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

Xin Zhao **Zehui Jiang**
The University of Tokyo
{xzhao,zjiang}@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science,
The University of Tokyo
ynaga@iis.u-tokyo.ac.jp

Abstract

While feed-forward neurons in pre-trained language models (PLMs) can encode knowledge, past research targeted a small subset of neurons that heavily influence outputs. This leaves the broader role of neuron activations unclear, limiting progress in areas like knowledge editing. We uncover a global linear relationship between neuron activations and outputs using neuron interventions on a knowledge probing dataset. The gradient of this linear relationship, which we call **the neuron empirical gradient (NEG)**, captures how changes in activations affect predictions. To compute NEG efficiently, we propose **NeurGrad**, enabling large-scale analysis of neuron behavior in PLMs. We also show that NEG effectively captures language skills across diverse prompts through skill neuron probing. Experiments on **MCEval8k**, a multi-genre multiple-choice knowledge benchmark, support NEG's ability to represent model knowledge. Further analysis highlights the key properties of NEG-based skill representation: efficiency, robustness, flexibility, and interdependency. The code and data are released.

xzhao-tkl/NEG iszhaoxin/MCEval8K

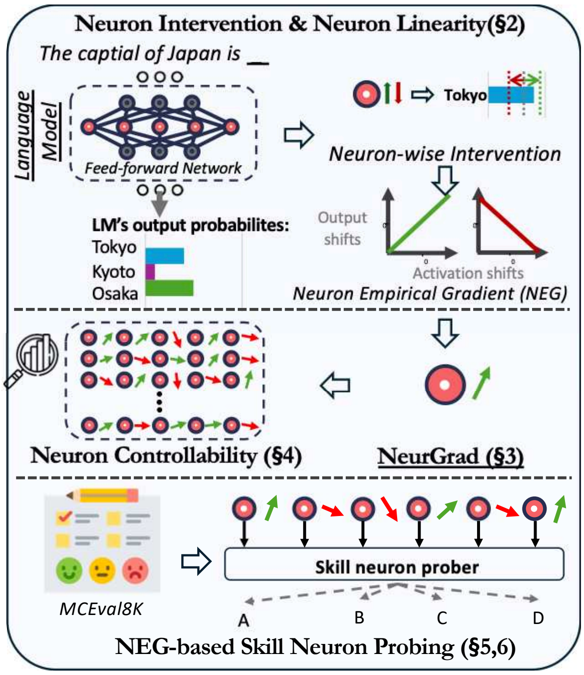


Figure 1: Our contributions: i) revealing linearity between activation and output shifts, ii) proposing NeurGrad, an efficient method to quantify it, iii) confirming that NEGs capture language skills on MCEval8K.

1 Introduction

Pre-trained language models (PLMs) based on Transformer architecture (Vaswani et al., 2017) effectively encode human knowledge, prompting efforts to understand their inner workings. While previous studies have shown that the feed-forward (FF) neurons play key roles in encoding factual knowledge (Dai et al., 2022; Yu and Ananiadou, 2024) and general language skills (Wang et al., 2022; Tan et al., 2024), they face two main challenges. First, current methods mostly rank neurons by importance without measuring the link between neuron activations and model outputs (Dai et al., 2022; Meng et al., 2022; Yu and Ananiadou, 2024), limiting use cases like knowledge editing (Zhang et al., 2024). Second, these methods are costly, involving repeated activation modifications (Dai et al., 2022; Meng et al., 2022; Goldowsky-Dill et al., 2023) or extensive tensor operations (Yu and Ananiadou, 2024), making them inefficient for analyzing all neurons on large models.

This study begins by quantitatively analyzing **how neuron activations influence model outputs (RQ1)** (Figrue 1). Using MyriadLAMA (Zhao et al., 2024), a factual knowledge probing dataset, on seven PLMs, including large language models (LLMs) like Llama2-70B (§ 2), we gradually modify random neuron activations and observe the changes in the token probabilities for correct knowledge (hereafter, *output shifts*). Notably, we find that within a certain range, shifts in neuron activations

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21446–21477
July 27 - August 1, 2025

(hereafter, *activation shifts*) have a linear impact on output shifts. We define and quantify the gradient of this linear relationship as *neuron empirical gradient (NEG)*, enabling quantitative neuron analysis.

Next, we explore **whether shifting neuron activations can precisely control PLMs' output probabilities (RQ2)**. Since computing NEGs requires costly inference, we propose *NeurGrad* (§ 3), an efficient method for estimating a single neuron's NEG. It builds on the empirical finding that computational gradients (§ 3.1) strongly correlate with NEG magnitudes, though less so with directions. Using MyriadLAMA, we validate NeurGrad's performance against ground-truth NEGs, showing it outperforms existing neuron-ranking methods (Dai et al., 2022; Yu and Ananiadou, 2024). With NeurGrad, we further examine multi-neuron control, finding that NEGs can accumulate across neurons (§ 4), while the effect weakens as more neurons are involved or larger activation shifts are involved.

Finally, we examine **whether NEGs can capture general language skills beyond factual knowledge (RQ3)**. To explore this, we perform skill neuron probing (Wang et al., 2022; Song et al., 2024) using NEGs. As the previous studies on skill neuron probing focus on limited language skills, we introduce *MCEval8K*, a benchmark covering six genres and 22 tasks for broad LLM evaluation. Our results show that NEGs can represent a wide range of language skills. Furthermore, our in-depth analysis highlights the properties of NEG-based skill neurons, including the efficiency in representing language skills, robustness when facing diverse contexts, substitutability of skills being represented by different neurons, and interdependency between neurons in representing skills.

Our contributions (Figrue 1) are as follows:

- We confirm that activation and output shifts are linearly correlated within a certain range through neuron intervention, defined as **neuron empirical gradient (NEG)** (§ 2).
- We introduce **NeurGrad**, an efficient method to NEG (§ 3), and conduct in-depth analysis about neuron controllability (§ 4).
- We show that NEGs can represent language skills through **skill neuron probing** (§ 5); skill neurons exhibit efficiency, robustness, inclusivity, and interdependency (§ 6).
- We develop **MCEval8K**, a multiple-choice benchmark covering six genres and 22 language understanding tasks (§ 5.2, § F).

2 Neuron Linearity to Model Output

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

2.1 Neuron-level Intervention Experiments

Models. To ensure the generality of our findings, we evaluate both masked and causal LMs with varying sizes and learning strategies. For masked LMs, we use two BERT (Devlin et al., 2019) models, $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$, and have them predict masked tokens. For causal LMs, we examine five LLMs with different model sizes and language families, including Llama2 (Touvron et al., 2023) (7B, 70B), Llama3.1 (8B), Llama3.2 (3B) (Grattafiori et al., 2024), and Qwen2.5 (7B) (Qwen et al., 2025). All of these models are instruction-tuned. Following Zhao et al. (2024), we use a zero-shot prompt to generate single-token answers. See § E for details.

Dataset. We use MyriadLAMA (Zhao et al., 2024), a multi-prompt knowledge probing dataset for neuron intervention. Its diverse prompts help reduce bias from specific phrasing. We focus on single-token probing, where the target answer is a single token. For each PLM, we randomly sample 1000 prompts from MyriadLAMA, where the model predicts the token representing the correct answer. Due to tokenizer and knowledge differences, the sampled prompts vary across PLMs.

Neuron intervention. We shift activations in the range of [-10, 10][1] with the step size of 0.2 to track changes in target token probabilities. For example, consider the following prompt:

> Predict the `[MASK]` in each sentence in one word.
> Q: `[MASK]` is the capital of Japan.
> A:

We modify the activations of specific neurons and observe how the probability of the target word *Tokyo* changes at the final position. Since conducting one intervention experiment for a neuron-token-prompt combination requires 100 inferences due to the shift values, we randomly sample[2] neurons to make computation tractable while ensuring broad coverage.

[1]The range mostly covers the distribution of activations; see § B.1 for details.

[2]The actual number of randomly sampled neurons will be given in each experiment.

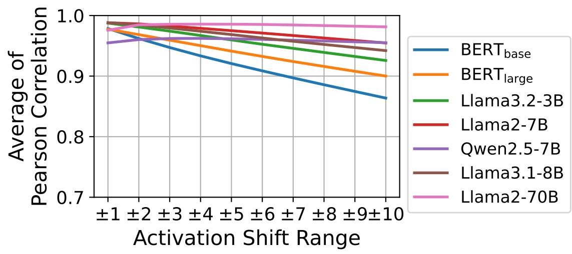


Figure 2: Average absolute correlation between activation and output shifts.

	Linear	Positive	Negative
$\text{BERT}_{\text{base}}$	.9201	.4981	.5004
$\text{BERT}_{\text{large}}$	.9209	.4938	.4981
Llama3.2-3B	.9483	.5000	.5000
Llama2-7B	.9659	.5157	.4843
Qwen2.5-7B	.6549	.5116	.4884
Llama3.1-8B	.9540	.5048	.4952
Llama2-70B	.9208	.4962	.5039

Table 1: Ratios of linear neurons and neuron polarity over 1000 prompts with 100 neurons.

Results. To understand how output shifts respond to neuron activation shifts, we compute the Pearson correlation coefficient (r) between activation shifts and output shifts of correct tokens, using absolute r values. To reduce computation, we average r over 10 randomly sampled prompts from the overall 1000 prompts, each with 1000 randomly sampled neurons at specific shift ranges (x-axis). Figure 2 shows a strong, nearly linear correlation across a wide ± 10 range, with stronger correlation at smaller shift ranges consistent across all models. This suggests predictable output changes from specific activation shifts. Since all PLMs behave similarly, we focus on BERT models and Llama2 LLMs in subsequent analysis.

2.2 Neuron Linearity

Based on the findings above, we ask **whether neurons generally show linearity with model output.** We define neurons as linear if their correlation (r) is at least 0.95[3] within a ± 2 shift range, as observed in Figure 2.

Then, we quantitatively analyze neuron linearity across prompts and Transformer layers by measuring the ratio of linear neurons from 1000 prompts and 100 randomly sampled neurons. [4] The ratios of linear neurons are reported in Table 1, showing that most neurons in LLMs exhibit linearity. This linearity is widespread across layers and prompts (§ A). We also define **polarity** as follows: neurons are *positive* if increasing activations boost target token probabilities, and *negative* otherwise. The analysis of neuron polarity is deferred to § 4.1.

Neuron empirical gradient. We quantify neuron linearity and polarity using the gradient of the linear relationship between activation and output shifts, termed *neuron empirical gradient* (NEG). To compute NEGs, we fit a zero-intercept linear regression between activation shifts and output shifts acquired through neuron intervention; the regression coefficient serves as the NEG for each neuron, prompt, and token.

[3] We assume $r \geq 0.95$ to be a strong linear relationship.

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

3 NeurGrad for NEG Estimation

Efficient and accurate computation of NEG is essential for quantitative neuron-level interpretability in PLMs. However, the neuron-wise intervention is costly. While prior knowledge attribution methods estimate neurons' influence on model outputs, they either require intensive computation or only provide relative importance, without directly measuring NEG (Dai et al., 2022; Geva et al., 2022; Meng et al., 2022; Yu and Ananiadou, 2024).

3.1 NeurGrad

In this section, we propose NeurGrad, an accurate and efficient method for estimating NEG, to support further analysis. This approach is based on preliminary findings using computational gradients[5] (**CG**) to approximate NEG. We compute CG and ground-truth NEG values for seven PLMs, including BERT variants and instruction-tuned Llama2, using 1000 prompts and 100 neurons per prompt with a shift range of ± 2. While CG values show a low correlation with NEG directly (average $r = -0.429$), their absolute values are highly correlated (average $r = 0.961$). Additionally, both CG and neuron activations determine the sign of NEG. Based on these results, we introduce NeurGrad:

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

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

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

	Correlation (r)				MAE		
	CG	**IG**	**LPI**	**NeurGrad**	**CG**	**IG**	**NeurGrad**
$\text{BERT}_{\text{base}}$	-.8909	.7360	-[6]	**.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**
Llama3.2-3B	-.1281	.5854	-	**.8094**	3.5e-06	1.8e-06	**1.6e-06**
Llama2-7B	.3023	.5377	.6469	**.8135**	1.7e-06	**1.2e-06**	1.3e-06
Qwen2-7B	.1043	.2939	-	**.5862**	1.1e-06	**7.6e-07**	7.7e-07
Llama3.1-8B	.0072	.5098	-	**.7286**	4.9e-06	**1.9e-06**	3.6e-06
Llama2-70B	.0283	n/a[7]	n/a	**1.000**	2.2e-04	n/a	**5.9e-07**
Avg. Runtime (Llama2-7B)	**0.149s**	19.349s	6.086s	0.161s	Same as left		

Table 2: Evaluation of NeurGrad and baselines in calculating NEGs, including two metrics: r and MAE.

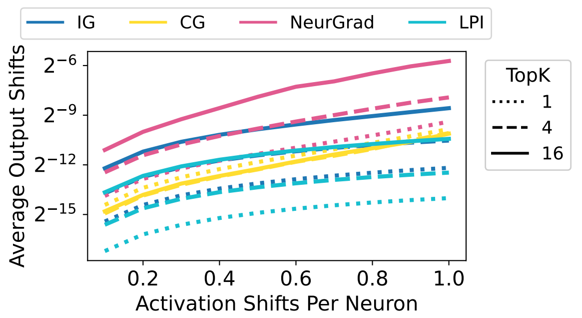


Figure 3: Comparison of neuron attribution methods in token probability enhancement. *X-axis*: activation shifts of selected neurons; *Y-axis*: average output shifts over 1000 factual prompts.

3.2 NEG Estimation Evaluation

We evaluate NeurGrad's ability to estimate NEG using the same setup as in the CG evaluation. Our experiment compares NeurGrad with three baselines: two gradient-based methods, CG and integrated gradients (IG) (Sundararajan et al., 2017; Dai et al., 2022), and one logit-based method (LPI) (Yu and Ananiadou, 2024). IG simulates NEG by small, repeated neuron interventions, while LPI (Log-Probability-Increase) estimates neuron importance based on increases in output probabilities.[8]

We assess NEG estimation using correlation (r) and mean absolute error (MAE) against ground-truth NEG. As show in Table 2, NeurGrad consistently achieves the highest r across the seven PLMs, capturing relative neuron importance well. NeurGrad also yields low MAE, indicating high accuracy. The average running time of Llama2-7B with an NVIDIA RTX A600 GPU (Table 2 (bottom)) demonstrates NeurGrad's superior efficiency.

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

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

[8] We exclude causal-tracing methods (Meng et al., 2022), which are too costly and do not meet our efficiency goals.

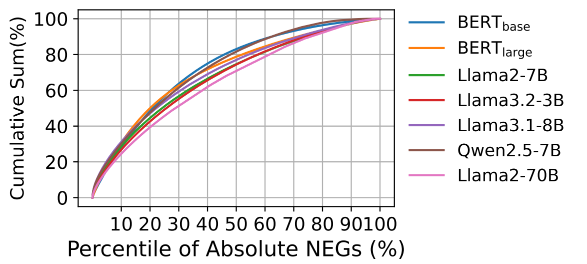


Figure 4: Cumulative distribution of NEG magnitudes. (*X-axis*: the percentiles of NEG magnitudes; *Y-axis*: the cumulative contribution of neurons to the total sum).

3.3 Knowledge Attribution Evaluation

We further evaluate NeurGrad's ability to identify important neurons. For 1000 prompts, we select the top-K neurons ($K = 1, 2^2, 2^4$) using CG, IG, LPI, and NeurGrad values, and then enhance their activations by increasing activations of positive neurons and decreasing those of negative neurons. The activation shift is conducted within the range of [0.1, 1] with a step size of 0.1. Figure 3 shows output shifts on Llama2-7B under these interventions. NeurGrad consistently outperforms baselines, due to its accurate NEG estimation and inclusion of both positive and negative neurons, unlike IG and LPI, which only consider positive ones, despite their equal distribution (Table 1). See § C for details.

4 Understanding Neurons' Controllability

This section explores neuron controllability: the ability to precisely adjust PLM output probabilities by modifying neuron activations. We use NEGs estimated by NeurGrad to achieve this.

4.1 How Are NEGs Distributed?

Do only a few neurons exhibit strong gradients? Figure 4 presents the cumulative NEG distribution

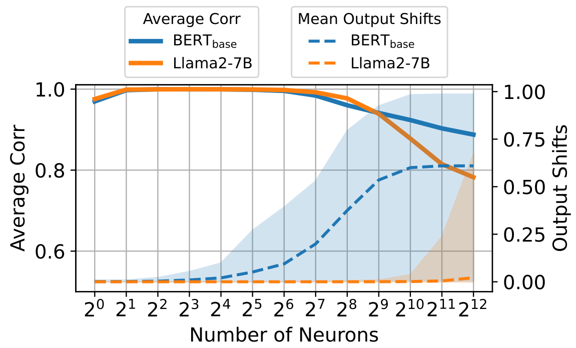


Figure 5: Multi-neuron enhancement with range [0,0.5] with different number of neurons.

for all neurons, showing steady growth that only converges when all neurons are included. This indicates that most neurons influence the model's output probabilities.

Do neurons have polarity preference? Table 1 (§ 2) shows the ratios of positive and negative neurons[9] across 1000 prompts, each with 100 randomly sampled neurons in seven PLMs, revealing nearly equal numbers. This suggests that PLMs have no polarity preference; interventions should consider gradient polarity rather than simply increasing or decreasing their activations (Dai et al., 2022). See more detailed analysis on NEGs' distribution in § B.2 and § B.3.

4.2 Does Linearity Hold for Multi-neuron?

We explore **whether output shifts can be predicted when intervening multiple neurons.** For each of 1000 prompts, we randomly sample N neurons and shift their activations on the basis of polarity measured by NeurGrad, applying positive/negative shifts to positive/negative neurons. We experiment on BERT$_{\text{base}}$ and Llama2-7B using neuron sizes of 2^N $(0 \leq N \leq 12)$.

Figure 5 shows the average correlation (r) between predicted and actual output shifts across all prompt-neuron pairs, using an enhancement range of [0, 0.5] and a 0.01 step size. While r decreases as more neurons are involved due to neuron interactions, it remains strong ($\geq$0.7) even with 2^{12} neurons in both PLMs. The figure also shows that larger neuron sets cause greater output shifts, suggesting an additive effect. Although Llama2 shows small average output shifts (due to the lower per-neuron NEG; see § B.2), significant changes still occur when over 2^{10} neurons are intervened. Experiments with larger enhancement ranges show similar trends: greater shifts load to lower r, reducing predictability (see § D for experiments with different ranges).

The analysis above shows that NeurGrad enables partial prediction of output shifts. However, both the number of modified neurons and the modification range require careful consideration.

4.3 Why Do Neurons Exhibit Linearity?

We propose the local linearity approximation hypothesis to explain neuron linearity, based on three observations: i) larger shift ranges reduce r, ii) larger PLMs show higher r due to the reduced influence of individual neurons (Figure 2), and iii) involving more neurons weakens linearity (Figure 5). This aligns with the first-order Taylor expansion under local differentiability.

Local Linear Approximation Hypothesis

Let $\mathbf{f}:\mathbb{R}^n \to \mathbb{R}^m$ denote the mapping from a model's FF activation layer $\mathbf{x} \in \mathbb{R}^n$ to its output probability vector $\mathbf{f}(\mathbf{x})$. We have:

$$f_j(\mathbf{x}') = f_j(\mathbf{x}) + \frac{\partial f_j}{\partial x_i}(\mathbf{x})\,\Delta x_i + o(\Delta x_i)$$

if we consider a perturbation Δx_i on i-th neuron, with perturbed vector $\mathbf{x}' = \mathbf{x} + \Delta x_i \mathbf{e}_i$. Here $(\mathbf{e}_i)_j = \delta_{ij}$ is a unit vector. Then, we define $\frac{\partial f_j}{\partial x_i}(\mathbf{x})$ as **NEG**.

5 Skill Neuron Probing using NeurGard

While NEG and NeurGrad allow neuron-level output control, variation in NEG across prompts limits their use for modifying specific language skills, which require handling diverse inputs. This section investigates **whether NEG can capture general language skills through skill neuron probing,** which identifies neurons linked to task-solving capabilities. Unlike prior work (Wang et al., 2022; Song et al., 2024) focusing on neuron activations, we utilize NEG for finding skill neurons.

5.1 Task Definition

Following Wang et al. (2022), we define skill neuron probing as follows. A skill dataset $\mathcal{D}$ consists of language sequence pairs: knowledge inquiries $\mathcal{Q} = \{q_1, ..., q_{|\mathcal{T}|}\}$ and corresponding answers $\mathcal{A} = \{a_1, ..., a_{|\mathcal{T}|}\}$, where each a_i is from

[9] We used ground-truth NEGs here to avoid flipping the polarity of neurons with NEGs near zero.

a candidate set $\hat{\mathcal{A}}_{\text{cands}}$. For example, in sentiment classification, Q is the set of documents and A the sentiment labels. We train classifiers using the behavior of a neuron subset $\mathcal{N}_s \subseteq \mathcal{N}$ as features to predict the correct answer a_i for each q_i, where $\mathcal{N}$ is the full neuron set. [10]

Our skill neuron probe finds the optimal neuron subset $\mathcal{N}_s^*$ that maximizes accuracy on the dataset $\mathcal{D}$.

$$\mathcal{N}_s^* = \underset{\mathcal{N}_s \subseteq \mathcal{N}}{\arg\max}\, \mathrm{Acc}(f(\mathcal{N}_s), D) \tag{2}$$

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

Here, $f(\mathcal{N}_s, q_i)$ is the classifer's prediction using neuron subset $\mathcal{N}_s$ for input q_i and $\mathbb{1}[X = Y]$ is 1 if $X = Y$, otherise 0.

5.2 Evaluation Benchmark: MCEval8K

In this section, we introduce a benchmark dataset for skill neuron probing. Since probing requires a fixed target token, previous work (Wang et al., 2022; Song et al., 2024) used on multiple-choice datasets with single-token labels (*e.g.*, A: positive, B: negative). While we follow a similar setup, earlier studies focused on small PLMs and probed them on basic tasks, which are insufficient for evaluating LLMs.

To address this limitation, we create MCEval8K, a diverse multiple-choice benchmark covering 22 language understanding tasks across six skill genres, incorporating most datasets from previous studies. To ensure consistency and reduce computational cost, we cap each task at 8K queries,[11] even for large datasets like cLang-8 (Mizumoto et al., 2011; Rothe et al., 2021). We also balance the number of correct options per task to avoid classification bias. Task and dataset details are provided below and in § F.

Linguistic: Part-of-speech tagging (POS), text chunking (CHUNK), named entity recognition (NER), and grammatical error detection (GED).

Content classification: Sentiment (IMDB), topic (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), knowledge probing (MyriadLAMA), commonsense QA (CSQA), and temporal fact probing (TempLAMA).

Self-reflection: Hallucination (HaluEval), toxicity (Toxic), and stereotype detections (Stereoset).

Multilinguality: Language identification (LTI), multilingual POS tagging (M-POS), sentiment classification (M-Amazon), knowledge probing (mLAMA), and textual entailment (XNLI).

5.3 NEG-based Skill Neuron Probe

We train skill neuron probes using NeurGrad's estimated NEG to examine how NEG encodes general language skills. Each skill dataset $\mathcal{D}$ is split into: training, validation, and test sets ($\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{valid}}$, and $\mathcal{D}_{\text{test}}$), with a ratio of 6:1:1. We compare the following three probes with different designs.

Polar-Probe is a majority-vote classifier, where each neuron votes based on its polarity (positive or negative). For each training pair (q_i, a_i) in $\mathcal{D}_{\text{train}}$ and neuron n_k, we record its polarity as feature $\boldsymbol{x}_{q_i,a_i}^{n_k}$. The dominant polarity across $\mathcal{D}_{\text{train}}$ defines the neuron's global polarity $\bar{\boldsymbol{x}}^{n_k}$, and neurons with more consistent polarity are ranked higher.

To predict q_i, we compare polarities $\boldsymbol{x}_{q_i,a_j}^{n_k}$ for each candidate $a_j \in \hat{\mathcal{A}}_{\text{cands}}$ and $n_j \in \mathcal{N}_s^*$:

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

The optimal size of $\mathcal{N}_s^*$ is selected using $\mathcal{D}_{\text{valid}}$.

Magn-Probe uses NEG magnitudes as features for a majority-vote classifier. For each training pair (q_i, a_i) in $\mathcal{D}_{\text{train}}$ and neuron n_k, we compare NEGs across $a \in \hat{\mathcal{A}}_{\text{cands}}$. Neurons that consistently show the highest or lowest NEGs for the correct answer a_i are marked as skill indicators, along with their preference for being highest or lowest. More consistent neurons are ranked higher. At inference, predictions follow the same voting rule as in Eq. 4. This probe evaluates whether NEG magnitudes can encode skill information.

Tree-Probe is designed to analyze the impact of accounting for interdependencies among skill neurons. We use the index (non-negative integers) of $a \in \hat{\mathcal{A}}_{\text{cands}}$ with the most significant NEGs as features for training a random forest classifier. The hyperparameters include the number of trees (`#n_trees`) and layers (`#n_layers`) used in each tree. See more details in § G.2.

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

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

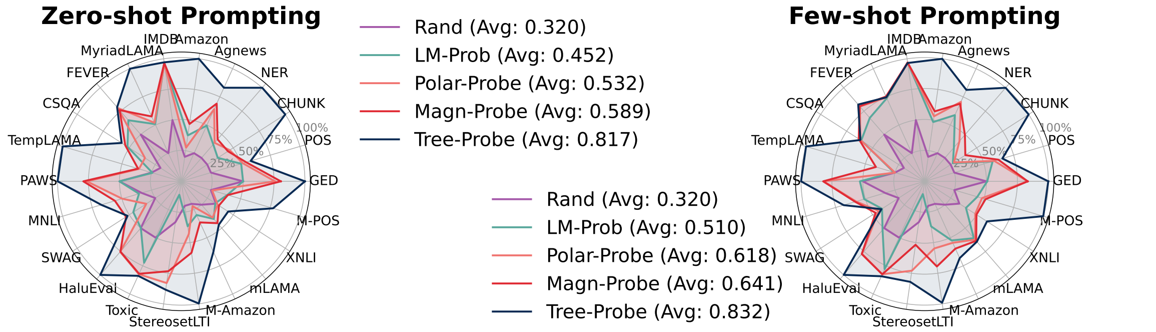


Figure 6: MCEval8K accuracies on Llama2-7B across tasks in zero-shot and few-shot settings, reported for Rand (random guess), LM-Prob (token probability), and three proposed probes. Legends show average accuracies.

Model	**NER**			**Agnews**			**PAWS**			**CSQA**			**HaluEval**			**mLAMA**		
	LM	Act	Mag	LM	Act	Mag	LM	Act	Mag	LM	Act	Mag	LM	Act	Mag	LM	Act	Mag
Llama3.2-3B	.598	**.633**	.618	.757	**.857**	.776	.500	**.843**	.831	.551	.681	**.691**	.650	**.822**	.780	.473	.563	**.570**
Llama2-7B	.361	.453	**.498**	.588	**.849**	.702	.524	**.825**	.815	.610	.613	**.639**	.520	**.788**	.783	.608	.622	**.637**
Qwen2.5-7B	.871	**.877**	**.877**	.751	**.858**	.755	.574	**.889**	.873	.768	**.816**	.814	.637	**.821**	.773	.784	**.795**	.785
Llama3.1-8B	.815	.826	**.833**	.770	**.883**	.812	.500	**.860**	.854	.674	.733	**.740**	.674	**.807**	.798	.762	.781	**.785**
Llama2-70B	.790	-	**.817**	.763	-	**.824**	.779	-	**.846**	.754	-	**.763**	.753	-	**.825**	.743	-	**.760**

Table 3: Accuracy of six tasks across five LLMs: **LM**-Prob, **Act**ivation-based, and **Magn**-Probe.

5.4 Experimental Setup

Dataset & Prompt. We split each MCEval8K dataset into train, validation, and test by the ratio of 6:1:1, ensuring balanced correct-answer tokens across subsets to avoid majority label bias (Zhao et al., 2021). We manually craft instructions and options for all MCEval8K tasks to ensure single-token outputs, considering both zero-shot and few-shot settings. See § I for all task instructions.

Probe. For majority-based probes, we select the optimal neuron size from $2^n(0 \leq n \leq 13)$ using $\mathcal{D}_{\text{valid}}$. For the Tree-Probe (random forest), we report accuracy using scikit-learn's default settings, where the optimal subset of features is selected automatically: 100 trees with no depth limitation. See § G.1, G.2 for details.

Models. We probe Llama2-7B with three NEG-based probes on all tasks in MCEval8K. For other LLMs, to reduce computational cost, we only probe one dataset per genre (NER, Agnews, PAWS, CSQA, HaluEval, and mLAMA) with 2^{10} training examples, focusing on majority-vote probes.

5.5 Result and Analysis

We compare our skill probes with three baselines: **Rand** (random guessing, **LM-Prob** (choosing the token with highest model probability), and **Act** (activation-based probes from Wang et al. (2022)).

Can NEG capture language skills? Figure 6 shows Llama2-7B's accuracy on MCEval8K tasks in zero- and few-shot settings. The results show that LM-Prob consistently outperforms random guessing, indicating that Llama2-7B can follow instructions to activate the relevant language skills. However, even the simple majority-vote probes can surpass the LM-Prob, suggesting that NEGs capture meaningful information about language skills in individual neurons. Furthermore, Tree-Probe significantly outperforms majority-vote probes by integrating NEGs from multiple neurons. This highlights that the interplay among neurons is crucial for accurately representing language skills. See Tables 8 and 9 in Appendix § G.1 for details on all tasks. Experiments on other LLMs further validate our findings as reported in Table 3.

Which encodes skills better: activation or gradient? Table 3 demonstrates that NEG-based probes have a different focus from the activation-based approach. Specifically, the gradient-based probes perform better on knowledge-intensive language tasks, such as factual knowledge and commonsense reasoning (mLAMA and CSQA). One possible explanation is that simple linguistic knowledge becomes saturated during pretraining, while complex knowledge remains undertrained, making it more effectively captured by NEG-based probes.

Furthermore, NEG-based probes outperform

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 4: The optimal neuron sizes for Mag-Probe.

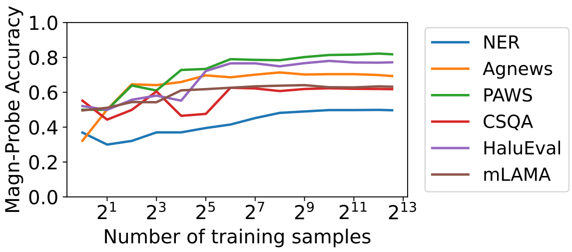


Figure 7: Accuracies with varying training sizes.

activation-based methods in efficiency by assigning distinct values to different target tokens per neuron, enhancing representational capacity. In contrast, activation-based approaches rely only on forward pass signals, limiting neurons to binary distinctions and making them less effective for multi-class tasks (Wang et al., 2022).

How do NEGs make predictions? We analyze NEGs associated with each option token to explore why simple majority-vote classifiers can achieve high accuracy. Using PAWS (binary classification) as an example, we acquire NEGs for target tokens ("*yes*/*no*") across 6K prompts. We find 97.21% of neurons show opposite signs for "*yes*/*no*" tokens, with a near-perfect negative correlation (-0.9996), indicating strong polarity across options. We further compare the NEGs given zero-shot and few-shot prompts. We find that NEG magnitudes of few-shot prompts over 22 tasks are 5.36 times larger than zero-shot, suggesting that adding demonstrations can effectively activate skill neurons.

6 Properties of Skill Neurons

To deepen our understanding of how language skills are represented within neurons, we analyze skill neurons' behavioral and structural properties through a series of focused studies.

6.1 Representation & Acquisition Efficiency

Representational efficiency: By tuning neuron size on the validation set, we find that skill-neuron probes achieve high accuracy with only a few neurons. The optimal neuron sizes for all tasks with Magn-Probe are reported in Table 4. Most tasks can achieve optimal accuracy within 256 neurons, demonstrating the efficiency of NEG in representing language skills. Notably, factuality tasks, such as MyriadLAMA, CSQA, and mLAMA, engage more neurons, suggesting that handling facts requires more diverse neurons, reflecting the complexity of factual understanding tasks.

Acquisition efficiency: Figure 7 reports the accuracy of skill-neuron probes with different training examples. While adding training examples can consistently increase the probes' accuracy, the earnings slow down after 128, indicating the efficiency of acquiring skill neurons with limited data.

6.2 How Robust Are Skill Neurons to Contextual Variations?

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 $\mathrm{ACC}^X_{\mathcal{N}_s^X}$. Then, we use the classifier built with $\mathcal{N}_s^X$ to evaluate the task by context Y as $\mathrm{ACC}^Y_{\mathcal{N}_s^X}$. We denote the robustness of $\mathcal{N}_s^X$ on context Y as $\frac{\max(\mathrm{ACC}^Y_{\mathcal{N}_s^X}-\alpha,0)}{\max(\mathrm{ACC}^Y_{\mathcal{N}_s^Y}-\alpha,0)}$, 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 robustness across combinations, we find that it remains very high (near 1) for varying instructions and demonstrations but drops significantly when target tokens change. The results show that skill neurons remain highly robust to changes in instructions and demonstrations but lose robustness when output tokens change. See § J for details of experimental settings and results.

6.3 Are Neurons Substitutable in Representing Language Skills?

We investigate whether skill neurons uniquely represent skills or can be substituted by different neurons. For investigation, we build Magn-Probes using different neuron sets. Specifically, we select 64 consecutive neurons from the ranked list, ordered by their importance as skill indicators (§ 5.3).[12]

[12]We use 64-neuron units that maintain high accuracy across tasks (§ A). With 352,256 neurons in Llama2-7B's FF layers, this produces 5,504 unique neuron sets.

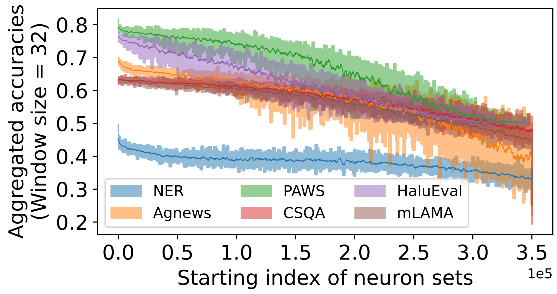


Figure 8: Accuracies of Magn-Probe with different neuron sets, plotting the mean accuracy within each window, along with the accuracy ranges, as the envelope. Neuron sets are selected from all neurons in Llama2-7B in groups of 32, ranked by importance as skill indicators.

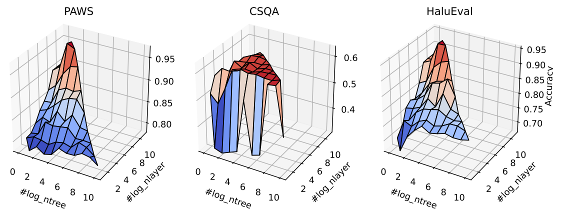


Figure 9: Accuracies of Tree-Probes with varying depths and trees. *X-axis*: logarithm of trees' number; *Y-axis*: logarithm of tree depths; *Z-axis*: Accuracy.

Figure 8 depicts the accuracies across six tasks with different neuron sets, showing that representational accuracy declines only gradually with less important neurons. Even using the least important neurons still outperforms Rand. This suggests that numerous neurons can act as skill indicators and skill neurons are broadly distributed and substitutable (see § H.2 for full results).

6.4 Do Skill Neurons Depend on Each Other?

The majority-vote probes assume neuron independence, while the Tree-Probe model their interdependencies via hierarchical classifiers. Figure 6 shows that Tree-Probles better represents language skills by accounting for neuron interdependency. To see how important interdependency is in encoding language skills, we train Tree-Probes with varying `#n_trees` and `#n_layers`.[13]

Figure 9 reports the resulting accuracies of Tree-Probe on PAWS, CSQA, and HaluEval. Their different shapes indicate the interdependency style between neurons for diverse skills differs. Some tasks (PAWS) prefer deep layers, while some (CSQA) prefer more trees, and some (HaluEval) require a balance. See § H.3 for details.

[13] `#n_trees` and `#n_layers` are set to 2^N and 2^M, where $0 \leq N \leq 10$, $1 \leq M \leq 11$, and $N + M < 12$.

7 Related Work

Neuron-level knowledge attribution methods. Previous work links neurons to knowledge (Oba et al., 2021) by measuring their impact on model predictions. Some use causal interventions (Geva et al., 2021; Meng et al., 2022; Chen et al., 2023; Wang et al., 2024), while others rely on heavy tensor calculations (Geva et al., 2022; Lee et al., 2024; Yu and Ananiadou, 2024). These methods are computationally costly, limiting their use for large-scale probing across diverse LLM prompts. They also focus on relative neuron rankings rather than the precise, quantitative neuron-output relationships, reducing their applicability in use cases like knowledge editing (Zhang et al., 2024) and bias mitigation (Gallegos et al., 2024). Gradient-based approaches (Lundstrom et al., 2022; Dai et al., 2022) similarly face high computational costs.

Skill neuron probing. Neurons in feed-forward layers can encode specific language skills, enabling the tasks to be solved using only their activations; such neurons are called skill neurons (Wang et al., 2022; Song et al., 2024). Prior work showed that neurons represent semantic skills like sentiment classification (Wang et al., 2022; Song et al., 2024) and complex skills such as style transfer (Lai et al., 2024) and translation (Tan et al., 2024). However, these studies focus on activations as knowledge indicators, overlooking neuron gradients' potential to represent language skills, which limits neuron-level model adjustments.

8 Conclusions

This is the first study to provide a global quantitative link between feed-forward layer neurons and model output, and discuss the potential of precise LLM control via neuron modifications. Through neuron intervention experiments, we reveal the linear relationship between neurons and model outputs and quantify such linearity by "neuron empirical gradients" (NEG) and propose NeurGrad, an efficient NEG estimation method. Finally, we verify NEG's ability to represent various language skills associated with diverse prompts through skill neuron probing.

As future work, we plan first to explore the tradeoff between precision and intensity in neuron modification based on NeurGrad, and then leverage these insights to develop a broader neuron-level modification methodology for applications such as knowledge editing and bias mitigation.

9 Limitations

Our research establishes a framework to quantitatively measure neuron influence on model output and shows that empirical gradients effectively represent language skills, linking language skill representation to outputs via neuron empirical gradients. However, tuning neuron values for skill-level output adjustment remains unexplored and could offer a more efficient alternative to weight-level tuning. We also plan to assess NEG's role in language generation skills, aiming for dynamic behavior changes without altering LLM parameters, reducing costs and improving flexibility in model adaptation.

Currently, our neuron linearity and gradient analysis focus on single-token factual prompts. Future work will expand to diverse domains and multi-token contexts to support wider use cases with language generation.

Acknowledgments

This paper is based on results obtained from a project, JPNP22007, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Ralf 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, Doha, Qatar. Association for Computational Linguistics.

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 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*, pages 2475–2485, Brussels, Belgium. 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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, and Others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

D. Heath, S. Kasif, and S. Salzberg. 1993. k-dt: A multi-tree learning method. In *Proceedings of the Second Intl. Workshop on Multistrategy Learning*, pages 138–149.

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. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

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 (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

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. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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 Validad 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 Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus Irawan, 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, Tai Ngee Chia, Ayu Purwarianti, Sebastian Ruder, William Chandra 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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Daniel D Lundstrom, Tianjian Huang, and Meisam Razaviyayn. 2022. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14485–14508. PMLR.

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. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

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 *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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.

Daisuke Oba, Naoki Yoshinaga, and Masashi Toyoda. 2021. Exploratory model analysis using data-driven neuron representations. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 518–528, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and Others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

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.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

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 *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

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:

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

Table 5: 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 ± 2.

$$\mathbf{LG} \triangleq \text{coverage}_{\text{layer}} \times \text{distribution}_{\text{layer}}, \tag{5}$$

$$\mathbf{PG} \triangleq \text{coverage}_{\text{prompt}} \times \text{distribution}_{\text{prompt}}, \tag{6}$$

where coverage and distribution are defined as:

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

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

where x refers to either layer or prompt, $\mathrm{maxVar}(\cdot)$ denotes the max possible variance. High coverage and distribution are desirable; a perfect generality then achieves coverage of one and distribution of one. The statistics in Table 5 suggest that linearity is a general property of neurons, largely independent of specific prompts or Transformer layers.

B Neurons' Statistics

B.1 Distribution of Neuron Activations

In this section, we analyze the distribution of neuron activations across seven PLMs, illustrated in Figure 10. The models include three BERT-based PLMs and three instruction-tuned LLaMA-2 LLMs. Figure 10 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 Section § 2.

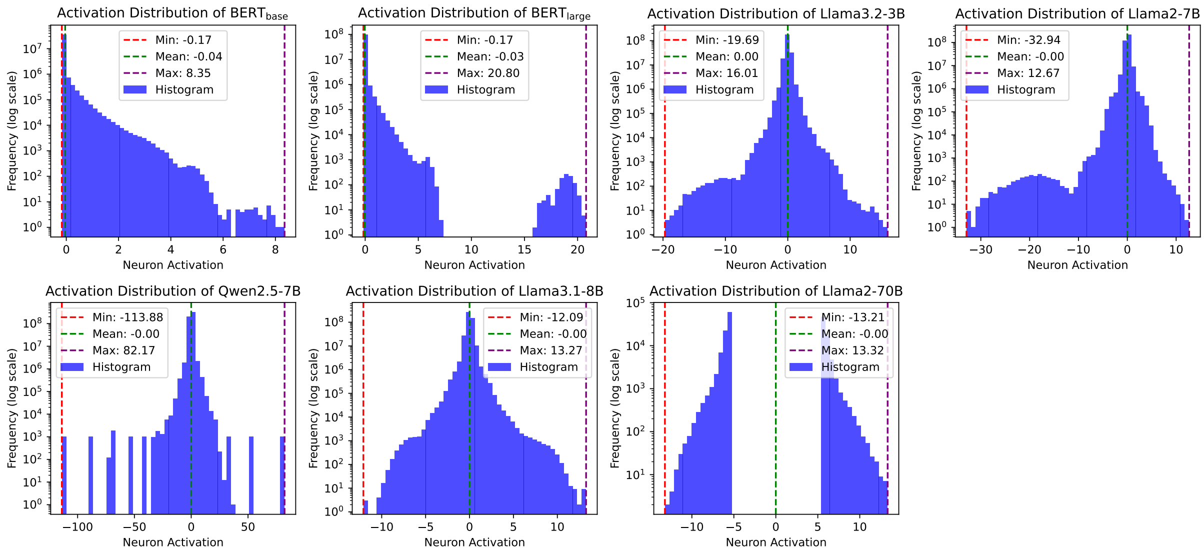


Figure 10: Histograms of neuron activations for seven models across 1,000 prompts, displayed on a logarithmic y-axis. The figure includes three BERT models and three LLaMA-2 models, with each subplot showing the distribution of activations for one model.

B.2 Distribution of Neuron NEGs

In this section, we report the distribution of neurons' NEG across the seven PLMs, as illustrated in Figure 11. Similar to our discussion in Section § 4.1, neurons capable of altering model output are not rare. For instance, in 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 exhibit zero gradients when averaging the NEGs across all neurons.

B.3 How are neurons distributed across layers?

We examine the variation in the absolute value of NEGs across Transformer layers to understand the distribution of neuron controllability. The Figure 12 illustrates the means and confidence interval 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 between variances and means,[14] 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 two BERT PLMs. We exclude LPI from the following experiments as LPI cannot be applied to BERT models. We follow a similar experiment setup to Section § 3.3.

The evaluation results are illustrated in Figure 13. 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.

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

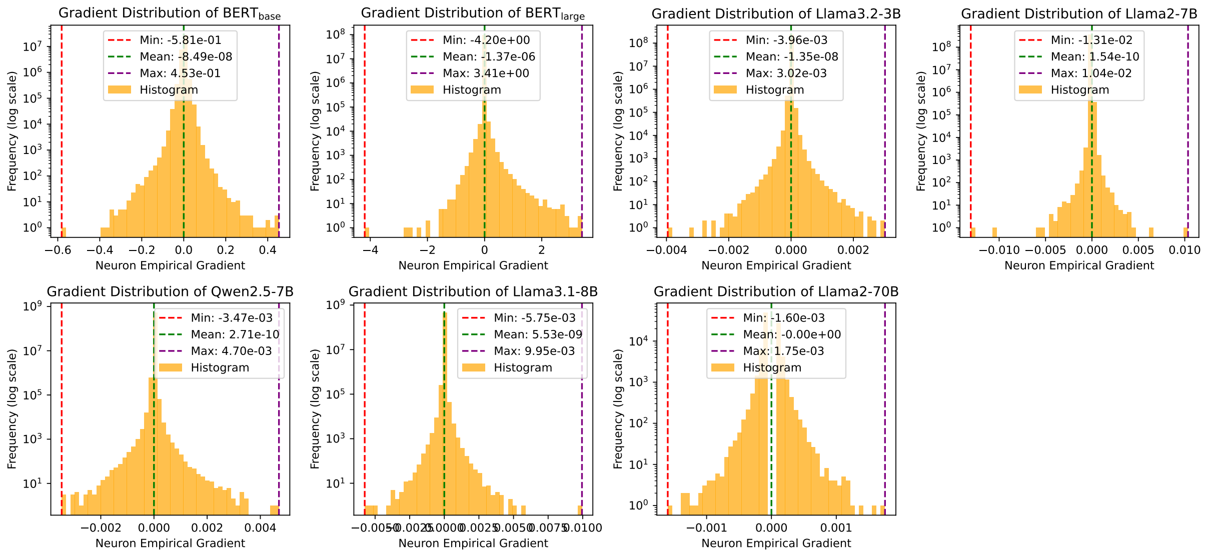


Figure 11: Histograms of neuron NEGs for seven models across 1,000 prompts, displayed on a logarithmic y-axis.

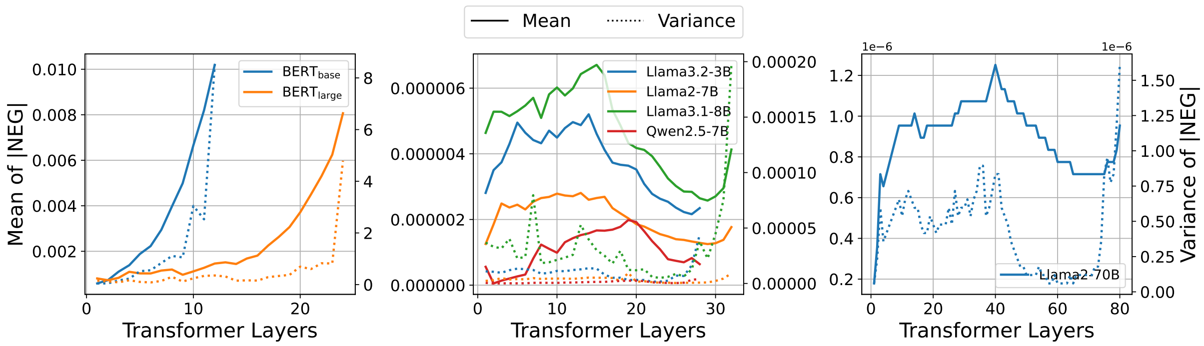


Figure 12: Means and variances of NEG magnitudes across Transformer layers on seven models. The data is calculated from the average of 1000 prompts.

D Multi-neuron Intervention: Supplementary Experiments

In this section, we report the multi-neuron intervention experiments conducted with different enhancement ranges on BERT$_{\text{base}}$ and Llama2-7B, following the similar experiment setup to Section § 4.2. 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 14. Figure 14 demonstrates that a larger enhancement range, involving more neurons, can vastly reduce the Corr, suggesting the output shift is less predictable. However, we can observe that BERT$_{\text{base}}$ consistently achieves strong Corr. (>0.8) for any scenario. While Llama2-7B is less stable than BERT$_{\text{base}}$, it can still maintain moderate positive Corr. (>0.5) for 4096 neurons with an enhancement range of [0,2].

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`

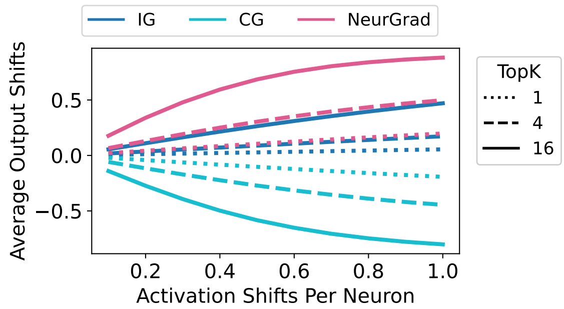
Llama3.2-3B: `https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct`

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

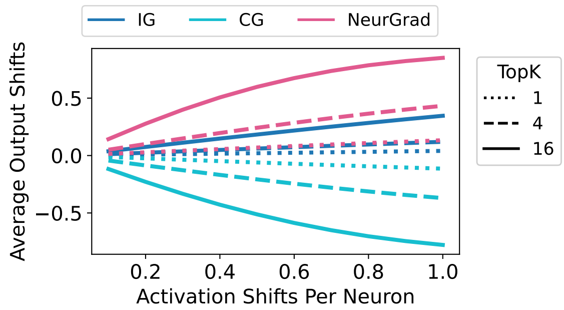
Qwen2.5-7B: `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`

Llama3.1-8B: `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

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



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



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

Figure 13: Knowledge attribution evaluation by comparing CG, IG, and NeurGrad on two BERT.

Model	#n_layers	#neurons_per_layer
$\mathrm{BERT}_{\mathrm{base}}$	12	3,072
$\mathrm{BERT}_{\mathrm{wwm}}$	24	4,096
Llama3.2-3B	28	11,008
Llama2-7B	32	11,008
Qwen2.5-7B	28	18,944
Llama3.1-8B	32	14,336
Llama2-70B	80	28,672

Table 6: Number of Layers and Intermediate Neurons per Layer for BERT and Llama2 Models

The statistics of these seven PLMs, including the number of layers (#n_layers) and neurons per layer (#neurons_per_layer) are listed in Table 6.

F Construction of MCEval8K

The motivation behind creating MCEval8K is to establish a comprehensive benchmark that spans diverse knowledge genres and language skills. Since we aim 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 convey 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.

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 7. 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.

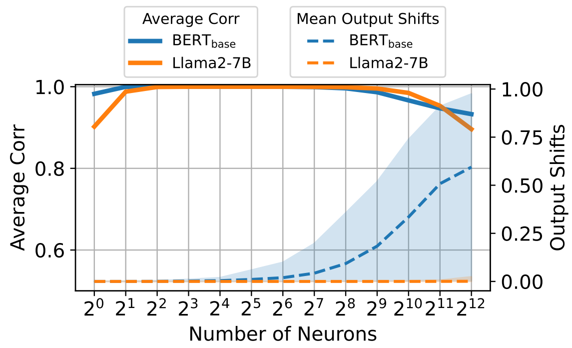
MyriadLAMA For factual inquiries formed from $<\mathrm{sub}_i, \mathrm{rel}_j>$, we collect all objects that appear as the target of the rel_j within the dataset and perform sampling to select three additional objects alongside the ground-truth tag.

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

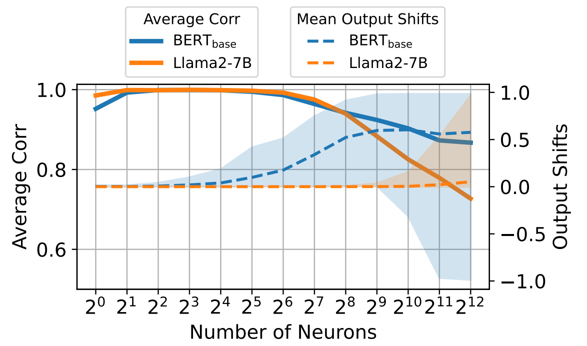
M-POS The process is similar to POS, applied separately for each language.

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

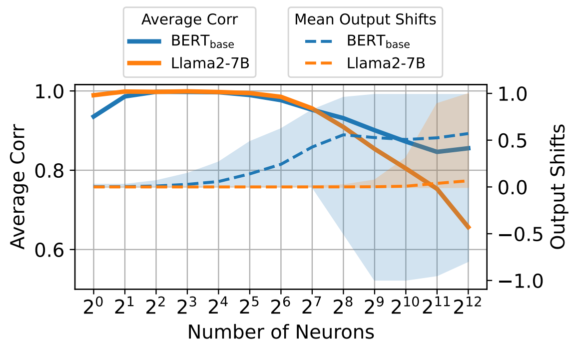
Balanced Options: Most datasets, except for Stereoset, contain over 8000 data points. To ensure balance across all options, we perform balanced sampling so that each option has an equal number of examples. From these datasets, we split 8000 examples into training, validation, and test sets, allocating 6,000, 1000, and 1000 examples,



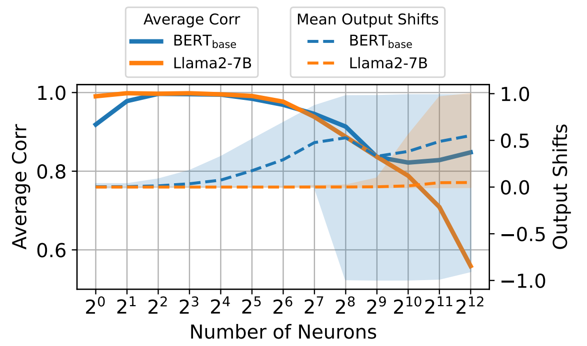
(a) Enhancement range of [0, 0.1].



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



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



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

Figure 14: Multi-neuron intervention experiment results with different enhancement ranges.

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 data 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

G.1 Per-task Probing Result

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 8) and few-shot prompt setting (Table 9). For two majority vote probes, optimal accuracies are acquired by performing a hyperparameter (optimal neuron size) search on the validation set and evaluating the test set. We report the optimal neuron sizes for all tasks and the table's accuracies. For the random-forest probing (Tree-Probe), we directly use the gradients of all neurons to train the random forest tree. As the random forest training algorithm only takes important features to construct the decision trees, we also report the number of neurons used to build the random forests. The details of the random-forest-based probe are introduced in § G.2.

G.2 Random Forest-based Probe

What is the random forest algorithm? The random forest is an ensemble learning algorithm (Heath et al., 1993) that works by creating a multitude of decision trees during training. For our multi-choice classification tasks in MCEval8K, the random forest's output is the option selected by most trees. A decision tree is a supervised learning model that makes predictions by recursively splitting data based on feature values. During training, the tree builds nodes by selecting features that best separate the data according to a chosen metric, such as Gini impurity. Splitting continues until the data in each leaf node is sufficiently pure or a maximum depth is reached. During inference, a new input is passed through the tree by following the feature-based decisions from the root to a leaf, where the final prediction is determined by the majority label or average value of samples in that leaf.

Feature design: Our study's objective is to explore the effectiveness of using NEG as features for knowledge representation and conduct further analysis. Therefore, the inputs for training and inference in the random forest model are constructed solely based on gradients estimated by NeurGrad. Specifically, each neuron is assigned an integer value for a given prompt. In our classification tasks, a neuron's feature is set to i if the gradient associated with the i-th token for that neuron is the largest among the gradients computed for all other candidate tokens (options). We ignore information on the smallest gradients to reduce the size of feature spaces.

Implementation details: For the implementation, we directly use *RandomForestClassifier* in scikit-learn (Pedregosa et al., 2011) for training and inference. We use the default parameters of RandomForestClassifier besides the number of trees (#n_trees) and layers (#n_layers) used in each tree. The number of trees refers to the number of decision trees used to ensemble the random forest. The number of layers refers to the layer depth for each tree. Noted that RandomForestClassifier constructs binary trees; thus, the number of features used in each tree is equal to or less than $2^{\#\text{n\_layers}} - 1$.

Visualization: We present an example of a single-tree random forest model learned from the PAWS dataset in the few-shot setting, illustrated in Figure 15. For learning this decision tree, the number of trees and layers is set to 1 and 8. The PAWS dataset is a binary classification task with candidate tokens "yes" and "no." To construct features for each neuron, we compare the NEGs computed by NeurGrad for the "yes" and "no" prompt pairs. If the gradient estimated for the prompt-yes pair exceeds that of the prompt-no pair, we assign a feature value of 1; otherwise, we assign 0.

H Additional Analysis on Probing Results

H.1 More Data about Efficiency

We report the accuracies of majority-vote probes with different neuron sizes for all tasks to provide additional evidence for the discussion about the representation and acquisition efficiency of skill neurons in § 6.1. The results are demonstrated in Figure 18 and Figure 19 for zero-shot and few-shot prompting settings.

H.2 Probing With Varying Neuron Sets

We report the aggregated accuracies across all 22 tasks in MCEval8K in Figure 20 to provide additional evidence for discussion in § 6.3. It demonstrates that many neurons can construct the classifiers in solving the language tasks, showing their ability to represent language skills and knowledge.

H.3 Tree-Probe: Flatteness vs. Hierarchy

To investigate the balance between hierarchy and independence of skill neurons, we train Tree-Probe with fixed neuron features (2^{10}) but with different depths and trees. For each task, we train 10 Tree-Probes, varying the number of trees (#n_tree $\in$ ($2^0 \sim 2^9$)) and the tree depth (#n_layer $\in$ ($2^{10} \sim 2^1$)), which fewer trees with deeper layers indicate a more hierarchical structure. All tasks show a camel curve given stronger hierarchies. We report the optimal #n_layer for different tasks as follows: CSQA(4), MNLI(16), SWAG(16), Stereoset(16), Agnews(32), Myriad-LAMA(32), mLAMA(32), XNLI(32), POS(64), FEVER(64), Toxic(64), LTI(64), GED(128), IMDB(128), M-Amazon(128), CHUNK(256), NER(256), Amazon(256), PAWS(256), HaluEval(256), M-POS(256), TempLAMA(1024). This demonstrates that Different language skills require different hierarchy levels. For instance, factual tasks benefit from flatter structures, while linguistic tasks prefer deeper hierarchies.

For all tasks in MCEval8K, we plot the accuracies of trained models with varying hyperparameters, including the number of trees and layers per tree. The number of trees is set to 2^N, where N ranges from 0 to 10, and the number of layers is set to 2^M, where M ranges from 1 to 11. Training is conducted only for configurations where

$N + M < 12$. The results are visualized as 3D surfaces, where the x-axis represents the logarithm of the number of trees (#log_ntree), the y-axis shows the logarithm of the number of layers (#log_nlayer), and the z-axis indicates the accuracy evaluated on the test set. We display the results for all tasks under the zero-shot setting in Figure 22 and those under the few-shot setting in Figure 23.

I 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 § 5.4, 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
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, 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:
```

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? [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 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:

J 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 § 6.2. 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 probes. We use 32 as the optimal neuron size of PAWS with the few-shot setting is 32. Finally, we report the pairwise generality values among different prompting settings in Figure 21.

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 7: Details of datasets in MCEval8K.

Tasks	Rand	LM-Prob	Polar-Probe (#n_neurons)	Magn-Probe (#n_neurons)	Tree-Probe (#n_neurons)
GED	.5000	.5000	.7580 (16)	.8050 (1024)	1.000 (54644)
POS	.2500	.5050	.5190 (16)	.5470 (4)	.5850 (91290)
CHUNK	.2500	.3510	.4660 (8)	.4490 (16)	1.000 (93282)
NER	.2500	.3950	.4120 (32)	.4490 (8)	1.000 (97185)
Agnews	.2500	.4950	.6410 (32)	.6900 (2)	.8310 (49369)
Amazon	.2000	.3750	.2750 (256)	.4680 (128)	1.000 (85696)
IMDB	.5000	.9660	.9630 (8192)	.9650 (1024)	.9710 (15892)
MyriadLAMA	.2500	.5080	.5200 (4)	.5760 (4)	1.000 (80167)
FEVER	.5000	.6530	.7830 (32)	.7610 (32)	.7920 (45564)
CSQA	.2000	.5170	.3490 (1)	.5380 (16)	.5730 (96696)
TempLAMA	.2500	.2430	.3560 (4096)	.3640 (16)	1.000 (113786)
PAWS	.5000	.5000	.7640 (128)	.7920 (128)	1.000 (58200)
MNLI	.3333	.3560	.4980 (4)	.5590 (128)	.6740 (79711)
SWAG	.2500	.4610	.3360 (512)	.5310 (2)	.5160 (96955)
HaluEval	.5000	.4990	.7540 (1024)	.7510 (32)	1.000 (58987)
Toxic	.5000	.7230	.8250 (1024)	.8210 (16)	.8390 (32263)
Stereoset	.3333	.1096	.8299 (16)	.7335 (16)	.8847 (29242)
M-Amazon	.2000	.2990	.2350 (4096)	.3740 (2)	.6260 (97623)
LTI	.2000	.3670	.4300 (4)	.5830 (8)	.9970 (12068)
mLAMA	.2500	.4020	.3880 (128)	.4470 (4)	.4640 (79839)
XNLI	.3333	.3270	.3500 (256)	.3620 (16)	.4510 (79212)
M-POS	.2500	.3890	.2610 (1024)	.3930 (4)	.7740 (90001)

Table 8: Optimal accuracies across all MCEval8K tasks in the zero-shot setting on Llama2-7B, with the neuron sizes achieving these accuracies.

Tasks	Rand	LM-Prob	Polar-Probe (#n_neurons)	Magn-Probe (#n_neurons)	Tree-Probe (#n_neurons)
GED	.5000	.5060	.8330 (16)	.8330 (64)	1.000 (43465)
POS	.2500	.5730	.5870 (4)	.6210 (16)	.6550 (80695)
CHUNK	.2500	.2710	.2820 (8192)	.3910 (64)	1.000 (101539)
NER	.2500	.3610	.4300 (4)	.4970 (64)	1.000 (93577)
Agnews	.2500	.5880	.7060 (64)	.6890 (512)	.8120 (42846)
Amazon	.2000	.4840	.5310 (1)	.5680 (128)	1.000 (84055)
IMDB	.5000	.9700	.9700 (64)	.9690 (64)	.9660 (13823)
MyriadLAMA	.2500	.7380	.7450 (256)	.7530 (4096)	.7460 (70446)
FEVER	.5000	.6780	.8000 (1)	.8030 (4)	.8210 (38943)
CSQA	.2000	.6100	.6180 (32)	.6340 (8192)	.6180 (94246)
TempLAMA	.2500	.2600	.2500 (1)	.4110 (4)	1.000 (106140)
PAWS	.5000	.5240	.8180 (16)	.8210 (32)	1.000 (44060)
MNLI	.3333	.5100	.5780 (32)	.5860 (64)	.6830 (67771)
SWAG	.2500	.4100	.4430 (256)	.4710 (64)	.4160 (95311)
HaluEval	.5000	.5200	.7750 (2048)	.7770 (256)	1.000 (51411)
Toxic	.5000	.7800	.8250 (8)	.8260 (4)	.8430 (29766)
Stereoset	.3333	.1040	.7297 (128)	.5180 (16)	.8204 (29774)
M-Amazon	.2000	.5250	.5470 (1024)	.5880 (128)	.6820 (87424)
LTI	.2000	.3680	.5480 (64)	.6950 (8)	.9910 (28362)
mLAMA	.2500	.6080	.6230 (8192)	.6360 (512)	.6450 (75439)
XNLI	.3333	.3970	.4860 (32)	.4980 (32)	.5990 (80886)
M-POS	.2500	.4440	.4830 (4)	.5130 (8)	1.000 (95537)

Table 9: Optimal accuracies across all MCEval8K tasks in the few-shot setting on Llama2-7B, with the neuron sizes achieving these accuracies. The number of demonstrations is set to the same number of options for each task.

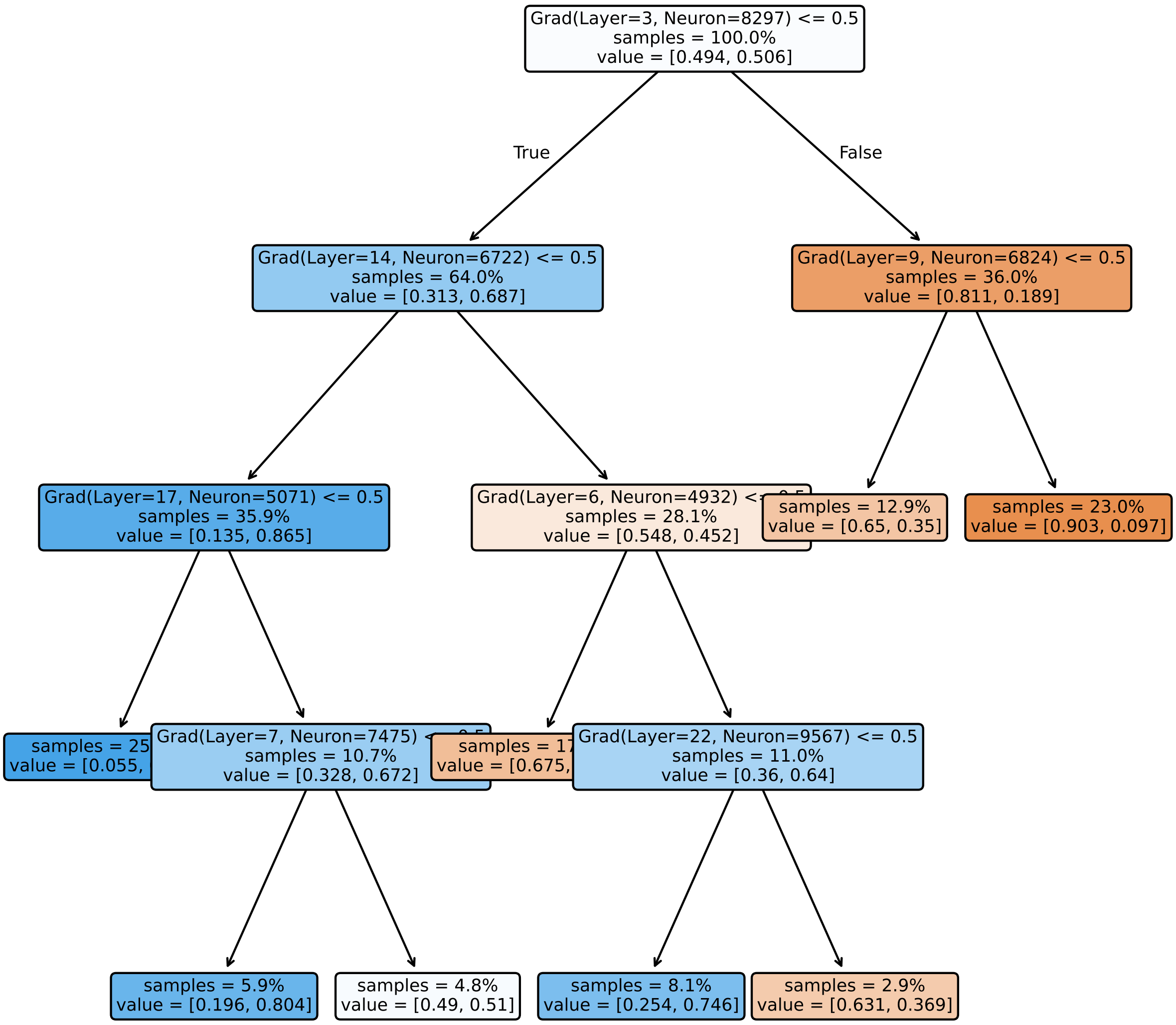


Figure 15: Visualization of a decision tree learned for the PAWS dataset with the few-shot setting on Llama2-7B model. The “samples” in each node refers to the percentage of samples reaching this node. The “value” shows the class distribution of samples in the node.

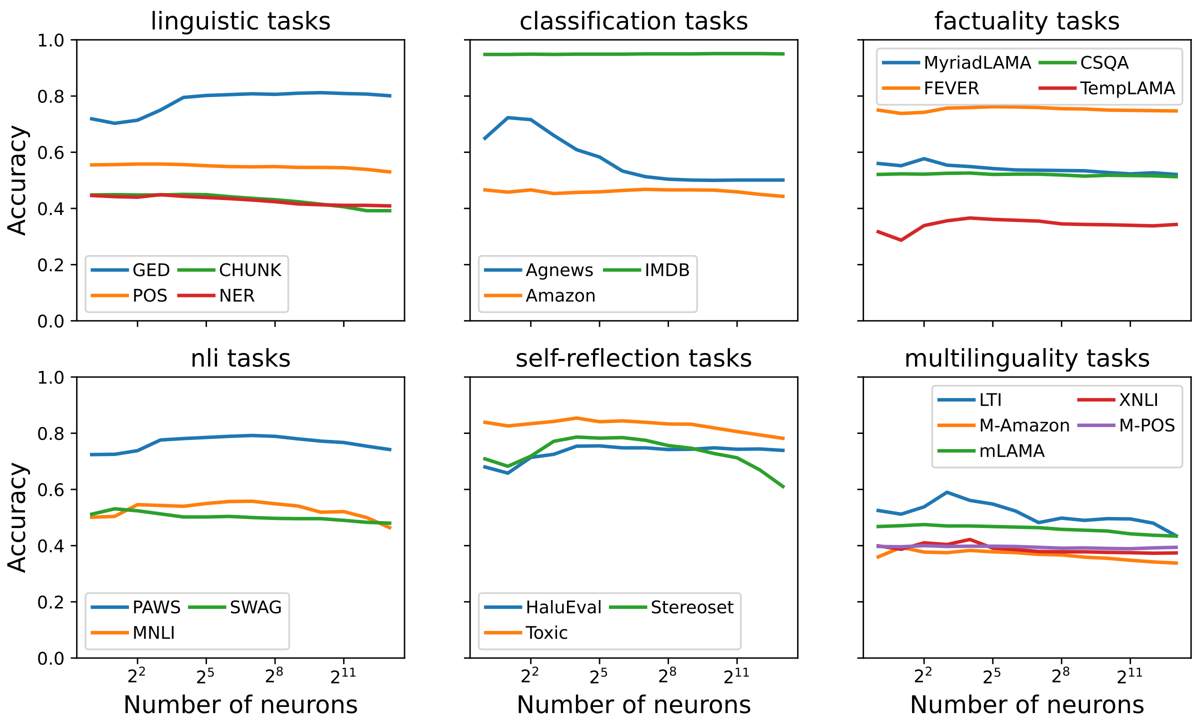


Figure 16: Per-task accuracies with varying neuron sizes on Llama2-7B, zero-shot prompt setting.

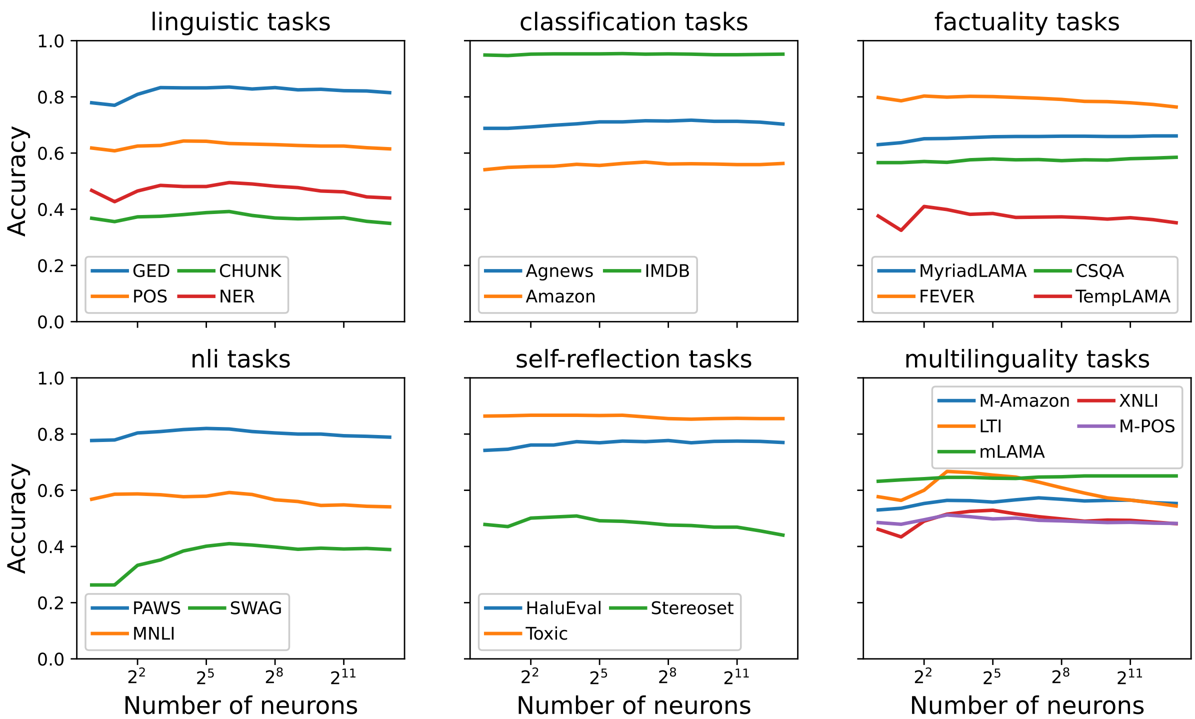


Figure 17: Per-task accuracies with varying neuron sizes on Llama2-7B, few-shot prompt setting.

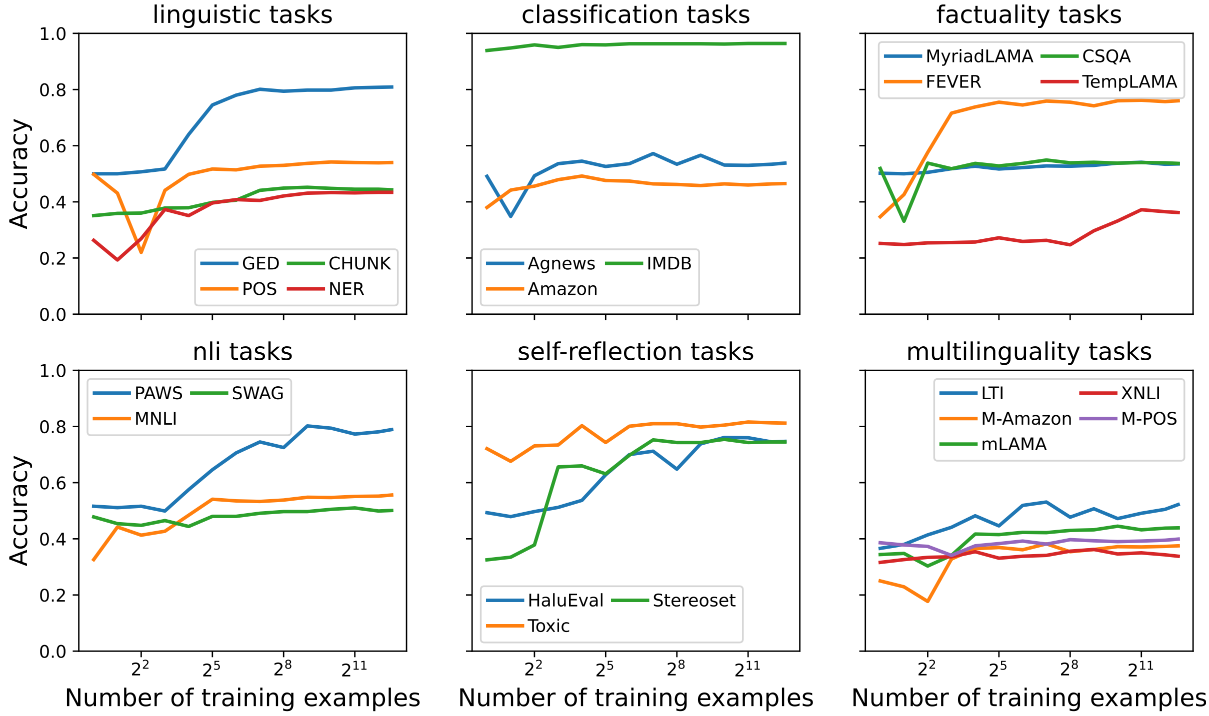


Figure 18: Per-task accuracies with the varying number of training examples on Llama2-7B, zero-shot prompt setting.

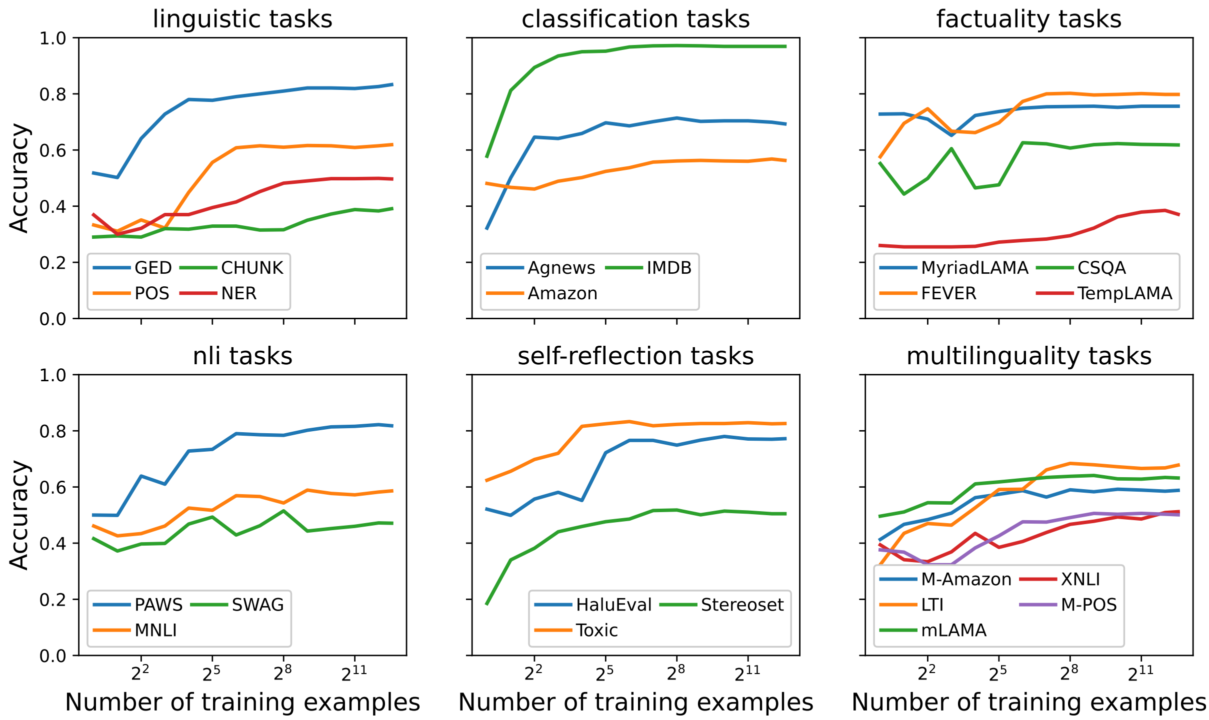


Figure 19: Per-task accuracies with the varying number of training examples on Llama2-7B, few-shot prompt setting.

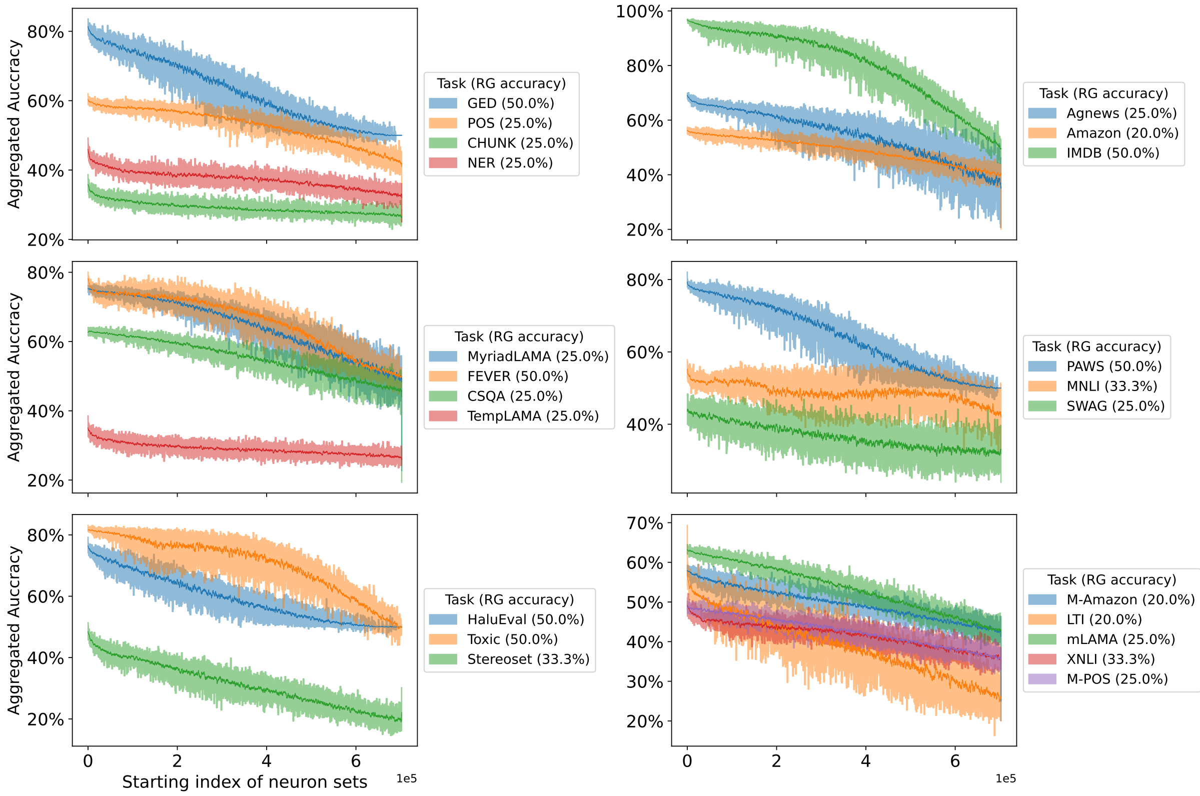


Figure 20: Per-task accuracies with varying neuron sets per with 64 neurons. We report the aggregated accuracies with a window size of 64 for better visualization, plotting the mean accuracy within each window, along with the corresponding accuracy ranges (minimum to maximum) as the envelope.

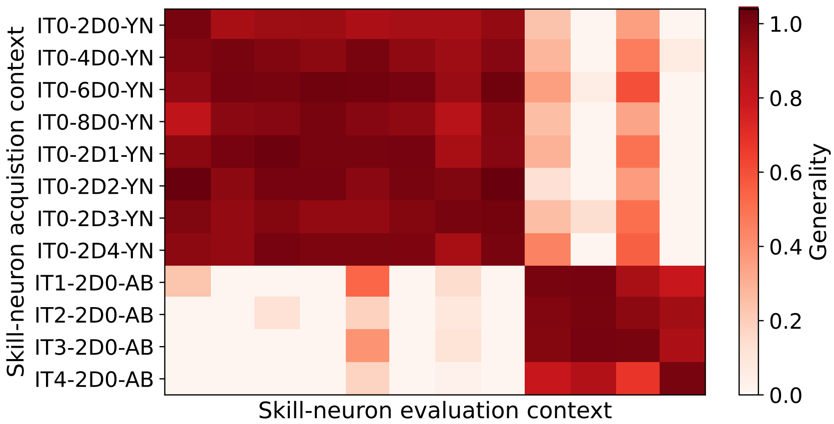


Figure 21: Generality of skill neurons across different contexts. **X-axis**: the context used to acquire skill neurons. **Y-axis**: evaluation context. The contexts on the x-axis are in the same order as on the y-axis. The context using the i-th instruction, k-th set of j-shot demonstrations, and yes/no answers is denoted as IT(i)-(j)D(k)-YN. “AB” refers to the a/b style options.

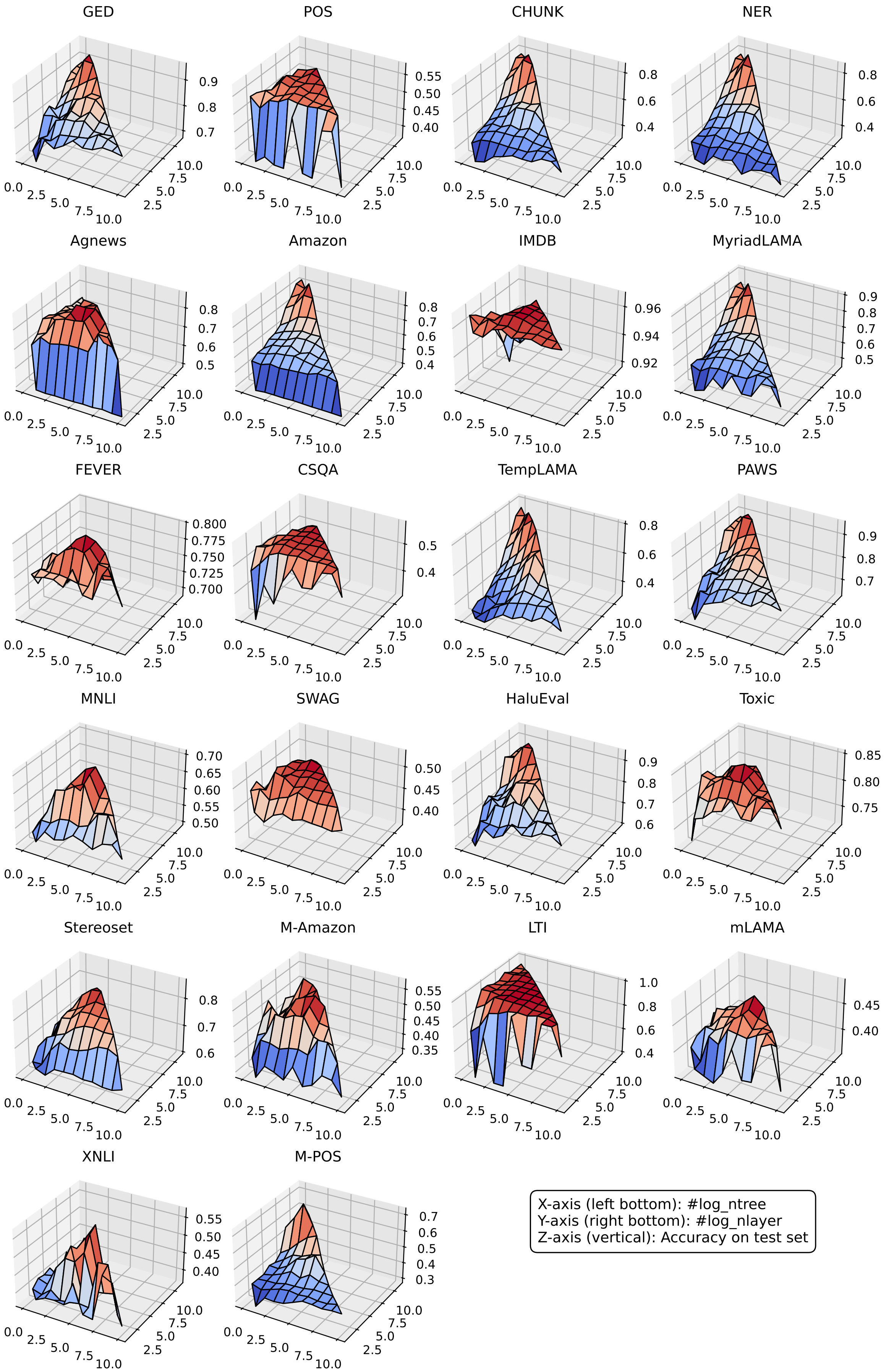


Figure 22: Accuracies of trained random forest models with the zero-shot setting on Llama2-7B.

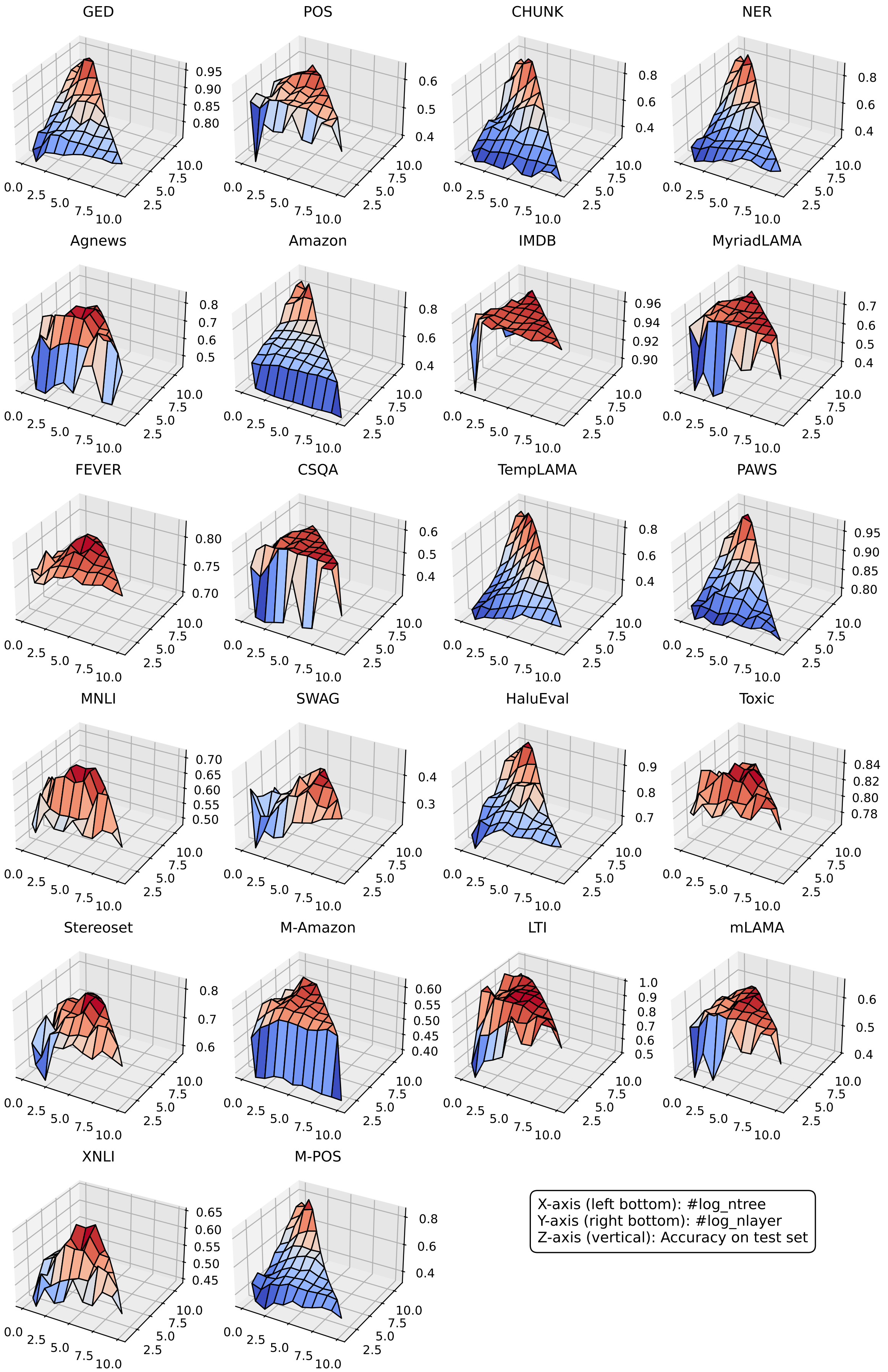


Figure 23: Accuracies of trained random forest models with the few-shot setting on Llama2-7B.