# DECODABLE BUT NOT DETACHABLE: TRAINING DATA GRANULARITY DETERMINES PARAMETRIC MODULARITY IN LARGE LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Do large language models contain domain-specific parametric shells: concentrated, causally necessary neuron populations whose removal selectively degrades a target domain while sparing others? We apply a uniform causal methodology across two domain granularities, three model families (1.5B to 7B parameters), and eight domains. At the academic subject level, zero neurons exceed 60% domain selectivity across 939,008 combined FFN neurons and causal damage matrices are flat, despite domain identity being linearly decodable above 85% accuracy. At the language and modality level, 0.65–1.14% of neurons exceed 60% selectivity, damage matrices are near-perfectly diagonal (ratios up to 595:1), and shell neuron sets are essentially disjoint (IoU $< 0.003$). Masking code-selective neurons reduces mathematical reasoning accuracy by 16–24 percentage points across all models; chain-of-thought prompting recovers up to 75% of this damage, revealing that the code shell carries both an answer-structuring function and an irreducible reasoning function. Masking Spanish or Chinese neurons leaves mathematical reasoning at or below random. Shell causal strength increases monotonically with model scale, shells are spatially interleaved in a pattern that precludes group-level selective quantization, and amplification of shell neurons degrades performance monotonically, confirming that shells operate at a precisely calibrated functional point. Parametric shells form precisely where training data was modular at the token level, and our evidence across three model families and eight domains suggests they do not form otherwise.

## 1 INTRODUCTION

Large language models (LLMs) process dozens of languages and thousands of subject domains within a single set of parameters, yet how this capacity is parametrically organized remains poorly understood. Prior work has attempted to identify language-specific neurons in multilingual LLMs: Tang et al. (2024) propose LAPE to locate neurons by differential activation likelihood, but Le and Li (2026) subsequently show that activation preference does not imply functional necessity — LAPE-identified neurons achieve LangSpec-F1 near zero under direct causal intervention. Le and Li (2026) address this with CRANE, a relevance-attribution framework with substantially stronger causal effects, but evaluate a single model family across three languages without a null condition against which to assess when parametric shells exist and when they do not.

This paper addresses that open question. We apply a uniform causal methodology across two domain granularities, three model families (1.5B to 7B parameters), and eight domains: four academic subject categories and four language and modality domains (English, code, Spanish, Chinese). For each domain, we capture mean pre-activation magnitudes across the FFN intermediate dimension, identify domain-selective neurons via a share-based metric with backbone exclusion, and measure causal damage matrices net of a count-matched random baseline. The results reveal a sharp granularity boundary: zero neurons exceed 60% domain selectivity at the subject level across 939,008 combined FFN neurons, while 0.65–1.14% exceed this threshold at the language level with near-perfectly diagonal damage matrices and shell IoU below 0.003. Parametric shells form where and only where training data was modular at the token level.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We make the following contributions:

1. **A granularity boundary for parametric modularity.** We establish that language and modality domains produce functional parametric shells while academic subject domains do not, despite equivalent linear decodability, contextualizing prior work by identifying the conditions under which shells form.
2. **Shell concentration and functional load.** We provide the first systematic Pareto analysis of language shell structure, showing that 2% of neurons carries approximately 94% of maximum domain-specific capability loss across all four language domains and all three model families.
3. **Scale analysis of shell effects.** Diagonal dNLL increases monotonically with model scale (1.82 to 3.37 nats across 1.5B to 7B parameters), indicating that larger models develop more functionally load-bearing shells independent of the proportion of selective neurons.
4. **Functional and spatial characterization.** Code-selective neurons carry mathematical reasoning capability across all tested models; language-selective neurons concentrate in final network layers for non-English domains; and functional shells are spatially interleaved throughout the weight matrices, precluding group-level selective quantization.

## 2 Related Work

**Neuron specialization in language models.** Prior work has established that individual FFN neurons encode interpretable features (Bau et al., 2018), that specific neurons correlate with factual associations and their suppression degrades factual recall (Dai et al., 2022), and that the FFN intermediate dimension functions as a population of sparse feature detectors rather than a monolithic block (Voita et al., 2024). Subsequent work localizes task-specific (Wang et al., 2022), culture-specific (Song et al., 2024), and syntactic (Mueller et al., 2022) neurons within this population. Our work applies the same causal methodology to ask not which domains have dedicated neurons but under what conditions any domain acquires them.

**Language-specific neurons in multilingual LLMs.** Tang et al. (2024) identify language-specific neurons via activation entropy (LAPE) and show that deactivating them degrades target-language performance. However, Le and Li (2026) demonstrate that activation preference does not imply functional necessity: LAPE-identified neurons achieve LangSpec-F1 near zero under direct causal intervention. Le and Li (2026) address this with CRANE, a relevance-attribution framework with substantially stronger causal effects, though limited to a single model family and three languages without a null condition. Kojima et al. (2024) and Gurgurov et al. (2025b) report complementary findings, noting that language-selective neuron sets are nearly disjoint and sufficient to steer output language. Our work differs in three respects: we apply a uniform methodology across two domain granularities, producing the controlled null result that contextualizes the language-positive finding; we provide Pareto and scale analyses quantifying shell concentration and growth; and our causal effect magnitudes are substantially larger, with diagonal dNLL of 1.82–3.37 nats versus the sub-unit perplexity changes reported by Tang et al. (2024). Recent work identifies task-relevant and multimodal domain-specific neuron populations through alternative attribution methods (Chen et al., 2025; Huo et al., 2024); our subject-level null result holds under activation-magnitude identification, though task-specific partitions may surface weaker structure at finer granularities. Concurrent work on transfer neurons (Tezuka and Inoue, 2025) and language arithmetic (Gurgurov et al., 2025a) identifies shared transitional parameters; our IoU finding applies to the top-2% most selective neurons and does not exclude partial sharing at lower thresholds.

**Multilingual representation structure.** Wendler et al. (2024) show that LLMs process non-English inputs by mapping them toward English representational space in intermediate layers before projecting back to the target language at output, providing a mechanistic account of the U-shaped layer distribution of language-specific neurons reported by Tang et al. (2024). Our layer distribution findings complicate this picture: in instruction-tuned models, language-selective neurons concentrate primarily in final layers rather than at both ends, consistent with instruction tuning absorbing input-side language mapping into the shared representation.

# 3 METHODOLOGY

## 3.1 MODELS AND DOMAINS

We evaluate three instruction-tuned decoder-only language models spanning distinct architectures and scales: Qwen2.5-1.5B-Instruct (Qwen et al., 2025), Llama-3.2-3B-Instruct (Grattafiori et al., 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). These models contain $N = 250{,}880$, $229{,}376$, and $458{,}752$ FFN intermediate neurons respectively, for a combined total of $939{,}008$ neurons across all three model families.

We study two domain granularities. **Subject domains** consist of four labeled categories derived from MMLU (Hendrycks et al., 2021), ARC-Challenge (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018): quantitative reasoning, biomedical, humanities and social sciences, and elementary science (full subject lists in Appendix E). Evaluation uses letter-logprob multiple-choice scoring. **Language and modality domains** consist of four corpora: English (WikiText-2 (Merity et al., 2016)), code (MBPP (Austin et al., 2021)), Spanish, and Chinese (both from OPUS-100 (Tiedemann and Thottingal, 2020)). Evaluation uses mean negative log-likelihood (NLL) over 512-token segments. We additionally evaluate mathematical reasoning using GSM8K (Cobbe et al., 2021) to test whether domain-selective neurons carry generalizable capability.

## 3.2 NEURON ACTIVATION CAPTURE

For each domain $d$ and each model layer $l \in \{1, \ldots, L\}$, we capture the mean absolute pre-activation of each FFN intermediate neuron. Concretely, for a forward pass on input $x$, let $\mathbf{a}^{(l)}(x) \in \mathbb{R}^{d_f}$ denote the input to the down-projection matrix $\mathbf{W}_{\text{down}}^{(l)} \in \mathbb{R}^{d_h \times d_f}$ at layer $l$, averaged over all token positions. We define the mean activation of neuron $n = (l, j)$ for domain $d$ as:

$$M_{d,n} = \mathbb{E}_{x \sim \mathcal{D}_d} \left[ \frac{1}{T} \sum_{t=1}^{T} \left| a_j^{(l)}(x)_t \right| \right] \tag{1}$$

where $\mathcal{D}_d$ is the set of inputs for domain $d$ and $T$ is the sequence length. For subject domains, inputs consist of the question text alone without answer choices, computed via a mean-pooled forward pass. Omitting answer choices is essential: a pilot experiment using full multiple-choice prompts revealed that final-token representations cluster by the model's predicted answer letter rather than question content, confounding domain identification. For language domains, inputs are 512-token text segments. All captures use a maximum input length of 1024 tokens with item-level norm logging; items in the bottom 2% by $\ell_2$ norm of the resulting representation are excluded from subspace fitting to prevent degenerate near-zero vectors from distorting domain statistics.

## 3.3 DOMAIN SELECTIVITY

Let $K$ denote the number of domains ($K = 4$ in all experiments, for both subject and language granularities) and $N = L \cdot d_f$ the total number of FFN neurons. We define the *domain share* of neuron $n$ for domain $d$ as:

$$\sigma_{d,n} = \frac{M_{d,n}}{\sum_{d'=1}^{K} M_{d',n} + \epsilon} \tag{2}$$

where $\epsilon = 10^{-12}$ is a small constant for numerical stability.

Under uniform domain activity, $\sigma_{d,n} = 1/K$ for all $d$. The *max-share* of neuron $n$ is $s_n = \max_d \sigma_{d,n}$, which equals $1/K$ for a neuron with no domain preference and approaches 1 for a neuron active exclusively for one domain. We report the distribution of $s_n$ across all neurons as the selectivity histogram, using $s_n > 0.60$ as our primary threshold for identifying neurons with strong domain preference.

We exclude a backbone set $\mathcal{B}$ consisting of the top 0.1% of neurons by global mean activation $\bar{M}_n = \frac{1}{K} \sum_d M_{d,n}$ from all subsequent analyses. These neurons exhibit high activity across all

domains and correspond to the massive-activation or super-weight phenomenon documented in prior work (Sun et al., 2024); including them inflates apparent domain damage in masking experiments because their removal degrades all domains uniformly rather than selectively.

For each domain $d$, we define the domain shell $\mathcal{S}_d$ as the top-$m$ neurons ranked by excess share $\sigma_{d,n} - 1/K$, excluding $\mathcal{B}$, where $m$ is determined by a fixed mask fraction $\rho \in [0, 1]$ applied to the total neuron count:

$$\mathcal{S}_d = \underset{n \notin \mathcal{B}}{\text{top-}m} \left( \sigma_{d,n} - \frac{1}{K} \right), \quad m = \lfloor \rho N \rfloor \tag{3}$$

We use $\rho = 0.02$ as our primary mask fraction and report Pareto curves over $\rho \in \{0.005, 0.01, 0.02, 0.05, 0.10, 0.20\}$. A count-matched random baseline $\mathcal{S}_{\text{rand}}$ is sampled uniformly from $\{1, \ldots, N\} \setminus \mathcal{B}$ at each fraction.

### 3.4 Causal Masking and Damage Measurement

To assess functional necessity, we intervene on the model by zeroing the intermediate activations at positions in a target shell during inference. Specifically, for a shell $\mathcal{S}_{d'}$ and an input $x$, the masked forward pass sets:

$$\tilde{a}_j^{(l)}(x) = \begin{cases} 0 & \text{if } (l, j) \in \mathcal{S}_{d'} \\ a_j^{(l)}(x) & \text{otherwise} \end{cases} \tag{4}$$

for all layers $l$ and all token positions. This is implemented via registered forward hooks on the down-projection pre-activation without modifying model weights.

For language domains, we measure the NLL change under masking. The damage of masking domain $d'$'s shell on language $d$ is reported net of the random baseline to isolate shell-specific effects from general disruption:

$$\Delta_{d' \to d} = \underbrace{\text{NLL}_d(\mathcal{S}_{d'})}_{\text{shell mask}} - \underbrace{\text{NLL}_d(\mathcal{S}_{\text{rand}})}_{\text{random mask}} \tag{5}$$

This yields a $K \times K$ damage matrix where diagonal entries indicate how much masking domain $d$'s shell degrades domain $d$ beyond random, and off-diagonal entries indicate cross-domain interference. For subject domains, accuracy replaces NLL as the evaluation metric and the same net-of-random normalization applies.

### 3.5 Shell Overlap and Spatial Organization

To assess the disjointness of domain shells, we compute the Jaccard similarity (IoU) between the top-$m$ shell sets for each domain pair:

$$\text{IoU}(d, d') = \frac{|\mathcal{S}_d \cap \mathcal{S}_{d'}|}{|\mathcal{S}_d \cup \mathcal{S}_{d'}|} \tag{6}$$

To assess spatial organization, we record the layer-wise distribution of shell neurons by counting, for each layer $l$ and domain $d$, the number of neurons $(l, j) \in \mathcal{S}_d$.

To assess whether functional shell structure permits selective quantization, we apply per-group asymmetric INT4 fake quantization (Frantar et al., 2023) to the down-projection weight matrix $\mathbf{W}_{\text{down}}^{(l)}$ under three conditions: uniform quantization of all weight groups, the core, and the inverse. We define the *core* as $(\bigcup_d \mathcal{S}_d)^c$, the complement of the union of all domain shells. Group size is $g = 128$ following standard deployment practice (Lin et al., 2025). NLL is evaluated after weight replacement without any calibration data, isolating the native quantization compatibility of the weight configuration.

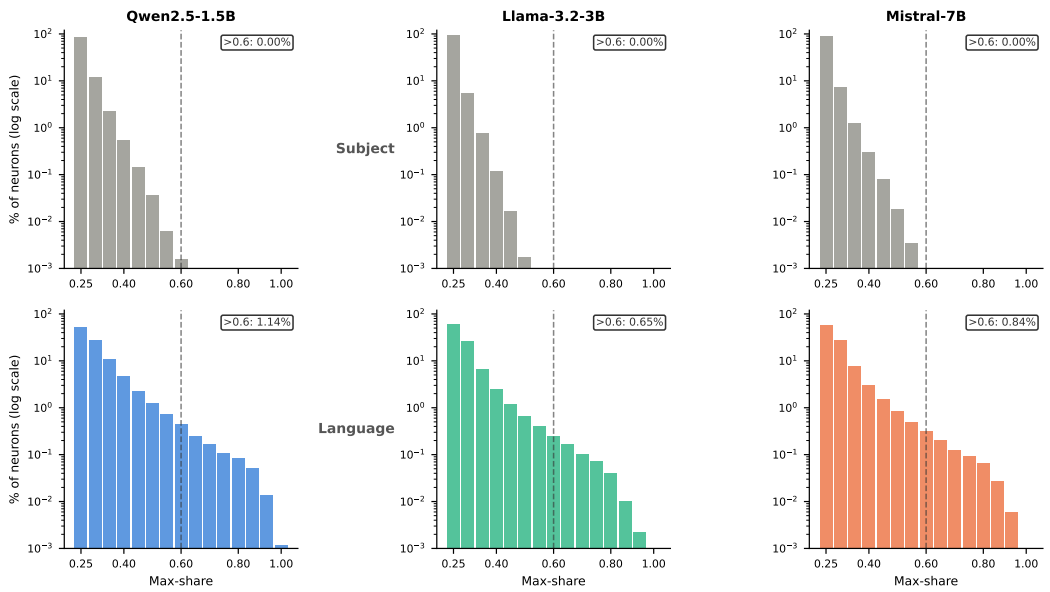


Figure 1: Selectivity histograms for subject-level domains (top row) and language/modality domains (bottom row) across three model families. The y-axis is log-scaled percentage of total FFN neurons. The dashed vertical line marks the 0.60 max-share threshold. At the subject level, no neurons exceed 0.60 across 939,008 combined neurons. At the language level, 0.65–1.14% exceed 0.60 in all three models.

# 4 Results

## 4.1 Subject-Level Domains: Zero Parametric Modularity

Figure 1 (top row) shows the selectivity histogram for subject-level domains across all three models. The distribution is unimodal and rapidly decaying in all cases: no neuron exceeds 0.60 max-share across a combined 939,008 FFN neurons, and the 99th percentile reaches at most 0.388.

Causal masking confirms functional indifference. The subject-level damage matrix has a diagonal mean of 0.121–0.181 accuracy points and an off-diagonal mean of 0.031–0.076, with no model exceeding a diagonal-to-off-diagonal ratio of 2.4. The elem_sci row is an exception whose off-diagonal entries reveal general disruption rather than a selective shell. These results hold despite domain identity being linearly decodable above 85% accuracy via a logistic regression probe, providing a clean dissociation: domain information is present in activation geometry but not supported by a dedicated parametric substrate.

## 4.2 Language-Level Domains: Concentrated Parametric Shells

Figure 1 (bottom row) shows the selectivity histogram for language and modality domains. The fraction of neurons exceeding 0.60 max-share is 1.14%, 0.65%, and 0.84% for Qwen2.5-1.5B, Llama-3.2-3B, and Mistral-7B respectively, compared to 0.00% at the subject level. Figure 2 shows the resulting damage matrices. Diagonal means are $\bar{\Delta}_{\text{diag}} = 1.82$, 2.19, and 3.37 nats; off-diagonal means are $-0.012$, $-0.060$, and $+0.006$ nats, with ratios from 37:1 to above 500:1. The largest single entry is the Chinese diagonal in Mistral-7B at $+8.58$ nats against an English off-diagonal of $+0.013$ nats under the same mask. Shell disjointness is near-complete: maximum IoU across all pairs is 0.0029 (Appendix C). Layer-wise, Spanish and Chinese neurons concentrate in final network layers while English and code distribute uniformly, consistent with instruction tuning consolidating non-English computation into the vocabulary projection stage (Appendix A).

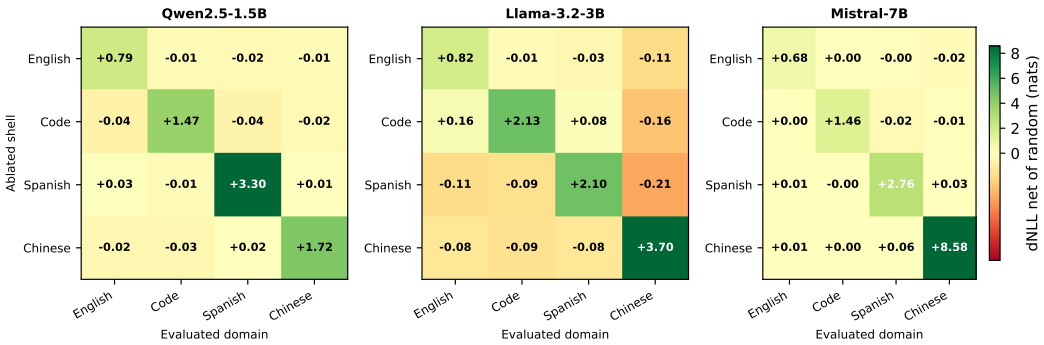


Figure 2: Language-level damage matrices (dNLL net of count-matched random baseline) for all three models at $\rho = 0.02$. Rows: ablated shell. Columns: evaluated domain. Diagonal means are 1.82, 2.19, and 3.37 nats; off-diagonal means are $-0.012$, $-0.060$, and $+0.006$ nats.

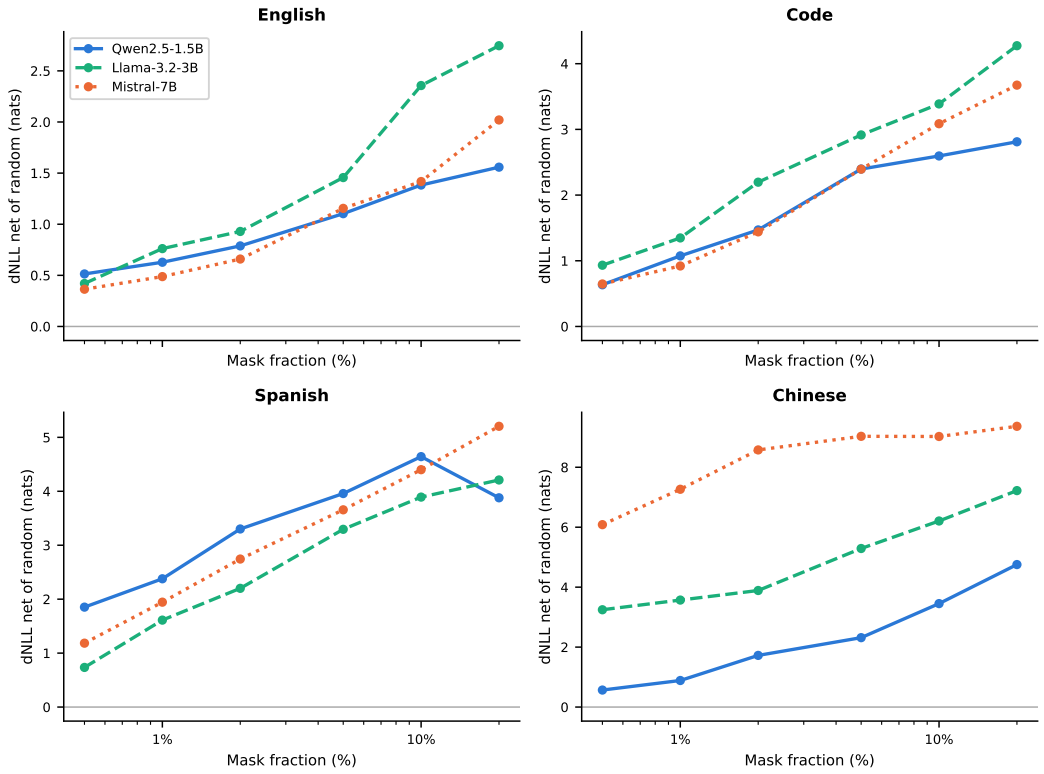


Figure 3: Diagonal dNLL net of random as a function of mask fraction $\rho$ for each language domain across all three models (log-scaled x-axis). Damage accumulates rapidly between $\rho = 0.005$ and $\rho = 0.02$ and saturates by $\rho = 0.05$ for all domains. The Chinese curve in Mistral-7B reaches $+6.09$ nats at $\rho = 0.005$ (2,293 neurons) and $+8.58$ nats at $\rho = 0.02$, representing 91.6% of the total damage observed at $\rho = 0.20$.

### 4.3 Pareto Analysis: Shell Concentration and Functional Load

Figure 3 shows damage accumulating rapidly between $\rho = 0.005$ and $\rho = 0.02$ and saturating by $\rho = 0.05$ for all language domains. At $\rho = 0.02$, damage represents on average 72% of the maximum observed; for Chinese in Mistral-7B, $\rho = 0.005$ (2,293 neurons) already produces $+6.09$ nats while the curve saturates at $+9.03$ nats by $\rho = 0.05$. At the subject level, damage curves are flat and undifferentiated from random throughout the $\rho$ range.

Table 1: GSM8K accuracy (%) under each masking condition. Net-of-random effects in parentheses. Bold: largest net effect per model.

| Condition | Qwen2.5-1.5B | Llama-3.2-3B | Mistral-7B |
|---|---|---|---|
| Baseline | 30.7 | 37.3 | 40.7 |
| Random mask | 36.0 | 29.3 | 34.0 |
| mask_english | 37.3 (+1.3) | 16.0 (−13.3) | 32.0 (−2.0) |
| mask_code | 18.7 (**−17.3**) | 5.3 (**−24.0**) | 18.0 (**−16.0**) |
| mask_spanish | 32.0 (−4.0) | 29.3 (0.0) | 30.7 (−3.3) |
| mask_chinese | 34.7 (−1.3) | 19.3 (−10.0) | 34.7 (+0.7) |

### 4.4 Code Shells Carry Mathematical Reasoning

Table 1 reports GSM8K accuracy under each masking condition net of the count-matched random baseline. Masking the code shell produces the largest and most consistent effect: $-17.3$, $-24.0$, and $-16.0$ percentage points for Qwen2.5-1.5B, Llama-3.2-3B, and Mistral-7B respectively. Mathematical reasoning, despite being expressed in English-language text, relies on the code-selective neuron population rather than the English-selective population. Masking Spanish produces effects at or below random in all three models. The Chinese result varies with pretraining composition: Mistral-7B shows near-zero effect ($+0.7$pp), Qwen2.5-1.5B shows $-1.3$pp, and Llama-3.2-3B shows $-10.0$pp, tracing a gradient that maps to presumed Chinese mathematical content during pretraining. Shell purity is a property of the training data, not of the language.

Chain-of-thought prompting partially recovers code shell damage in Mistral-7B: standard code masking costs 16 percentage points net of random while CoT masking costs only 4, recovering 75% of the damage. The code shell carries two functions: an answer-structuring function that CoT externalizes, and a core reasoning function that prompting cannot replace. Recovery fractions vary across models (14% for Qwen2.5-1.5B, $-15$% for Llama-3.2-3B), consistent with training differences in how these functions are distributed; we leave systematic characterization as future work.

### 4.5 Scale Analysis

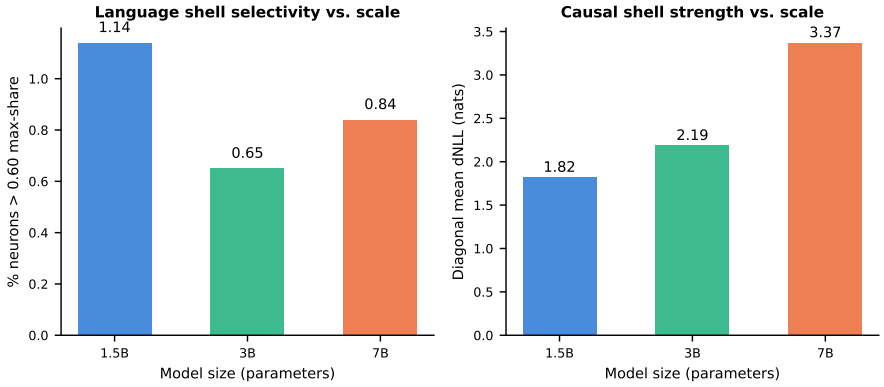


Figure 4: Language shell properties as a function of model scale. Left: fraction of neurons exceeding 0.60 max-share shows a non-monotonic relationship with parameter count. Right: diagonal mean dNLL at $\rho = 0.02$ increases monotonically with scale (1.82, 2.19, 3.37 nats), indicating that larger models develop more functionally load-bearing language shells.

Figure 4 shows language shell strength as a function of model scale. The fraction of neurons exceeding 0.60 max-share is non-monotonic (1.14%, 0.65%, 0.84%), reflecting differences in architecture and pretraining composition. In contrast, $\bar{\Delta}_{\text{diag}}$ increases monotonically: 1.82, 2.19, and 3.37 nats. At $\rho = 0.005$, Mistral-7B's Chinese shell produces 6.09 nats of targeted damage

versus 0.57 nats in Qwen2.5-1.5B. Larger models develop not more selective neurons proportionally, but more functionally load-bearing ones; absolute dNLL values reflect differences in tokenizer and architecture across families and should be interpreted as a directional finding.

### 4.6 Spatial Interleaving: Functional Shells Are Parametrically Dispersed

Under per-group asymmetric INT4 quantization with group size $g = 128$, a group qualifies for core-only quantization only if all 128 neurons fall below the selectivity threshold. At shell densities of 12–21%, the probability that any group of 128 consecutive neurons is entirely core evaluates to $2.2 \times 10^{-8}$ for the highest-density model. In practice, zero qualifying groups are found in any of the three models.

The zero qualifying groups directly implies that no weights are modified under either selective condition; int4_core and int4_inv are therefore identical to BF16 by construction. The informative finding is the uniform INT4 result: all languages degrade equally ($+0.028$–$+0.081$ nats), with no language showing systematically larger degradation than others, confirming that standard group-level quantization cannot differentially target language shells.

### 4.7 Shell Neurons Cannot Be Amplified

To test whether shell neurons can be exploited for performance gains — turning them up rather than zeroing them — we multiply the pre-activations of each domain shell by scalars $\alpha \in \{1.5, 2.0, 3.0, 5.0\}$ and measure the effect on GSM8K accuracy and domain NLL.

All scalar values $\alpha \in \{1.5, 2.0, 3.0, 5.0\}$ were applied to both shell-selective and count-matched random neuron conditions. At $\alpha = 1.5$, code shell amplification reduces GSM8K accuracy by 14 percentage points and increases code NLL by 0.10 nats. At $\alpha = 2.0$, the reduction reaches 30 percentage points. By $\alpha = 3.0$, accuracy collapses to zero and code NLL increases by 1.77 nats. No scalar produces improvement on any domain or model. Random amplification produces similar catastrophic effects at $\alpha = 5.0$, with large off-diagonal NLL increases across all languages, confirming that the damage is not shell-specific but reflects the residual stream being pushed outside the distribution of activation patterns the model was trained on.

Shell neurons are necessary but not sufficient: they operate at a precisely calibrated functional point, and downstream computation depends on this calibration rather than on absolute activation magnitude.

## 5 The Granularity Principle

The results converge on a single organizing principle: parametric shells form precisely where training data was modular at the token level, and our evidence across three model families and eight domains suggests they do not form otherwise.

**Why language domains produce shells.** English, Spanish, Chinese, and code occupy largely non-overlapping regions of the token distribution during pretraining. Gradient updates for Chinese inputs systematically push a subset of parameters toward configurations useful for Chinese processing while leaving parameters used for English largely undisturbed. Repeated across billions of tokens, this pressure produces a population of neurons concentrated on one partition of the training distribution. The shell IoU values below 0.003 are the parametric residue of this partitioned training signal.

**Why subject domains do not.** A question about cellular respiration and a question about differential equations are both written in English with similar syntactic structures and overlapping academic vocabulary. The gradient signal for biology questions and physics questions updates largely overlapping parameter sets, producing no systematic pressure toward domain-selective neurons. The 88% probe accuracy confirms domain information is present in activation geometry, but representational and parametric organization need not co-occur.

**Shell purity tracks training data composition.** Mistral-7B's Chinese shell carries essentially no mathematical reasoning capability ($+0.7$pp on GSM8K net of random), while Llama-3.2-3B's shows $-10.0$pp, and Qwen2.5-1.5B's shows $-1.3$pp. This gradient maps directly to the presumed gradient in Chinese mathematical content during pretraining. A shell learned on purely linguistic content carries only linguistic function; a shell learned on thematically diverse content carries a corresponding mixture. Shell purity is a property of the training data, not of the domain.

## 6 Discussion

**Implications for multilingual model design.** Parametric separability is a direct consequence of whether training data partitions by token distribution, not an emergent property of scale. Deliberate curation of partition sharpness across target languages, rather than simply maximizing token counts, may produce more predictable behavior under language-targeted interventions.

**Implications for editing and compression.** For language domains, the concentrated and disjoint shell structure makes neuron-level editing feasible: the Chinese shell in Mistral-7B is identifiable, functionally load-bearing, and nearly orthogonal to all other shells. For subject domains, no analogous structure exists. For compression, spatial interleaving means functional modularity cannot be directly converted into memory efficiency under standard group-level quantization; per-neuron precision assignment and weight-matrix reordering that clusters shell neurons contiguously are weight-equivalent directions for future work.

**Relationship to prior work.** Our results extend Tang et al. (2024) and Le and Li (2026) in three respects: causal effect magnitudes are substantially larger (LAPE produces sub-unit perplexity changes; our approach identifies neurons whose removal is functionally catastrophic); the subject-level null makes the language-level result interpretable as a boundary condition; and the final-layer concentration we observe in instruction-tuned models differs from the U-shaped distribution reported for base models (Tang et al., 2024), consistent with instruction tuning consolidating language-specific computation into the vocabulary projection stage.

**Limitations.** Our analysis covers three instruction-tuned families in the 1.5B–7B range; whether the granularity principle holds for base models, larger models, or different architectures is not established. Our activation-magnitude identification method may not surface all functionally relevant neurons; gradient or relevance-based approaches may identify complementary populations. The subject-level null covers four coarse categories; finer-grained partitions might reveal structure at intermediate granularities.

## 7 Conclusion

We investigated whether large language models contain domain-specific parametric shells: concentrated, causally necessary neuron populations whose removal selectively degrades a target domain while sparing others. Across two domain granularities, three model families, and eight domains, the answer depends entirely on granularity. At the subject level, zero neurons exceed 60% domain selectivity across 939,008 combined FFN neurons and causal damage matrices are flat, despite domain identity being linearly decodable above 85% accuracy. At the language and modality level, 0.65–1.14% of neurons exceed 60% selectivity, damage matrices are near-perfectly diagonal with ratios up to 595:1, shell neuron sets are essentially disjoint (IoU $< 0.003$), and masking code-selective neurons reduces mathematical reasoning accuracy by 16–24 percentage points while masking Spanish or Chinese neurons leaves it at or below random.

These results are unified by the granularity principle: parametric shells form precisely where training data was modular at the token level, and our evidence across three model families and eight domains suggests they do not form otherwise. The model carves its parameters the same way its training data was carved by script, vocabulary, and syntactic register.

## AI USE STATEMENT

This manuscript was prepared with the assistance of large language model tools for drafting and editing. All experimental design, data collection, analysis, and scientific claims are the work of the authors. All AI-assisted text was reviewed, edited, and verified by the authors.

## ETHICS STATEMENT

This work analyzes publicly available open-weight language models (Qwen2.5-1.5B-Instruct, Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3) using publicly available benchmark datasets (MMLU, ARC-Challenge, OpenBookQA, GSM8K, WikiText-2, MBPP, OPUS-100). No human subjects were involved. No personal data was collected or used. Code will be released upon acceptance.

## REPRODUCIBILITY STATEMENT

All hyperparameters are documented in Table 4 (Appendix E). Activation capture and causal masking implementations are described in detail in Appendix E. All experiments use a fixed random seed (42). Full model identifiers, dataset splits, and sampling procedures are specified in Section 3 and Appendix E. Code will be released as supplementary material.

## REFERENCES

J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, and C. Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

A. Bau, Y. Belinkov, H. Sajjad, N. Durrani, F. Dalvi, and J. Glass. Identifying and controlling important neurons in neural machine translation, 2018. URL https://arxiv.org/abs/1811.01157.

L. Chen, A. Dejl, and F. Toni. Identifying query-relevant neurons in large language models for long-form texts. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i22.34529. URL https://doi.org/10.1609/aaai.v39i22.34529.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, and F. Wei. Knowledge neurons in pretrained transformers. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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/.

E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2023. URL https://arxiv.org/abs/2210.17323.

A. Grattafiori, A. Dubey, A. Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

D. Gurgurov, K. Trinley, Y. Al Ghussin, T. Baeumel, J. van Genabith, and S. Ostermann. Language arithmetics: Towards systematic language neuron identification and manipulation. In K. Inui, S. Sakti, H. Wang, D. F. Wong, P. Bhattacharyya, B. Banerjee, A. Ekbal, T. Chakraborty, and D. P. Singh, editors, *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2911–2937, Mumbai, India, Dec. 2025a. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.156. URL https://aclanthology.org/2025.ijcnlp-long.156/.

D. Gurgurov et al. Targeted fine-tuning of language-specific neurons in multilingual language models. *arXiv preprint*, 2025b.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

J. Huo, Y. Yan, B. Hu, Y. Yue, and X. Hu. MMNeuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6816, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.387. URL https://aclanthology.org/2024.emnlp-main.387/.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

T. Kojima et al. Language-specific neurons in multilingual models. *arXiv preprint*, 2024.

Y. Le and Y. Li. Crane: Causal relevance analysis of language-specific neurons in multilingual large language models, 2026. URL https://arxiv.org/abs/2601.04664.

J. Lin, J. Tang, H. Tang, S. Yang, G. Xiao, and S. Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *GetMobile: Mobile Comp. and Comm.*, 28(4): 12–17, Jan. 2025. ISSN 2375-0529. doi: 10.1145/3714983.3714987. URL https://doi.org/10.1145/3714983.3714987.

S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843.

T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

A. Mueller et al. Coloring the black box: What syntactic probes learn. *arXiv preprint*, 2022.

Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

R. Song, S. He, S. Jiang, Y. Xian, S. Gao, K. Liu, and Z. Yu. Does large language model contain task-specific neurons? In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7101–7113, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.403. URL https://aclanthology.org/2024.emnlp-main.403/.

M. Sun, X. Chen, J. Z. Kolter, and Z. Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

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

H. Tezuka and N. Inoue. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual llms, 2025. URL https://arxiv.org/abs/2509.17030.

J. Tiedemann and S. Thottingal. OPUS-MT – building open translation services for the world. In A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, and M. L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, Nov. 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.61/.

E. Voita, J. Ferrando, and C. Nalmpantis. Neurons in large language models: Dead, n-gram, positional. pages 1288–1301, 01 2024. doi: 10.18653/v1/2024.findings-acl.75.

X. Wang, K. Wen, Z. Zhang, L. Hou, Z. Liu, and J. Li. Finding skill neurons in pre-trained transformer-based language models. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.765. URL https://aclanthology.org/2022.emnlp-main.765/.

C. Wendler, V. Veselovsky, G. Monea, and R. West. Do llamas work in English? on the latent language of multilingual transformers. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

## A Layer-wise Shell Distributions

## B Subject-Level Damage Matrices

Figure 6 shows the subject-level damage matrices (net of count-matched random baseline, measured in accuracy points) for all three models. In contrast to the language-level matrices in Figure 2, no model produces a consistently diagonal structure. For the quant, biomed, and humsoc rows, diagonal and off-diagonal entries are of comparable magnitude across all three models, consistent with the interpretation that masking any of these subject shells disrupts general question-answering capability rather than a domain-specific parametric substrate.

The elem_sci row is an exception, showing larger values across all columns rather than a concentrated diagonal entry. This pattern reflects the composition of the elementary science

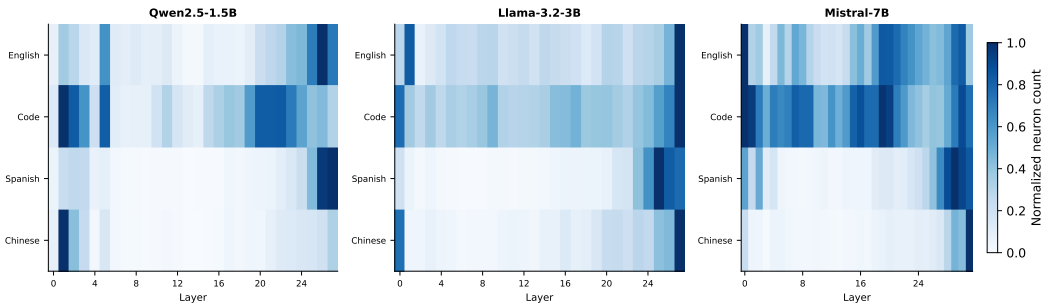


Figure 5: Layer-wise distribution of language shell neurons (normalized by the maximum count across layers per domain) for each model. Spanish and Chinese neurons concentrate strongly in final layers across all three model families. English and code neurons are distributed more uniformly. This final-layer concentration for non-English domains differs from the U-shaped distribution reported for base models (Tang et al., 2024), consistent with instruction tuning absorbing input-side language mapping into the shared representation.

evaluation set (ARC-Challenge and OpenBookQA), which draws on general English vocabulary and commonsense reasoning that overlaps substantially with all other subject domains. Masking the elem_sci shell therefore degrades all domains, a signature of disrupting shared linguistic processing rather than a domain-specific shell. This interpretation is supported by the subject-level selectivity histogram (Figure 1, top row): the elem_sci shell contains no neurons with max-share above 0.60 and therefore cannot constitute a parametrically selective substrate for elementary science in particular.

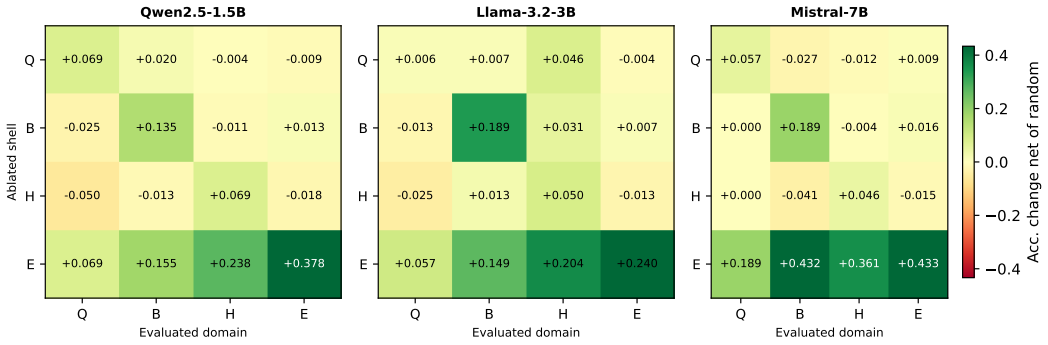


Figure 6: Subject-level damage matrices (accuracy points net of random baseline) for all three models. Rows correspond to the ablated shell; columns to the evaluated subject domain. Unlike the language-level matrices, no model produces a consistently diagonal structure. The elem_sci row shows large off-diagonal values consistent with disruption of shared general question-answering capability rather than a domain-selective shell. Domain labels: Q = quantitative, B = biomedical, H = humanities and social sciences, E = elementary science.

## C Shell Overlap Matrices

Tables 2 and 3 report the Jaccard similarity (IoU) between top-2% shell sets for all domain pairs across all three models.

Language shell overlap is near-zero across all pairs and all models (Table 2). The maximum observed IoU is 0.0029 (Mistral-7B, Spanish and Chinese), and the majority of pairs report IoU below 0.001. The slight elevation in the Spanish–Chinese pair relative to other pairs is consistent with the observation that Spanish and Chinese both require final-layer vocabulary mapping neurons that partially overlap in the intermediate dimension, though the overlap remains negligible in absolute terms.

Subject shell overlap is similarly near-zero across most pairs (Table 3), with the exception of the biomed–elem_sci pair, which reaches 0.0115–0.0162 across models. This modest elevation reflects the shared biological science content between biomedical MMLU questions and ARC elementary science questions. Even at its maximum, the overlap across all pairs and all models remains below 1.7%, confirming that the absence of functional shells at the subject level is not attributable to excessive overlap between shell definitions.

Table 2: Jaccard similarity (IoU) between top-2% language shell sets for each domain pair and model. Values below 0.003 across all pairs and all models confirm near-complete shell disjointness.

| Domain pair | Qwen2.5-1.5B | Llama-3.2-3B | Mistral-7B |
|---|---|---|---|
| English – Code | 0.0001 | 0.0015 | 0.0001 |
| English – Spanish | 0.0000 | 0.0004 | 0.0001 |
| English – Chinese | 0.0001 | 0.0005 | 0.0001 |
| Code – Spanish | 0.0000 | 0.0001 | 0.0000 |
| Code – Chinese | 0.0000 | 0.0004 | 0.0001 |
| Spanish – Chinese | 0.0000 | 0.0019 | 0.0029 |

Table 3: Jaccard similarity (IoU) between top-2% subject shell sets for each domain pair and model. The biomed–elem_sci elevation reflects shared biological science content between MMLU biomedical questions and ARC elementary science items.

| Domain pair | Qwen2.5-1.5B | Llama-3.2-3B | Mistral-7B |
|---|---|---|---|
| Quant – Biomed | 0.0002 | 0.0009 | 0.0010 |
| Quant – Humsoc | 0.0000 | 0.0015 | 0.0005 |
| Quant – Elem_sci | 0.0005 | 0.0009 | 0.0014 |
| Biomed – Humsoc | 0.0003 | 0.0012 | 0.0013 |
| Biomed – Elem_sci | 0.0115 | 0.0136 | 0.0162 |
| Humsoc – Elem_sci | 0.0006 | 0.0012 | 0.0009 |

## D Per-Model Pareto Curves

Figure 7 shows the Pareto curves reorganized by model rather than by language domain, allowing direct comparison of shell concentration across domains within each model. Within each model, Chinese and Spanish consistently show the steepest initial curves, saturating earliest. English shows the shallowest curve in all three models, consistent with English capability being distributed throughout the network rather than concentrated in a dedicated shell. The ordering of domain steepness is consistent across all three models despite differences in absolute dNLL magnitude, supporting the interpretation that shell concentration is a property of the domain partition rather than of any particular model architecture.

## E Experimental Details

**Models.** All models are loaded in bfloat16 precision via HuggingFace Transformers (Wolf et al., 2020) with `device_map=auto`. Full model identifiers are `Qwen/Qwen2.5-1.5B-Instruct`, `meta-llama/Llama-3.2-3B-Instruct`, and `mistralai/Mistral-7B-Instruct-v0.3`.

**Hardware.** All experiments were conducted on NVIDIA A100 80GB GPUs via Google Colab Pro. Qwen2.5-1.5B and Llama-3.2-3B experiments are compatible with a single T4 16GB GPU. Mistral-7B requires at minimum an A100 40GB GPU in bfloat16 precision.

**Dataset construction.** Subject domain evaluation pools are constructed from MMLU validation splits (600 items total, stratified across all available subjects with a minimum of 1 item per subject),

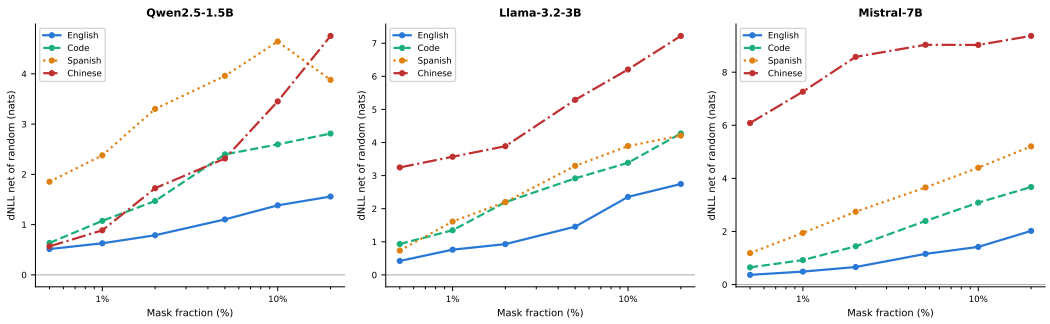


Figure 7: Diagonal dNLL net of random as a function of mask fraction $\rho$, organized by model. Each panel shows all four language domains for one model. Chinese and Spanish curves are consistently steeper and saturate earlier than English and code curves within each model, regardless of architecture or scale.

ARC-Challenge test split (225 items, randomly sampled), and OpenBookQA test split (225 items, randomly sampled). Subject assignment follows the taxonomy in Section 3, with all sampling using a fixed random seed. Language domain corpora use 96 segments of 512 tokens each, drawn from WikiText-2 test split (English), MBPP test split (code), and OPUS-100 training splits (Spanish, Chinese) with streaming access. GSM8K evaluation uses 150 randomly sampled test items with greedy decoding and a maximum of 256 new tokens.

Table 4 summarizes all primary hyperparameters used across experiments.

Table 4: Experimental hyperparameters.

| Parameter | Value |
|---|---|
| Model precision | bfloat16 |
| Activation capture depth | layer $\lfloor 0.75 \cdot L \rfloor$ where $L$ = total layers |
| Norm exclusion threshold | bottom 2% by $\ell_2$ norm |
| Backbone exclusion | top 0.1% by global mean activation |
| Primary mask fraction $\rho$ | 0.02 |
| Pareto fractions | $\{0.005, 0.01, 0.02, 0.05, 0.10, 0.20\}$ |
| INT4 group size | 128 |
| INT4 calibration data | none (calibration-free) |
| Max input length | 1024 tokens |
| GSM8K decoding | greedy, max 256 new tokens |
| Random seed | 42 |
| Subject pool size | 1017 items (600 MMLU + 225 ARC + 225 OBQA) |
| Language segments | 96 $\times$ 512 tokens per domain |
| GSM8K pool size | 150 items |

**Activation capture implementation.** Neuron activations are captured via PyTorch forward hooks registered on the `down_proj` pre-activation at each FFN layer. For subject domains, hooks are applied to question-only forwards without answer choices or the `Answer:` scaffold token, and activations are mean-pooled across all non-padding token positions. For language domains, hooks are applied to teacher-forced forwards over fixed 512-token segments. Hook registration and removal are performed immediately before and after each forward pass to prevent cross-example contamination. All captures accumulate into float64 accumulators to prevent precision loss during averaging over large item counts.

**Causal masking implementation.** Neuron masking is implemented via the same forward hook mechanism, zeroing the pre-activation of selected neurons at all token positions across all forward passes in the evaluation set. Masks are defined once per condition and reused across all items; no per-item mask computation occurs at evaluation time. The random baseline mask is sampled once

per experiment with a fixed random seed and held constant across all model and domain comparisons to ensure that random baseline variation does not confound cross-condition comparisons.