Rethinking Relation-Specific Neurons in Large Language Models

Lea Hirlimann [1 2] **Sebastian Gerstner** [1 2] **François Yvon** [3 †] **Hinrich Schütze** [1 2 †]

Abstract

Previous work has identified *relation-specific neurons* that selectively activate on specific semantic relations in factual knowledge tasks. However, the conclusions we draw about these representations depend heavily on the methodological assumptions underlying this procedure. We systematically reflect on three such assumptions, showing that (i) the number of relevant neurons varies across relations; (ii) the choice of internal signal for neuron identification shapes the results; (iii) cross-relation entanglement is structural rather than an artifact of subject overlap. We additionally present a preliminary investigation into the mismatch between benchmark-defined relation categories and model-internal organization. For instance, we show that the absence of a strong expert set for the `product_company` relationship reflects conceptual heterogeneity within the category rather than localization failure, and that targeted ablation of the subrelation `car_company` yields substantially stronger results. Together, our findings show that the apparent structure of relational representations is jointly shaped by the model's internal organization and the methodological lens applied to study it.

1. Introduction

A central goal of mechanistic interpretability is to identify the internal components of neural networks responsible for specific computations. For large language models (LLMs), this includes localizing where and how factual knowledge is stored. A growing body of work suggests that factual associations are not distributed uniformly across all parameters, but are partially localized within the MLP layers of transformer models (Geva et al., 2021; Meng et al., 2022). Recently, this picture was further refined by Liu et al. (2025), showing that knowledge is organized not merely at the level of individual facts, but as semantic relations, such as `person_mother` or `star_constellation`, represented through dedicated groups of neurons identifiable via expertise ranking of MLP column vector activations.

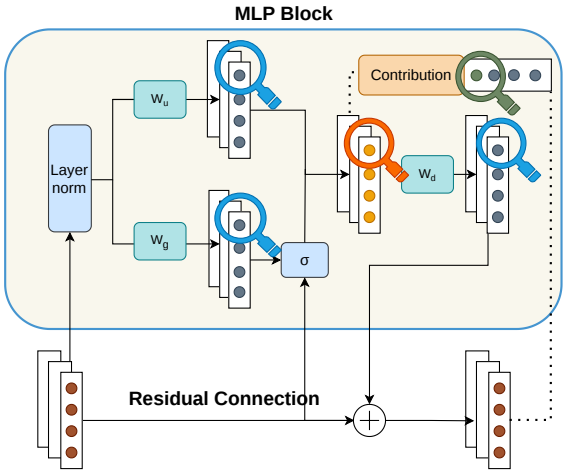


Figure 1. Which internal signal best identifies relation-specific neurons? We investigate different candidate expertise indicators at different stages of the MLP computation: column vector activations (blue lens), full neuron activations (orange lens), and residual stream contributions (green lens), evaluating each by its ability to produce strong and reliable expert sets.

We build on this foundation and ask a more uncomfortable question: to what extent do our conclusions about internal representations reflect genuine model structure, rather than the methodological assumptions embedded in the analysis? If the tools we use to localize knowledge systematically shape what we find, then the identified patterns and components cannot be taken as ground truth. By testing multiple new variants of the method presented by Liu et al. (2025) and evaluating them employing a custom metric, we challenge four assumptions implicit in prior work:

- **Challenged assumption 1:** *Column vector activations constitute the only informative signal for relation specific neuron identification.* Instead, experimenting with

[1]Center for Information and Language Processing, LMU Munich, Germany [2]Munich Center for Machine Learning (MCML), Germany [3]Sorbonne Université, CNRS, ISIR, France. [†]Equal advising. Correspondence to: Lea Hirlimann <hirliman@cis.lmu.de>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. *Copyright 2026 by the author(s).*

alternative expertise indicators, visualized in Figure 1, reveals that **multiple internal signals independently identify strong expert sets**. Moreover, combining complementary signals yields an even more reliable and compact set of relation-specific neurons than any single signal alone.

- **Challenged assumption 2:** *A single fixed neuron count adequately captures all relational representations.* Instead, we show that **the optimal number of neurons varies wildly across relations**. Ablating a fixed amount of relation specific neurons either does not fully resolve all contributors of a certain relations, or damages general language modeling capability.
- **Challenged assumption 3:** *Cross-relation entanglement is caused by data overlap.* Ablating the expert set of one relation frequently disrupts other relations. We find that this pattern persists even after controlling for subject overlap through reversed-relation controls, confirming that **entanglement reflects genuine coexistence of representations within MLP layers**.
- **Challenged assumption 4:** *The predefined relation categories correspond to coherent units of internal representations.* For some relations, no expert set yields strong ablation performance. We explore the possibility that the predefined relation category aggregates over multiple internally distinct representations. In preliminary results, we show that replacing `product_company` with the t-SNE-derived subcluster `car_company` leads to substantially stronger results, suggesting that **relation-specific neurons exist at a conceptual granularity different from that assumed by the dataset**.

Together, these findings advance our understanding of how relational knowledge is organized in LLMs; they also call for greater methodological reflexivity in mechanistic interpretability research.

2. Related Work

Localizing factual knowledge in large language models has attracted substantial research interest in recent years. Methods such as causal tracing (Meng et al., 2022), knowledge attribution (Dai et al., 2022) or sparse probes (Gurnee et al., 2023) have enabled increasingly precise identification of the components and circuits responsible for specific knowledge and functionalities.

In this paper we focus on the method of *contrastive activations*. By comparing activations across positive and negative examples of a concept, neurons that consistently discriminate the two can be identified, regardless of how the concept is lexically realized. This approach was first introduced by Cuadros et al. (2022) for activation steering to mitigate bias. It has since been adopted to identify language-specific and language-shared MLP components (Tang et al., 2024; Wang et al., 2024), treating MLP layer activations as discrimination signals in the spirit of the key-value memory view (Geva et al., 2021). The approach was originally developed for attention and MLP activations in GPT-2. Kargaran et al. (2025) and Liu et al. (2025) built on it and extended it to cover additional model families and architectures, including the SwiGLU nonlinearity (Shazeer, 2020) employed in the models used in this work (Touvron et al., 2023; Team et al., 2024).

For the specific case of relation-specific neurons, Liu et al. (2025) focused on column vector activations of the W_{gate}, W_{up}, and W_{down} projections, subsequently finding that down-projection columns do not contribute meaningfully to ablation success. The relation-specific neurons identified through this method were shown to possess three notable properties: they are *cumulative*, in that ablating progressively larger subsets degrades relational knowledge incrementally; *polysemantic*, in that the same neurons can contribute to multiple relations; and *competitively organized*, in that deactivating the expert neurons of one relation can incidentally improve performance on others. These properties imply a highly entangled internal structure in which relational representations coexist and interact within shared MLP components. We build directly on this framework, interrogating the methodological assumptions that shape what such analyses can and cannot reveal about this structure.

3. Methodology

We reuse the relation-specific neuron identification framework of Liu et al. (2025), extending it along three axes: we experiment with various internal signals to estimate neuron expertise; we select different neuron counts for different relations; and we apply stricter criteria to both data validation and ablation evaluation.

3.1. Dataset

We use the factual knowledge dataset of Hernandez et al. (2024), which provides factual triples (s, r, o) of subject s, relation r, and object o. Following Liu et al. (2025), we select the $R = 12$ semantic relations that each contain more than 300 data points. Each relation $r \in \mathcal{R}$ is associated with one or two prompt templates t_r, which are filled with subject–object pairs $(s, o) \in \mathcal{D}_r$ to form input prompts $t_r(s)$ with expected answer o. The data is split into training and test partitions $\mathcal{D}_r^{\text{train}}$ and $\mathcal{D}_r^{\text{test}}$.

Validation We only retain training datapoints where the model demonstrates prior knowledge of the answer. Whereas Liu et al. (2025) validated samples by checking

whether the first two generated tokens form a prefix of the correct answer o, we apply a stricter[1] criterion: a sample (s, o) is retained if the full token sequence of o appears within the first $k = 10$ generated tokens. This eliminates false positives introduced by incidental single-character prefix matches from the validated partition $\mathcal{D}_r^{\text{known}}$. Following Liu et al. (2025), $\mathcal{D}_r^{\text{test}}$ is left unfiltered, preserving prompts the model cannot answer prior to ablation and allowing any knowledge newly expressed after intervention to be captured.

Contrastive training data For each relation r, up to $N = 500$ samples are drawn uniformly at random from $\mathcal{D}_r^{\text{known}}$. These are combined with $4N$ contrastive samples drawn uniformly from all other relations $\mathcal{R} \setminus \{r\}$, yielding a balanced training set $\mathcal{T}_r$. The positive-to-contrastive ratio of 1:4 ensures that the expertise ranking is informed by a broad representation of competing relations.

3.2. Models

We investigate relation specific neurons in three open-weight language models, namely `llama-2-7b` and `llama-2-13b` from the Llama-2 family (Touvron et al., 2023), and `gemma-7b` (Team et al., 2024). This choice allows for comparisons across sizes and model families. All models share a GLU-based MLP architecture (Shazeer, 2020), as is usual in modern open-weights LLMs: Llama uses SwiGLU and Gemma uses GEGLU.

3.3. Neuron Features

The MLP of each transformer layer consists of a gate projection $W_g \in \mathbb{R}^{d \times d_{\text{ff}}}$, an up-projection $W_u \in \mathbb{R}^{d \times d_{\text{ff}}}$, and a down-projection $W_d \in \mathbb{R}^{d_{\text{ff}} \times d}$,[2] operating on hidden state $x \in \mathbb{R}^d$. For a token sequence of length T, let x_t denote the hidden state at position t. The **full neuron activation** is defined as:

$$h_i(x) = \sigma\big(x^\top w_{g,i}\big) \cdot \big(x^\top w_{u,i}\big) \tag{1}$$

where $w_{g,i}$ and $w_{u,i}$ denote the i-th column vectors of W_g and W_u respectively. The σ represents Swish in SwiGLU models such as Llama, and GELU in GEGLU models such as Gemma. The MLP as a whole outputs

$$\text{MLP}(x) = h(x) W_d \in \mathbb{R}^d, \tag{2}$$

where $h(x) \in \mathbb{R}^{d_{\text{ff}}}$ is the vector of all $h_i(x)$. We refer to the scalar pre-nonlinearity values $x^\top w_{g,i}$, $x^\top w_{u,i}$, and the individual components of $\text{MLP}(x)$, as **column vector** activations.

We cache activations across all prompts in $\mathcal{T}_r$ for each relation r, considering two positional aggregations: the *last token* value x_T and the *mean* over all positions $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$.

Additionally, we compute the **contribution** of each full neuron i to the residual stream following the formulation by Ferrando & Voita (2024), defined as:

$$c_i = \frac{\text{prox}(\text{out}_i, x_{\text{post}})}{\text{prox}(x, x_{\text{post}}) + \sum_l \text{prox}(\text{out}_l, x_{\text{post}})} \tag{3}$$

$$\text{prox}(a, b) = \max(-||a - b||_1 + ||b||_1, 0) \tag{4}$$

where $\text{out}_i = h_i(x_{in}) \cdot W_d^i$ is the vector that neuron i writes to the residual stream,[3] and $x_{\text{post}} = x + \text{MLP}(x)$ is the residual stream state after the MLP.

While column vector activations have been described as effective before, their ranking introduces redundancy, as gate and up projection ablations produce equivalent downstream effects on $h_i(x)$, and Liu et al. (2025) showed that down-projection columns are uninformative. Full neuron activations and residual stream contributions collapse these redundant signals into a single value per neuron, motivating our comparison of multiple internal signals as expertise indicators. This yields five distinct neuron feature signals for the expertise ranking: column vector and full neuron activations each under last-token and mean-token aggregation, and full neuron contribution scores at the last token position.

3.4. Neuron Ranking

For each neuron i and relation r, Liu et al. (2025) rank neurons by their ability to discriminate target samples $\mathcal{T}_r^+$ from contrastive samples $\mathcal{T}_r^-$ using **Average Precision (AP)**. Given the activation values $\{a_i(x_k), k = 1 \ldots N\}$ for neuron i across all samples, AP is defined as:

$$\text{AP}_i^r = \sum_n (\text{Recall}_n - \text{Recall}_{n-1}) \cdot \text{Precision}_n \tag{5}$$

where precision and recall are computed at each of n thresholds by treating higher activation as indicative of target membership. A perfect score $\text{AP}_i^r = 1$ is achieved when $a_i(x) > a_i(x')$ for all $(x, x') \in \mathcal{T}_r^+ \times \mathcal{T}_r^-$, i.e. when the neuron's activations are strictly higher on target than on contrastive samples.

This formulation, however, only rewards strictly higher target activations and fails to capture neurons with strictly lower activations on target samples. Full neuron activations in particular can take large negative values, as the sign of $h_i(x)$ is jointly determined by both gate and up projections.

Note as well, that AP_i^r measures discriminative separation of activation values across classes, but does not condition on activation magnitude. A neuron with near-zero activations

[1] see discussion of *strictness* in Appendix A.2

[2] We use left-to-right notation of the type xW rather than Wx.

[3] W_d^i denotes the i-th *row* of W_d

on both target and counter-relation samples can achieve a perfect score if the ordering is consistent.

Combined ranking. To leverage complementary information from column vector activations and residual stream contributions, we define an additional unified score that prefers neurons that are both strongly activated and contribute meaningfully to downstream computation. For each neuron i, strong relation discrimination is sufficiently indicated by either the gate or up projection column vector, as both jointly determine $h_i(x)$ and a strong discriminative signal in either is informative. We therefore take $\mathrm{AP}_i^{r,\mathrm{col}} = \max\big(\mathrm{AP}_i^{r,\mathrm{gate}}, \mathrm{AP}_i^{r,\mathrm{up}}\big)$ as the column vector expertise score, selecting whichever projection provides the stronger separation. This is then combined with the contribution AP score $\mathrm{AP}_i^{r,\mathrm{con}}$, which captures the neuron's functional impact on downstream computation, preferring neurons that are both strongly activated and meaningfully write to the residual stream. The combined expertise score is then:

$$\hat{\mathrm{AP}}_i^r = \mathrm{AP}_i^{r,\mathrm{col}} + \mathrm{AP}_i^{r,\mathrm{con}} \tag{6}$$

In all variations, neurons are ranked in descending order of the respective AP values, and the top-K neurons form the expert set $\mathcal{E}_r^K$ for relation r.

We use the following abbreviations: ColVec: column vector activations; FullNeu: full neuron activations after SwiGLU/GEGLU; Contrib: residual stream contribution scores; Combined: sum of avg. ColVec and Contrib AP scores.

3.5. Neuron Ablation

To validate expert neuron selection, we perform zero-ablation experiments in which the activations of all neurons in $\mathcal{E}_r^K$ are set to zero during inference. We then measure both the degree to which relational knowledge in the target relation is suppressed and the extent to which other relations are affected.

Representation Size Liu et al. (2025) evaluated ablation success using a single fixed expert count K across all relations, implicitly assuming uniform representation size or density. This assumption is unlikely to hold, as the number of relevant components and their individual contribution to a relational mapping depend on factors such as relation complexity, training frequency, and the cardinality of the subject and object spaces. We therefore treat K as a per-relation hyperparameter and select it via grid search over a logarithmically spaced candidate set:

$$K = \{100,\ 178,\ 316,\ 562,\ 1000,\ 1778,\ 3162,\ 5623,\ 10000,\ 17783\} \tag{7}$$

with the optimal $K_r^* = \arg\max_{K\in\mathcal{K}} F_1(r_{\text{target}}, K)$ selected on the training partition according to the objective described below.

Ablation Success Evaluation Let a_p be the expected answer to prompt p, and $o(p)$, $o_{\text{abl}}(p)$ the first k=10 tokens generated by the original and ablated model respectively. Knowledge of answer a_p is considered present if the full answer token sequence appears within the generation, i.e. $a_p \in o(p)$ in the same fashion as described for training data validation in section 3.1. Let P_r denote the set of evaluation prompts for relation r that the original model answers correctly according to this criterion.

We measure ablation success via a custom F_1 score that jointly rewards suppression of knowledge in the target relation and preservation of knowledge in all other relations. The *ablation recall* for the target relation r_{target} measures the fraction of prompts for which the model loses the correct answer after intervention:

$$\text{abl_success}(p) = \mathbf{1}[a_p \in o(p) \ \wedge\ a_p \notin o_{\text{abl}}(p)] \tag{8}$$

$$\text{Recall}(r_{\text{target}}) = \frac{1}{|P_{r_{\text{target}}}|} \sum_{p\in P_{r_{\text{target}}}} \text{abl_success}(p) \tag{9}$$

The *ablation precision* measures the average fraction of correct answers preserved across all non-target relations, penalizing collateral damage to other relational knowledge:

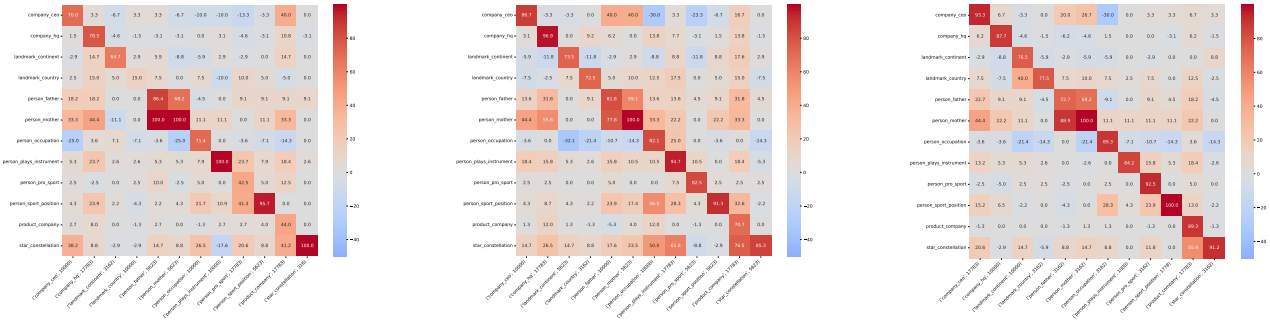
$$\text{stable}(p) = \mathbf{1}[a_p \in o(p) \ \wedge\ a_p \in o_{\text{abl}}(p)] \tag{10}$$

$$\text{Prec}(r_{\text{target}}) = \frac{1}{|\mathcal{R}|-1} \sum_{\substack{r\in\mathcal{R}\\ r\neq r_{\text{target}}}} \frac{1}{|P_r|} \sum_{p\in P_r} \text{stable}(p). \tag{11}$$

The combined F_1 score rewards expert sets that strongly suppress target-relation knowledge while leaving general language modeling ability intact. The potential positive effects of the ablation experiments, i.e., the correct answers that were not expressed before ablation in the relations $r \neq r_{target}$, do not affect the score.

4. Results and Discussion

Table 1 reports average F_1 scores across all models and method variants on the held-out test partition. Figure 2 provides a relation-level view of these results, showing the accuracy drop on each relation when its corresponding expert set $\mathcal{E}_r^{K^*}$ is ablated (diagonal), as well as the collateral effect on all other relations (columns). A perfect F_1 score for the experts of one relation matches 100% accuracy drop on the relation itself, and either no or positive effects on remaining relations within the same column. The per-relation F_1 scores on the training partition used for K^* selection are reported in the Appendix A.



(a) COLVEC (avg) *(b)* CONTRIB *(c)* COMBINED

Figure 2. Accuracy drop results of ablations made on the test data with `gemma-7b` of the best-performing method variants. x-axis: ablated neurons; y-axis: effect on relation.

		gemma-7b			llama-2-7b			llama-2-13b			all
Signal	**Pos.**	$F_1^{\text{avg}} \pm$ std	K^*_{avg}	K_{min}	$F_1^{\text{avg}} \pm$ std	K^*_{avg}	K_{min}	$F_1^{\text{avg}} \pm$ std	K^*_{avg}	K_{min}	$\overline{F_1}$
COLVEC	avg.	0.79 ± 0.18	9475	316	0.84 ± 0.11	10691	1778	0.80 ± 0.17	10815	1778	0.81
	last	0.73 ± 0.19	8143	178	0.69 ± 0.22	13125	1778	0.58 ± 0.25	15473	5623	0.67
FULLNEU	avg.	0.53 ± 0.27	13346	178	0.62 ± 0.12	16486	10000	0.55 ± 0.19	14481	562	0.57
	last	0.48 ± 0.19	15679	316	0.59 ± 0.17	16770	5623	0.43 ± 0.24	17783	17783	0.50
CONTRIB	last	0.85 ± 0.05	9552	3162	0.84 ± 0.08	8627	1778	0.70 ± 0.16	14824	5623	0.80
COMBINED	—	$\mathbf{0.90 \pm 0.05}$	7842	1000	$\mathbf{0.88 \pm 0.06}$	8436	1000	$\mathbf{0.84 \pm 0.11}$	13774	1778	**0.87**

Table 1. Comparison of neuron feature signals. K^*_{avg} and K_{min} denote the mean and minimum optimal expert count across relations. The average and standard deviation of F_1 scores are computed across relations.

4.1. Signal Quality for Expertise Estimation

Combined Ranking Is Best. Table 1 shows that COLVEC (avg.) and CONTRIB achieve strong ablation performance, with $\overline{F_1}$ of 0.81 and 0.80 across all models, compared to 0.57 and 0.50 for FULLNEU variants. The combined ranking, summing $\hat{\text{AP}}_i^{r,\text{col}}$ and $\text{AP}_i^{r,\text{con}}$, achieves the highest overall performance at 0.87, outperforming either signal in isolation. Notably, CONTRIB and COMBINED show low variance across relations. This suggests the combined ranking is stronger and more consistent. Figure 2 confirms that combined expert sets produce strong diagonal suppression of the target relation with limited cross-relation interference. This finding is reminiscent of that of Mondorf et al. (2025), who found that ensembling across circuit detection methods improves performance. This suggests a general pattern that model interpretations are more robust when there are several methods to support them.

Different Methods Find Different Neurons. Despite complementary strengths, expert sets $\mathcal{E}_r^{K^*}$ show $< 10\%$ overlap on all models between COLVEC and CONTRIB (Table 3). These two signals therefore constitute methodologically distinct yet equally valid perspectives on relation representation: column vectors capture activation patterns upstream of the nonlinearity, while contributions reflect the functional impact on downstream computation. Their combination leverages both, selecting neurons that are strongly activated and contribute meaningfully to the residual stream.

Full Neuron Activations Are a Consistently Weak Expertise Signal. FULLNEU activations perform substantially worse than COLVEC and CONTRIB. Beyond lower F_1^{avg}, the high variance of FULLNEU (avg.) indicates that performance is unreliable across relations. Furthermore, the optimal expert counts K^*_{avg} for FULLNEU are the largest of all variants, suggesting that grid search selects larger, less focused expert sets to achieve even modest recall. The poor quality of this signal directly follows the ambiguity in $h_i(x)$ discussed in Section 3.4: the sign and magnitude of full neuron activation is jointly determined by W_u and W_g, and thus possibly negative; this causes relation-specific neurons to receive low expertise scores under the standard AP ranking.

Last Token Position Is Consistently Inferior to Mean Token Aggregation Across both COLVEC and FULLNEU, last-token activations underperform mean-token aggregation in Table 1. This is consistent with the expectation that the last token position captures a single, potentially noisy activation snapshot, while averaging over all positions produces

a more stable estimate of the neuron's typical response to the relation.

4.2. Relation-Specific Representation Size

The optimal expert count K_r^* varies substantially across relations but proves relatively stable across different sets of randomly selected training prompts (see Appendix B). This supports the interpretation that it reflects a genuine property of how each relation is represented rather than an artifact of the specific training partition. Knowledge of relations such as `star_constellation` in `gemma-7b` can be fully ablated (Recall of 1.0) with just 316 neurons (see Figure 9l), while similar results can be observed with 1778 neurons in `person_sport_position` (see Figure 9j) or the upper end of the tested range with 17,783 neurons for `person_plays_instrument` (see Figure 9h), suggesting a more distributed or compositionally complex internal representation. The appropriate representation size of a relation also varies across models. Evaluating all relations under a single fixed count systematically over-localizes dense representations and under-localizes sparse ones (see Appendix A.1 for per-relation details).

5. Analysis of Internal Relation Representation Structure

Even under the best-performing combined ranking, ablation results remain poor for some relations, and the cross-relation interference patterns visible in Figure 2 reveal structural asymmetries. Certain relations are largely unaffected by ablation of their own expert set $\mathcal{E}_r^{K^*}$, while others are consistently disrupted by the ablation of experts identified for unrelated relations. In this section we describe three independent follow-up experiments aiming to address potential methodological limitations.

5.1. Beyond Strictly Higher: Bidirectional Expertise Ranking

As discussed in Section 3.4, FULLNEU activations are particularly susceptible to limitations of standard AP_i^r. To account for informative negative activations, we define a *bidirectional ranking* $\widetilde{\mathrm{AP}}_i^r = \max(\mathrm{AP}_i^r,\ \mathrm{AP}_i^{\neg r})$ where $\mathrm{AP}_i^{\neg r}$ is the AP computed with inverted class labels. This score rewards both higher and lower separation, selecting the stronger discriminative direction.

Table 2 reports ablation results for `gemma-7b` using the bidirectional score $\widetilde{\mathrm{AP}}_i^r$ for COLVEC and FULLNEU. The extension yields degrading performance for COLVEC and slight improvement over standard AP for FULLNEU on `gemma-7b`, though gains vary across relations and do not close the gap to COLVEC or CONTRIB. Figure 3, shows a neuron selected by FULLNEU using $\widetilde{\mathrm{AP}}_i^r$ with positive-relation samples below -1.0 and negative samples mostly above. This shows that relational specificity can manifest as strongly negative FULLNEU values, which standard AP_i^r fails to reward.

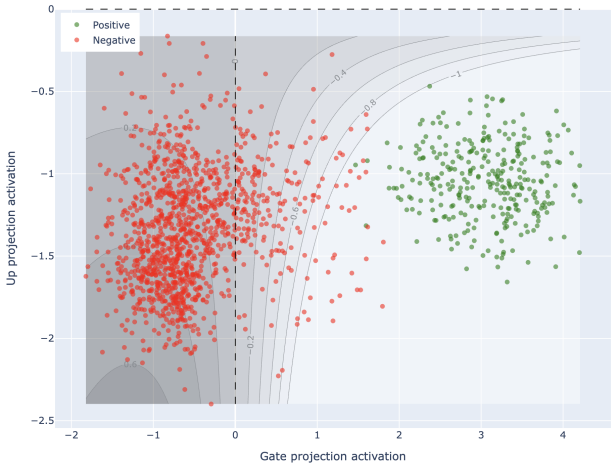


Figure 3. Examples of column vector and full neuron (gradient) activations of a neuron selected as a top expert for the relation `person_plays_instrument` based on the bidirectional AP ranking using full neuron activations.

However, the bidirectional ranking exacerbates a more fundamental problem. Figure 12 in the Appendix shows another top neuron selected by $\widetilde{\mathrm{AP}}_i^r$ with positive-relation samples concentrated near zero, indicating this neuron barely fires for the target relation. Such neurons satisfy the strictly-lower criterion not because they carry an informative signal, but because activations on both target and counter-relation samples are uniformly near zero. Since inactive neurons contribute nothing to downstream computation, assigning them high expertise is methodologically unsound.

This ambiguity is not unique to the bidirectional variant: standard AP_i^r fails symmetrically when counter-relation activations are negative, allowing near-zero target activations to score highly by exceeding negative counter-relation values, despite non-activation. In both cases, the core issue is the lack of conditioning on activation magnitude, making both susceptible to high expertise scores for neurons that do not meaningfully participate in the computation.

This failure is also reflected in the increased K^*_{avg} values and the AP value distributions in Figure 13: Under standard ranking (Figure 13a), expertise scores span 0.2 to 1.0 with a long tail toward perfect separation, providing a meaningful gradient between informative from uninformative neurons. Under the bidirectional ranking (Figure 13b), scores compress into a narrow band above 0.6, collapsing the variation that makes the ranking useful.The corrupted expertise distribution causes the grid search to select larger expert sets

Signal	Position	F_1^{avg}	F_1^{avg} Diff.	K^*_{avg}	K^*_{avg} Diff.
ColVec	average	0.65 ± 0.22	-0.14	15019	+5544
	last	0.58 ± 0.23	-0.16	14175	+6032
FullNeu	average	0.67 ± 0.17	+0.14	15303	+1957
	last	0.59 ± 0.19	+0.11	16486	+807

Table 2. Bidirectional AP ranking results for `gemma-7b` showing the average F_1 scores, along with the difference to the original average F_1 scores in 1 and the difference in the average number of neurons ablated. **The bidirectional $\widetilde{\text{AP}}$ ranking improves scores slightly for FullNeu, decreases scores for ColVec, and consistently causes higher K^* on average.**

$\mathcal{E}_r^{K^*}$ — most notably $+5{,}544$ more neurons on average for ColVec (avg.) in Table 2 — since the ranking no longer reliably reflects relational specificity.

The bidirectional AP variant thus trades marginal gains in isolated cases for a systematic distortion of the expertise distribution.

5.2. Reliance on Subject Overlap

Cross-relation interference observed in ablation experiments has been interpreted as evidence of representational entanglement (Liu et al., 2025), but a confounding explanation exists: relations such as `person_mother` and `person_father` share many subjects, meaning these patterns might not occur from relation entanglement, but from neurons responding to shared subject-level features. To distinguish these explanations, we construct reversed versions of three reversible relations, replacing `company_ceo`, `person_mother` and `person_father` with `ceo_company`, `mother_person` and `father_person` with new templates (see Appendix D). This eliminates subject overlap while preserving the relational mapping objects.

Figure 4 shows cross-relation ablation patterns for the original and reversed relations. Two observations hold across both conditions. First, ablating one relation's expert set frequently affects other relations regardless of their semantic similarity, whether or not the relations share subjects. Second, only 13–17% of neurons from the original relations are reselected for the reversed relations in this Contrib experiment. Compared to 88–95% overlap across different training datasets of the same relation (see Appendix B), this indicates that the reversed relations are represented mostly separately from the original ones. However, the overall cross-affection pattern remains strikingly similar to the original experiment. Together, these results rule out subject overlap as the primary driver of cross-relation interference and support the interpretation that entanglement between relational representations is a structural property of how concepts coexist within the model's MLP layers.

5.3. Category Mismatch or Localization Failure

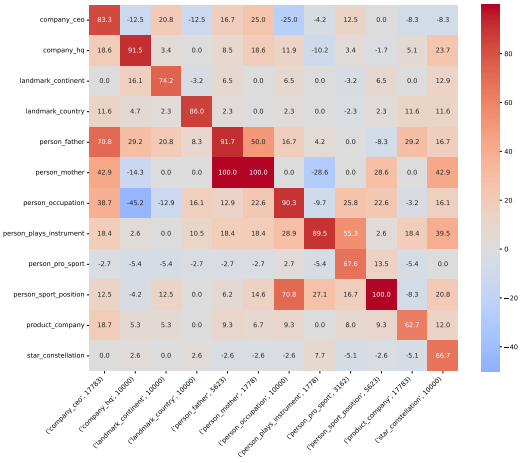
Across signals and model sizes, some relations resists ablation: no $K \in \mathcal{K}$ yields a satisfactory F_1 score. This may reflect category mismatch rather than failed localization.

Figure 5 displays a t-SNE projection of Contrib values in `llama-2-13b`, a consistently difficult relation. Rather than forming a single cluster, the activations separate into manually inspected semantic subcategories: including `car_company`, `video_game_company`, and `jet_company`. The benchmark category `product_company` thus aggregates multiple internally distinct representations, forcing the expertise ranking to treat a heterogeneous set as a single class. We replace `product_company` with the subrelation `car_company` using the same prompt templates, which retains enough data samples for reliable evaluation. This substantially improves ablation success (Figure 6), reaching training $F_1 > 0.8$ with only 3162 experts. Failure to ablate a relation is not unambiguous evidence that the relevant components cannot be found, but may reflect a mismatch between benchmark relation granularity and the model's internal organization of knowledge. Neural footprints are therefore conditional on conceptual resolution.

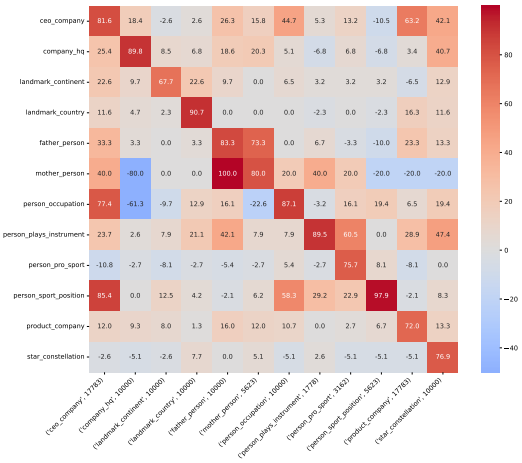
6. Conclusion

This paper examines methodological assumptions underlying relation-specific neuron identification, demonstrating that conclusions about internal representations are jointly shaped by the model's computation and the analytical choices brought to the investigation. On one hand, the neuron count representing a relation varies strongly across relations, reflecting differences in relational complexity, training frequency, and subject-object space cardinality, and further shows that fixed-count evaluations misrepresent at least part of this spectrum. The ranking used to determine these neurons on the other hand can be refined and compressed by adding information based on contribution scores. This new Combined ranking reveals the strongest and most consistent ablation performance across models.

Additional preliminary experiments emphasize the chal-



(a) CONTRIB



(b) CONTRIB with reversed relations

Figure 4. Accuracy drop results (analogous to Figure 2) with relation reversal on Model `llama-7b` using contribution-based expert selection.

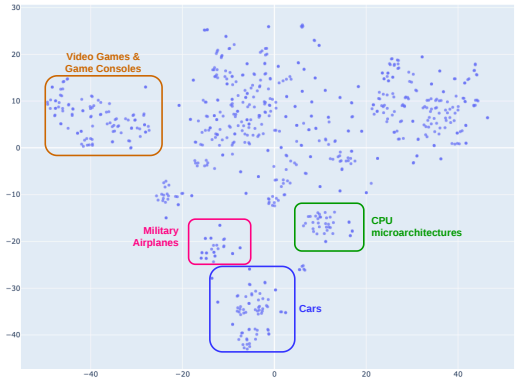


Figure 5. t-SNE of the contibution values for all prompts of `product_company` in `llama-2-13b-model` with selected annotated clusters.

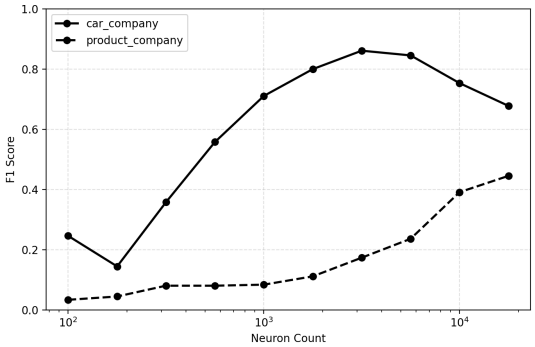


Figure 6. Ablation F_1 scores for `llama-2-13b` across all neuron counts for `product_company` vs. `car_company` on the training data using CONTRIB. For `product_company` the scores remain comparatively low across all variants (see Figure 11k), indicating the absence of a coherent expert set. Because of time constraints we only tried the *contribution* method in this experiment.

lenge of adjusting this method to effectively process information from FULLNEU activations, sharpen our understanding of how relational knowledge is organized by demonstrating that cross-relation entanglement is not an artifact of subject overlap and that poor ablation performance might indicate conceptual mismatch with the predefined categories.

Together, these findings suggest that accurately characterizing relational knowledge in large language models requires not only careful probing but equally careful scrutiny of the analysis itself as these choices determine what can and cannot be concluded about the model's internal structure.

Limitations

All experiments use 12 relations from the LRE dataset (Hernandez et al., 2024) with fixed templates, which constrains both relation diversity and linguistic naturalness in which knowledge is probed. Whether these patterns generalize to more naturalistic formulations remains open. Additionally, the t-SNE analysis is manual and limited to `product_company` using CONTRIB, so further category-granularity mismatches remain undetected.

The contrastive expertise ranking by design favors neurons that activate *selectively* for the target relation over counter-relations. Polysemantic neurons, active across multiple relations, receive low scores and may be excluded from $\mathcal{E}_r^{K^*}$. If relational knowledge is partially carried by such shared, high-utilization neurons, expert sets may represent only exclusive portion of the underlying representation, affecting both the true size and the functional importance of the full neural substrate.

References

Cuadros, X. S., Zappella, L., and Apostoloff, N. Self-conditioning pre-trained language models. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4455–4473. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/cuadros22a.html.

Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022.

Ferrando, J. and Voita, E. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17432–17445, 2024.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., and Bertsimas, D. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Kargaran, A. H., Liu, Y., Yvon, F., and Schütze, H. How programming concepts and neurons are shared in code language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26905–26917, 2025.

Liu, Y., Chen, R., Hirlimann, L., Hakimi, A. D., Wang, M., Kargaran, A. H., Rothe, S., Yvon, F., and Schütze, H. On relation-specific neurons in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 992–1022, 2025.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Mondorf, P., Wang, M., Gerstner, S., Hakimi, A. D., Liu, Y., Veloso, L., Zhou, S., Schutze, H., and Plank, B. Blackboxnlp-2025 mib shared task: Exploring ensemble strategies for circuit localization methods. *ArXiv*, abs/2510.06811, 2025. URL https://api.semanticscholar.org/CorpusID:281891963.

Shazeer, N. GLU variants improve transformer. ArXiv, 2020. URL https://arxiv.org/pdf/2002.05202.

Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, X., Wei, F., and Wen, J.-R. Language-specific neurons: The key to multilingual capabilities in large language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wang, W., Haddow, B., Peng, W., and Birch, A. Sharing matters: Analysing neurons across languages and tasks in LLMs. *ArXiv*, abs/2406.09265, 2024. URL https://api.semanticscholar.org/CorpusID:270440181.

A. Additional Results

We provide additional analyses supporting the main findings. Table 3 reports the overlap between COLVEC and CONTRIB (avg.) expert sets, showing consistently low agreement across relations and models. Figures 9, 10 and 11 present per-relation training F_1 scores of each signal for `gemma-7b`, `llama-2-7b` and `llama-2-13b-hf` respectively, highlighting substantial variation in ablation performance across relations and methods. Finally, Figures 7, 8a and 8b show the selected optimal expert counts K^* per relation and setting for each model, illustrating the strong dependence of representation size on both relation and signal choice.

Model	**Overlap (%)**
`gemma-7b`	7.3 ± 4.2
`llama-2-7b`	7.8 ± 3.1
`llama-2-13b`	9.8 ± 3.5

Table 3. Average percentage of neurons selected by COLVEC (avg.) that are also selected by CONTRIB, reported as mean $\pm$ standard deviation across relations. **The consistently low overlap confirms that the two signals identify largely disjoint expert sets despite comparable ablation performance.**

Method	**Pos**	company_ceo	company_hq	landmark_continent	landmark_country	person_father	person_mother	person_occupation	person_plays_instrument	person_pro_sport	person_sport_position	product_company	star_constellation
COLVEC	average	0.80	0.81	0.78	0.36	0.87	**0.94**	0.82	**0.97**	0.57	0.94	0.58	**0.99**
	last	0.72	.60	0.82	0.84	0.69	0.93	0.80	0.90	0.34	0.74	0.41	0.99
FULLNEU	average	0.66	0.17	0.30	0.26	0.79	0.83	0.47	0.58	0.22	0.82	0.27	0.98
	last	0.59	0.47	0.16	0.44	0.68	0.81	0.58	0.54	0.18	0.65	0.25	0.42
CONTRIB	last	0.87	0.86	0.83	0.82	0.80	0.90	0.81	0.87	0.88	0.91	0.73	0.90
COMBINED	—	0.88	**0.90**	**0.83**	**0.87**	0.79	0.94	**0.92**	0.90	**0.91**	**0.98**	**0.84**	0.94

Table 4. F1-Score for ablation experiments on the test data across relations using `gemma-7b`.

A.1. Per-Relation Optimal Neuron Counts

The differences in K_r^* across relations are not minor variations around a shared mean but span almost the full range of $\mathcal{K}$, from a lower bound of 178 to an upper bound of 17,783. This variation also increases with model size: under COLVEC (avg.), mean optimal counts increase from 9,475 on `gemma-7b` to 10,691 on `llama-2-7b` and 10,815 on `llama-2-13b` while K_{min} rises from 316 to 1,778 to 5,623, suggesting that representations become more distributed as model capacity grows. The combined ranking partially counteracts this trend, maintaining lower K^*_{avg} values of 7,842 and 8,436 for `gemma-7b` and `llama-2-7b`, consistent with its ability to identify more focused expert sets through the joint activation-contribution criterion.

A.2. Notes on Answer Validation

There are multiple perspectives towards the design of the definition of correct matches in the validation and evaluation. That the gold answer "appears within" the generated output may not be more restrictive than alternative formulations, but rather a different matching criterion. An alternative proposal would be to demand the correct answer to be at the start of the generated tokens.

In practice, the adopted evaluation allows flexibility in surface realization, e.g., permitting responses that embed the correct relation within longer sentences ("his father is ...") while still requiring the full answer content (e.g., complete entity names

Method	Position	company_ceo	company_hq	landmark_continent	landmark_country	person_father	person_mother	person_occupation	person_plays_instrument	person_pro_sport	person_sport_position	product_company	star_constellation
COLVEC	average	**0.94**	0.82	0.79	0.82	0.81	0.92	**0.93**	0.88	0.70	**0.97**	0.57	**0.90**
	last	0.75	0.81	0.52	0.78	0.69	0.84	0.71	0.90	0.10	0.92	0.53	0.78
FULLNEU	average	0.73	0.63	0.31	0.62	0.67	0.74	0.58	0.64	0.72	0.72	0.54	0.50
	last	0.72	0.51	0.35	0.56	0.75	0.74	0.71	0.70	0.53	0.85	0.30	0.40
CONTRIB	last	0.80	0.91	0.81	**0.87**	**0.85**	0.91	0.88	**0.93**	0.76	0.93	0.72	0.71
COMBINED	—	0.92	**0.91**	**0.81**	0.86	0.84	**0.94**	0.89	0.91	0.88	0.96	**0.73**	0.88

Table 5. F1-Score for ablation experiments on the test data across relations using `llama-2-7b`.

Method	Position	company_ceo	company_hq	landmark_continent	landmark_country	person_father	person_mother	person_occupation	person_plays_instrument	person_pro_sport	person_sport_position	product_company	star_constellation
COLVEC	average	**0.96**	0.57	0.84	**0.72**	**0.89**	**0.96**	0.86	0.90	0.56	**0.97**	0.46	0.87
	last	0.48	0.46	0.77	0.41	0.44	0.92	0.41	0.86	0.10	0.94	0.36	0.77
FULLNEU	average	0.67	0.57	0.66	0.20	0.58	0.78	0.80	0.55	0.51	0.67	0.14	0.53
	last	0.55	0.17	0.29	0.08	0.70	0.82	0.61	0.32	0.62	0.63	0.16	0.18
CONTRIB	last	0.72	0.54	0.82	0.53	0.76	0.90	0.74	0.89	0.45	0.91	0.45	0.69
COMBINED	—	0.92	**0.61**	**0.88**	0.68	0.84	0.92	**0.90**	**0.93**	**0.85**	0.96	**0.66**	**0.88**

Table 6. F1-Score for ablation experiments on the test data across relations using `llama-2-13b`.

rather than partial forms). This is intended to avoid over-penalizing variations in phrasing while still enforcing relational correctness and ensure that the observed accuracy drops during ablation are not a mere artifact of changes in word order.

A potential limitation is that such flexibility may occasionally leave room for multiple possible answers in the generated tokens, though this was not observed to be an issue in the evaluated outputs. Overall the selected perspective reflects a trade-off between strict token-level matching and robustness to natural language variation in model generations.

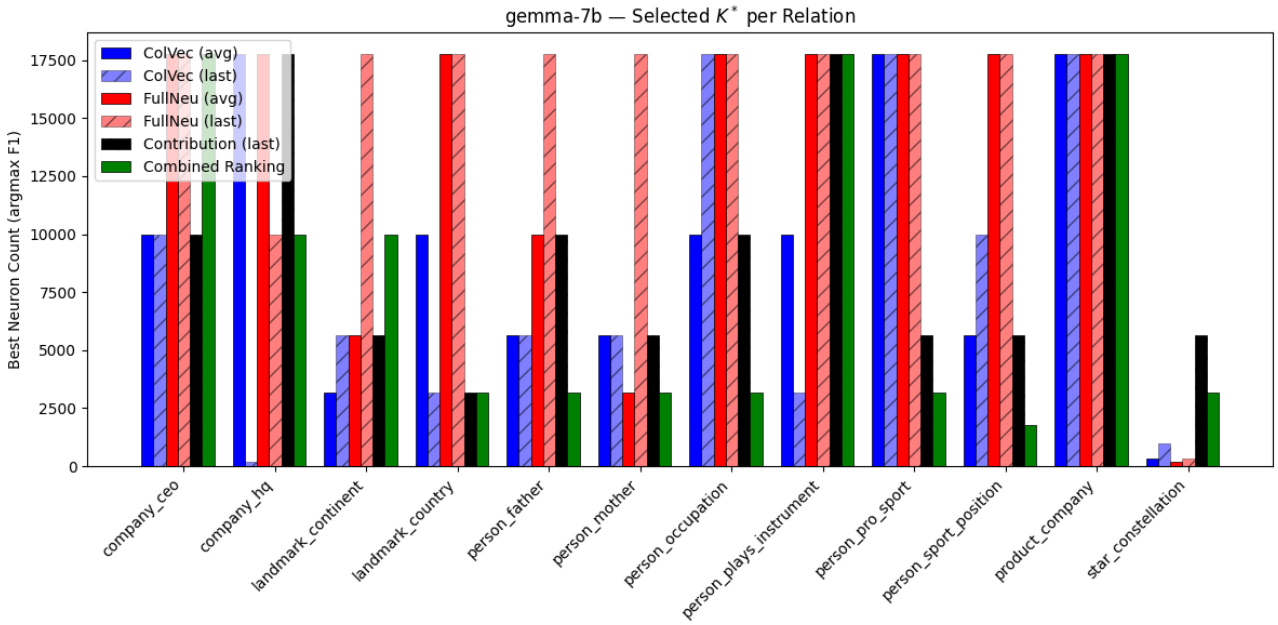


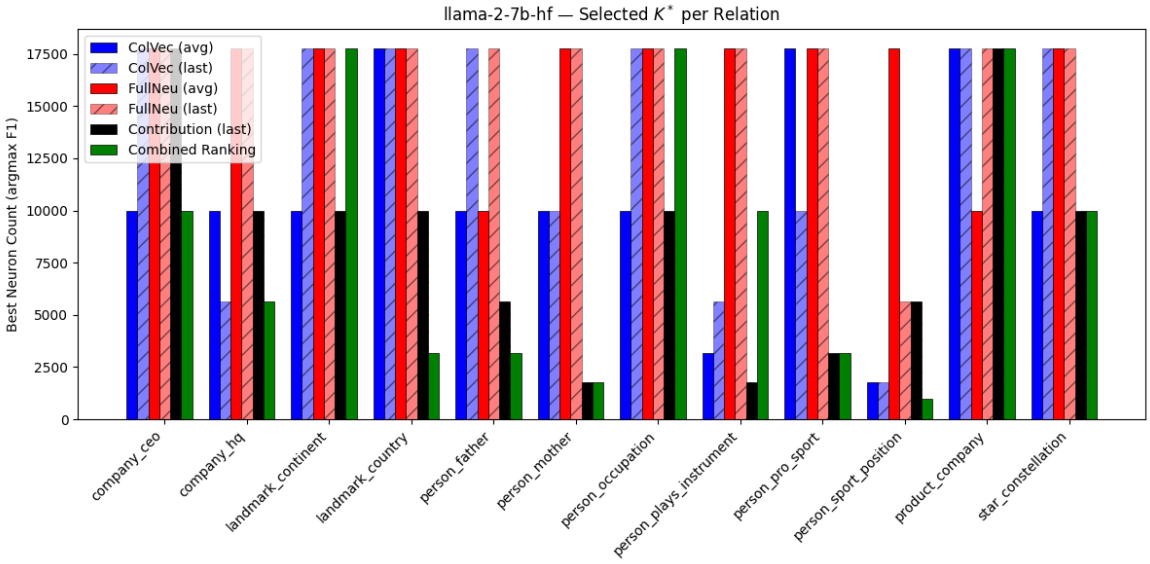
Figure 7. Selected K_r^* for all relations and method variants in `gemma-7b`.

B. Results using a different random seed

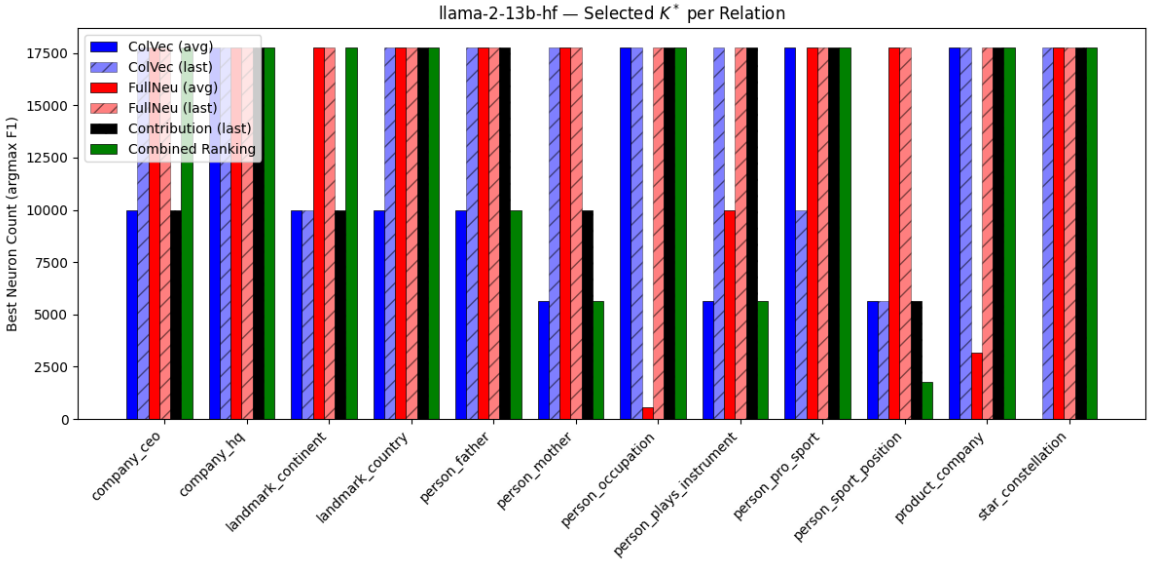
To assess whether the identified expert sets reflect stable properties of the model rather than artifacts of a particular data sample, we repeat a subset of experiments using a different random seed for the initial relation data sampling, producing a distinct contrastive training set $\mathcal{T}_r$. We evaluate whether the F_1 score results remain consistent across seeds (see table 7) and measure the overlap between the resulting expert sets $\mathcal{E}_r^{K^*}$ (table 8). Results confirm that both the selected neuron counts and the composition of the expert sets are largely stable across seeds, with high overlap between independently sampled runs, supporting the interpretation that the identified neurons reflect genuine model-internal structure rather than properties of the specific training partition.

		gemma-7b			**llama-2-7b**			**llama-2-13b**		
Signal	**Position**	$F_1^{\text{avg}} \pm \text{std}$	K^*_{avg}	K_{min}	$F_1^{\text{avg}} \pm \text{std}$	K^*_{avg}	K_{min}	$F_1^{\text{avg}} \pm \text{std}$	K^*_{avg}	K_{min}
ColVec	average	0.80 ± 0.15	10714	562	0.84 ± 0.10	11660	3162	0.79 ± 0.17	12148	5623
Contrib	last	0.84 ± 0.06	9063	3162	0.83 ± 0.08	7434	1000	0.70 ± 0.17	15188	10000
Combined	—	$\mathbf{0.90 \pm 0.05}$	7167	1778	$\mathbf{0.88 \pm 0.05}$	8720	562	$\mathbf{0.83 \pm 0.13}$	13490	1778

Table 7. Comparison of neuron feature signals with different random seed. ColVec: column vector activations; FullNeu: full neuron activations; Contrib: residual stream contribution scores; Combined: sum of avg. ColVec and Contrib AP scores. K^*_{avg} and K_{min} denote the mean and minimum optimal expert count across relations.

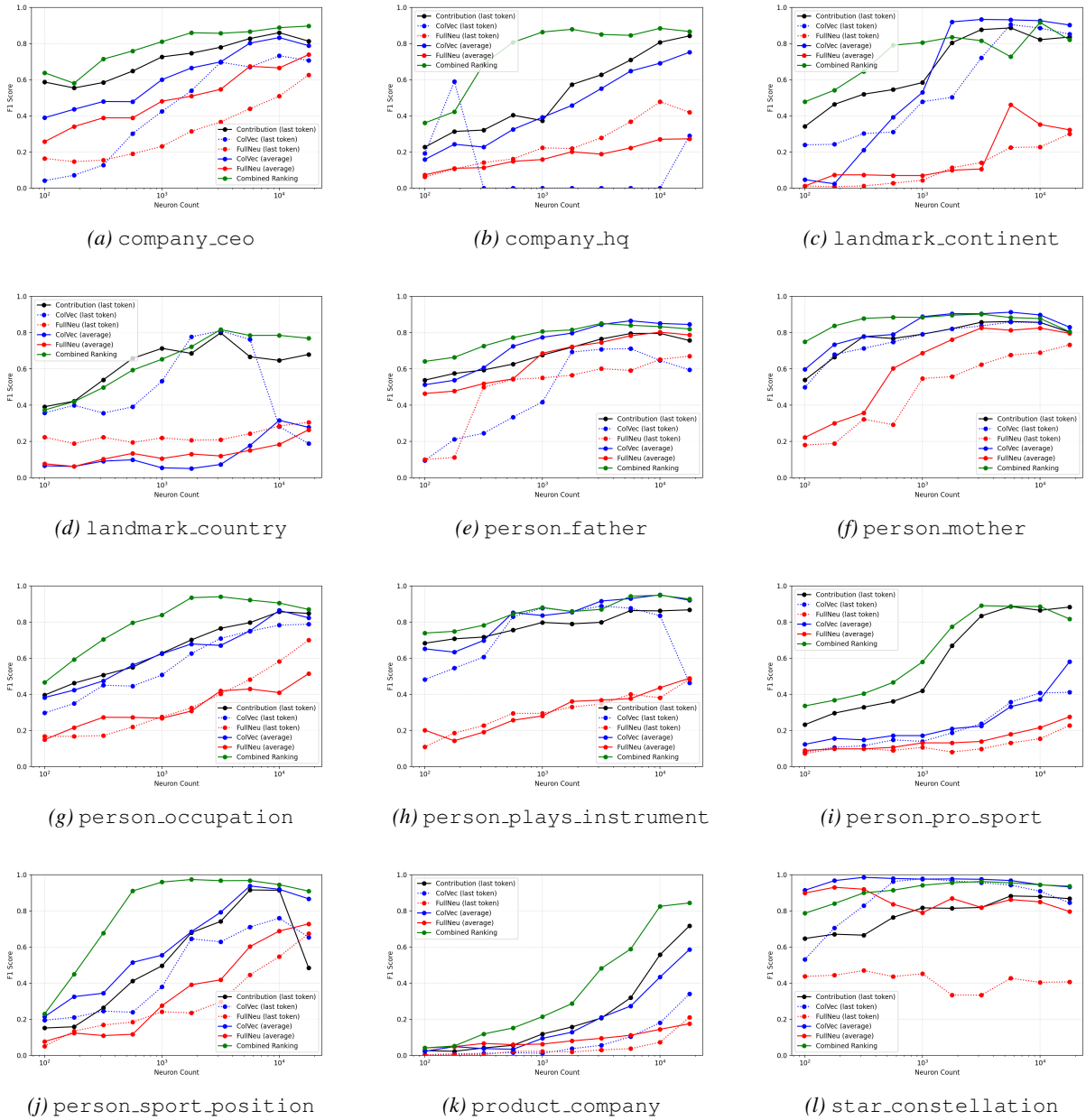


(a)



(b)

Figure 8. Selected K^* across all relations and method variants for `llama-2-7b-hf` in 8a and `llama-2-7b-hf` in 8b.



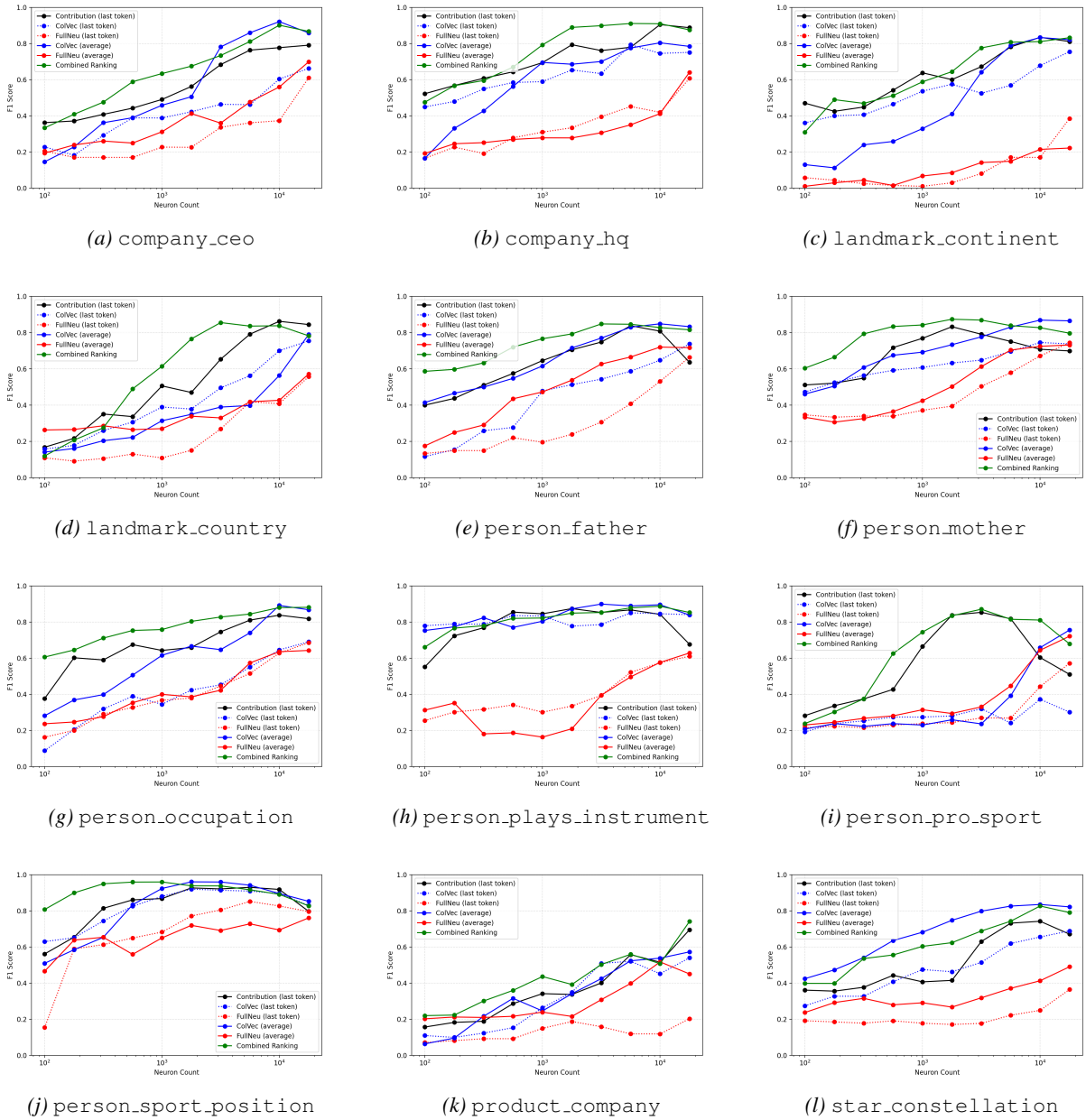
(a) `company_ceo` *(b)* `company_hq` *(c)* `landmark_continent`

(d) `landmark_country` *(e)* `person_father` *(f)* `person_mother`

(g) `person_occupation` *(h)* `person_plays_instrument` *(i)* `person_pro_sport`

(j) `person_sport_position` *(k)* `product_company` *(l)* `star_constellation`

Figure 9. Evaluation results of all relations and method variants across different expert set sizes for `gemma-7b`.



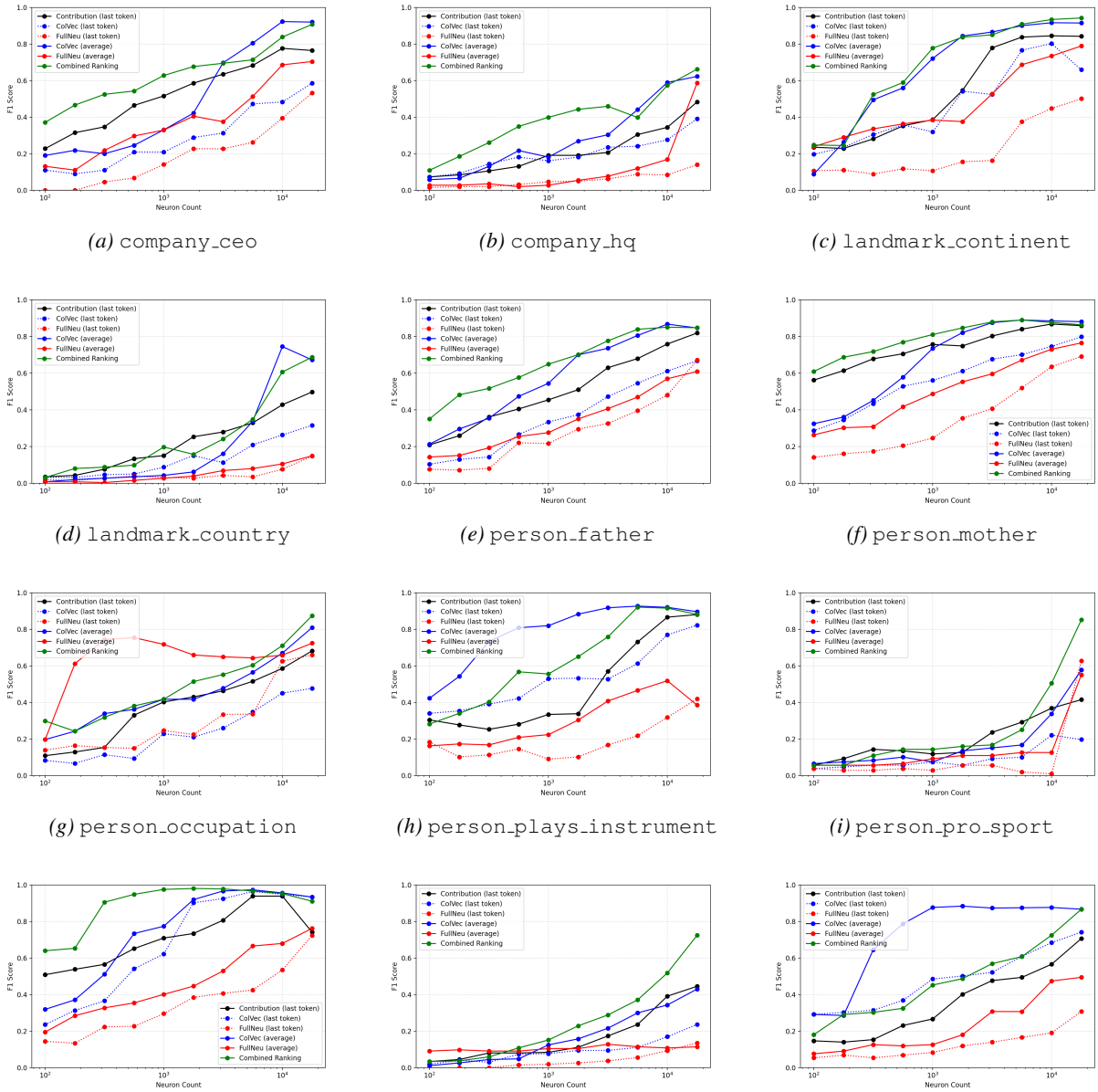
(a) `company_ceo` *(b)* `company_hq` *(c)* `landmark_continent`

(d) `landmark_country` *(e)* `person_father` *(f)* `person_mother`

(g) `person_occupation` *(h)* `person_plays_instrument` *(i)* `person_pro_sport`

(j) `person_sport_position` *(k)* `product_company` *(l)* `star_constellation`

Figure 10. Evaluation results of all relations and method variants across different expert set sizes for `llama-2-7b-hf`.



(a) `company_ceo` *(b)* `company_hq` *(c)* `landmark_continent`

(d) `landmark_country` *(e)* `person_father` *(f)* `person_mother`

(g) `person_occupation` *(h)* `person_plays_instrument` *(i)* `person_pro_sport`

(j) `person_sport_position` *(k)* `product_company` *(l)* `star_constellation`

Figure 11. Evaluation results of all relations and method variants across different expert set sizes for `llama-2-13b-hf`.

Signal	Pos.	Gemma-7B	LLaMA-2-7B	LLaMA-2-13B
ColVec	avg.	0.88	0.91	0.90
	last	0.88	0.92	0.92
FullNeu	avg.	0.92	0.94	0.93
	last	0.92	**0.95**	**0.95**
Contrib	last	0.91	0.93	0.95
Combined	—	**0.92**	0.93	0.93

Table 8. Average overlap between expert sets $\mathcal{E}_r^{K^*}$ identified on two independently sampled contrastive training sets $\mathcal{T}_r$, across all relations and all signal variants. High overlap across all models and signals confirms that identified neurons reflect stable model-internal structure rather than properties of a specific data sample.

C. Additional Analysis on the Extended AP

Figure 13 compares the score distributions induced by standard and extended AP ranking. The extended variant substantially compresses the distribution, reducing the separation between high- and low-expertise neurons and thereby weakening the effective ranking signal.

Figure 12 provides an example of a top-ranked neuron under the extended $\widetilde{\text{AP}}$ criterion exhibiting near-zero full neuron activations in both target and counter-relation samples, which reflects the lack of activation-magnitude conditioning in both AP formulations.

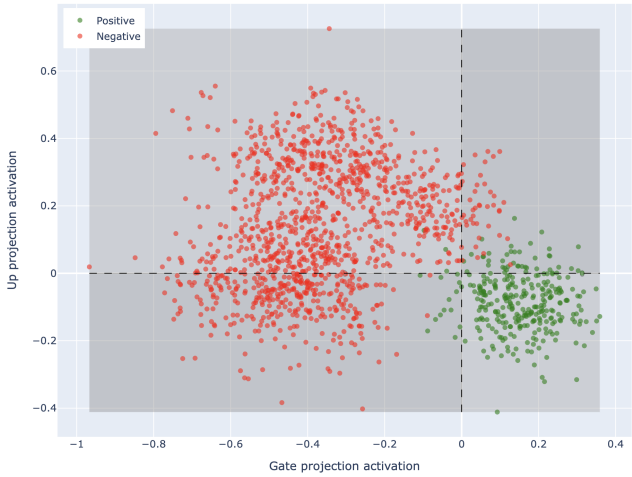
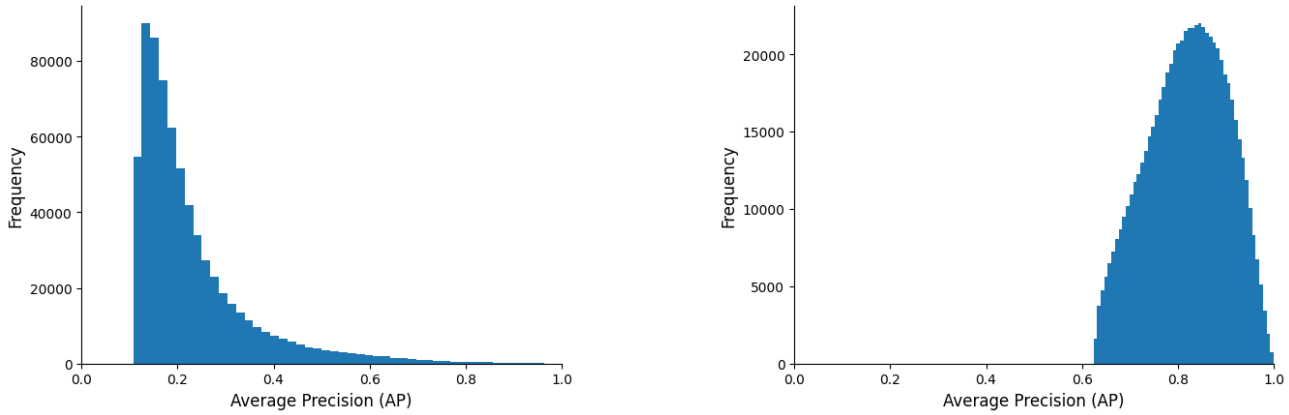


Figure 12. Top expert neuron selected based on FULLNEU using extended AP ranking for `gemma-7b`.



(a) *(b)*

Figure 13. Comparison of full neuron Average Precision (AP) score distributions of the relation person_sport_position for the original ranking in 13a and the Extended AP variant in 13b.

D. Reversed Relations

To create prompts for the newly introduced reversed relations `ceo_company`, `father_person` and `mother_person` the following templates were created:

- `ceo_company`
  "{} is the Ceo of? Answer:"
  What company is {} leading? Answer:"
- `father_person`
  "{} is the father of? Answer:"
  "What is the name of {}'s child? Answer:"
- `mother_person`
  "{} is the mother of? Answer:"
  "What is the name of {}'s child? Answer:"

These reversed prompts, mapping objects to subjects, replace the original subject–object pairs of the corresponding relations, while all remaining non-reversible relations retain their original data. The full pipeline from data validation onward, including contrastive training set construction, expertise ranking, and ablation evaluation, is applied identically to the reversed relations.