Llama2-7B-PT: https://huggingface.co/meta-llama/Llama-2-7B

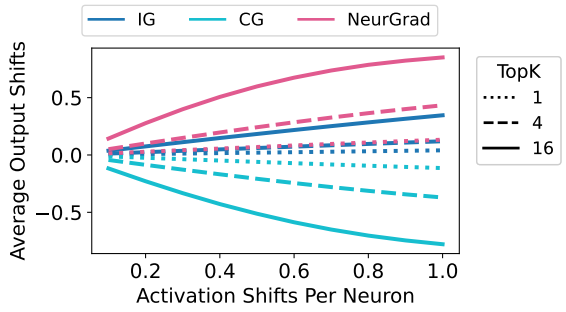
Llama2-13B-PT: https://huggingface.co/meta-llama/Llama-2-13B

Phi3-mini: https://huggingface.co/microsoft/Phi-3-mini-128k-instruct

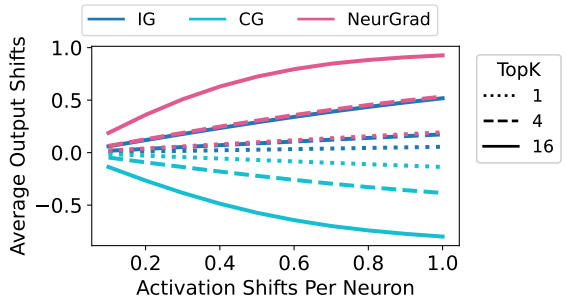
The statistics of these six PLMs, including the number of layers (#n_layers) and neurons per layer (#neurons_per_layer) are listed in Table 5.



(a) Evaluation results on $\text{BERT}_{\text{base}}$.



(b) Evaluation results on $\text{BERT}_{\text{large}}$.



(c) Evaluation results on $\text{BERT}_{\text{large}}$.

Figure 10: Knowledge attribution evaluation by comparing CG, IG, and NeurGrad on Three BERT.

F Construction of MCEval8K

The motivation behind creating MCEval8K is to establish a comprehensive benchmark that spans diverse knowledge genres and language skills. Since our goal is to facilitate skill neuron probing experiments where a single token must represent answers, we adopt a multi-choice task format. Additionally, we aim for the benchmark to be adaptable while avoiding redundancy for effective evaluation. In summary, we adhere to several guiding principles to design MCEval8K.

1. All datasets must be in multi-choice format.
2. Avoid including datasets that covey similar language skills.
3. To eliminate potential bias from imbalanced classifications, we ensure that the number of correct options is evenly distributed across all answer choices. This balance helps maintain fairness and accuracy in the analysis results.
4. We use a unified number (8000) of data to avoid high computational costs.

Model	#n_layers	#neurons_per_layer
$\text{BERT}_{\text{base}}$	12	3,072
$\text{BERT}_{\text{large}}$	24	4,096
BERT_{wwm}	24	4,096
LLama2-7B(-PT)	32	11,008
Llama2-13B(-PT)	40	13,824
Llama2-70B	80	28,672
Phi3-mini	32	8,192

Table 5: Number of Layers and Intermediate Neurons per Layer for BERT and Llama2 Models

Multi-choice format: We created MCEval8K to include six different genres with 22 tasks, which are linguistic, content classification, natural language inference (NLI), factuality, self-reflection, and multilingualism. All the genres and tasks are listed in Table 6. For datasets that are not multi-choice tasks, we create options for each inquiry following rules. These datasets include POS, CHUNK, NER, MyriadLAMA, TempLAMA, Stereoset, M-POS, and mLAMA. The rules we adhere to create options are listed below:

POS We use weighted sampling across all POS tags to select three additional tags alongside the ground-truth tag.

CHUNK The process is analogous to POS.

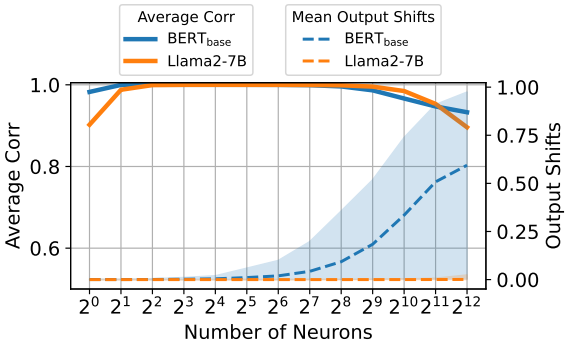
NER The process is analogous to POS.

MyriadLAMA For factual inquiries formed from $< \text{sub}_i, \text{rel}_j >$, we collect all objects that appear as the target of the rel_j within the dataset and perform sampling to select three additional objects alongside the ground-truth tag.

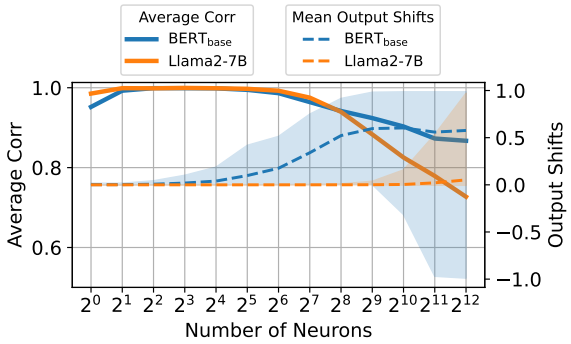
TempLAMA We randomly sample three additional candidate years from the range 2009 to 2020, alongside the ground-truth tag.

M-POS The process is similar to POS, applied separately for each language.

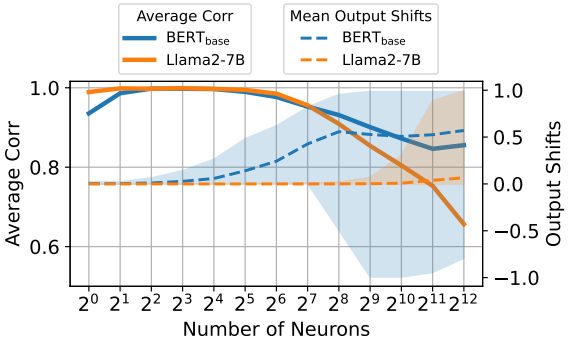
mLAMA The process is similar to MyriadLAMA, applied separately for each language.



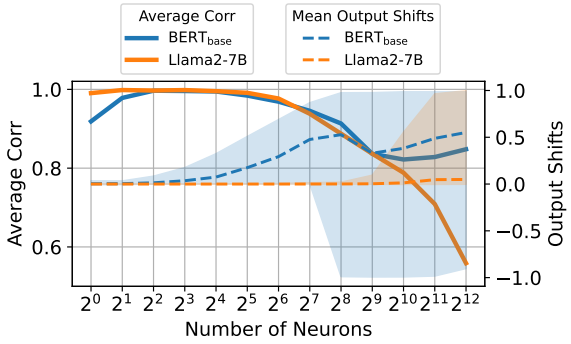
(a) Enhancement range of [0, 0.1].



(b) Enhancement range of [0, 1].



(c) Enhancement range of [0, 1.5].



(d) Enhancement range of [0, 2].

Figure 11: Multi-neuron intervention experiment results with different enhancement ranges.

Balanced Options: Most datasets, except for Stereoset, contain more than 8000 data points. To ensure balance across all options, we perform balanced sampling so that each option has an equal number of examples. From these datasets, we split 8000 examples into training, validation, and test sets, allocating 6,000, 1000, and 1000 examples, respectively. For instance, in the case of mLAMA, where each inquiry has four options, we ensure that the correct answer is represented equally across all four positions. This results in 1,500 occurrences (6,000/4) per position in the training set and 250 occurrences per position in both the validation and test sets.

Creation of multilingual tasks: For multilingual datasets, we focus on five languages: English (en), German (de), Spanish (es), French (fr), and Chinese (zh). These languages vary significantly in linguistic distance, with English being closer to German, French closer to Spanish, and Chinese being distant from all of them. This selection allows for a deeper analysis considering linguistic distances between languages. We ensure that 5 languages have the same number of datAn examples in each dataset (1,600 per language). Furthermore, for datasets like mLAMA, XNLI, and M-AMAZON, we ensure that each piece of knowledge is expressed in all five languages. This consistency enables direct comparisons of language understanding abilities across different languages.

G Details of Skill Neuron Probing

In this section, we report the details of our skill neuron probing evaluation, including the full optimal accuracies on all tasks with zero-shot prompt setting (Table 7), few-shot prompt setting (Table 8). For two major vote probers, optimal accuracies are acquired by performing a hyper-parameter (optimal neuron size) search on the validation set and evaluating the test set. We report the optimal neuron sizes for all tasks along with the accuracies in the table.

H Properties of Skill Neurons

H.1 Representation & Acquisition Efficiency

Neuron sizes	Tasks
$2^0 \sim 2^3$	Toxic, LTI, M-POS, FEVER, TempLAMA
$2^4 \sim 2^8$	GED, POS, CHUNK, NER, Amazon, IMDB, PAWS, MNLI, SWAG, HaluEval, XNLI, M-Amazon
$2^9 \sim 2^{13}$	Agnews, MyriadLAMA, CSQA, mLAMA

Table 9: Optimal number of skill neurons in Magn-prober.

Representational efficiency: By finding the optimal neuron size on the validation set, we observe that skill-neuron prober can achieve high accuracy with a few neurons. We summarize optimal neuron sizes for all tasks with Magn-prober in Table 9. Most tasks achieved optimal accuracy within 256 neurons, demonstrating the efficiency of NEG in representing language skills. Notably, factuality tasks, such as MyriadLAMA, CSQA, and mLAMA, engage a larger number of neurons, suggesting that handling facts requires more diverse neurons, reflecting the complexity of factual understanding tasks.

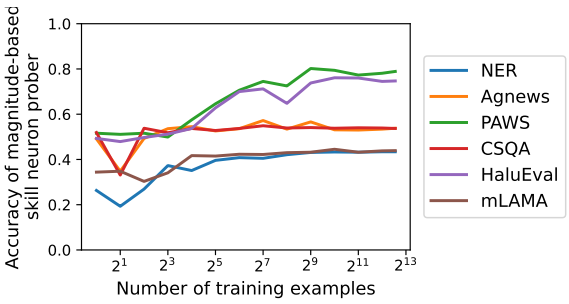


Figure 12: Accuracies with varying training sizes.

Acquisition efficiency: We report the accuracy of skill-neuron probers with different training examples in Figure 12. While adding training examples can consistently increase the probers' accuracy, the earnings slow down after 128, indicating the efficiency of acquiring skill neurons with limited data.

H.2 Generality Across Diverse Contexts

We investigate how skill neurons change when we provide different contexts, including instructions, demonstrations, and options for the same task. Given context X, we first acquire the skill neurons $\mathcal{N}_s^X$ and the accuracy $\text{ACC}^X_{\mathcal{N}_s^X}$. Then, we use the classifier built with $\mathcal{N}_s^X$ to evaluate the task by context Y as $\text{ACC}^Y_{\mathcal{N}_s^X}$. We denote the generality of $\mathcal{N}_s^X$ on context Y as $\frac{\max(\text{ACC}^Y_{\mathcal{N}_s^X} - \alpha, 0)}{\max(\text{ACC}^Y_{\mathcal{N}_s^Y} - \alpha, 0)}$, where α is the accuracy by Rand.

Using PAWS as an example, we create 12 distinct contexts by varying the instructions, the selection of demonstrations, and the output token styles. By measuring the generality for different combinations, we observe that the generality for prompting settings with different instructions and demonstrations is very high (close to 1), while the generality largely decreases if target tokens are changed. The results indicate that skill neurons maintain strong generality across different inputs, including variations in instructions and demonstrations. However, this generality diminishes when the output tokens are changed. See § K for details of experimental settings and results, including 12 designed contexts and generality results.

H.3 Are Neurons Exclusive in Skill Representations?

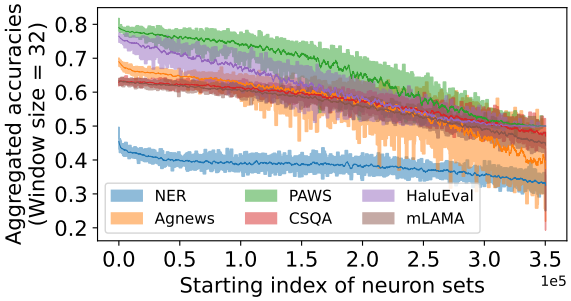


Figure 13: Accuracies of Magn-prober probers with different neuron sets, plotting the mean accuracy within each window, along with the accuracy ranges (min to max), as the envelope. Neuron sets are selected from all neurons in Llama2-7B in groups of 64, ranked by importance to be used as skill indicators.

We investigate whether skill neurons exclusively represent specific skills or can be substituted by different neuron sets. We thus build Magn-probers using various neuron sets. Specifically, we select 64 consecutive neurons from the ranked list, ordered by their importance as skill indicators (§ 7.1).[12]

Figure 13 depicts the accuracies across six tasks. The result suggests skill neurons are broadly distributed, with numerous neurons acting as skill

[12]We use 64-neurons units, which maintain high accuracies across tasks (§ A). With 352,256 neurons in Llama2-7B's FF layers, this yields 5,504 accuracy values per task.

indicators. Even when relying on less important neurons, the model's representational ability only gradually declines. Moreover, using only the least important neurons (end of each line) still yields better performance than random guesses, underscoring the inclusivity of skill neurons (See § I.3 for inclusivity evaluation on all datasets).

I Additional Analysis on Probing Results

I.1 Interpreting in-context learning with empirical gradients.

To understand why simple majority-vote classifiers achieve high accuracy, we analyze the gradients associated with each answer choice. Using PAWS (binary classification) as an example, we inspect the gradient pairs for target tokens (yes/no) across all training prompts. We find that 97.21% of neurons display opposite signs for yes/no tokens. Moreover, the Corr between yes/no gradients is -0.9996. This pronounced inverse Corr suggests that empirical gradients are sharply polarized, making it easier for a majority-vote approach to distinguish between the target tokens. Furthermore, we examine how zero-shot and few-shot prompting differ from the perspective of empirical gradients. Our analysis reveals that the total gradient magnitudes in few-shot scenarios over 22 tasks are 5.36 times greater than in zero-shot. This indicates that demonstrations in context can effectively activate skill neurons, leading to better task understanding.

I.2 More Data about Efficiency

We report the accuracies of major-vote probers with different neuron sizes for all tasks to provide additional evidence for the discussion about the representation and acquisition efficiency of skill neurons in § H.1. The results are demonstrated in Figure 16 and Figure 17 for zero-shot and few-shot prompting settings.

I.3 Probing With Varying Neuron Sets

We report the aggregated accuracies across all 22 tasks in MCEval8K in Figure 18 to provide additional evidence for discussion in § H.3. It demonstrates that many neurons can construct the classifiers in solving the language tasks, showing their ability to represent language skills and knowledge.

J Prompting Setups

In this subsection, we list all the instructions we use for each task in MCEval8K. It includes design instructions, options, and a selection of few-shot examples. As mentioned in § 7.2, we adopt two instruction settings, zero-shot and few-shot. For few-shot prompting, we set the number of examples to the same number as the number of options and ensure each option only appears once to prevent majority label bias (Zhao et al., 2021). All the few-shot examples are sampled from the training set. Finally, we list all the instructions and options we used for skill neuron probing examples by showing one zero-shot prompt.

GED

```
### Instruction: Which of the sentence below is linguistically acceptable?
### Sentences:
a.I set the alarm for 10:00 PM but I could n't wake up then .
b.I set the alarm for 10:00PM but I could n't wake up then .
### Answer:
```

POS

```
### Instruction: Determine the part-of-speech (POS) tag for the highlighted target word in the given text. Choose the correct tag from the provided options.
### Input text:One of the largest population centers in pre-Columbian America and home to more than 100,000 people at its height in about 500 CE, Teotihuacan was located about thirty miles northeast of modern Mexico City.
### Target word:'pre-Columbian'
### Options:
a.DET
b.ADJ
c.PRON
d.PUNCT
### Answer:
```

CHUNK

```
### Instruction: Identify the chunk type for the specified target phrase in the sentence and select the correct label from the provided options.
### Input text:B.A.T said it purchased 2.5 million shares at 785 .
### Target phrase:'said'
### Options:
a.PP
b.VP
```

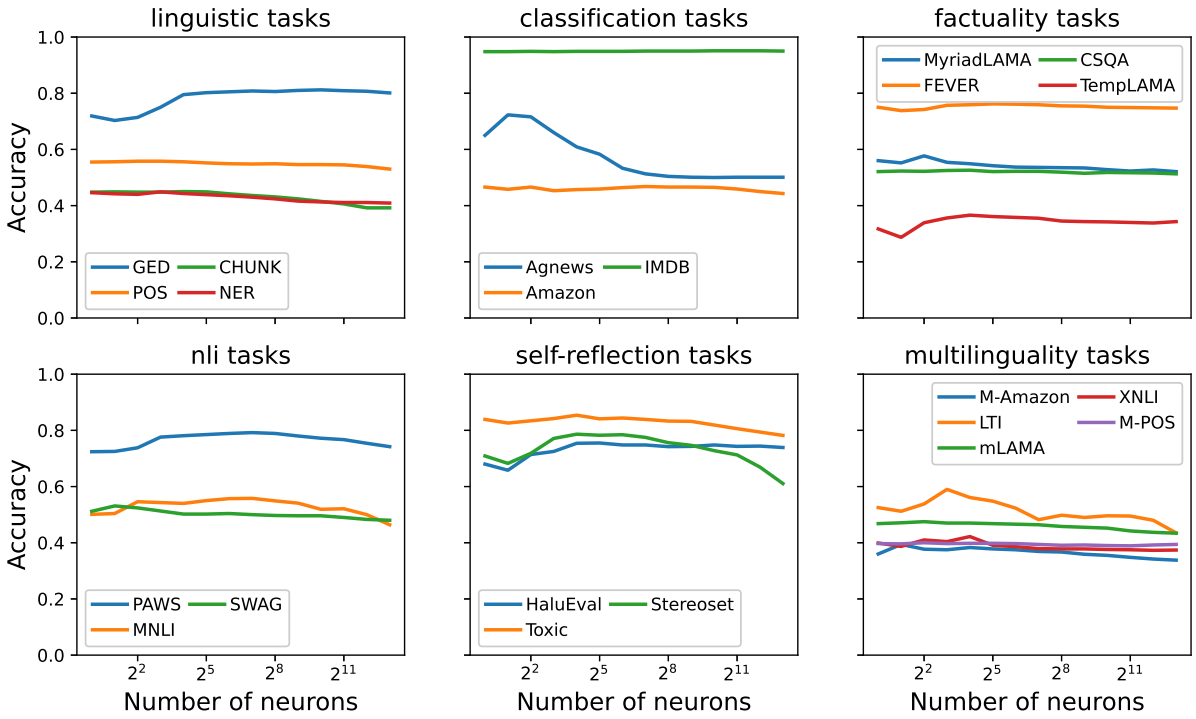


Figure 14: Per-task accuracies with varying neuron sizes on Llama2-7B, zero-shot prompt setting.

c.NP
d.ADVP
Answer:

NER

Identify the named entity type for the specified target phrase in the given text. Choose the correct type from the provided options
Input text:With one out in the fifth Ken Griffey Jr and Edgar Martinez stroked back-to-back singles off Orioles starter Rocky Coppinger (7-5) and Jay Buhner walked .
Target phrase:'Orioles'
Options:
a.LOC
b.ORG
c.MISC
d.PER
Answer:

Agnews

Instruction: Determine the genre of the news article. Please choose from the following options: a.World b.Sports c.Business d.science. Select the letter corresponding to the most appropriate genre.
Text:Context Specific Mirroring
"Now, its not that I dont want to have this content here. Far from it. Ill always post everything to somewhere on this site. I just want to treat each individual posting as a single entity and place it in as fertile a set of beds as possible. I want context specific mirroring. I want to be able to newlinechoose
multiple endpoints for a post, and publish to all of them with a single button
click."

Genres:
a.World
b.Sports
c.Business
d.Science
Answer:

Amazon

Instruction: Analyze the sentiment of the given Amazon review and assign a score from 1 (very negative) to 5 (very positive) based on the review. Output only the score.
Input Review:I never write reviews,

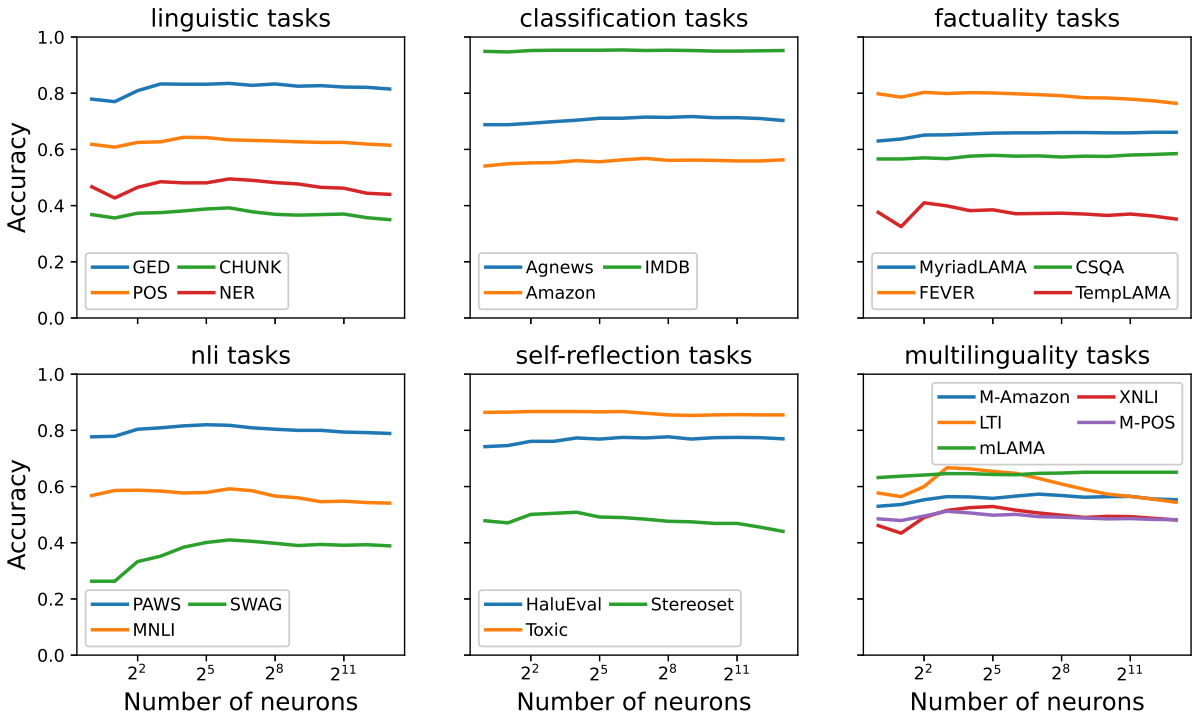


Figure 15: Per-task accuracies with varying neuron sizes on Llama2-7B, few-shot prompt setting.

but this one really works, doesn't float up, is clean and fun. Kids can finally take a bath!
Output Score:

IMDB

Instruction: Based the review, is the movie good or bad?
Review:Stewart is a Wyoming cattleman who dreams to make enough money to buy a small ranch in Utah ranch <...abbreviation...>. In spontaneous manner, Stewart is lost between the ostentatious saloon owner and the wife-candidate...
Answer:

MyriadLAMA

Instruction: Predict the [MASK] in the sentence from the options. Do not provide any additional information or explanation.
Question:What is the native language of Bernard Tapie? [MASK].
Options:
a.Dutch
b.Telugu
c.Russian
d.French
Answer:

CSQA

Instruction: Please select the most accurate and relevant answer based on the context.
Context: What does a lead for a journalist lead to?
Options:
a.very heavy
b.lead pencil
c.store
d.card game
e.news article
Answer:

TempLAMA

Instruction: Select the correct year from the provided options that match the temporal fact in the sentence. Output the index of the correct year.
Question:Pete Hoekstra holds the position of United States representative.
Options:
a.2013
b.2014
c.2018
d.2011
Answer:

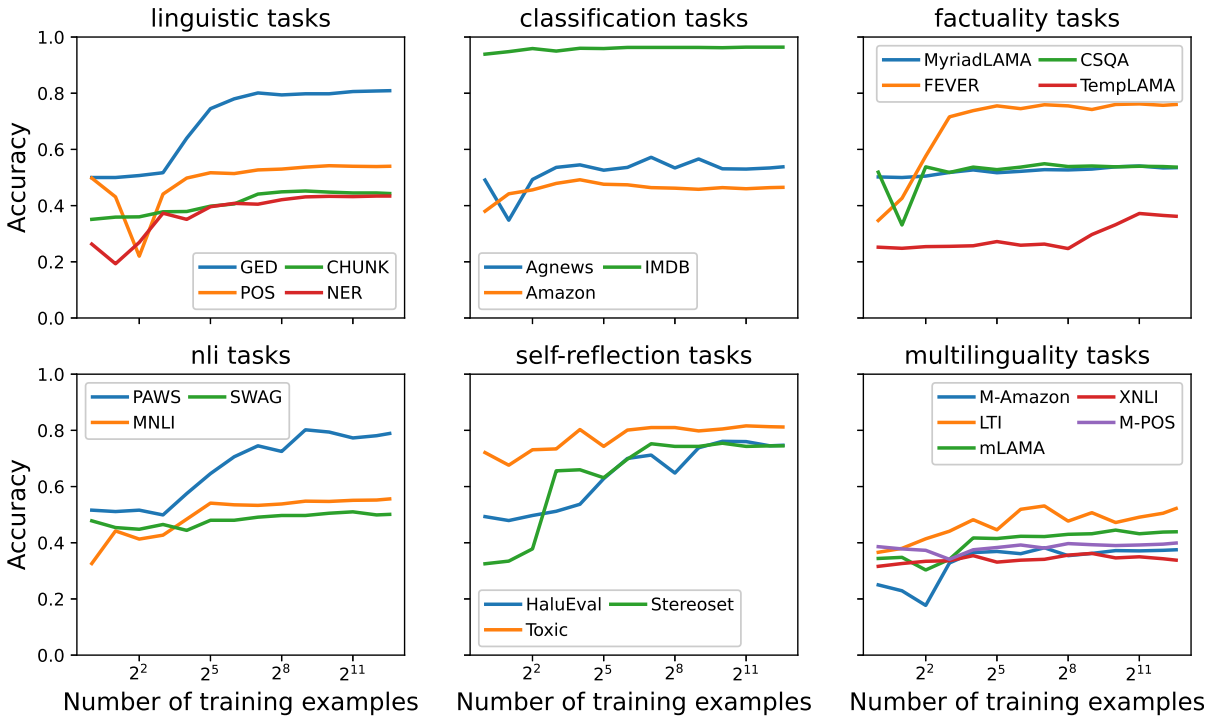


Figure 16: Per-task accuracies with the varying number of training examples on Llama2-7B, zero-shot prompt setting.

PAWS

```
### Instruction: Is the second sentence a paraphrase of the first? Answer exactly 'yes' or 'no'.
### Sentence 1: It is directed by Kamala Lopez and produced by Cameron Crain , Richard Shelgren and Kamala Lopez .
### Sentence 2: It was produced by Cameron Crain , Richard Shelgren and Kamala Lopez and directed by Kamala Lopez .
### Answer:
```

MNLI

```
### Instruction: Given a premise and a hypothesis, determine the relationship.
### Premise: easily yeah yeah and then if you want popcorn and stuff it's just i mean uh it's incredible
### Hypothesis: It's anti-incredible, very ordinary and unimpressive.
### Question: What is the relationship between the two sentences?.
### Options:
a.Entailment
b.Neutral
c.Contradiction
### Answer:
```

SWAG

```
### Instruction: Given the context, select the most likely completion from the following choices. Please exactly answer the label.
### Context: He looks back at her kindly and watches them go. In someone's dark bedroom, someone
### Options:
a.paces with the bandage, his back to someone.
b.spies a framed photo of a burmese soldier on a black horse.
c.blinks covers the apartment's couch.
d.lays her sleeping niece down gently onto the bed.
### Answer:
```

HaluEval

```
### Instruction: Given the knowledge context, dialogue histroy and response, determine if any hallucination is present. Provide a response of either 'yes' or 'no' only.
### Context:Kim Edwards wrote The Memory Keeper's Daughter
### Dialogue history:[Human]: Could you recommend something by Kim Edwards?
```

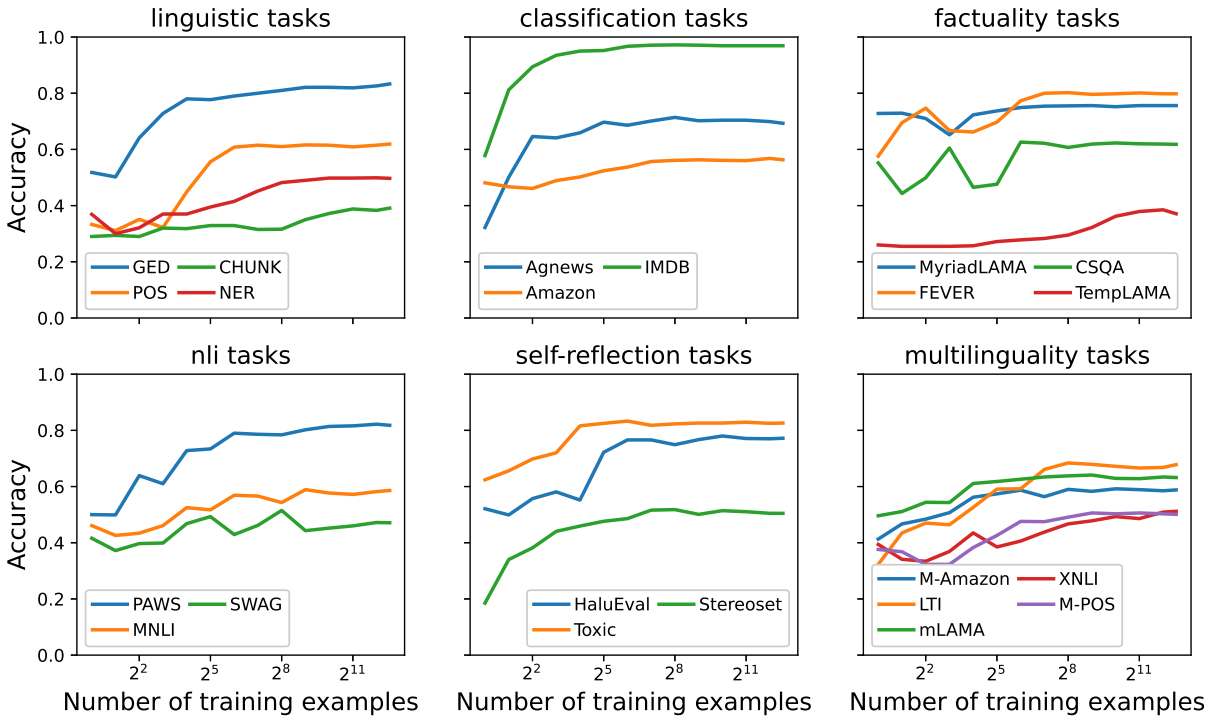


Figure 17: Per-task accuracies with the varying number of training examples on Llama2-7B, few-shot prompt setting.

[Assistant]: Absolutely, The Memory Keeper's Daughter is good. Do you like Fiction? [Human]: Yes, I do. I really love Sci Fi. [Assistant]: OK. Some Sci Fi choices are Blade Runner, The Matrix and Alien. Do you like any of these? [Human]: I wasn't fond of Alien or Blade Runner but I loved The Matrix.
Response:Actually, Kim Edwards is a director and her best work so far is The Memory Keeper's Daughter. For Sci-Fi, I suggest you watch the 1989 classic Field of Dreams. Would you like to give it a try?
Hallucination (yes or no):

Toxic

Instruction: Determine if the provided text contains toxic content.Provide a response of either 'yes' or 'no' only.
Text:Dear SchfiftyThree
I would just like to say that i hate you and the rest of the wikipedia editors for deleting my creations. fuck you.
Toxicity (yes or no):

Stereoset

Instruction: Given the context, identify and select the sentence that does not convey the stereotype related to the context.
Context:A performing artist is often up late at night.
Options:
a.Many artists are very likeable people.
b.Having children is a momentous occasion.
c.Any who performs is a diva.
Answer:

M-Amazon

Anweisung: Analysieren Sie die Stimmung der gegebenen Amazon-Bewertung und vergeben Sie eine Punktzahl von 1 (sehr negativ) bis.
Eingabebewertung:Produkt ist gefährlich. Hat den Fi rausgehen. Man hat das gefühl es fällt auseinander. Billigste Qualität!! Man fühlt sich einfach betrogen!!!
Ausgabewertung:

LTI

Instruction: Identify the language of

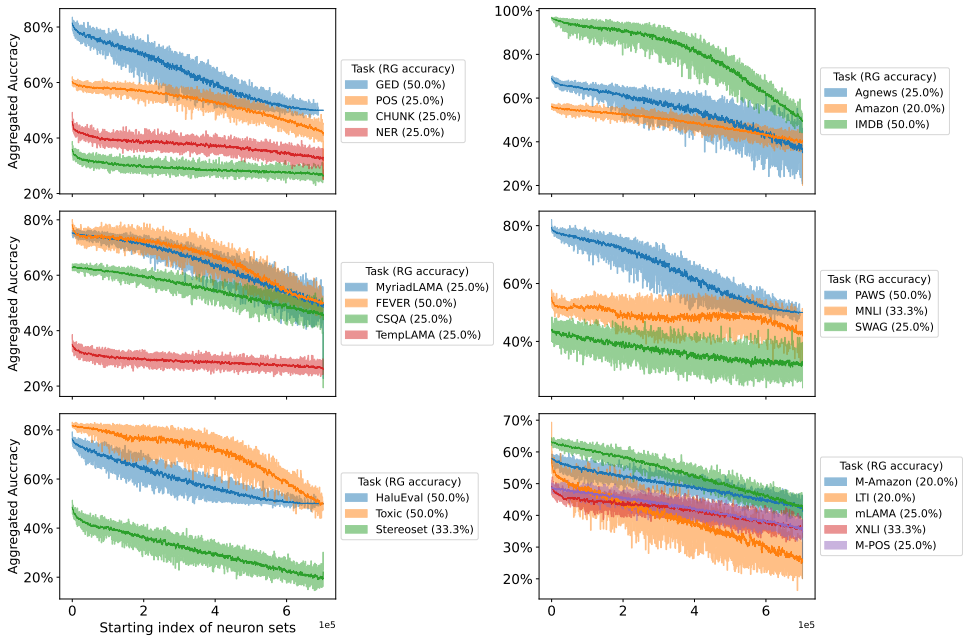


Figure 18: Per-task accuracies with varying neuron sets per with 64 neurons. We report the aggregated accuracies with a window size of 64 for better visualization, plotting the mean accuracy within each window, along with the corresponding accuracy ranges (minimum to maximum) as the envelope.

```
the given sentence.
### Text:S'en retournait, et assis sur
son chariot, lisait le prophète Ésaïe.
### Options:
a.English
b.French
a.German
a.Chinese
a.Spanish
### Answer:
```

mLAMA

```
### Instrucción: Prediga el [MASK] en la
oración a partir de las opciones. No
proporcione información ni explicaciones
adicionales.
### Respuesta:La capital de Irán es
[MASK].
### Opciones:
a.Indianápolis
b.Génova
c.Teherán
d.París
### Pregunta:
```

XNLI

```
### Instruction: Étant donné une prémisse
et une hypothèse, déterminez la relation.
### Prémisse: Ouais nous sommes à environ
km au sud du lac Ontario en fait celui qui
a construit la ville était un idiot à mon
avis parce qu' ils l' ont construit ils l'
ont construit assez loin de la ville qu'
il ne pouvait pas être une ville portuaire
### Hypothèse: Nous sommes à 10 km au sud
du lac Ontario en bas i-35 .
### Options:
a.Implication
b.Neutre
c.Contradiction
### Réponse:
```

M-POS

```
### 指令：确定给定文本中高亮目标词的词
性。从提供的选项中选择正确的词性标签。
### 文本:但是，有一個全面的人口統計數據
分析，對象包括婦女，特是有養育孩子的那
些。
### 目标词:'一'
### 选项:
a.NUM
b.AUX
c.ADJ
d.VERB
### 问题:
```

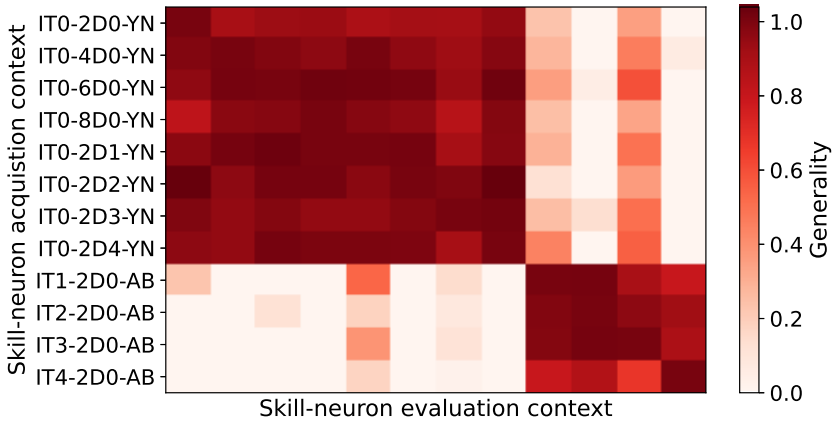


Figure 19: Generality of skill neurons across different contexts. **X-axis**: the context used to acquire skill neurons. **Y-axis**: evaluation context. The contexts on the x-axis are in the same order as on the y-axis. The context using the i-th instruction, k-th set of j-shot demonstrations, and yes/no answers is denoted as IT(i)-(j)D(k)-YN. "AB" refers to the a/b style options.

K Diverse Contexts for Skill Neuron Generality Evaluation

In this section, we report the instructions we used for experiments to measure the generality of skill neurons in § H.2. We report five types of instruction settings with 2-shot, IT0, IT1, IT2, IT3, IT4, where IT0 use yes/no as it candidate target tokens while others use a/b.

We fix the number of skill neurons to 32 when training the skill-neuron-based probers. We use 32 as the optimal neuron size of PAWS with the few-shot setting is 32. Finally, we report the pair-wise generality values among different prompting settings in Figure 19.

An example of IT0

```
### Instruction: Is the second sentence a paraphrase of the first? Answer exactly 'yes' or 'no'.
### Sentence 1: The canopy was destroyed in September 1938 by Hurricane New England in 1938 , and the station was damaged but repaired .
### Sentence 2: The canopy was destroyed in September 1938 by the New England Hurricane in 1938 , but the station was repaired .
### Answer:no
### Sentence 1: Pierre Bourdieu and Basil Bernstein explore , how the cultural capital of the legitimate classes has been viewed throughout history as the " most dominant knowledge " .
### Sentence 2: Pierre Bourdieu and Basil Bernstein explore how the cultural capital of the legitimate classes has been considered the " dominant knowledge " throughout history .
### Answer:yes
### Sentence 1: It is directed by Kamala Lopez and produced by Cameron Crain , Richard Shelgren and Kamala Lopez .
### Sentence 2: It was produced by Cameron Crain , Richard Shelgren and Kamala Lopez and directed by Kamala Lopez .
### Answer:
```

An example of IT1

```
### Instruction: Given two sentences, determine if they are paraphrases of each other.
### Sentence 1: The canopy was destroyed in September 1938 by Hurricane New England in 1938 , and the station was damaged but repaired .
### Sentence 2: The canopy was destroyed in September 1938 by the New England Hurricane in 1938 , but the station was repaired .
### Options:
a.not paraphrase
b.paraphrase
### Answer:a
### Sentence 1: Pierre Bourdieu and Basil Bernstein explore , how the cultural capital of the legitimate classes has been viewed throughout history as the " most dominant knowledge " .
```

Sentence 2: Pierre Bourdieu and Basil Bernstein explore how the cultural capital of the legitimate classes has been considered the “ dominant knowledge ” throughout history .
Options:
a.not paraphrase
b.paraphrase
Answer:b
Sentence 1: It is directed by Kamala Lopez and produced by Cameron Crain , Richard Shelgren and Kamala Lopez .
Sentence 2: It was produced by Cameron Crain , Richard Shelgren and Kamala Lopez and directed by Kamala Lopez .
Options:
a.not paraphrase
b.paraphrase
Answer:

An example of IT2

Instruction: Review the two given sentences and decide if they express the same idea in different words.
Sentence 1: The canopy was destroyed in September 1938 by Hurricane New England in 1938 , and the station was damaged but repaired .
Sentence 2: The canopy was destroyed in September 1938 by the New England Hurricane in 1938 , but the station was repaired .
Options:
a.non-equivalent
b.equivalent
Answer:a
Sentence 1: Pierre Bourdieu and Basil Bernstein explore , how the cultural capital of the legitimate classes has been viewed throughout history as the “ most dominant knowledge ” .
Sentence 2: Pierre Bourdieu and Basil Bernstein explore how the cultural capital of the legitimate classes has been considered the “ dominant knowledge ” throughout history .
Options:
a.non-equivalent
b.equivalent
Answer:b
Sentence 1: It is directed by Kamala Lopez and produced by Cameron Crain , Richard Shelgren and Kamala Lopez .
Sentence 2: It was produced by Cameron Crain , Richard Shelgren and Kamala Lopez and directed by Kamala Lopez .
Options:
a.non-equivalent
b.equivalent
Answer:

An example of IT3

Instruction: Examine the two sentences provided. Determine if the second sentence is a valid paraphrase of the first sentence.
Sentence 1: The canopy was destroyed in September 1938 by Hurricane New England in 1938 , and the station was damaged but repaired .
Sentence 2: The canopy was destroyed in September 1938 by the New England Hurricane in 1938 , but the station was repaired .
Options:
a.different
b.similar
Answer:a
Sentence 1: Pierre Bourdieu and Basil Bernstein explore , how the cultural capital of the legitimate classes has been viewed throughout history as the “ most dominant knowledge ” .
Sentence 2: Pierre Bourdieu and Basil Bernstein explore how the cultural capital of the legitimate classes has been considered the “ dominant knowledge ” throughout history .
Options:
a.different
b.similar
Answer:b
Sentence 1: It is directed by Kamala Lopez and produced by Cameron Crain , Richard Shelgren and Kamala Lopez .
Sentence 2: It was produced by Cameron Crain , Richard Shelgren and Kamala Lopez and directed by Kamala Lopez .
Options:
a.different
b.similar

Answer:

An example of IT4

Instruction: You are provided with two sentences. Identify whether they convey identical ideas or differ in meaning.
Sentence 1: The canopy was destroyed in September 1938 by Hurricane New England in 1938 , and the station was damaged but repaired .
Sentence 2: The canopy was destroyed in September 1938 by the New England Hurricane in 1938 , but the station was repaired .
Options:
a.The sentences convey different idea.
b.The sentences convey the same ideas.
Answer:a
Sentence 1: Pierre Bourdieu and Basil Bernstein explore , how the cultural capital of the legitimate classes has been viewed throughout history as the “ most dominant knowledge ” .
Sentence 2: Pierre Bourdieu and Basil Bernstein explore how the cultural capital of the legitimate classes has been considered the “ dominant knowledge ” throughout history .
Options:
a.The sentences convey different idea.
b.The sentences convey the same ideas.
Answer:b
Sentence 1: It is directed by Kamala Lopez and produced by Cameron Crain , Richard Shelgren and Kamala Lopez .
Sentence 2: It was produced by Cameron Crain , Richard Shelgren and Kamala Lopez and directed by Kamala Lopez .
Options:
a.The sentences convey different idea.
b.The sentences convey the same ideas.
Answer:

Genres	Task	Language skills	Dataset	#n_choices	#n_examples
Linguistics	POS	Part-of-speech tagging	Universal Dependencies (Nivre et al., 2017)	4	8000
	CHUNK	Phrase chunking	CoNLL-2000 (Tjong Kim Sang and Buchholz, 2000)	4	8000
	NER	Named entity recognition	CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003)	4	8000
	GED	Grammatic error detection	cLang-8 (Rothe et al., 2021; Mizumoto et al., 2011)	2	8000
Content classification	IMDB	Sentiment classification	IMDB (Maas et al., 2011)	2	8000
	Agnews	Topic classification	Agnews (Zhang et al., 2015)	4	8000
	Amazon	Numerical sentiment classification	Amazon Reviews (Hou et al., 2024)	5	8000
Natural language inference (NLI)	MNLI	Entailment inference	MNLI (Williams et al., 2018)	3	8000
	PAWS	Paraphrase identification	PAWS (Zhang et al., 2019)	2	8000
	SWAG	Grounded commonsense inference	SWAG (Zellers et al., 2018)	4	8000
Factuality	FEVER	Fact checking	FEVER (Thorne et al., 2018)	2	8000
	MyriadLAMA	Factual knowledge question-answering	MyriadLAMA (Zhao et al., 2024)	4	8000
	CSQA	Commonsense knowledge question-answering	CommonsenseQA (Talmor et al., 2019)	4	8000
	TempLAMA	Temporary facts question-answering	TempLAMA (Dhingra et al., 2022)	4	8000
Self-reflection	HaluEval	Hallucination detection	HaluEval-diag (Li et al., 2023)	2	8000
	Toxic	Toxicity post identification	Toxicity prediction (cjadams et al., 2017)	2	8000
	Stereoset	Social stereotype detection	Stereoset (Nadeem et al., 2021)	3	4230
Multilinguality	LTI	Language identification	LTI LangID corpus (Brown, 2014; Lovenia et al., 2024)	5	8000
	M-POS	Multilingual POS-tagging	Universal Dependencies (Nivre et al., 2017)	4	8000
	M-Amazon	Multilingual Amazon review classification	Amazon Reviews Multi (Keung et al., 2020)	5	8000
	mLAMA	Multilingual factual knowledge question-answering	mLAMA (Kassner et al., 2021)	4	8000
	XNLI	Multilingual entailment inference	XNLI (Conneau et al., 2018)	3	8000

Table 6: Details of datasets in MCEval8K.

Tasks	**Rand**	**LM-Prob**	**Polar-prober (#n_neurons)**	**Magn-prober (#n_neurons)**
GED	.5000	.5000	.7580 (16)	.8050 (1024)
POS	.2500	.5050	.5190 (16)	.5470 (4)
CHUNK	.2500	.3510	.4660 (8)	.4490 (16)
NER	.2500	.3950	.4120 (32)	.4490 (8)
Agnews	.2500	.4950	.6410 (32)	.6900 (2)
Amazon	.2000	.3750	.2750 (256)	.4680 (128)
IMDB	.5000	.9660	.9630 (8192)	.9650 (1024)
MyriadLAMA	.2500	.5080	.5200 (4)	.5760 (4)
FEVER	.5000	.6530	.7830 (32)	.7610 (32)
CSQA	.2000	.5170	.3490 (1)	.5380 (16)
TempLAMA	.2500	.2430	.3560 (4096)	.3640 (16)
PAWS	.5000	.5000	.7640 (128)	.7920 (128)
MNLI	.3333	.3560	.4980 (4)	.5590 (128)
SWAG	.2500	.4610	.3360 (512)	.5310 (2)
HaluEval	.5000	.4990	.7540 (1024)	.7510 (32)
Toxic	.5000	.7230	.8250 (1024)	.8210 (16)
Stereoset	.3333	.1096	.8299 (16)	.7335 (16)
M-Amazon	.2000	.2990	.2350 (4096)	.3740 (2)
LTI	.2000	.3670	.4300 (4)	.5830 (8)
mLAMA	.2500	.4020	.3880 (128)	.4470 (4)
XNLI	.3333	.3270	.3500 (256)	.3620 (16)
M-POS	.2500	.3890	.2610 (1024)	.3930 (4)

Table 7: Optimal accuracies across all MCEval8K tasks in the zero-shot prompt setting on Llama2-7B, along with the neuron sizes achieving these accuracies.

Tasks	**Rand**	**LM-Prob**	**Polar-prober (#n_neurons)**	**Magn-prober (#n_neurons)**
GED	.5000	.5060	.8330 (16)	.8330 (64)
POS	.2500	.5730	.5870 (4)	.6210 (16)
CHUNK	.2500	.2710	.2820 (8192)	.3910 (64)
NER	.2500	.3610	.4300 (4)	.4970 (64)
Agnews	.2500	.5880	.7060 (64)	.6890 (512)
Amazon	.2000	.4840	.5310 (1)	.5680 (128)
IMDB	.5000	.9700	.9700 (64)	.9690 (64)
MyriadLAMA	.2500	.7380	.7450 (256)	.7530 (4096)
FEVER	.5000	.6780	.8000 (1)	.8030 (4)
CSQA	.2000	.6100	.6180 (32)	.6340 (8192)
TempLAMA	.2500	.2600	.2500 (1)	.4110 (4)
PAWS	.5000	.5240	.8180 (16)	.8210 (32)
MNLI	.3333	.5100	.5780 (32)	.5860 (64)
SWAG	.2500	.4100	.4430 (256)	.4710 (64)
HaluEval	.5000	.5200	.7750 (2048)	.7770 (256)
Toxic	.5000	.7800	.8250 (8)	.8260 (4)
Stereoset	.3333	.1040	.7297 (128)	.5180 (16)
M-Amazon	.2000	.5250	.5470 (1024)	.5880 (128)
LTI	.2000	.3680	.5480 (64)	.6950 (8)
mLAMA	.2500	.6080	.6230 (8192)	.6360 (512)
XNLI	.3333	.3970	.4860 (32)	.4980 (32)
M-POS	.2500	.4440	.4830 (4)	.5130 (8)

Table 8: Optimal accuracies across all MCEval8K tasks in the few-shot prompt setting on Llama2-7B, along with the neuron sizes achieving these accuracies. The number of demonstrations is set as the same number of options for each task.