Grammatical "grandmother neurons" are rare in LLMs

Linyang He **Nima Mesgarani**
Zuckerman Mind Brain Behavior Institute, Columbia University
linyang.he@columbia.edu

Abstract

Understanding how Large Language Models (LLMs) encode linguistic structures remains a fundamental challenge in interpretability research. While diagnostic classifiers (or "probes") are widely used for this task, they face significant methodological criticism: training auxiliary classifiers introduces capacity confounds and calibration issues, often making it difficult to distinguish the model's intrinsic representations from the probe's ability to learn the task. To address these limitations, we introduce a probe-free framework for localizing linguistic selectivity at the individual neuron level. Leveraging the controlled contrasts of linguistic minimal pairs, we propose a Neuron Separability Index (NSI), a metric that directly quantifies how reliably single neurons differentiate grammatical from ungrammatical constructions without parameter updates. Applying NSI across 68 linguistic paradigms and seven checkpoints reveals three main patterns: 1) raw separability reaches near-peak levels earlier for morphological and syntactic distinctions than for syntax–semantics interface and conceptual distinctions. 2) after permutation normalization, single-unit selectivity is sparse, weak, and narrowly tuned: only a small fraction of units are sensitive to an average paradigm, and strongly selective "grandmother neurons" are rare. 3) whole-vector linear separability, single-neuron selectivity, and behavioral competence are largely dissociated, and targeted ablations further separate activation selectivity from causal reliance.

1 Introduction

Understanding how large language models (LLMs) encode linguistic structure remains a central challenge for interpretability. A recurring question concerns *granularity*: is a grammatical distinction carried by a small number of highly selective units, in the spirit of the long-debated "grandmother neuron" hypothesis in neuroscience (Kanwisher et al., 1997; Gross, 2002; Quiroga et al., 2005; Posani et al., 2025), or is it spread thinly across many units, each contributing only weakly? Work on *superposition* and *polysemanticity* suggests that neither extreme is guaranteed: single units can mix multiple features, so apparent selectivity may depend on which contrasts one happens to test (Elhage et al., 2022; Bills et al., 2023; Huang et al., 2024). Sparse feature discovery can partially disentangle such mixtures, but the recovered directions are not automatically complete, atomic, or tied to model behavior, which makes unit-level functional claims difficult to establish (Bricken et al., 2023; Cunningham et al., 2024; Makelov et al., 2025; Leask et al., 2025). Settling the granularity question therefore requires a measurement that is defined at the level of individual units, comparable across many linguistic phenomena, and free of auxiliary training.

Existing evidence comes from three lines of work, none of which supplies such a measurement. Targeted minimal-pair evaluation established that language models capture a wide range of grammatical dependencies (Linzen et al., 2016b; Marvin & Linzen, 2018; Warstadt et al., 2020; Misra et al., 2023; Jumelet et al., 2025), but it treats the model as a black box and says nothing about internal organization. Representation analyses open the box and map where linguistic information is decodable across depth (Tenney et al., 2019; Starace et al., 2023; He et al., 2024), yet a trained probe reads out a *whole activation vector* and

contributes its own capacity, so its success is silent about whether any individual unit carries the distinction (Hewitt & Liang, 2019; Pimentel et al., 2020; Belinkov, 2022). Mechanistic interpretability does reach individual units (Lakretz et al., 2019; Geva et al., 2021; Finlayson et al., 2021; Meng et al., 2022), but each study is built around one or two phenomena, leaving open whether strong single-unit selectivity is the normal case or a rare exception across the grammar as a whole. We discuss all three lines in detail in Section A.

The open question, then, is quantitative and comparative: *if a model behaves grammatically, how much of that competence is expressed in individual neurons, and is such selectivity broad (domain-general) or narrow (domain-specific)?* We address it with a *probe-free* neuron-level framework that exploits the controlled contrasts of minimal pairs. Combining BLiMP (Warstadt et al., 2020) and COMPS (Misra et al., 2023), we organize the suite into a three-level hierarchy of 4 domains, 13 phenomena, and 68 paradigms, so that selectivity can be compared on a common scale from a single agreement contrast up to entire linguistic domains. For each paradigm we feed matched grammatical and ungrammatical sentences into a language model, extract last-token activations, and, for each neuron independently, compute a correlation-derived *raw separability* score between its paired positive and negative activation vectors. Because minimal pairs differ in as little as one token, this contrast is tightly controlled, and because the score is read directly off the activations, it requires no parameter updates and no auxiliary classifier.

Raw separability can still be inflated by lexical overlap or sampling noise. We therefore normalize it against a null distribution obtained by randomly swapping the grammatical labels within pairs, which preserves lexical content while destroying the grammatical contrast. The resulting *Neuron Separability Index* (NSI) expresses each neuron's discrimination in units of its own null, making values comparable across neurons, layers, and paradigms: a high NSI indicates discrimination beyond what lexical content or noise explains, whereas a near-zero NSI indicates insensitivity to the targeted contrast.

Applying this framework reveals a consistent but more nuanced picture. Raw separability reaches near-peak levels earlier for morphological and syntactic distinctions than for syntax–semantics interface and conceptual ones. After permutation normalization, however, the same activations yield a far more restrictive picture of single-unit selectivity: grammaticality-sensitive neurons are sparse, their effects are generally weak, and units meeting a strong-selectivity criterion are rare. Selectivity is also narrow, with even poly-selective neurons retaining a dominant within-domain preference rather than acting as general grammar detectors. This pattern replicates across seven checkpoints from four model families and survives our threshold and pairing controls.

Neuron-level selectivity also proves distinct from both whole-vector decodability and model behavior: probe accuracy is essentially unrelated to minimal-pair behavioral accuracy across paradigms, NSI is likewise only weakly associated with behavior, and targeted ablations of the rare above-threshold units and their highest-scoring same-layer neighbors are no more damaging than matched random ablations. Information can therefore be linearly decodable from a representation without being strongly localized to individual units, and neither property alone establishes that the model behaviorally relies on it. Our contributions are correspondingly threefold: a probe-free, permutation-normalized measure of unit-level selectivity; a systematic map of that selectivity across a three-level linguistic hierarchy; and an empirical dissociation between representational decodability, single-unit localization, and behavioral reliance.

2 Minimal pair-based Neuron Separability

We use minimal pairs from BLiMP (Warstadt et al., 2020) and COMPS (Misra et al., 2023) as a controlled testbed for grammatical and conceptual contrasts; full dataset details are provided in Appendix B. Across the combined suite, the targeted contrasts are organized into a three-level hierarchy (Domain → Phenomenon → Paradigm), spanning four domains, 13 phenomena, and 68 paradigms in total. A domain is the broadest grouping, a phenomenon groups related minimal-pair paradigms, and a paradigm is one specific benchmark contrast.

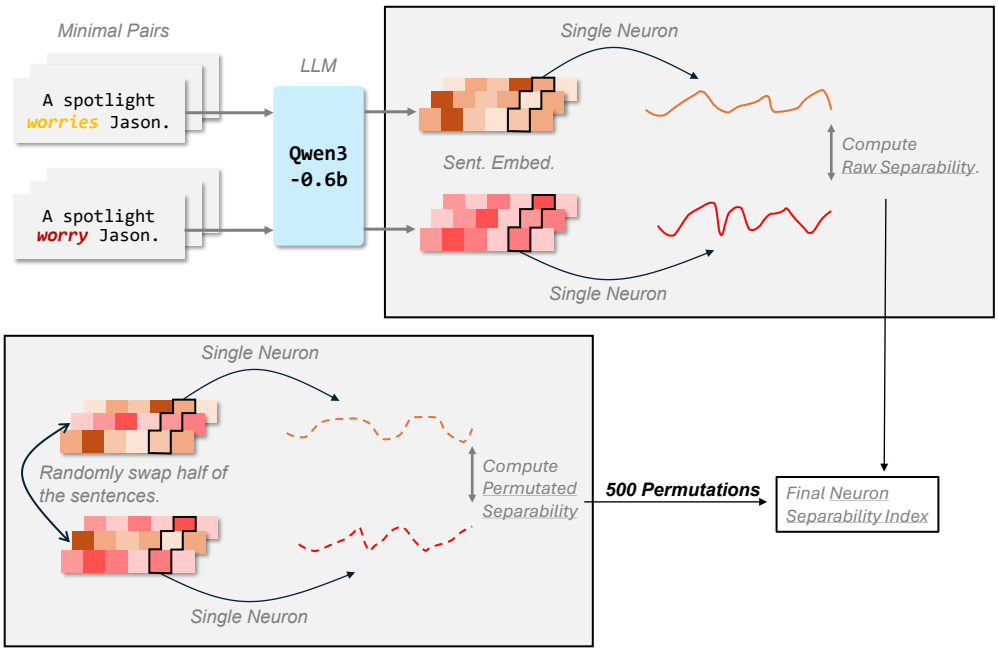

Figure 1: Minimal-pair neuron separability pipeline. Minimal pairs from BLiMP/COMPS are fed into a transformer LM (here, Qwen3-0.6B). For each minimal pair, we extract last-token activations, then focus on a single neuron and assemble paired activation vectors across items for the grammatical (positive) and ungrammatical (negative) sentences. We compute a raw separability score from the correlation between the two paired vectors. To control for lexical effects, we construct a null distribution by randomly swapping the grammatical/ungrammatical labels within half of the pairs and recomputing separability across many permutations; the observed raw score is then converted into a neuron separability index (NSI) value.

Let $\mathcal{X}_p = \{(x_i^+, x_i^-)\}_{i=1}^{M_p}$ denote the minimal pairs for a BLiMP/COMPS paradigm p (e.g., one specific subject–verb agreement contrast). For a transformer with L layers and hidden width d_ℓ at layer ℓ, let $h_\ell(x) \in \mathbb{R}^{d_\ell}$ be the last-token hidden state (for causal LMs). For neuron $j \in \{1, \ldots, d_\ell\}$ define scalar activation $a_{\ell j}(x) = h_\ell(x)_j$.

2.1 Raw separability score.

For neuron (ℓ, j) in paradigm p, form paired activation vectors

$$\mathbf{v}^+_{\ell j,p} = [a_{\ell j}(x_1^+), \ldots, a_{\ell j}(x_{M_p}^+)],$$
$$\mathbf{v}^-_{\ell j,p} = [a_{\ell j}(x_1^-), \ldots, a_{\ell j}(x_{M_p}^-)].$$

We first define the raw separability score $S^{\text{raw}}_{\ell j}(p)$ based on the correlation between the paired activation vectors:

$$S^{\text{raw}}_{\ell j}(p) = \frac{1 - \operatorname{corr}\left(\mathbf{v}^+_{\ell j,p}, \mathbf{v}^-_{\ell j,p}\right)}{2}. \tag{1}$$

Intuitively, $S^{\text{raw}}_{\ell j}(p)$ is an *effect-size* metric: it measures the degree of *within-item contrast consistency* between grammatical and ungrammatical sentences in paradigm p, independent of any explicit null model. Pearson correlation performs its own centering and scale normalization, so no separately normalized activation variable is required. If a neuron responds differently to grammatical versus ungrammatical members across items, the correlation is low, yielding $S^{\text{raw}}_{\ell j}(p)$ close to 1.

To interpret Eq. (1), note that $S^{\text{raw}}_{\ell j}(p) \approx 0$ corresponds to nearly identical activation patterns across the paired sentences (corr$\approx$ 1), $S^{\text{raw}}_{\ell j}(p) \approx 0.5$ corresponds to weak or inconsistent separation (corr$\approx$ 0), and $S^{\text{raw}}_{\ell j}(p) \to 1$ corresponds to a strong, consistent contrast where the two conditions vary in opposite directions (corr$\to -1$).

2.2 Permuted neuron separability index.

To control for lexical effects and background noise, we employ a permutation test. For each paradigm, we generate a null distribution by randomly swapping the grammatical/ungrammatical labels *within* exactly half of the minimal pairs ($M_p/2$), while keeping the paired sentences themselves intact. This procedure disrupts the systematic grammatical contrast while preserving the lexical content of each pair. We repeat this process $N_{\text{perm}} = 500$ times to obtain a distribution of permutation scores $\{S^{\text{perm}}_k\}_{k=1}^{N_{\text{perm}}}$.

In addition to the raw effect size $S^{\text{raw}}_{\ell j}(p)$, we define a neuron separability index metric that measures how strongly the observed separability exceeds the null distribution:

$$\text{NSI}_{\ell j}(p) = \max\left(0, \frac{S^{\text{raw}}_{\ell j}(p) - \mu_{\text{perm}}}{\sigma_{\text{perm}} + \epsilon}\right), \tag{2}$$

where μ_{perm} and σ_{perm} are the mean and standard deviation of the permutation scores, and ϵ is a small constant for numerical stability. We refer to $\text{NSI}_{\ell j}(p)$ as *Neuron Separability Index*. In contrast to the raw effect size, $\text{NSI}_{\ell j}(p)$ is a *noise-controlled* metric: it quantifies how strongly the observed separability exceeds the paradigm-matched null obtained by within-pair label swapping (i.e., beyond what can be explained by lexical content and sampling noise). In our later analyses, we retain *both* separability metrics: $S^{\text{raw}}_{\ell j}(p)$ captures the direct separability effect size, while $\text{NSI}_{\ell j}(p)$ provides a permutation-normalized, comparable measure of selectivity intensity across paradigms. Appendix D characterizes what this score measures in terms of a neuron's item-wise response profile across matched items.

Because $\text{NSI}_{\ell j}(p)$ is expressed in units of the permutation-null standard deviation, we use two operational thresholds in our analyses: neurons with $\text{NSI}_{\ell j}(p)>0$ are treated as *sensitive* (showing positive separability), while neurons with $\text{NSI}_{\ell j}(p)>2$ are treated as *strongly sensitive*. We use the latter as a descriptive strong-selectivity threshold rather than a Gaussian significance claim; Appendix G reports a threshold sweep and empirical-null diagnostics. This criterion operationalizes a **"grandmother neuron"**: a single neuron whose activity alone can reliably discriminate grammatical vs. ungrammatical minimal pairs for paradigm p.

3 Experimental Setup

Our primary analysis uses Qwen3-0.6B (Yang et al., 2025). To test whether the main pattern is specific to this checkpoint, we repeat the analysis across Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Pythia-410M, TinyLlama-1.1B, and Llama-3.1-8B (Dubey et al., 2024). All models are decoder-only causal LMs, and we read the hidden state at the final non-padding token so that every representation has access to the complete sentence. Please refer to Appendix F.1 for architectures and coverage statistics.

We additionally evaluate robustness to the NSI threshold and the construction of sentence pairs, including random-pairing controls and critical-token deletion; please refer to Appendices G and H for the full protocols and results.

Finally, we compare neuron-level NSI with both model behavior and whole-vector linear probing on the 67 BLiMP paradigms, treating probing as a representation-level control rather than evidence of behavioral reliance. We also conduct a targeted ablation sanity check on the three paradigms containing an NSI>2 unit: the candidates identified by the observational analysis are frozen, and at each candidate's own layer we zero the top

$k \in \{5, 10, 20\}$ residual-stream dimensions at every non-padding position, comparing the resulting change in the minimal-pair log-probability margin, $\log p(x^+) - \log p(x^-)$, with bottom signed-score sets and 100 same-layer random sets. Please refer to Appendices I and J for the complete setups, statistics, and results.

3.1 Quantification of Hierarchical Tuning Breadth

To characterize the functional specialization of neurons across network depths, we define "tuning breadth" (B) as the number of distinct categories to which a neuron responds at a specified level of the hierarchy. We analyze this metric at three levels of abstraction: fine-grained paradigms ($\mathcal{P}$), intermediate phenomena ($\mathcal{H}$), and broad domains ($\mathcal{D}$).

Let $z_{u,p}$ denote the selectivity score (NSI) of unit u for paradigm p. We define the binary responsiveness indicator $r_{u,p} = \mathbb{1}(z_{u,p} > 0)$, where $\mathbb{1}(\cdot)$ is the indicator function. The tuning breadth for unit u at each hierarchical level is calculated as follows:

Paradigm level. The raw count of responsive paradigms is defined as:

$$B^{(u)}_{\text{paradigm}} = \sum_{p \in \mathcal{P}} r_{u,p} \tag{3}$$

Phenomenon and domain levels. To account for the nested taxonomy, we apply a logical disjunction (Boolean OR) aggregation. A unit is considered responsive to a phenomenon h or domain d if it responds to *at least one* constituent paradigm:

$$B^{(u)}_{\text{phenomenon}} = \sum_{h \in \mathcal{H}} \left(\max_{p \in \mathcal{P}_h} r_{u,p} \right) \tag{4}$$

$$B^{(u)}_{\text{domain}} = \sum_{d \in \mathcal{D}} \left(\max_{p \in \mathcal{P}_d} r_{u,p} \right) \tag{5}$$

where $\mathcal{P}_h$ and $\mathcal{P}_d$ denote the paradigms belonging to phenomenon h and domain d, respectively.

Finally, the layer-wise expected tuning breadth $\bar{B}_k$ for layer k is computed by averaging over all N_k units in the layer:

$$\bar{B}_k = \frac{1}{N_k} \sum_{u \in \text{Layer}_k} B^{(u)} \tag{6}$$

This metric serves as a proxy for neuronal polysemanticity: higher values indicate broad generalization, while lower values indicate functional specialization.

3.2 Clustering neurons by linguistic phenomenon selectivity

To analyze the functional organization of neurons, we move beyond individual paradigms and aggregate selectivity at the phenomenon level. Let $\mathcal{P}$ denote the set of tested paradigms and $\mathcal{H}$ the set of linguistic phenomena. Each phenomenon $h \in \mathcal{H}$ is associated with a subset $\mathcal{P}_h \subset \mathcal{P}$; for example, the Subject–Verb Agreement phenomenon contains several specific BLiMP paradigms. For a fixed layer ℓ (or after flattening neurons across the network), let $z_j(p)$ be the selectivity score of neuron j for paradigm p. We define its phenomenon-level selectivity as the mean across constituent paradigms:

$$v_{h,j} = \frac{1}{|\mathcal{P}_h|} \sum_{p \in \mathcal{P}_h} z_j(p).$$

Neuron filtering via tuning breadth. Rather than selecting neurons based on peak magnitude alone, we retain neurons engaged across multiple phenomena. The phenomenon-level tuning breadth of neuron j is

$$B_j = \sum_{h \in \mathcal{H}} \mathbb{1}(v_{h,j} > 0).$$

We select $\mathcal{N}^* = \{j \mid B_j \geq 3\}$, focusing the clustering analysis on neurons with broader linguistic responsiveness rather than noise or paradigm-idiosyncratic tuning.

Clustering and matrix construction. We construct a phenomenon-by-neuron matrix $\mathbf{X} \in \mathbb{R}^{|\mathcal{H}| \times |\mathcal{N}^*|}$ from the aggregate scores of the selected neurons. Rows (phenomena) are grouped a priori by domain in the order *Concept*, *Syntax–Semantics Interface*, *Syntax*, and *Morphology*. We cluster the columns (neurons) to identify groups with similar tuning profiles. Specifically, we compute the pairwise correlation distance

$$D(j, j') = 1 - \mathrm{corr}(\mathbf{X}_{:,j}, \mathbf{X}_{:,j'}),$$

and apply agglomerative hierarchical clustering with average linkage (UPGMA) (Schlee, 1975). The resulting dendrogram determines the column order in the heatmap, revealing groups of neurons with preferences for particular domains or phenomena.

4 Results

4.1 Raw Separability Patterns

Figure 2 shows that Syntax and Morphology concentrate around layer 14, whereas Concept peaks around layer 19, suggesting a shift from local grammatical constraints toward higher-level distinctions with depth. A complementary 90% saturation analysis, computed separately for each of the 68 paradigms as the earliest layer reaching 90% of its own peak and then averaged by domain, yields the same late-to-early ordering: Concept (19.0), Syntax–Semantics Interface (10.3), Syntax (9.5), and Morphology (8.4; Appendix Figure 8). At the paradigm level, raw separability and whole-vector probing saturation layers show a positive correspondence (Spearman $\rho = 0.23$; Appendix Figure 9), providing qualitative support for a shared coarse progression while showing substantial measure-specific variation.

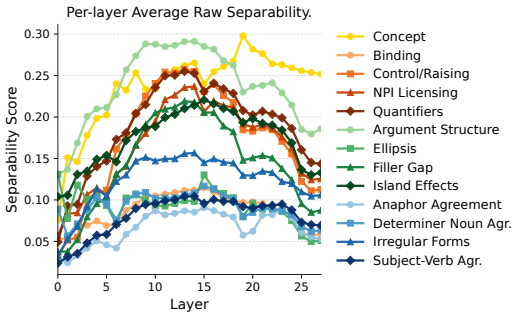

Figure 2: Colors indicate linguistic domains: *Concept*, *Syntax–Semantics Interface*, *Syntax*, and *Morphology*. Layer-wise average separability curves across linguistic phenomena. Each curve reports mean neuron separability for one phenomenon, averaged across its constituent paradigms.

4.2 Neuron Separability Index Patterns

Grammaticality sensitive neurons are sparse in LLMs. Figure 3-(a) quantifies how many neurons exhibit positive sensitivity to grammatical contrasts (positive NSI) across layers.

Overall, such neurons constitute only a small fraction of the model at any layer, indicating that grammaticality is not broadly encoded by most neurons. This sparsity is also domain-dependent: the mean proportion of sensitive neurons is approximately 8.27% for Concept, 3.31% for Syntax–Semantics Interface, 2.50% for Syntax, and 3.88% for Morphology. Thus, Concept recruits the largest pool, whereas the three grammatical domains remain much sparser. In addition, the layer-wise trends reveal a clear decay: the proportion of grammar-sensitive neurons is highest in early to middle layers and gradually diminishes toward deeper layers. This pattern suggests a functional redistribution across depth. As representations become more abstract, fewer neurons are dedicated to encoding local grammatical well-formedness, and more capacity is likely devoted to higher-level information such as semantic integration, reasoning, and discourse context.

"Grandmother neurons" are rare in LLMs. One natural question is whether a given linguistic phenomenon is localized to a small set of "grandmother" neurons that respond

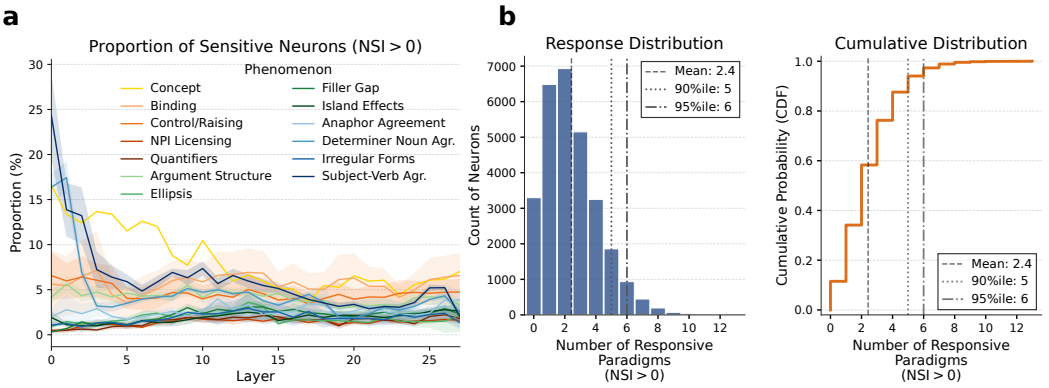

Figure 3: **(a)** Proportion of sensitive neurons (NSI>0) across layers, broken down by linguistic phenomenon. Lines show the mean proportion across paradigms within each phenomenon; shaded bands indicate variability across paradigms. Overall, neurons with positive sensitivity to the targeted contrasts are sparse throughout the network. Mean recruitment differs across the four domains: Concept ($\approx$ 8.27%), Syntax–Semantics Interface ($\approx$ 3.31%), Syntax ($\approx$ 2.50%), and Morphology ($\approx$ 3.88%). Sensitivity also tends to decrease in deeper layers, suggesting that later layers allocate relatively less capacity to local grammatical well-formedness and more to higher-level information (e.g., semantics, reasoning, and discourse). The right panel of Appendix Figure 11 provides the paradigm-level distributions. **(b)** Distribution of *neuron breadth* across paradigms. For each neuron (a unit at a specific layer), we count the paradigms for which it is responsive (NSI>0). **Left:** histogram of the number of responsive paradigms per neuron. **Right:** empirical CDF of the same quantity. Vertical lines mark the mean and the 90th/95th percentiles, showing that most responsive neurons participate in only a small number of paradigms and that broad, multi-paradigm responsiveness is rare. Corresponding response distributions at the domain and phenomenon levels appear in Appendix Figures 13 and 14.

strongly and selectively to that phenomenon. Using NSI, we find little evidence for such extreme localization. Figure 4a shows that even after restricting attention to sensitive neurons (NSI>0), their average separability remains low (typically around $\sim$ 0.5). Consistently, for 65 of 68 paradigms, even the most selective neuron fails to reach NSI>2, our operational threshold for strong selectivity. For each of the remaining three paradigms, Appendix Figure 12 shows that the threshold is exceeded by only one neuron out of the full network (28 layers $\times$ 1024 neurons).

Together, these results suggest that grammatical contrasts are typically encoded in a distributed manner: instead of a single neuron reliably separating grammatical from ungrammatical minimal pairs, separability is spread across neurons whose individual effects are modest.

Most neurons are domain-specific and narrowly tuned. While Figure 3a shows that only a small fraction of neurons are sensitive to any given paradigm, a complementary question is how many different paradigms a single neuron responds to. We define each neuron's breadth as the number of paradigms for which it has NSI>0. Figure 3b shows that most neurons respond to only a small number of paradigms and that broadly responsive neurons are rare.

Figure 5 shows the average number of responsive paradigms, phenomena, and domains per neuron. Breadth declines sharply in the early layers, remains low through the middle of the network, and increases modestly near the output. Together with the scarcity of NSI>2 units, this pattern supports a representation in which grammatical information is distributed across neurons with narrow, weak-to-moderate effects rather than concentrated in a few universally responsive units.

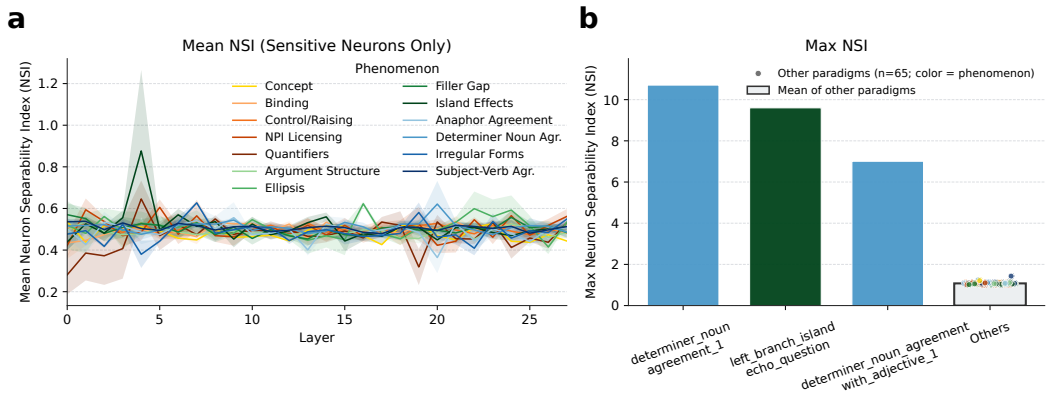

Figure 4: **(a)** Mean NSI across layers computed over sensitive neurons only (NSI>0), grouped by linguistic phenomenon. Even within this subset, the average NSI remains small, indicating that sensitivity is generally weak rather than sharply selective. **(b)** For each BLiMP/COMPS paradigm, we compute the maximum NSI across neurons. Bar and point colors indicate each paradigm's parent phenomenon; the outlined *Others* bar shows the mean over the remaining 65 paradigms. Only three paradigms contain a neuron with NSI>2. Appendix Figure 12 gives paradigm- and neuron-level detail.

Even broadly tuned neurons remain selective. Some neurons are poly-selective in the sense that they respond to at least three linguistic phenomena, but Figure 6 shows that this does not make them general-purpose grammar neurons. Clustering reveals clear blocks in which a unit has comparatively strong mean NSI for one phenomenon or a small group of related phenomena, while its responses elsewhere are weaker. Cross-phenomenon co-activation therefore resembles spillover around a dominant preference rather than uniformly strong selectivity across domains.

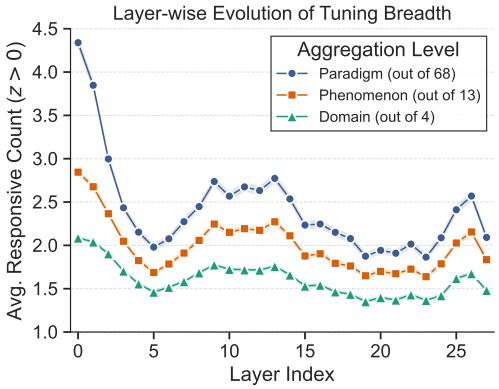

Figure 5: Layer-wise average tuning breadth at the paradigm, phenomenon, and domain levels. A neuron is counted as responsive to a phenomenon or domain when it has positive selectivity for at least one constituent paradigm. Shaded regions show standard error.

4.3 Robustness and Generalization

Generalization across models. The sparse, narrow-tuning pattern is not specific to Qwen3-0.6B. Across Qwen3-0.6B/1.7B/4B/8B, Pythia-410M, TinyLlama-1.1B, and Llama-3.1-8B, an average paradigm recruits only 2.52–3.94% of units with positive NSI (Figure 7a). Viewed from the complementary per-neuron perspective, mean tuning breadth ranges from only 1.72 to 2.68 responsive paradigms out of 68, and the 95th percentile is 5–7 paradigms (Figure 7b). Each checkpoint has only 2–6 paradigms with any NSI>2 unit. Across model scales and families, this sparse-selectivity pattern remains consistent. Complete model statistics and breadth distributions are in Appendix F.1.

Robustness to analysis choices. The conclusion does not depend on the exact cutoff. At NSI>1, all paradigms retain at least one above-threshold unit, yet these units comprise only 0.094% of the network on average. From NSI $\geq$ 1.645 through NSI>3, only 3 of 68 paradigms contain any above-threshold unit, with one such unit per paradigm (Figure 7c). Pairing controls clarify why both item matching and null normalization matter: random good–good and bad–bad pairs have raw separability near 0.50 but yield no NSI>2 units, whereas breaking item matching in cross-item good–bad pairs creates sporadic high-NSI

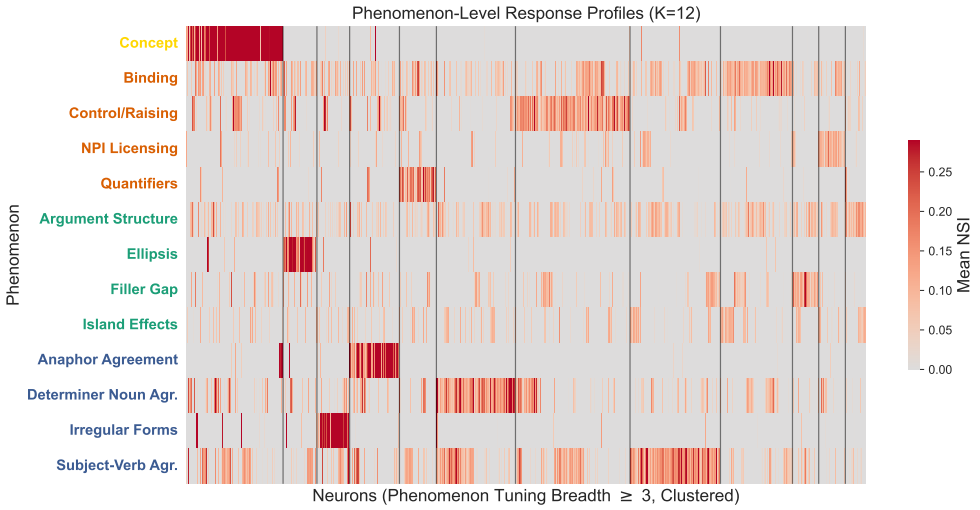

Figure 6: Functional modularity in tuning across linguistic phenomena. The heatmap displays mean phenomenon-level NSI for neurons responsive to at least three phenomena. Rows are grouped by domain, and columns are hierarchically clustered neurons. The block structure shows that even poly-selective neurons retain dominant within-domain preferences rather than behaving as uniform cross-domain grammar detectors.

artifacts. In the clean one-prefix deletion subset, removing the critical contrast token reduces mean raw separability from 0.0753 to 0.0011, while deleting a matched non-critical token leaves it at 0.0746. Appendix G reports the complete threshold analysis, and Appendix H reports the complete pairing controls.

Representation is not behavioral reliance. Whole-vector representations are highly linearly separable (mean layer-14 probe accuracy 0.917), while the same model's mean BLiMP behavioral accuracy is 0.758. Across paradigms the two are effectively unrelated (Pearson $r = -0.032$; Figure 7d). For example, `sentential_subject_island` has perfect probe accuracy (1.000) but behavioral accuracy of only 0.238. Thus, whole-vector linear separability, single-neuron NSI, and behavioral preference are empirically distinct quantities; neither probing nor NSI alone establishes causal use. Appendix I provides the full correlation, domain-level, and paradigm-level results.

Targeted group ablation does not reveal NSI-specific effects. We use ablations of $k \in \{5, 10, 20\}$ dimensions as the primary intervention, with each top group containing the frozen NSI>2 candidate and its highest-scoring same-layer neighbors (Table 1). Across all nine group comparisons, no top set is more damaging than 100 same-size random sets at $p < 0.05$ (all empirical $p \geq 0.109$), and absolute accuracy changes remain at or below 0.009. Both agreement paradigms show near-zero or positive margin changes across group sizes. The largest targeted drop occurs for the left-branch island paradigm at $k = 20$ ($\Delta M = -0.543$, paired 95% CI $[-0.630, -0.455]$), but it is not selective relative to random groups ($p = 0.109$), and the corresponding bottom-20 control also reduces the margin (-0.439). The group intervention therefore does not establish that high-NSI dimensions form a behaviorally privileged subset. This pattern is consistent with distributed or redundant encoding, although the null selective result does not prove that the selected dimensions are irrelevant. Appendix J reports the random intervals, top-1 auxiliary check, and all conditions.

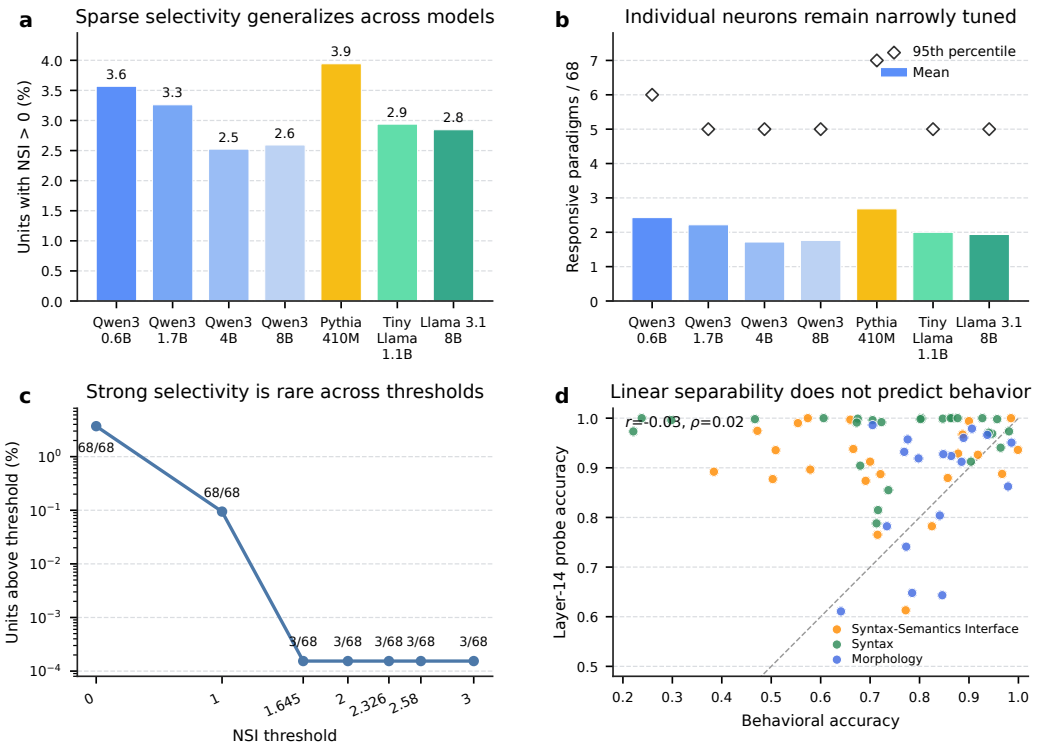

Figure 7: Robustness and generalization of neuron-level selectivity. **(a)** Across four Qwen3 scales and three additional checkpoints, only 2.5–3.9% of units have positive NSI for an average paradigm. **(b)** The average neuron responds to only 1.7–2.7 of 68 paradigms; diamonds mark the 95th percentile. **(c)** For Qwen3-0.6B, the fraction of units above threshold falls sharply, and only 3 of 68 paradigms contain any unit from NSI ≥ 1.645 through NSI > 3. **(d)** Whole-vector probe accuracy is essentially uncorrelated with minimal-pair behavior across 67 BLiMP paradigms.

Table 1: Primary targeted group-ablation results. ΔM is the change in mean $\log p(x^+) - \log p(x^-)$; brackets give paired-bootstrap 95% CIs, and p_{rand} compares each top group with 100 same-size random groups.

Paradigm	k	Top ΔM [95% CI]	Random mean	Bottom ΔM	p_{rand}
Det.–noun agr.	5	-0.005 $[-0.014, 0.003]$	-0.003	0.147	0.446
	10	0.000 $[-0.011, 0.011]$	-0.009	0.054	0.634
	20	0.111 $[0.084, 0.138]$	-0.034	-0.022	0.911
Det.–noun agr.+adj.	5	-0.006 $[-0.015, 0.004]$	-0.005	0.161	0.545
	10	-0.005 $[-0.020, 0.012]$	-0.012	0.177	0.604
	20	0.026 $[0.004, 0.048]$	-0.009	0.351	0.733
Left-branch island	5	0.009 $[-0.019, 0.038]$	-0.077	-0.229	0.752
	10	0.001 $[-0.039, 0.040]$	-0.166	-0.488	0.693
	20	-0.543 $[-0.630, -0.455]$	-0.148	-0.439	0.109

5 Conclusion

We present NSI, a minimal-pair, probe-free neuron-level measure that localizes linguistic selectivity across layers and characterizes it consistently at the domain, phenomenon, and paradigm levels. The approach is simple, statistically principled, and complementary to mechanistic analyses. Targeted ablations of groups containing the three rare NSI>2 candidates and their highest-scoring same-layer neighbors yield no effect distinguishable from same-size random controls, reinforcing the distinction between activation selectivity and causal reliance. Future work will move beyond residual-coordinate ablation to connect neuron clusters to concrete causal circuits through patching and mediation analysis (Vig et al., 2020), and extend to multilingual minimal pairs.

References

Omer Antverg and Yonatan Belinkov. On the pitfalls of analyzing individual neurons in language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=8uz0EWPQIMu.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Identifying and controlling important neurons in neural machine translation. *arXiv preprint arXiv:1811.01157*, 2018.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. 2023.

Tamera Bricken, Nelson Elhage Chan, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits (tech report)*, 2023. https://transformer-circuits.pub/2023/monosemantic-features.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6309–6317, 2019.

Wietse De Vries, Andreas Van Cranenburgh, and Malvina Nissim. What's so special about bert's layers? a closer look at the nlp pipeline in monolingual and multilingual models. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 4339–4350, 2020.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart M Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural 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)*, pp. 1828–1843, 2021.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 70–76, 2020.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 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.

Charles G Gross. Genealogy of the "grandmother cell". *The Neuroscientist*, 8(5):512–518, 2002.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In *NAACL*, 2018. URL https://aclanthology.org/N18-1108/.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R Brennan. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4488–4497, 2024.

Linyang He, Ercong Nie, Sukru Samet Dindar, Arsalan Firoozi, Van Nguyen, Corentin Puffay, Riki Shimizu, Haotian Ye, Jonathan Brennan, Helmut Schmid, et al. Xcomps: A multilingual benchmark of conceptual minimal pairs. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pp. 75–81, 2025a.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schütze, Nima Mesgarani, and Jonathan Brennan. Large language models as neurolinguistic subjects: Discrepancy between performance and competence. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19284–19302, 2025b.

Linyang He, Qiaolin Wang, Xilin Jiang, and Nima Mesgarani. Layer-wise minimal pair probing reveals contextual grammatical-conceptual hierarchy in speech representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35326–35341, 2025c.

Linyang He, Tianjun Zhong, Richard Antonello, Gavin Mischler, Micah Goldblum, and Nima Mesgarani. Far from the shallow: Brain-predictive reasoning embedding through residual disentanglement. *arXiv preprint arXiv:2510.22860*, 2025d.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. 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)*, pp. 4129–4138, 2019.

Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5040–5060, 2023.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*, 2020.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. Ravel: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, 2024.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. How large language models encode context knowledge? a layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8235–8246, 2024.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*, 2025.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11): 4302–4311, 1997.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3849–3864, 2021.

Jenny Kunz and Marco Kuhlmann. Where does linguistic information emerge in neural language models? measuring gains and contributions across layers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4664–4676, 2022.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. The emergence of number and syntax units in lstm language models. *arXiv preprint arXiv:1903.07435*, 2019.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, volume 2025, pp. 53617–53642, 2025.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the ACL*, 2016a. URL https://aclanthology.org/Q16-1037/.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016b.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, et al. Zhoblimp: a systematic assessment of language models with linguistic minimal pairs in chinese. *arXiv preprint arXiv:2411.06096*, 2024.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 2024.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *International Conference on Learning Representations*, volume 2025, pp. 33588–33636, 2025.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1151. URL https://aclanthology.org/D18-1151.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2920–2941, 2023.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. Cross-linguistic syntactic evaluation of word prediction models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5523–5539, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.490. URL https://aclanthology.org/2020.acl-main.490/.

Aaron Mueller, Yu Xia, and Tal Linzen. Causal analysis of syntactic agreement neurons in multilingual language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pp. 95–109, 2022.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, 2020.

Lorenzo Posani, Shuqi Wang, Samuel P Muscinelli, Liam Paninski, and Stefano Fusi. Rarely categorical, always high-dimensional: how the neural code changes along the cortical hierarchy. *biorxiv*, pp. 2024–11, 2025.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866, 2020.

Dieter Schlee. Numerical taxonomy. the principles and practice of numerical classification, 1975.

Giulio Starace, Konstantinos Papakostas, Rochelle Choenni, Apostolos Panagiotopoulos, Matteo Rosati, Alina Leidinger, and Ekaterina Shutova. Probing llms for joint encoding of linguistic categories. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7158–7179, 2023.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*, 2018.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

Limitations

NSI is a probe-free lens on neuron-level selectivity, but the results should be interpreted within the following scope.

Models and data. We study seven decoder-only checkpoints from the Qwen3, Pythia, TinyLlama, and Llama families; differences in corpora, tokenizers, architectures, objectives, or post-training could change sparsity and apparent localization. BLiMP and COMPS are controlled, templated English resources and under-represent natural discourse, pragmatics, gradient acceptability, and typologically diverse languages. The conclusions are therefore robust within the tested models and benchmarks, not universal.

Separability is not causality. NSI quantifies activation separability, while our intervention tests whether selected residual coordinates have uniquely strong causal effects. These within-benchmark tests cover three paradigms in one checkpoint; held-out phenomena and circuit-level interventions such as activation patching and mediation would provide stronger generalization and causal tests.

Representation choices. We use last-token layer outputs and treat each scalar dimension as a neuron. This can miss information at earlier positions, distributed over time, or expressed in attention and MLP signals; neuron definitions also vary across architectures. Alternative positions, components, and readouts may yield different selectivity patterns.

Calibration and cost. We use a matched within-pair permutation null that preserves lexical content while disrupting grammatical labels. Future work can model remaining lexical, tokenization, template, or item dependencies more explicitly. Robust normalization requires extensive permutation testing, but the independent permutations can be parallelized. Different similarity metrics emphasize distinct properties of neuronal responses. Although the sparse unit-level pattern persists under the Spearman and cosine variants (Appendix G), future work could more systematically examine what metric-dependent differences reveal about neuronal response structure and linguistic selectivity.

A Related Work

Targeted syntactic evaluation and minimal pairs. Evaluating the grammatical competence of language models has evolved from calculating overall perplexity to using targeted diagnostic datasets. Early work introduced small-scale, hand-crafted test suites to verify specific syntactic generalizations (Linzen et al., 2016a; Gulordava et al., 2018; Marvin & Linzen, 2018; Wilcox et al., 2018). This methodology was significantly scaled up with benchmarks like BLiMP (Warstadt et al., 2020) and automated platforms such as SyntaxGym (Gauthier et al., 2020), which use minimal pairs to isolate grammatical phenomena and reveal systematic gaps in LM performance, and was subsequently broadened to systematic syntactic generalization suites and to conceptual and multilingual coverage (Hu et al., 2020; Mueller et al., 2020; Liu et al., 2024; He et al., 2025a). High benchmark accuracy is nonetheless difficult to interpret on its own, as it can coexist with instability, prompt sensitivity, and divergence from human grammatical judgments (Dentella et al., 2023; Hu & Levy, 2023; Mahowald et al., 2024). More fundamentally, while these behavioral metrics effectively diagnose *what* linguistic rules a model violates, they treat the model as a black box, offering limited insight into *where* and *how* these distinctions are represented internally (He et al., 2024; 2025b).

Representational structure and probing. To understand internal representations, the community turned to diagnostic classifiers, or "probes." Seminal layer-wise analyses demonstrated that classical NLP pipeline steps, such as part-of-speech tagging and parsing, are naturally rediscovered in the hierarchical geometry of transformer representations (Tenney et al., 2019; Hewitt & Manning, 2019; Manning et al., 2020; De Vries et al., 2020; Koto et al., 2021; He et al., 2025c;d; Ju et al., 2024), although the reported ordering varies with architecture, task, and metric (Belinkov & Glass, 2019; Rogers et al., 2020). Despite these insights, the probing paradigm faces significant methodological criticism. A core debate concerns whether a probe reveals the model's intrinsic knowledge or merely exploits the probe's own capacity to learn the task from the embeddings (Hewitt & Liang, 2019). Theoretical work using information-theoretic criteria further suggests that probing results can be confounded by the ease of extracting information rather than its explicit presence (Pimentel et al., 2020; Voita & Titov, 2020), and related analyses show that probes can succeed on information the model does not actually use and that probe rankings are sensitive to design choices (Ravichander et al., 2021; Kunz & Kuhlmann, 2022). These limitations motivate the need for *probe-free* diagnostics, like our proposed framework, which directly measure selectivity without the interference of auxiliary training.

Mechanistic interpretability and neuron-level analyses. Mechanistic interpretability instead attributes computations to concrete components, such as attention heads, feed-forward key–value memories, and circuits supporting factual recall (Elhage et al., 2021; Geva et al., 2021; Meng et al., 2022). At the finest granularity, neuron-level studies have linked individual units to translation and morphology features (Bau et al., 2018; Dalvi et al., 2019) and to factual knowledge (Dai et al., 2022), while methodological work cautions that probe-based neuron ranking conflates information that is encoded with information the model actually uses (Antverg & Belinkov, 2022). Closest to our setting, causal mediation has localized neuron-level contributions to subject–verb agreement in English (Finlayson et al., 2021) and across multilingual models (Mueller et al., 2022), although the reliability of such interventions depends on patching design choices and their documented failure modes (Zhang & Nanda, 2023). These studies show that unit-level analysis is feasible, but each targets one or two phenomena in isolation, leaving open how typical their findings are. Our study is systematic in both coverage and measurement: a single classifier-free index is applied uniformly to 68 paradigms nested under 13 phenomena and four domains, computed identically for every neuron in every layer, and replicated across seven checkpoints from four model families under a common set of pairing and threshold controls. To our knowledge, this is the first survey of neuron-level linguistic selectivity at this scale, and it shifts the question from whether selective neurons exist for a given construction to how common and how broad such selectivity is across the grammar as a whole.

B Dataset

We leverage two complementary minimal-pair resources.

COMPS. The **COMPS** dataset (Misra et al., 2023) extends minimal-pair evaluation to conceptual and semantic compositional phenomena, focusing on whether models (and individual neurons) track meaning-sensitive distinctions beyond surface form. COMPS pairs are constructed to preserve as much lexical overlap as possible while flipping a compositional or conceptual requirement. We pool the selected COMPS contrasts into one paradigm, `comps_base`, assigned to the Concept phenomenon within the Concept domain. One example is:

(i) **Domain: Concept; phenomenon: Concept**
a) A kettle is used for boiling.
*b) *A hammer is used for boiling.*

BLiMP. The **BLiMP** benchmark (Warstadt et al., 2020) contains 67 paradigms of automatically generated minimal pairs, where each pair differs minimally but flips grammatical acceptability. We group these paradigms into 12 phenomena nested within three domains: **Syntax–Semantics Interface** (Binding, Control/Raising, NPI Licensing, and Quantifiers), **Syntax** (Argument Structure, Ellipsis, Filler Gap, and Island Effects), and **Morphology** (Anaphor Agreement, Determiner–Noun Agreement, Irregular Forms, and Subject–Verb Agreement). Examples from paradigms in each BLiMP domain include:

(ii) **Domain: Syntax–Semantics Interface; phenomenon: NPI Licensing**
a) Even Suzanne has really joked around.
*b) *Even Suzanne has ever joked around.*

(iii) **Domain: Syntax; phenomenon: Filler Gap**
a) Mark figured out that most governments appreciate Steve.
*b) *Mark figured out who most governments appreciate Steve.*

(iv) **Domain: Morphology; phenomenon: Subject–Verb Agreement**
a) The hospital appreciates Claire.
*b) *The hospitals appreciates Claire.*

Together, the combined hierarchy contains four domains, 13 phenomena, and 68 paradigms: 67 BLiMP paradigms plus one pooled COMPS paradigm. The two resources provide 116,300 English minimal pairs generated from linguist-crafted templates, with 96.4% (BLiMP) and 93.1% (COMPS) human agreement.

C Models

We conduct our main analysis on the recent **Qwen3-0.6B** model (Yang et al., 2025), a 0.6-billion-parameter causal transformer released by Alibaba Group. Qwen3 adopts a decoder-only architecture with rotary position embeddings, multi-head self-attention, and feed-forward layers following the standard transformer design. Despite its relatively modest size, Qwen3-0.6B achieves strong performance across a wide range of language modeling and reasoning benchmarks, making it a suitable testbed for fine-grained interpretability studies. Its compact scale allows us to efficiently extract and analyze neuron-level activations across all layers.

To study the generality of neuron-level selectivity, we additionally analyze larger Qwen3 checkpoints (**Qwen3-1.7B**, **Qwen3-4B**, and **Qwen3-8B**) and three checkpoints from other model families: **Pythia-410M**, **TinyLlama-1.1B**, and **Llama-3.1-8B** (Dubey et al., 2024). Architecture, calibration, sparsity, and tuning-breadth statistics are reported in Appendix F.1.

D What does NSI measure?

Because S^{raw} is a function of the Pearson correlation between the paired activation vectors $\mathbf{v}^+$ and $\mathbf{v}^-$, it characterizes a neuron's *item-wise response profile*, that is, the pattern of relative responses across the matched items of a paradigm. It is therefore informative about how the grammatical contrast reorganizes that profile, and is by construction invariant to the overall level and scale of the neuron's response.

The affine case makes this behavior explicit. Suppose the grammatical manipulation acts on a neuron as

$$\mathbf{v}^- = a\,\mathbf{v}^+ + c\,\mathbf{1}, \tag{7}$$

with gain $a \neq 0$, constant offset c, and $\mathbf{1}$ the all-ones vector. Then $\operatorname{corr}(\mathbf{v}^+, \mathbf{v}^-) = \operatorname{sign}(a)$, so

$$S^{\text{raw}} = \begin{cases} 0, & a > 0, \\ 1, & a < 0. \end{cases}$$

A uniform offset ($a = 1$, $c \neq 0$) and a pure gain change ($a > 0$, $c = 0$) both leave the relative ordering of items intact and receive the minimum score, whereas an order-reversing response ($a < 0$) receives the maximum. Departures from an order-preserving affine relation fall between these limits, with the score increasing as the item-wise profile under the ungrammatical condition departs further from its grammatical counterpart.

The permutation normalization then asks whether the observed reorganization exceeds what within-pair label swapping produces. A high NSI therefore indicates that the grammatical contrast systematically reshapes a neuron's relative response profile across matched linguistic items, beyond the permutation null.

This is the quantity the minimal-pair design is best suited to identify. Because the two members of a pair differ by as little as one token, some change in a neuron's overall activation level is expected for a large fraction of units and reflects the lexical substitution as much as the grammatical contrast itself. A change in the relative ordering across items is more specific, in that it indicates that the neuron's response depends on the grammatical status of each individual item. NSI is accordingly a measure of item-specific profile selectivity; mean-level effects constitute a complementary aspect of the paired contrast, which the uncentered cosine variant in Appendix G retains.

E Raw-Score Layer Localization

We first report localization patterns from the raw separability score, following the same raw-score-first order as the main results. Figure 8 summarizes the paradigm-first saturation analysis, and Figure 9 provides a cross-method comparison with whole-vector probing.

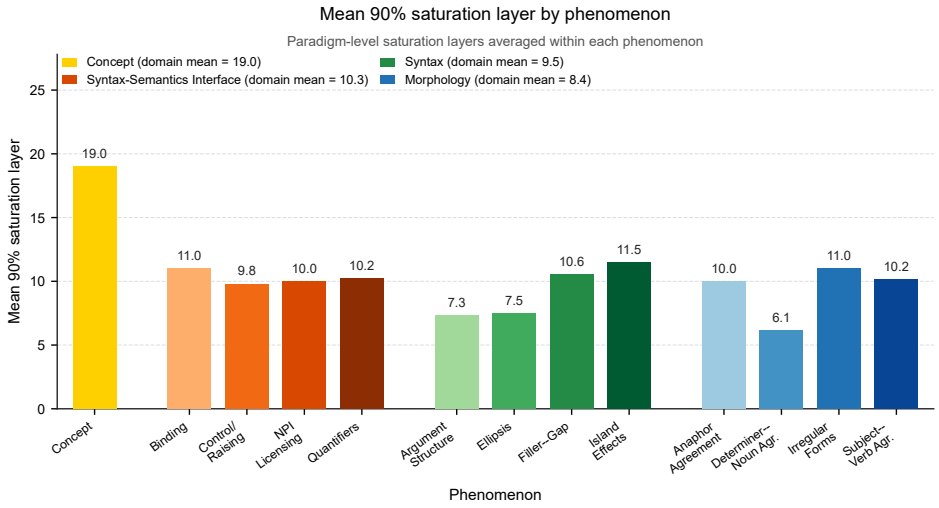

Figure 8: Paradigm-first 90% saturation layers summarized by phenomenon. For each paradigm, we identify the earliest layer at which its median raw-separability score reaches 90% of its own maximum. Bars show the mean of these paradigm-level saturation layers within each phenomenon. The legend reports means computed directly over all constituent paradigms in each domain: 19.0 for Concept, 10.3 for Syntax–Semantics Interface, 9.5 for Syntax, and 8.4 for Morphology.

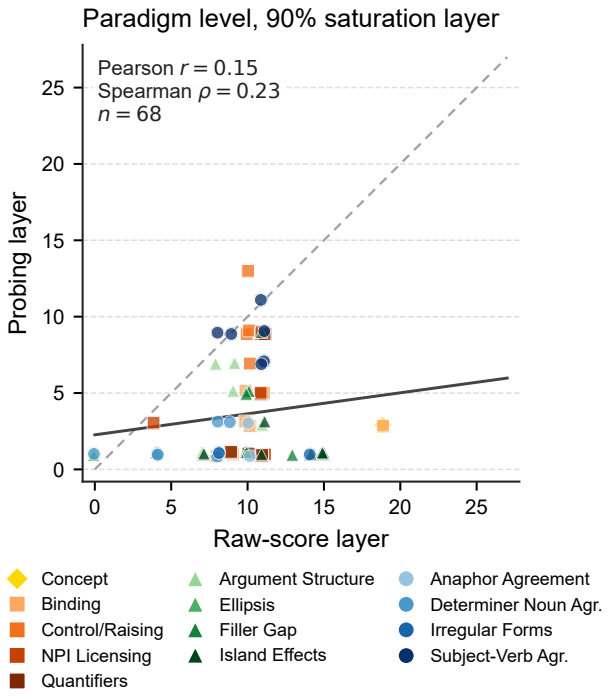

Figure 9: Correspondence between raw-separability and probing 90% saturation layers across the 68 paradigms. For each paradigm and measure, saturation is the earliest layer reaching 90% of that paradigm's own maximum. Colors encode phenomena, marker shapes encode domains, the dashed diagonal indicates equal layers, and the solid line is the least-squares fit. The association is weakly positive (Pearson $r = 0.15$; Spearman $\rho = 0.23$), indicating limited paradigm-level agreement between the two localization measures. Probing embedding index 0 is excluded, and sampled hidden-state indices $2, 4, \ldots, 28$ are mapped to model layers $1, 3, \ldots, 27$.

F NSI Selectivity and Tuning Breadth

F.1 Cross-model robustness

Table 2 reports architecture, calibration, sparsity, and tuning-breadth statistics across the seven checkpoints. Figure 10 shows the corresponding breadth distributions.

Table 2: Cross-model selectivity and tuning breadth over 68 paradigms. "Paradigms > 2" is the number of paradigms containing any NSI>2 unit. Breadth B counts the number of paradigms with NSI>0 for each neuron.

Model	Layers × width	NSI>0 (%)	NSI>2 (%)	Paradigms > 2	Mean B	P95/max B	$B \geq 10$ (%)
Qwen3-0.6B	28 × 1024	3.568	1.54×10^{-4}	3	2.43	6/13	0.18
Qwen3-1.7B	28 × 2048	3.261	2.82×10^{-4}	6	2.22	5/13	0.09
Qwen3-4B	36 × 2560	2.524	1.12×10^{-4}	2	1.72	5/13	0.06
Qwen3-8B	36 × 4096	2.593	1.92×10^{-3}	4	1.76	5/15	0.24
Pythia-410M	24 × 1024	3.942	5.39×10^{-4}	2	2.68	7/16	0.86
TinyLlama-1.1B	22 × 2048	2.939	6.20×10^{-4}	2	2.00	5/17	0.35
Llama-3.1-8B	32 × 4096	2.846	8.98×10^{-5}	4	1.94	5/16	0.11

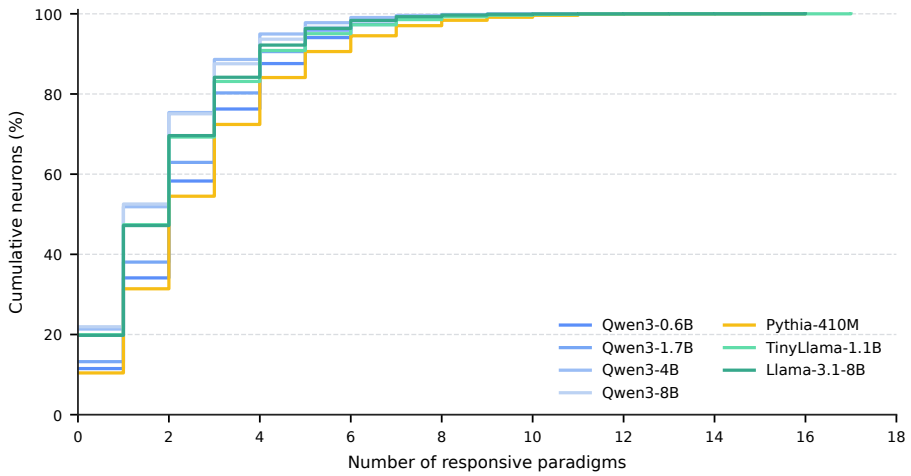

Figure 10: Cumulative distribution of paradigm-level tuning breadth across models. Even under the lenient NSI>0 definition, the high-breadth tail is small: the 95th percentile is 5–7 paradigms and fewer than 1% of neurons respond to 10 or more paradigms in every model.

F.2 Selectivity distributions and high-selectivity cases

The three paradigms containing an NSI>2 unit in the main 500-permutation Qwen3-0.6B analysis are illustrated below. As Figure 12 shows, each paradigm contains only one such unit across all 28 × 1024 layer-neuron coordinates.

(i) **Domain:** Morphology; **Phenomenon:** Determiner–Noun Agreement
Paradigm: Determiner–Noun Agreement 1
a) Raymond is selling this sketch.
*b) *Raymond is selling this sketches.*

(ii) **Domain:** Syntax; **Phenomenon:** Island Effects
Paradigm: Left-Branch Island Echo Question
a) Benjamin was researching whose books?
*b) *Whose was Benjamin researching books?*

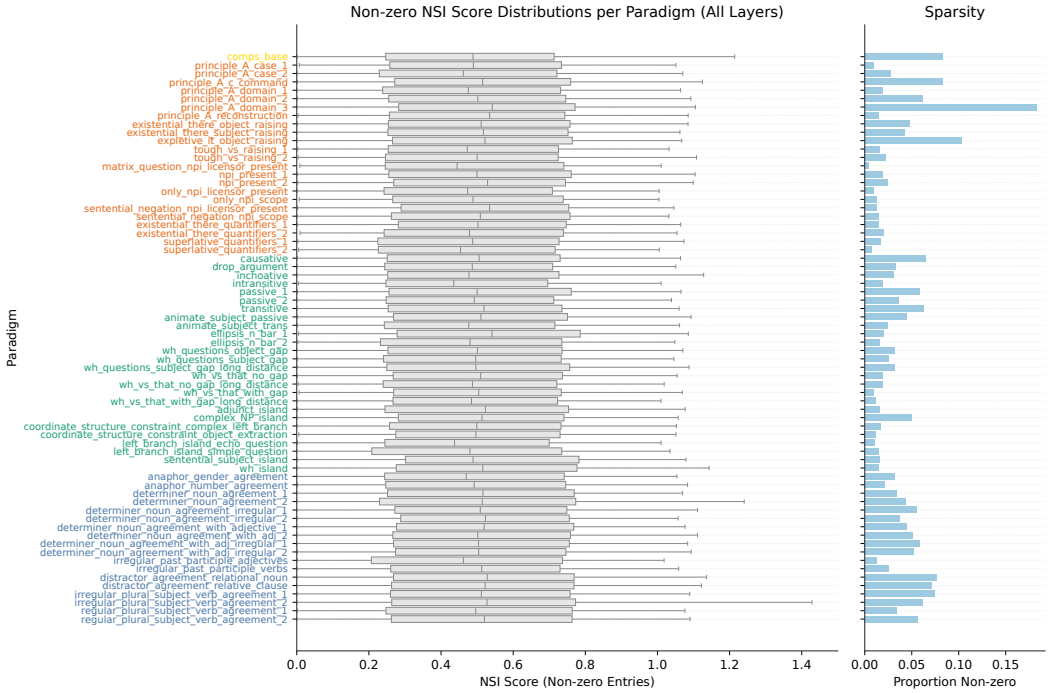

Figure 11: **Left**: Distributions of non-zero NSI scores (i.e., over sensitive neurons with NSI>0) aggregated across all layers, shown separately for each paradigm. Across paradigms, the mean NSI is typically around ~ 0.5, indicating that most sensitive neurons exhibit only weak separability. **Right**: Paradigm-wise proportion of sensitive neurons across all 68 BLiMP/COMPS paradigms for Qwen3-0.6B.

(iii) **Domain:** Morphology; **Phenomenon:** Determiner–Noun Agreement
Paradigm: Determiner–Noun Agreement with Adjective 1
a) Rebecca was criticizing those good documentaries.
*b) *Rebecca was criticizing those good documentary.*

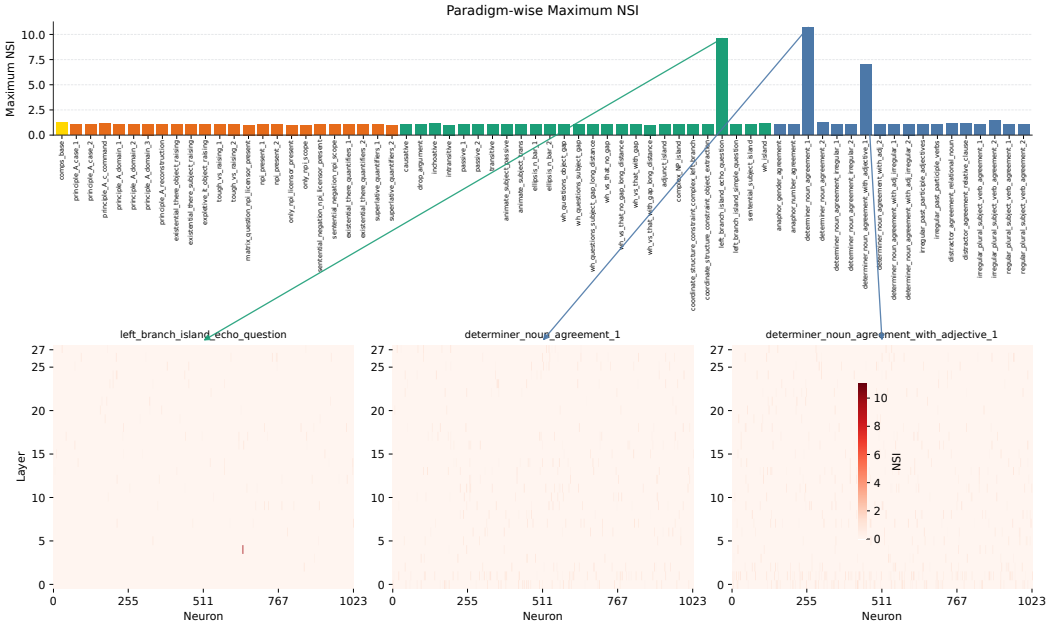

Figure 12: Paradigm-wise maximum NSI scores across all 68 BLiMP/COMPS paradigms for Qwen3-0.6B. Only three paradigms contain a unit with NSI>2; the bottom panels show that each case is realized by one unit across the network.

F.3 Tuning breadth across domains and phenomena

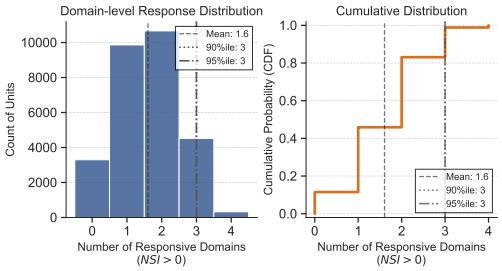

Figure 13: Domain-level tuning breadth. Most neurons respond to only one or two of the four domains.

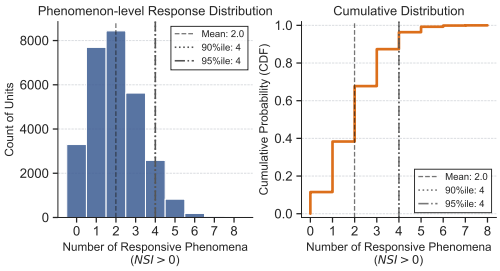

Figure 14: Phenomenon-level tuning breadth. More than 90% of neurons respond to four or fewer phenomena.

G Similarity Metrics and Threshold Calibration

Threshold sweep and empirical null. Table 3 reports the complete threshold sweep for Qwen3-0.6B. Weak positive effects occur in every paradigm, but the above-threshold population contracts rapidly. At NSI>1, an average paradigm contains only 27.1 units out of 28,672. At every tested threshold from 1.645 to 3, only three paradigms contain any above-threshold unit, with one unit in each paradigm.

Table 3: Threshold sweep over 68 paradigms and 28 × 1024 units per paradigm.

Threshold	Units above (%)	Paradigms with any	Mean count	Max count
0	3.7010	68	1061.15	5229
1	0.0944	68	27.06	120
1.645	1.54×10^{-4}	3	0.044	1
2	1.54×10^{-4}	3	0.044	1
2.326	1.54×10^{-4}	3	0.044	1
2.58	1.54×10^{-4}	3	0.044	1
3	1.54×10^{-4}	3	0.044	1

For empirical calibration, we compute $p = (1 + \#\{S^{\text{perm}} \geq S^{\text{raw}}\})/(1 + N_{\text{perm}})$ and apply Benjamini–Hochberg correction within each paradigm. No unit survives at $q < 0.05$ or $q < 0.01$. This result is deliberately treated as a conservative sanity check: with $N_{\text{perm}} = 200$, the minimum attainable empirical p-value is $1/201 = 0.004975$, which is too coarse for a powerful correction over 28,672 units. The permutation nulls are also non-Gaussian (mean absolute skewness 1.99; mean paradigm-level median excess kurtosis 4.45). These diagnostics motivate describing NSI>2 as an operational strong-selectivity threshold rather than a literal Gaussian significance cutoff.

Alternative similarity metrics. We recompute the score with Spearman rank correlation and cosine similarity while keeping the 68 paradigms, Qwen3-0.6B checkpoint, layer–neuron grid, and permutation-normalization procedure fixed. At the operational strong-selectivity threshold $z > 2$, only 0.222% of all tested layer–neuron units exceed the threshold under Spearman and 0.277% under cosine. Thus, even these more permissive variants leave more than 99.7% of units below threshold, preserving the central conclusion that strong single-unit selectivity is sparse.

Neither alternative is as well matched to our estimand. Each vector coordinate in our analysis is a repeated observation of the same scalar neuron across matched items, rather than a distinct representation feature. Spearman replaces activation values with ranks, making it invariant to monotonic rescaling but discarding graded response magnitudes and becoming sensitive to ties. Cosine does not center each neuron's responses across items, so baseline and mean-level activation effects can appear as separability; its standardized scores can also become unstable when the permutation variance is extremely small. Pearson correlation instead centers each neuron's item-wise response vector and directly measures whether the relative response pattern is preserved across the two members of a minimal pair. We therefore retain Pearson for the primary analysis and treat Spearman and cosine as complementary metric checks.

H Pairing and Deletion Controls

All controls in this section are run on Qwen3-0.6B with 200 permutations per paradigm.

H.1 Random-pair controls

We compare the original item-matched grammatical–ungrammatical pairs with three deranged pairings, so no sentence remains paired with itself. Table 4 separates raw decorrelation from null-normalized selectivity. Random same-label pairs have raw scores near 0.5 because unrelated sentences are weakly correlated, but they produce no NSI>2 units. Conversely, cross-item good–bad pairs generate many sporadic outliers, showing that item alignment is necessary for interpreting the contrast. The original row here comes from the 200-permutation control run; its four above-threshold paradigms should not be conflated with the three paradigms in the 500-permutation main analysis.

Table 4: Pairing controls over 68 paradigms.

Condition	Mean raw	NSI>0 (%)	NSI>2 (%)	Paradigms > 2	Median max
Matched good–bad	0.1438	3.469	6.67×10^{-4}	4	1.128
Random good–good	0.5008	100.000	0	0	0.980
Random bad–bad	0.5009	100.000	0	0	0.989
Cross-item good–bad	0.5004	16.464	0.0454	50	3.325

H.2 Critical-token deletion

For agreement paradigms, we compare the original pair with deletion of either the annotated critical token or a matched non-critical token. The clean test is the 10-paradigm one-prefix subset: the critical word directly realizes the good–bad contrast, so deleting it makes the pair identical or nearly identical. Table 5 shows that critical deletion nearly eliminates raw separability, whereas random deletion does not.

Table 5: Critical-token deletion controls. We emphasize raw separability because deletion changes the geometry and variance of the permutation null.

Subset	Condition	Paradigms	Mean raw	Median raw
One-prefix	Original	10	0.0753	0.0757
	Critical deleted	10	0.0011	0.0007
	Random deleted	10	0.0746	0.0772
Two-prefix	Original	6	0.1007	0.1069
	Shared word deleted	6	0.1802	0.1500
	Random deleted	6	0.0971	0.0933

For example, deleting the critical verb from *Paula references Robert* versus *Paula reference Robert* leaves the same fragment, *Paula Robert*, on both sides. In two-prefix paradigms, however, the annotated shared target is not the contrasting word: deleting it can leave prefixes such as *The students* versus *The student*. Those six paradigms therefore do not constitute a clean contrast-removal test and are reported separately rather than used in the main conclusion. Normalized NSI after critical deletion is not emphasized because near-identical pairs also collapse the null variance, making the resulting z-score unstable.

I Probe, NSI, and Behavior

The behavioral analysis covers the 67 standard BLiMP paradigms and excludes COMPS. For each minimal pair, behavior is correct when the grammatical sentence receives higher mean token log-probability. Whole-vector results use a logistic-regression probe on layer-14 final-token representations with five-fold stratified cross-validation. Mean behavioral accuracy is 0.758, whereas mean probe accuracy is 0.917.

Table 6: Correlations across 67 BLiMP paradigms.

Comparison	n	Pearson r	Spearman ρ
Behavior vs. layer-14 probe	67	-0.032	0.023
Behavior vs. peak probe	67	-0.030	0.008
Behavior vs. mean NSI	67	-0.021	-0.022
Behavior vs. maximum NSI	67	0.169	0.177
Layer-14 probe vs. mean NSI	67	-0.557	-0.578
Layer-14 probe vs. maximum NSI	67	0.092	-0.178

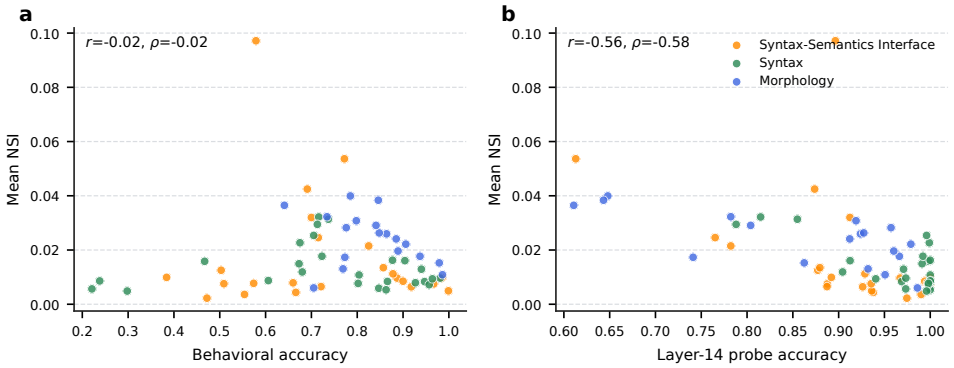

Figure 15: Additional comparisons among behavior, whole-vector probing, and neuron-level NSI. **(a)** Mean NSI is uncorrelated with behavioral accuracy. **(b)** Whole-vector probe accuracy and mean NSI are negatively correlated, further showing that distributed linear separability and average single-neuron selectivity capture different properties.

Table 7: Domain-level averages over the 67 BLiMP paradigms. The legacy `semantics`, `syntax_semantics`, and `syntax/semantics` source labels are merged into the Syntax–Semantics Interface domain.

Domain	Paradigms	Behavior	Probe	Mean NSI	Max NSI
Syntax–Semantics Interface	23	0.727	0.907	0.018	1.133
Syntax	26	0.736	0.964	0.014	1.424
Morphology	18	0.831	0.861	0.024	1.932

Table 8: Representative paradigms ranked by behavioral accuracy. Near-perfect linear probing can coexist with poor model behavior.

Paradigm	Domain	Behavior	Probe	Mean NSI	Max NSI
`principle_A_case_1`	Syntax–Semantics Interface	0.999	0.936	0.005	1.134
`anaphor_number_agreement`	Morphology	0.986	0.951	0.011	1.118
`principle_A_domain_1`	Syntax–Semantics Interface	0.985	1.000	0.010	1.224
`wh_vs_that_with_gap_long_distance`	Syntax	0.221	0.973	0.006	1.098
`sentential_subject_island`	Syntax	0.238	1.000	0.009	1.105
`wh_vs_that_with_gap`	Syntax	0.298	0.996	0.005	1.064
`npi_present_1`	Syntax–Semantics Interface	0.384	0.892	0.010	1.165
`drop_argument`	Syntax	0.467	0.998	0.016	1.127

J Targeted Ablation

We use ablation as a targeted causal follow-up rather than as a second neuron-selection procedure. The observational NSI analysis is frozen before intervention and identifies one NSI>2 coordinate in each of three Qwen3-0.6B paradigms: layer 20, neuron 389 for `determiner_noun_agreement_1` (NSI 10.073); layer 20, neuron 389 for `determiner_noun_agreement_with_adjective_1` (NSI 6.037); and layer 4, neuron 646 for `left_branch_island_echo_question` (NSI 8.927). Selection and evaluation use the same 1,000 minimal pairs, so this is a within-benchmark intervention test, not an independent replication.

At the frozen target layer, we zero the top $k \in \{1, 5, 10, 20\}$ residual-stream dimensions at every non-padding token; each top group contains the above-threshold candidate together with its highest-scoring same-layer neighbors. The main analysis treats $k \in \{5, 10, 20\}$ as group interventions; $k = 1$ is an auxiliary necessity check. Signed null-normalized Pearson scores determine the top and bottom sets. Random controls comprise 100 unique same-layer sets per k, sampled from a pool that excludes the union of the top-20 and bottom-20 dimensions. All full runs use bfloat16. The primary outcome is the change in the mean total log-probability margin, $M = \log p(x^+) - \log p(x^-)$. We obtain item-level 95% CIs from 2,000 paired bootstrap resamples. Separately, the one-sided random-control statistic is

$$p_{\text{rand}} = \frac{1 + \#\{\Delta M_{\text{random}} \leq \Delta M_{\text{top}}\}}{101}, \tag{8}$$

where a more negative ΔM is considered more damaging. Bootstrap resampling is used only for confidence intervals, not for this empirical p-value.

Figure 16 and Table 9 summarize the intervention results. Across the nine primary group comparisons ($k \in \{5, 10, 20\}$), no targeted set is more damaging than same-size random controls at $p < 0.05$. Although the left-branch top-20 intervention has a clear negative item-level effect, it is not extreme relative to the random interventions, and its bottom control is similarly negative. The auxiliary top-1 checks are also small and non-selective, but we do not treat single-coordinate ablation as the main causal test. We therefore conclude only that this intervention does not identify a behaviorally privileged high-NSI group, not that the selected dimensions are causally irrelevant. Redundancy, superposition, and the coarseness of residual-coordinate zeroing remain possible explanations for the lack of selective effects.

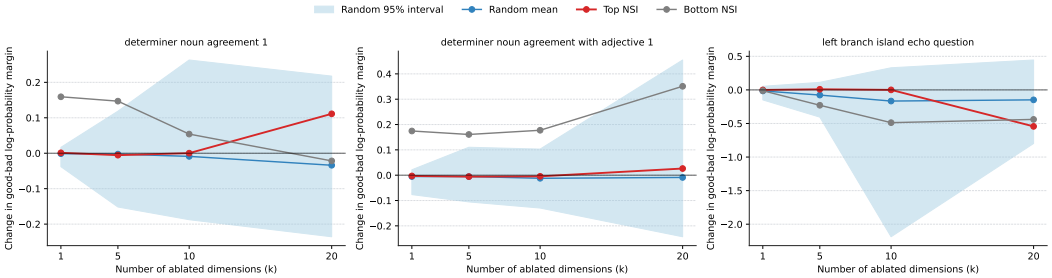

Figure 16: Targeted ablation at each frozen candidate's actual layer. Red and gray lines show the top-score and bottom signed-score sets; the blue line and band show the mean and 2.5th–97.5th percentile interval across 100 same-layer random sets. Negative values indicate a reduced grammatical-over-ungrammatical log-probability margin.

Table 9: Complete targeted-ablation results. Top CI is the paired-bootstrap 95% interval over items; Random interval is the 2.5th–97.5th percentile interval across 100 random sets. ΔA is the change in total-log-probability minimal-pair accuracy.

Paradigm	k	Top ΔM	Top CI	Random mean	Random interval	Bottom ΔM	ΔA	p_{rand}
Det.–noun agr.	1	0.001	$[-0.004, 0.006]$	-0.001	$[-0.038, 0.017]$	0.159	0.001	0.554
	5	-0.005	$[-0.014, 0.003]$	-0.003	$[-0.152, 0.119]$	0.147	-0.001	0.446
	10	0.000	$[-0.011, 0.011]$	-0.009	$[-0.188, 0.263]$	0.054	0.002	0.634
	20	0.111	$[0.084, 0.138]$	-0.034	$[-0.236, 0.218]$	-0.022	-0.002	0.911
Det.–noun agr.+adj.	1	-0.003	$[-0.008, 0.001]$	-0.005	$[-0.077, 0.021]$	0.174	0.001	0.396
	5	-0.006	$[-0.015, 0.004]$	-0.005	$[-0.106, 0.110]$	0.161	0.000	0.545
	10	-0.005	$[-0.020, 0.012]$	-0.012	$[-0.130, 0.103]$	0.177	-0.003	0.604
	20	0.026	$[0.004, 0.048]$	-0.009	$[-0.243, 0.455]$	0.351	-0.003	0.733
Left-branch island	1	-0.0003	$[-0.018, 0.019]$	-0.015	$[-0.149, 0.054]$	-0.012	-0.003	0.614
	5	0.009	$[-0.019, 0.038]$	-0.077	$[-0.406, 0.113]$	-0.229	-0.004	0.752
	10	0.001	$[-0.039, 0.040]$	-0.166	$[-2.187, 0.330]$	-0.488	-0.006	0.693
	20	-0.543	$[-0.630, -0.455]$	-0.148	$[-0.799, 0.446]$	-0.439	-0.009	0.109