DV-LENS: REVEALING THE FUNCTIONAL ORGANIZATION OF LANGUAGE MODEL PARAMETERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the functional roles of learnable parameters in model computation is important for elucidating the internal mechanisms of large language models (LLMs). However, how to connect parameters from different modules to verifiable output effects and further characterize the relationship between their functional organization and model capability remains to be explored. To this end, we introduce the downstream vocabulary lens (DV-Lens), a parameter-level interpretability framework that links native parameter directions to their downstream vocabulary responses. Specifically, we first estimate module-specific downstream Jacobians over a reference prompt set for attention query, key, value, and output (Q/K/V/O) projections and feed-forward networks (FFNs). Second, we use these mappings to project native parameter columns into the final vocabulary space, obtaining signed readouts that characterize their average local output responses. Third, we group parameter columns by their vocabulary readouts and introduce downstream vocabulary complexity (DV-Complexity), which quantifies within-group structural variation using normalized reconstruction residuals of the original weights. At the parameter level, randomized controls and finite-difference tests show that DV-Lens readouts capture non-random vocabulary structure and predict local logit changes with 98.0% coordinate-orientation agreement across 720 cases from nine models. These readouts further guide parameter ablation, steering, and swapping across 21 models, shifting target-token probabilities in the predicted directions under controlled conditions. At the model level, the joint-parameter score of DV-Complexity achieves a Spearman correlation of 0.904 with benchmark-based capability rankings across 48 language models. Together, these results provide intervention-based evidence for DV-Lens interpretations and reveal an association between DV-Complexity and model capability.

1 INTRODUCTION

Large language models (LLMs) excel at language understanding, generation, and reasoning (Suzgun et al., 2022; Wang et al., 2024), yet how their learned parameters support these capabilities remains unclear (Belrose et al., 2023; Ghandeharioun et al., 2024). Understanding parameter functions requires examining how directions in different Transformer modules influence outputs through downstream computation (Geva et al., 2021; Katz et al., 2024). Beyond individual parameters, their functional organization may reveal structural properties associated with model capability (Eilertsen et al., 2020; Dravid et al., 2024). This motivates two connected questions: how can we characterize and verify the downstream effects of native parameter directions, and how is the resulting parameter organization related to model capability?

Existing interpretability research provides different views of these questions. Activation-based approaches analyze input-dependent representations using vocabulary readouts, contextual probing, and sparse feature decompositions (Belrose et al., 2023; Ghandeharioun et al., 2024; Huben et al., 2024; Dunefsky et al., 2024), while activation patching tests their behavioral roles through interventions (Zhang & Nanda, 2023). Parameter-oriented studies interpret feed-forward weights as key-value memories and investigate the parameters involved in factual associations (Geva et al., 2021; Meng et al., 2022). For example, WeightLens (Golimblevskaia et al., 2026) derives feature descriptions from learned transcoder weights, whereas ROTATE (Avrahamy et al., 2026) decomposes FFN weights into vocabulary-aligned directions. These advances demonstrate that vocabulary readouts can

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.

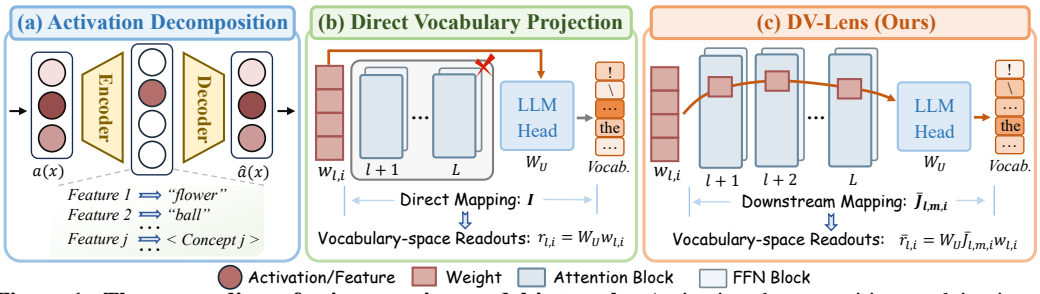


Figure 1: Three paradigms for interpreting model internals. Activation decomposition explains input-dependent activations through learned features; direct vocabulary projection maps a parameter direction $w_{\ell,m,i}$ into vocabulary space through the LLM head alone; DV-LENS first propagates the direction through the downstream mapping $\bar{J}_{\ell,m}$ and then reads it out in vocabulary space.

expose meaningful structure in model internals. However, they usually interpret parameters without considering how their effects pass through later computations to the model output. This motivates us to study how semantic information is organized across parameter dimensions and whether this organization is related to model capability.

To address these questions, we introduce the downstream vocabulary lens (DV-LENS), which uses native parameter columns as the unit of analysis. For attention Q/K/V/O projections and FFN modules, we estimate downstream Jacobian mappings over a reference prompt set and propagate parameter directions into the final vocabulary space. The resulting signed readouts characterize the average local logit response along each direction, placing parameters from different modules in a common output representation; Figure 1 contrasts this with activation decomposition and direct vocabulary projection. Randomized controls and finite-difference tests assess these readouts, while parameter ablation, steering, and source-to-target swapping across 21 models provide behavioral evidence for predicted changes in target-token probabilities under controlled settings. To extend the analysis from individual parameter functions to parameter organization, we further introduce downstream vocabulary complexity (DV-COMPLEXITY). This metric groups parameters by their normalized vocabulary readouts and measures the normalized reconstruction residual of the original weights within these groups, quantifying structural variation not captured by coarse functional roles. Across 48 language models, its joint-parameter score achieves a Spearman correlation of 0.904 with benchmark-based capability rankings. The main contributions are summarized as follows:

- We introduce DV-LENS, a parameter-level interpretability framework that uses module-specific downstream mappings to characterize native parameter directions across attention and FFN modules in a common vocabulary space.
- We provide complementary statistical and behavioral evidence for DV-LENS through matched random controls, finite-difference tests, and direct parameter interventions, linking its readouts to observed output changes under controlled settings.
- We introduce DV-COMPLEXITY to quantify structural variation within parameter groups defined by downstream readouts, and show that the resulting model-level score is strongly associated with benchmark-based capability rankings across 48 language models.

2 RELATED WORK

Interpretability Methods for Language Models. Existing interpretability methods study internal model structure from several complementary perspectives, including decoding intermediate representations, intervening on activations, and identifying structured features or directions. Tuned Lens and Patchscopes interpret intermediate representations through learned decoders or alternative model computations (Belrose et al., 2023; Ghandeharioun et al., 2024), while activation patching measures the behavioral effect of replacing internal activations (Zhang & Nanda, 2023). Sparse autoencoders, transcoders, ICA-based methods, and analyses of verbalizable representations characterize internal model structure through sparse features, distinctive directions, and functionally meaningful representations (Huben et al., 2024; Dunefsky et al., 2024; Liu & Han, 2026; Gurnee et al., 2026). In contrast, DV-LENS focuses directly on LLMs' parameter columns and characterizes their functional roles through module-specific mappings and prompt-averaged responses.

Parameter Semantics, Localization, and Editing. Parameter-oriented studies examine the information represented in learned weights and its relationship to model behavior. Feed-forward layers have been interpreted as key-value memories, where input-dependent coefficients combine value directions that admit vocabulary projections (Geva et al., 2021). Backward Lens projects gradients into vocabulary space to inspect how information is written into model parameters during learning (Katz et al., 2024). WeightLens derives feature descriptions from learned transcoder weights, whereas ROTATE extracts vocabulary-aligned directions from feed-forward weights without forward passes (Golimblevskaia et al., 2026; Avrahamy et al., 2026). Causal tracing localizes states that mediate factual predictions, and Rank-One Model Editing (ROME) uses this localization to perform rank-one edits to feed-forward weights (Meng et al., 2022). DV-LENS tests predicted output changes through direct interventions on the identified parameter signals.

3 DOWNSTREAM VOCABULARY LENS

This section presents Downstream Vocabulary Lens (DV-LENS), which traces learned parameter directions to model outputs. We derive vocabulary readouts and use them to guide ablation, steering, and source-to-target swapping. We then introduce DV-COMPLEXITY to summarize parameter organization, examine its relation to capability, and develop its theoretical interpretation.

3.1 MOTIVATION: TRACING PARAMETER EFFECTS TO THE OUTPUT

We begin with a local calculation that motivates DV-LENS. Given an input x, a Transformer module m at layer ℓ, and a learned parameter column $w_{\ell,m,i}$, we ask how increasing the activation associated with this parameter changes the model's final output logits.

Learned Parameters and Input-Dependent Activations. Consider a linear projection $h_{\ell,m}(x) = W_{\ell,m} f_{\ell,m}(x)$, where $W_{\ell,m} = [w_{\ell,m,1}, \ldots, w_{\ell,m,n_m}]$, $f_{\ell,m}(x)$ is the input activation, and n_m is its width. The module output can be written as

$$h_{\ell,m}(x) = \sum_{i=1}^{n_m} w_{\ell,m,i} f_{\ell,m,i}(x). \tag{1}$$

Thus, $w_{\ell,m,i}$ specifies the direction written by parameter column i, while $f_{\ell,m,i}(x)$ determines how strongly that direction participates in the current computation. This columnwise view applies uniformly to FFN and attention projection matrices.

Propagation through Subsequent Model Computation. Now perturb the corresponding activation by a small amount, $f_{\ell,m,i}(x) \to f_{\ell,m,i}(x) + \epsilon$. Equation 1 gives $h_{\ell,m}(x;\epsilon) = h_{\ell,m}(x) + \epsilon w_{\ell,m,i}$. Let $z(x) = F_{\ell,m\to L}(h_{\ell,m}(x); x)$ denote the final normalized representation before the LLM head, where $F_{\ell,m\to L}$ contains the remaining computation after module (ℓ, m) and L denotes the final layer. We define the downstream Jacobian as

$$J_{\ell,m\to L}(x) = \frac{\partial z(x)}{\partial h_{\ell,m}(x)}. \tag{2}$$

By the chain rule, we have $\frac{\partial z(x)}{\partial f_{\ell,m,i}} = J_{\ell,m\to L}(x) w_{\ell,m,i}$.

With $\text{logits}(x) = W_U z(x) + b_U$, where W_U is the unembedding matrix and b_U is the output bias, the logit response is

$$\frac{\partial \text{ logits}(x)}{\partial f_{\ell,m,i}} = W_U J_{\ell,m\to L}(x) w_{\ell,m,i}. \tag{3}$$

Equivalently, for a small perturbation ϵ,

$$\text{logits}(x;\epsilon) - \text{logits}(x;0) = \epsilon W_U J_{\ell,m\to L}(x) w_{\ell,m,i} + O(\epsilon^2). \tag{4}$$

Define the prompt-local unit-response readout $r_{\ell,m,i}(x) = W_U J_{\ell,m\to L}(x) w_{\ell,m,i}$. Equation 3 characterizes the response per unit change of $f_{\ell,m,i}(x)$. The first-order contribution already present in the prompt computation is instead $c_{\ell,m,i}(x) = f_{\ell,m,i}(x) r_{\ell,m,i}(x)$. The activation coefficient sets the vocabulary readout's orientation and scale; changing its sign reverses the coordinate ordering.

3.2 DOWNSTREAM VOCABULARY LENS

The local response in Equation 3 depends on the computation induced by a particular input x. DV-LENS turns this input-dependent response into a common parameter representation by estimating the downstream mapping over a fixed prompt set $\mathcal{D}$.

Estimating the Downstream Mapping. For each module (ℓ, m), the mean downstream Jacobian is

$$\bar{J}_{\ell,m} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} J_{\ell,m \to L}(x). \tag{5}$$

For sequence-valued module outputs, the Jacobian is first aggregated over the fixed set of valid prediction positions and then averaged equally across prompts. This gives a common downstream mapping from the output of module (ℓ, m) to the final model representation.

DV-LENS Readout. The DV-LENS *readout* of column $w_{\ell,m,i}$ is

$$\bar{r}_{\ell,m,i} = W_U \bar{J}_{\ell,m} w_{\ell,m,i} \in \mathbb{R}^{|\mathcal{V}|}. \tag{6}$$

Here $\mathcal{V}$ is the output vocabulary. The entries of $\bar{r}_{\ell,m,i}$ describe the prompt-set average local response to a unit change of the corresponding activation coefficient. Their magnitudes identify vocabulary coordinates most affected along this parameter direction. The averaged readout is a parameter footprint, not the activation-weighted contribution realized on any individual prompt. Vocabulary-space readouts have been widely used to interpret intermediate representations and model parameters (Belrose et al., 2023; Geva et al., 2021; Katz et al., 2024; Golimblevskaia et al., 2026; Avrahamy et al., 2026). Unlike a direct projection of an intermediate representation or weight, DV-LENS first propagates the parameter through the subsequent model computation via $\bar{J}_{\ell,m}$ and only then reads its effect in the final output vocabulary.

Readouts across a Parameter Matrix. For an entire weight matrix $W_{\ell,m} = [w_{\ell,m,1}, \ldots, w_{\ell,m,n_m}]$, the DV-LENS readouts of all parameter columns can be written jointly as

$$R_{\ell,m} = W_U \bar{J}_{\ell,m} W_{\ell,m} \in \mathbb{R}^{|\mathcal{V}| \times n_m}. \tag{7}$$

Its i-th column is exactly the DV-LENS readout of $w_{\ell,m,i}$, $R_{\ell,m}[:, i] = \bar{r}_{\ell,m,i}$. We estimate $\bar{J}_{\ell,m}$ once per module using automatic differentiation and read out all columns jointly as $W_U \bar{J}_{\ell,m} W_{\ell,m}$. We apply the same procedure independently to each eligible Q/K/V/O and FFN weight matrix, placing their learned parameters in a common representation space defined by the final output vocabulary.

3.3 PARAMETER SIGNALS FOR TARGETED TOKEN CONTROL

The vocabulary readout maps learned parameters to output tokens. For a target concept, we identify its associated parameter signals and use them to control the corresponding tokens.

Representing a Concept in the Output Vocabulary. Let $\mathcal{T}_c$ denote the token set associated with concept c. For intervention analysis, we incorporate the learned final-normalization scale into the unembedding as $\widetilde{W}_U = W_U \Gamma$, where Γ is its diagonal scale matrix. We represent concept c by $u_c = |\mathcal{T}_c|^{-1} \sum_{v \in \mathcal{T}_c} \widetilde{W}_U[v, :]^\top$. For C concepts, we distinguish each from the others using the one-versus-rest readout

$$s_c = u_c - \frac{1}{C-1} \sum_{c' \neq c} u_{c'}. \tag{8}$$

These readouts satisfy $\sum_{c=1}^{C} s_c = 0$ and therefore span at most $C - 1$ dimensions.

Mapping Output Concepts to Learned Parameters. Using a fixed prompt set $\mathcal{D}_J$ for Jacobian estimation, let $\bar{J}_{\ell,m} = |\mathcal{D}_J|^{-1} \sum_{x \in \mathcal{D}_J} J_{\ell,m \to L}(x)$. We map concept c to module (ℓ, m) through

$$d_{\ell,m,c} = \bar{J}_{\ell,m}^\top s_c \in \mathbb{R}^{d_m}, \tag{9}$$

which we call the *parameter signal* for concept c; d_m is the module output width. For a parameter column $w_{\ell,m,i}$, its association with this concept is $\beta_{\ell,m,i,c} = d_{\ell,m,c}^\top w_{\ell,m,i} = s_c^\top \bar{J}_{\ell,m} w_{\ell,m,i}$. The last expression has the same form as the DV-LENS readout: it traces a learned parameter through subsequent model computation and evaluates its effect with respect to the target concept.

Ablation and Steering. Let $\widehat{d} = d/\|d\|_2$ denote the unit parameter signal and λ the intervention strength (denoted α in the experiments). For a source concept c, we reduce its associated signal in each selected parameter as $w'_{\ell,m,i} = w_{\ell,m,i} - \lambda \widehat{d}_{\ell,m,c} \left(\widehat{d}^{\top}_{\ell,m,c} w_{\ell,m,i} \right)$. For a target concept, we increase its associated signal as $w'_{\ell,m,i} = w_{\ell,m,i} + \lambda \|w_{\ell,m,i}\|_2 \widehat{d}_{\ell,m,c}$.

Source-to-Target Swapping. Swapping combines source removal and target addition. We select the source-removal set $\mathcal{S}_{\mathrm{src}}$ and target-addition set $\mathcal{S}_{\mathrm{tgt}}$ independently and apply

$$w'_{\ell,m,k} = \begin{cases} w_{\ell,m,k} - \lambda_{\mathrm{src}} \widehat{d}_{\ell,m,\mathrm{src}} \left(\widehat{d}^{\top}_{\ell,m,\mathrm{src}} w_{\ell,m,k} \right), & k \in \mathcal{S}_{\mathrm{src}}, \\ w_{\ell,m,k} + \lambda_{\mathrm{tgt}} \|w_{\ell,m,k}\|_2 \widehat{d}_{\ell,m,\mathrm{tgt}}, & k \in \mathcal{S}_{\mathrm{tgt}}. \end{cases} \tag{10}$$

3.4 DV-Complexity and Model Capability

Beyond individual parameters, we test whether DV-Lens readouts organize learned weights in a way that relates to model capability. DV-Complexity groups parameters by their readouts and measures the corresponding structure in the original weights.

DV-Lens-Guided Parameter Grouping. For a weight matrix $a = (\ell, m)$, write $W_a = [w_{a,1}, \ldots, w_{a,n_a}]$ and let $R_a = W_U J_a W_a$ denote its DV-Lens readout matrix. Its i-th column $r_{a,i} = R_a[:, i]$ is the vocabulary readout associated with parameter $w_{a,i}$. We normalize each readout as $\widetilde{r}_{a,i} = r_{a,i}/(\|r_{a,i}\|_2 + \varepsilon)$, where $\varepsilon > 0$ stabilizes normalization. With k clusters and readout centroids $\mu_{a,c}$, the assignments are

$$c^{\mathrm{DVL}}_{a,i} = \arg \min_{c \in \{1,\ldots,k\}} \|\widetilde{r}_{a,i} - \mu_{a,c}\|_2^2, \qquad k = 32. \tag{11}$$

Parameter Organization in Learned Weights. The cluster assignments are determined by DV-Lens readouts, while their organization is evaluated in the original learned parameter space. For each cluster c, let $\bar{w}_{a,c} = |\{i : c^{\mathrm{DVL}}_{a,i} = c\}|^{-1} \sum_{i: c^{\mathrm{DVL}}_{a,i} = c} w_{a,i}$. We define the matrix-level reconstruction error as

$$E_a^{\mathrm{DVL} \to \mathrm{raw}} = \frac{1}{\sum_{j=1}^{n_a} \|w_{a,j}\|_2^2} \sum_{i=1}^{n_a} \left\| w_{a,i} - \bar{w}_{a, c^{\mathrm{DVL}}_{a,i}} \right\|_2^2. \tag{12}$$

Smaller values indicate tighter grouping of learned weights with similar DV-Lens readouts.

Model-Level Parameter Organization. For model M and module scope $\mathcal{S}$, let $\mathcal{A}_{\mathcal{S}}(M)$ denote its eligible weight matrices. We define

$$\text{DV-Complexity}_{\mathcal{S}}(M) = \frac{1}{|\mathcal{A}_{\mathcal{S}}(M)|} \sum_{a \in \mathcal{A}_{\mathcal{S}}(M)} E_a^{\mathrm{DVL} \to \mathrm{raw}}. \tag{13}$$

This aggregates the matrix-level organization into a single structural score for each model.

Full-Cohort Relation to Model Capability. Let $\mathcal{C} = \{M_j\}_{j=1}^{48}$ be the fixed checkpoint cohort and let y_j denote the capability score of M_j. For every module scope $\mathcal{S} \in \{\mathrm{Joint}, Q, K, V, O, \mathrm{FFN}\}$, define the descending capability and structural ranks

$$R_y(j) = \mathrm{rank}_{\downarrow}(y_j), \qquad R_{\mathcal{S}}(j) = \mathrm{rank}_{\downarrow}(\text{DV-Complexity}_{\mathcal{S}}(M_j)), \tag{14}$$

where rank 1 denotes the largest value and ties receive their standard average rank. We report Spearman association as the Pearson correlation of these two complete rank vectors,

$$\rho_{\mathcal{S}} = \mathrm{Corr}_{\mathrm{Pearson}} \left(\{R_{\mathcal{S}}(j)\}_{j=1}^{48}, \{R_y(j)\}_{j=1}^{48} \right). \tag{15}$$

The capability variable, cohort membership, and rank convention are shared across all six scopes; no scope-specific checkpoint filtering or cohort selection is performed. This estimand tests cross-sectional association between the organization exposed by DV-Lens and the fixed capability measure.

3.5 Theoretical Interpretation of DV-Complexity

We interpret DV-Complexity as measuring functional parameter structure that cannot be compressed into a small set of roles without affecting the predictive objective. Specifically, the score

quantifies the normalized parameter energy removed by collapsing weights within each DV-LENS role, which can be related to the corresponding increase in predictive loss. This establishes DV-COMPLEXITY as a measure of loss-supported functional parameter richness. Appendix E derives this interpretation and provides the proofs.

4 EXPERIMENTS

We test DV-LENS readouts with matched random controls and finite differences, then assess their behavioral effects through parameter ablation, steering, and swapping. Model-level analyses examine capability association and DV-COMPLEXITY stability under benchmark exposure.

4.1 STATISTICAL AND FUNCTIONAL VALIDATION

We test non-random vocabulary structure and local perturbation prediction. The original validation uses five models: Llama-3.2-3B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), OLMo-2-7B (Team OLMo et al., 2025), Granite-3.3-8B (Granite Team, IBM, 2024), and Qwen2.5-0.5B (Yang et al., 2025b), with Qwen2.5-0.5B used only during discovery. We further extend the statistical comparison to 21 models and the finite-difference test to nine models. CI denotes confidence interval. Complete matched-null results, including module-level and family-level comparisons, are provided in Appendix B.1. Figure labels abbreviate model names by family initial, family version, and parameter count (Q1.5-14B, Q2.5-14B, and Q3-14B are Qwen1.5, Qwen2.5, and Qwen3 models (Bai et al., 2023; Yang et al., 2025a); Y1.5, G3.3, O2, M, and L3.2 stand for Yi-1.5 (01.AI et al., 2024), Granite-3.3, OLMo-2, Mistral, and Llama-3.2).

4.1.1 NON-RANDOM VOCABULARY STRUCTURE

The Gaussian controls match DV-LENS directions in dimension, covariance treatment, and norm. The extended analysis covers 21 models from 13 families, three depths, and Q/K/V/O/FFN. Each model-depth-module cell compares 12 concept-mean DV-LENS directions with 192 matched Gaussian directions. Table 1 shows the eight models that also appear in the finite-difference study; complete results are reported in Appendix B.1. Table 1 shows that all eight displayed model means are above the matched-random baseline, although the size of the difference varies across models. Across 21 models, the mean excess-kurtosis difference is $+2.363$ (model-bootstrap 95% CI $[1.911, 2.869]$). The original prompt-level validation shows the same direction: the excess-kurtosis difference is $+2.128$ with 95% CI $[0.609, 3.671]$, and the Top-1% ℓ_2-energy difference is $+0.0191$ with 95% CI $[0.0119, 0.0263]$. Together, these results show that DV-LENS directions have stronger heavy-tail and concentration structure than matched random directions. The 21-model concept-level analysis and the original prompt-level analysis use different aggregation units and are therefore reported separately.

Table 1: Model-level matched-null comparison. $\Delta K = K_{\mathrm{DVL}} - K_{\mathrm{null}}$ measures the excess-kurtosis advantage of DV-LENS over matched Gaussian directions. Positive values favor DV-LENS.

Model	ΔK	95% CI
Random baseline	0.000	–
OLMo-2-7B	+0.630	[0.284, 1.054]
Yi-1.5-9B	+1.128	[0.561, 1.735]
Qwen2.5-0.5B	+1.488	[0.787, 2.316]
Qwen2.5-14B	+2.534	[1.128, 4.283]
Qwen1.5-14B	+2.858	[0.594, 6.522]
Llama-3.2-3B	+3.107	[1.372, 5.190]
Mistral-7B	+3.263	[1.224, 5.771]
Granite-3.3-8B	+5.391	[1.714, 10.192]
Mean	**+2.363**	**[1.911, 2.869]**

4.1.2 FINITE-DIFFERENCE PREDICTION

We next test whether Equation 4 predicts the actual response to small parameter perturbations. The evaluation contains 720 fixed cases from nine models, with 80 cases per model. For a score vector s, let U_s^K and L_s^K denote the token IDs with the K largest and smallest scores. We measure coordinate-orientation agreement as the fraction of matching predicted and observed coordinate signs. With FD denoting finite differences, define $R_{32\to K} = \left(|U^{32}_{\mathrm{DVL}} \cap U^{K}_{\mathrm{FD}}| + |L^{32}_{\mathrm{DVL}} \cap L^{K}_{\mathrm{FD}}|\right)/64$ for $K \in \{32, 128\}$. For $K = 32$, this is the exact two-tail Top-32 overlap. For $K = 128$, it becomes

Top-32 Recall@128, which allows small rank changes while keeping the positive and negative tails separate. The random expectations are $1/2$ for orientation and $32/|\mathcal{V}|$ and $128/|\mathcal{V}|$ for overlap.

Figure 2 shows that coordinate orientation is close to one for all nine models. Exact Top-32 overlap is lower and varies more across models, while every model improves under Recall@128. This indicates that many exact-rank misses remain within the same high-impact vocabulary region rather than moving to unrelated coordinates. Thus, DV-LENS predicts the direction and broader coordinate structure of the local response more reliably than the exact ordering within each tail. The three metrics distinguish coordinate signs, exact tail membership, and recovery within a wider tail. Their joint pattern makes the source of disagreement visible: close orientation agreement can coexist with changes in the ranking of high-impact tokens. Keeping upper and lower tails separate preserves the direction of the comparison: each predicted token is matched against the corresponding tail of the finite-difference response.

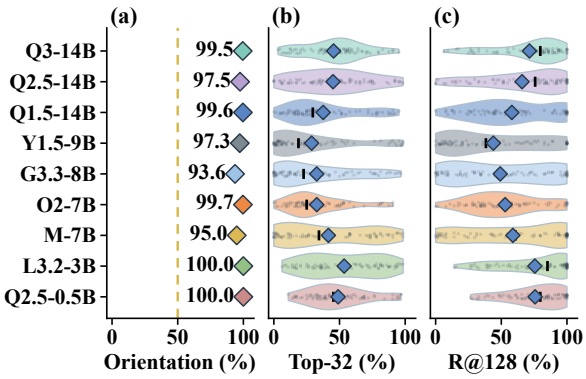


Figure 2: Exact DV-LENS readouts recover finite-difference coordinate orientation and high-impact vocabulary regions over 720 frozen cases from nine models, with 80 cases per model. **(a)** Model-mean coordinate-orientation agreement; the gold dashed line marks the 50% random-orientation baseline. **(b)** Exact two-tail Top-32 overlap. **(c)** Two-tail Top-32 Recall@128.

Table 2 gives the corresponding module-level results. Pooled orientation agreement is 0.980, exact Top-32 overlap is 0.407, and Recall@128 is 0.613. Orientation remains high across all modules, while coordinate recovery is lowest for K and highest for O and FFN. The gain from Top-32 overlap to Recall@128 further shows that many misses come from local rank changes within the same high-impact tail.

Table 2: Finite-difference validation. Orientation agreement, exact two-tail Top-32 overlap, and two-tail Top-32 Recall@128 are reported by module.

Module	Orient.	Top-32	R@128
Q	0.976	0.313	0.525
K	0.947	0.209	0.375
V	0.988	0.412	0.650
O	0.992	0.537	0.743
FFN	0.997	0.565	0.770
Pooled	**0.980**	**0.407**	**0.613**

Overall, these experiments provide two complementary checks: DV-LENS directions show non-random vocabulary structure and predict the local response to small parameter changes. Additional structural and token-level analyses are provided in Appendix B. We next test whether these local predictions translate into controlled changes in model behavior in Section 4.2.

Prediction across Layer Depths. Figure 3 complements the model and module summaries by grouping the four sampled layers according to their relative depth. Each depth contains the same nine models, five modules, and 180 cases. From the earliest to the latest sampled layer, orientation agreement increases from 94.5% to 99.96%, exact Top-32 overlap from 21.4% to 67.2%, and Recall@128 from 38.3% to 86.3%. All nine models improve between these endpoints for each metric. The early layers therefore already recover most coordinate signs, while their high-impact token rankings are less precise. Later layers improve exact and wider-tail recovery, locating the larger ranking discrepancies in earlier layers.

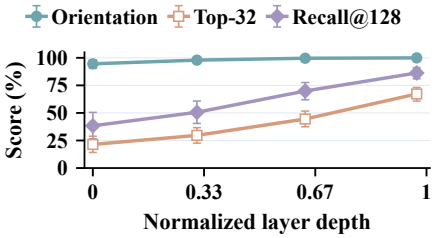


Figure 3: Finite-difference prediction by layer depth. The same 720 cases are grouped by the four sampled layer positions, with 180 cases per depth. Curves average nine models; error bars show 95% model-bootstrap CIs. Horizontal positions use mean normalized depths.

4.2 BEHAVIORAL INTERVENTION

We evaluate **ablate**, **steer**, and **swap** on 21 models and ten factual concepts across Q/K/V/O/FFN and their joint pool, using eight perturbed-parameter fractions (0.05%–5%). The remaining abbreviations in Figure 4 are DS-R1 for DeepSeek-R1-

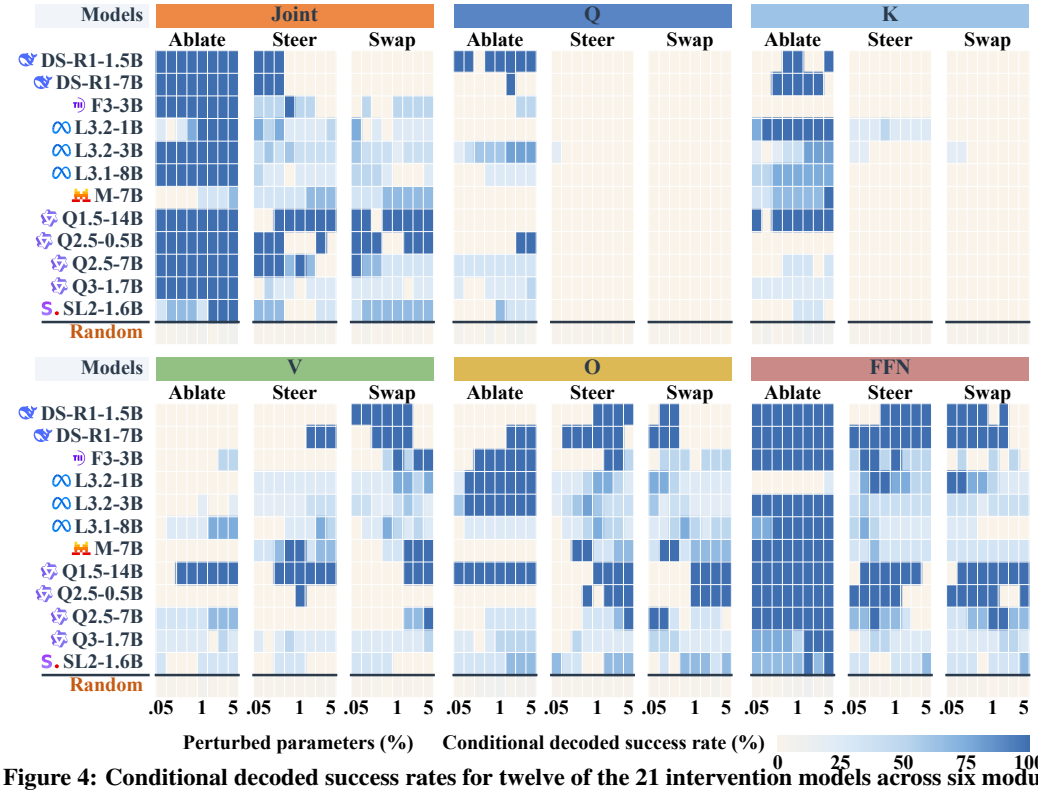


Figure 4: Conditional decoded success rates for twelve of the 21 intervention models across six module scopes and three operations. The eight columns in each bank are the perturbed fractions from 0.05% to 5%; the bottom row is the matched random baseline. All 21 models are shown in Appendix C.2, and the matched random controls in Appendix C.4.

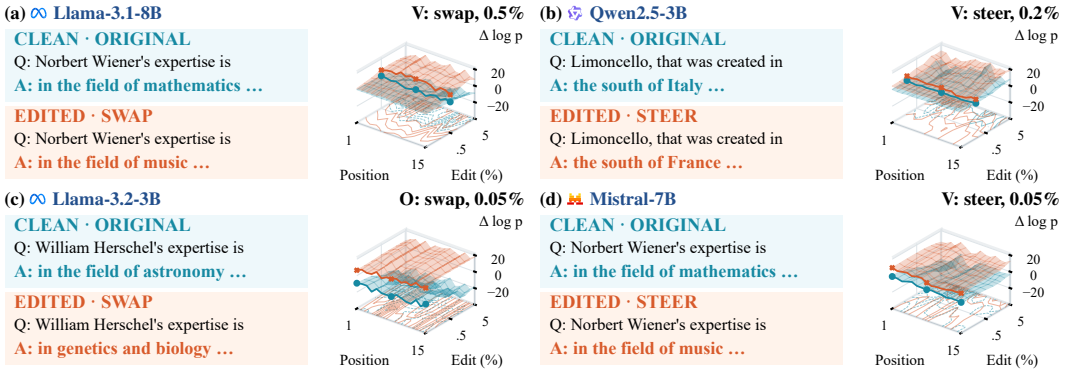


Figure 5: Decoded intervention cases. Blue/orange cards show clean/edited answers; headers give model, module, operation, and fraction. Blue/orange surfaces show source/target log-probability changes (edited minus clean); positive/negative height indicates promotion/suppression. Edit (%) spans eight fractions logarithmically. Bold traces mark the header's fraction; floor contours project the surfaces.

Distill (DeepSeek-AI, 2025), F3 for Falcon3 (Technology Innovation Institute, 2024), and SL2 for StableLM-2 (Bellagente et al., 2024).

Figure 4 reports the primary conditional decoded success rate for twelve models (Figure 10 shows all 21): whether the edited continuation satisfies the operation-specific answer-containment rule, among concept cells whose clean continuation contains the original gold answer. It is defined for 2,880 aggregate rows; 144 lack an eligible clean continuation. Appendix C.3 reports future-token probability changes for all ten concepts.

Pooling concept counts at Joint scope, peak conditional decoded success rates are 61/66 for ablation (5%), 28/66 for steering (0.1%–0.2%), and 23/66 for swapping (5%).

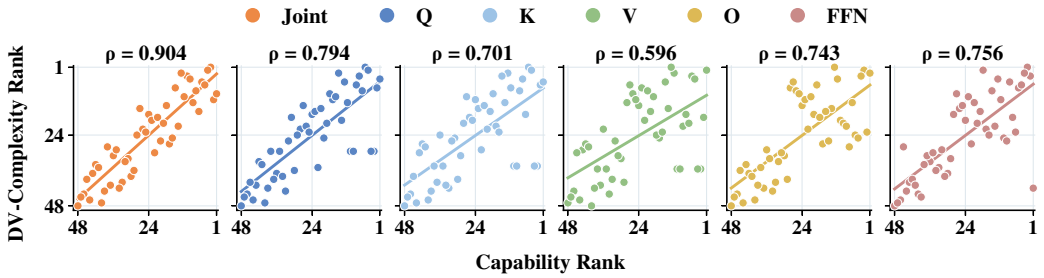


Figure 6: Uniform six-task capability rank and DV-Complexity rank on one fixed 48-checkpoint cohort. The equal-width grid is ordered Joint, Q, K, V, O, FFN; every figure uses all 48 paired rank positions.

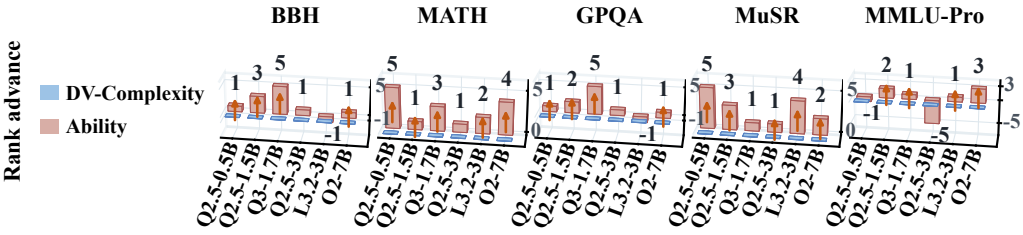


Figure 7: Benchmark-derived ability ranks advance while the primary DV-Complexity rank remains stationary under complete-benchmark exposure. Orange arrows mark positive non-ceiling joint responses.

Figure 5 connects decoded answer changes with source suppression and target promotion at the highlighted fraction. Detailed cases for six models appear in Appendix C, Tables 11–16, covering Q, K, V, O, FFN, and Joint with paired clean and edited answer excerpts.

4.3 Model-level Capability and Benchmark-Exposure Diagnostics

Cross-Sectional Capability Ranking. For 48 post-trained LLMs, DV-Complexity uses 2,000 Jacobian-estimation prompts, a rank-128 approximation, $k = 32$, and all eligible layers and six module scopes. Capability is the unweighted mean over six Open LLM Leaderboard v2 benchmarks: Instruction-Following Eval (IFEval), BIG-Bench Hard (BBH), MATH, Graduate-Level Google-Proof Q&A (GPQA), Multistep Soft Reasoning (MuSR), and MMLU-Pro, the enhanced Massive Multitask Language Understanding benchmark. Appendix A.3 specifies scoring, benchmark settings, and model versions. Figure 6 shows positive rank associations in all six scopes, strongest for Joint. Spearman correlations are $\rho = 0.904$, 0.794, 0.701, 0.596, 0.743, and 0.756 for Joint, Q, K, V, O, and FFN, respectively. Ties receive midranks. Fused QKV projections are split into their native Q/K/V output-row ranges before scoring. Appendix A.4 reports label, representation, and leave-one-family-out controls supporting the association across model families.

Stability under Complete-Benchmark Exposure. We expose eight models from six families to the complete BBH, MATH, GPQA, MuSR, and MMLU-Pro test splits, then rerank each against the other models' unchanged baseline scores. A DV-Complexity rank change of at most one is stable. Figure 7 shows improved ability ranks and largely unchanged DV-Complexity ranks under complete-test-split exposure. Appendix A.5 details the protocol and the omission of IFEval because its exposure training was unstable and prone to collapse.

5 Conclusion

In this paper, we introduce DV-Lens, which maps LLMs' parameter columns to vocabulary-space readouts, connecting parameter interpretation with targeted interventions. Statistical and finite-difference tests support the structure and local predictive accuracy of these readouts, while decoded cases illustrate their behavioral effects. DV-Complexity summarizes functional organization and tracks broad capability across 48 models. Its ranks remain largely stable under benchmark exposure, even when benchmark performance improves. Together, these results connect local parameter analysis with model-level comparison through a shared vocabulary-space representation.

Reproducibility Statement

The main text defines the score and evaluation procedure. Appendix A records the experimental protocols, Appendix B and Appendix C document the supporting checks, and Appendix E contains the assumptions and proofs for the theoretical results.

AI Use Statement

Generative AI tools were used to polish the grammar and wording of the manuscript. The authors reviewed the final text and take responsibility for its content.

References

01.AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, et al. Yi: Open foundation models by 01.AI. *arXiv preprint arXiv:2403.04652*, 2024. URL https://arxiv.org/abs/2403.04652.

Asaf Avrahamy, Yoav Gur-Arieh, and Mor Geva. Disentangling mlp neuron weights in vocabulary space. *arXiv preprint arXiv:2604.06005*, 2026.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. URL https://arxiv.org/abs/2309.16609.

Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, et al. Stable LM 2 1.6B technical report. *arXiv preprint arXiv:2402.17834*, 2024. URL https://arxiv.org/abs/2402.17834.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Amil Dravid, Yossi Gandelsman, Kuan-Chieh Wang, Rameen Abdal, Gordon Wetzstein, Alexei A. Efros, and Kfir Aberman. Interpreting the weight space of customized diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2b8f4db0464cc5b6e9d5e6bea4b9f308-Abstract-Conference.html.

Gabriel Eilertsen, Daniel Jönsson, Timo Ropinski, Jonas Unger, and Anders Ynnerman. Classifying the classifier: Dissecting the weight space of neural networks. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, volume 325, pp. 1119–1126. IOS Press, 2020. doi: 10.3233/FAIA200209.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of EMNLP*, 2021. URL https://aclanthology.org/2021.emnlp-main.446/.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 2024.

Elena Golimblevskaia, Aakriti Jain, Bruno Puri, Ammar Ibrahim, Wojciech Samek, and Sebastian Lapuschkin. Circuit insights: Towards interpretability beyond activations. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.14936.

Granite Team, IBM. Granite 3.0 language models. Technical report, IBM Research, 2024. URL https://github.com/ibm-granite/granite-3.0-language-models/blob/main/paper.pdf.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html. Published July 6, 2026; accessed September 10, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2103.03874.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Hugging Face Open LLM Leaderboard Team. Open LLM leaderboard: Tasks, metrics, and reproducibility. https://huggingface.co/docs/leaderboards/open_llm_leaderboard/about, 2024. Accessed 2026-09-01.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Shahar Katz, Yonatan Belinkov, Mor Geva, and Lior Wolf. Backward lens: Projecting language model gradients into the vocabulary space. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2390–2422. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.142.

Sida Liu and Feijiang Han. Ica lens: Interpreting language models without training another dictionary. *arXiv preprint arXiv:2606.11722*, 2026. URL https://arxiv.org/abs/2606.11722. Preprint.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. URL https://arxiv.org/abs/2311.12022.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.16049.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022. URL https://arxiv.org/abs/2210.09261.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025. URL https://arxiv.org/abs/2501.00656.

Technology Innovation Institute. Welcome to the Falcon3 family of open models! Hugging Face blog, 2024. URL https://huggingface.co/blog/falcon3.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2406.01574.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025b. URL https://arxiv.org/abs/2412.15115.

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

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. URL https://arxiv.org/abs/2311.07911.

APPENDIX

The appendix is organized as follows. Appendix A specifies the intervention and capability protocols, including the fixed model sets and controls. Appendix B reports matched-null coverage and structural inspection of the parameter directions. Appendix C provides the complete intervention grids, matched controls, and decoded transition cases. Appendix D records the computations used to obtain DV-LENS signals from model weights and prompts. Appendix E develops the theory of the DV-COMPLEXITY score, including its energy interpretation and relation to cross-entropy cost. Appendix F discusses limitations. Appendix G addresses questions about the interpretation and validation of the readouts and model-level score.

A EXPERIMENTAL PROTOCOLS

A.1 INTERVENTION SETUP

The intervention experiments cover 21 models, ten factual concepts, six scopes, three operations, and eight intervention fractions: 0.05%, 0.1%, 0.2%, 0.5%, 1%, 2%, 3%, and 5%. The six scopes are Joint, which combines Q+K+V+O+FFN, and Q, K, V, O, and FFN separately. For Joint, the requested fraction is selected independently within Q, K, V, O, and FFN, and the five selected sets are then edited together. The main decoded measure includes only clean continuations that contain the original gold answer.

A.2 CAPABILITY ANALYSIS

The analysis uses a fixed set of 48 models and the joint-layer output-metric matrix computed from 2,000 prompts at every eligible layer, with rank 128 and $k = 32$. Joint, Q, K, V, O, and FFN are all evaluated on the same 48 models.

Six Phi-3 and InternLM2 models use a fused QKV projection. For these models, we split both the frozen operator and the model weights according to the architecture's original Q, K, and V layout before computing separate scores. For Phi-3, Q, K, and V occupy contiguous row ranges determined by the numbers of query and key/value heads. For InternLM2, rows are grouped by key/value head, with the corresponding query group followed by its K and V rows. In both cases, the Q, K, and V parts are disjoint and together cover the full output dimension.

The six scopes therefore give $48 \times 6 = 288$ observed values. Correlation coefficients are recomputed directly from the displayed descending ranks. The fitted lines summarize the relation within this fixed set of models.

A.3 CAPABILITY SCORE AND FIXED MODEL SET

We use the same capability score y_j for all 48 models. The Open LLM Leaderboard v2 evaluates IFEval, BBH, MATH Level 5, GPQA, MuSR, and MMLU-Pro under fixed task-specific settings (Hugging Face Open LLM Leaderboard Team, 2024; Zhou et al., 2023; Suzgun et al., 2022; Hendrycks et al., 2021; Rein et al., 2023; Sprague et al., 2024; Wang et al., 2024). Let $s_{j,t}$ denote model j's percentage score on task t. We define the capability score as the unweighted mean of the six tasks:

$$y_j = \frac{1}{6}\left(s_{j,\mathrm{IFEval}} + s_{j,\mathrm{BBH}} + s_{j,\mathrm{MATH5}} + s_{j,\mathrm{GPQA}} + s_{j,\mathrm{MuSR}} + s_{j,\mathrm{MMLUPro}}\right). \tag{16}$$

The fixed settings are zero-shot strict instance- and prompt-level IFEval accuracy, three-shot normalized BBH accuracy, four-shot exact match on MATH Level 5, zero-shot normalized GPQA accuracy, zero-shot normalized MuSR accuracy across its three subtasks, and five-shot MMLU-Pro accuracy. For post-trained models, we use the released chat template and present few-shot examples as multiturn conversations, following the official Open LLM Leaderboard v2 protocol. Higher values are better for all six tasks, and all six scores are available for every model.

For each scope $\mathcal{S}$, the model-level DV-COMPLEXITY score averages all eligible matrices as defined in Equation 13. We rank the 48 capability scores and the corresponding 48 DV-COMPLEXITY scores in descending order. Rank 1 denotes the largest value, and ties use the standard midrank. The reported

Spearman coefficient is the Pearson correlation between these two 48-entry rank vectors, following Equations 14–15.

Joint, Q, K, V, O, and FFN therefore use the same capability score, the same 48 models, and all 48 paired ranks. No models are removed for individual modules. The set includes instruction-tuned, chat, and hybrid reasoning models. For the six fused-QKV models, the architecture-specific splits described above provide separate Q, K, and V measurements. Table 3 lists all models, while their released versions, architectures, and capability scores are provided in the supplementary material.

Table 3: The fixed set of 48 LLMs used with the uniform six-task capability definition. The table lists the LLM and its type. The released version, architecture, and capability score of every LLM are recorded with the supplementary material.

#	LLM repository	Type
1	tiiuae/Falcon3-1B-Instruct	instruct
2	tiiuae/Falcon3-7B-Instruct	instruct
3	NousResearch/Hermes-2-Pro-Llama-3-8B	chat
4	NousResearch/Hermes-3-Llama-3.1-8B	chat
5	unsloth/Llama-3.1-8B-Instruct	instruct
6	meta-llama/Llama-3.2-1B-Instruct	instruct
7	unsloth/Llama-3.2-3B-Instruct	instruct
8	mistralai/Ministral-8B-Instruct-2410	instruct
9	mistralai/Mistral-7B-Instruct-v0.2	instruct
10	mistralai/Mistral-7B-Instruct-v0.3	instruct
11	NousResearch/Nous-Hermes-2-Mistral-7B-DPO	chat
12	teknium/OpenHermes-2.5-Mistral-7B	chat
13	microsoft/Phi-3-medium-4k-instruct	instruct
14	microsoft/Phi-3-mini-4k-instruct	instruct
15	microsoft/Phi-3.5-mini-instruct	instruct
16	Qwen/Qwen1.5-0.5B-Chat	chat
17	Qwen/Qwen1.5-1.8B-Chat	chat
18	Qwen/Qwen1.5-14B-Chat	chat
19	Qwen/Qwen1.5-4B-Chat	chat
20	Qwen/Qwen2-0.5B-Instruct	instruct
21	Qwen/Qwen2-1.5B-Instruct	instruct
22	Qwen/Qwen2-7B-Instruct	instruct
23	Qwen/Qwen2.5-0.5B-Instruct	instruct
24	Qwen/Qwen2.5-14B-Instruct	instruct
25	Qwen/Qwen2.5-Coder-7B-Instruct	instruct
26	Qwen/Qwen2.5-Math-1.5B-Instruct	instruct
27	Qwen/Qwen2.5-Math-7B-Instruct	instruct
28	Qwen/Qwen3-0.6B	hybrid
29	Qwen/Qwen3-1.7B	hybrid
30	HuggingFaceTB/SmolLM-1.7B-Instruct	instruct
31	HuggingFaceTB/SmolLM-135M-Instruct	instruct
32	HuggingFaceTB/SmolLM-360M-Instruct	instruct
33	HuggingFaceTB/SmolLM2-1.7B-Instruct	instruct
34	berkeley-nest/Starling-LM-7B-alpha	chat
35	01-ai/Yi-1.5-6B-Chat	chat
36	01-ai/Yi-1.5-9B-Chat	chat
37	cognitivecomputations/dolphin-2.9.3-mistral-7B-32k	chat
38	ibm-granite/granite-3.0-2b-instruct	instruct
39	ibm-granite/granite-3.0-8b-instruct	instruct
40	ibm-granite/granite-3.1-8b-instruct	instruct
41	ibm-granite/granite-3.2-8b-instruct	instruct
42	ibm-granite/granite-3.3-8b-instruct	instruct
43	internlm/internlm2-chat-1_8b	chat
44	internlm/internlm2_5-1_8b-chat	chat
45	internlm/internlm2_5-7b-chat	chat
46	openchat/openchat-3.5-0106	chat
47	stabilityai/stablelm-2-1_6b-chat	chat

Continued on next page

Table 3 (continued)

#	LLM repository	Type
48	stabilityai/stablelm-2-zephyr-1_6b	chat

A.4 CAPABILITY CONTROLS

All controls use the same joint-layer output-metric representation, the same 48 models and capability ranks as Figure 6, the same layers and matrices, $k = 32$, seed 0, 12 k-means iterations, and the same architecture-specific Q/K/V splits. In every table, the DV-COMPLEXITY value is exactly the value reported in Figure 6.

Random Labels with Matched Cluster Sizes. For each matrix, we permute the DV-COMPLEXITY labels 128 times while keeping the sizes of the occupied clusters unchanged. Table 4 compares the original coefficient with the mean and standard deviation (SD) of these random-label results. The one-sided conditional randomization p-value is

$$p_{\text{cond}} = \frac{1 + \sum_{b=1}^{128} \mathbf{1}\left\{\rho_{\text{rand}}^{(b)} \geq \rho_{\text{DVL}}\right\}}{129}, \tag{17}$$

with a minimum possible value of $1/129 \approx 0.008$.

Table 4: Random-label control with matched cluster sizes on the fixed 48-model set. All six DV-COMPLEXITY correlations are larger than those from all 128 random-label permutations, giving the minimum possible conditional randomization value.

Scope	DV-COMPLEXITY ρ	Random mean ρ	Random SD	p_{cond}
Joint	0.904	0.216	0.005	0.008
Q	0.794	0.217	0.013	0.008
K	0.701	0.214	0.015	0.008
V	0.596	0.180	0.022	0.008
O	0.743	0.149	0.010	0.008
FFN	0.756	0.110	0.012	0.008

These results show that the observed capability correlations are not reproduced by random cluster assignments with the same cluster sizes. However, the model set and capability ranks remain fixed, so these values should not be read as significance tests for a broader model population or for unseen model families.

Raw-Weight Baseline. The raw-weight baseline clusters directions directly in the original weight-column space. It uses the same $k = 32$, seed 0, 12 iterations, models, layers, and matrices, and evaluates the same raw-space relative sum-of-squared-errors (SSE) score. Table 5 reports $\Delta\rho = \rho_{\text{DVL}} - \rho_{\text{baseline}}$, together with a 95% interval from 2,000 model-family-cluster bootstrap samples.

Table 5: Raw-weight baseline on the fixed 48-model set. Positive $\Delta\rho$ means that the DV-COMPLEXITY correlation is higher. The last column gives the model-family-cluster bootstrap interval for the difference.

Scope	DV-COMPLEXITY ρ	Baseline ρ	$\Delta\rho$	Bootstrap 95% CI
Joint	0.904	0.800	+0.104	[−0.023, 0.311]
Q	0.794	0.653	+0.141	[−0.073, 0.457]
K	0.701	0.519	+0.181	[−0.083, 0.532]
V	0.596	0.593	+0.003	[−0.293, 0.331]
O	0.743	0.689	+0.054	[−0.023, 0.210]
FFN	0.756	0.647	+0.109	[−0.034, 0.304]

The DV-COMPLEXITY correlation is higher in all six comparisons. However, every bootstrap interval includes zero, so this experiment supports a consistent direction in the point estimates rather than a clear difference under the bootstrap analysis.

Removing One Model Family at a Time. We remove each of the 17 model families in turn and recompute the original Spearman correlation using all remaining models. Table 6 reports the full-set value, the range and median across the 17 removals, and the largest absolute change.

Table 6: Capability correlations after removing one model family at a time. The range contains 17 values, each computed after removing one complete model family.

Scope	DV-COMPLEXITY ρ	ρ range after removal	Median after removal	Max. $\lvert\Delta\rho\rvert$
Joint	0.904	[0.888, 0.909]	0.903	0.016
Q	0.794	[0.755, 0.875]	0.792	0.082
K	0.701	[0.643, 0.824]	0.701	0.123
V	0.596	[0.535, 0.723]	0.600	0.127
O	0.743	[0.680, 0.812]	0.749	0.069
FFN	0.756	[0.717, 0.851]	0.748	0.095

All six correlations remain positive after every family removal. Joint changes the least, ranging from 0.888 to 0.909, with a maximum absolute change of 0.016.

A.5 COMPLETE-BENCHMARK EXPOSURE

Exposure Setup. The experiment covers eight models from six families and the complete BBH, MATH, GPQA, MuSR, and MMLU-Pro test splits. Each model–benchmark arm starts from the original checkpoint and uses that benchmark's reference answers as training targets. Thus, exposure to one benchmark is evaluated separately from exposure to another. After training, we recompute the corresponding benchmark score and DV-COMPLEXITY score. The structural score uses the same 2,000-prompt Jacobian estimation, rank-128 approximation, $k = 32$, and eligible layers as the baseline analysis.

Rank Comparison. For each benchmark, the exposed checkpoint replaces only its own baseline score; all other models retain their original scores. We apply this procedure separately to benchmark performance and DV-COMPLEXITY. Higher scores receive better ranks, rank 1 is the highest, and ties receive midranks. Rank advance is the baseline rank minus the post-exposure rank, so positive values indicate improvement. A DV-COMPLEXITY rank change with absolute value at most one is classified as stable. Models already at benchmark rank 1 have no room for further rank advance. Figure 7 displays the paired advances for six of the eight evaluated models.

Why IFEval Is Omitted. IFEval exposure training was unstable and prone to collapse in our runs. We therefore omit IFEval from this exposure comparison and report the five benchmarks above. IFEval remains part of the six-task capability score in Appendix A.3; its omission applies only to the exposure experiment.

B STATISTICAL BOUNDARIES AND STRUCTURAL INSPECTION

B.1 MATCHED-NULL COVERAGE AND VARIATION

We use the same concept-mean comparison across 21 models from 13 families. For each model, we evaluate three fixed depth fractions and five modules: Q, K, V, O, and FFN. This gives 315 model-depth-module cells. In each cell, 12 concept-mean DV-LENS directions are compared with 192 matched Gaussian directions, resulting in 3,780 concept-mean directions, 30,240 prompt references, and 60,480 Gaussian directions. Table 7 reports the model-level results, while Tables 8–10 summarize the results by module, family, and after removing one family at a time. All tables use the same model names and order as the intervention figures.

Figure 8 shows that every model has a positive mean difference in both panels, although the amount of variation differs across models. The overall mean also stays clearly above zero for both excess kurtosis and Top-1% ℓ_2-energy. Thus, the positive pattern is shared across models, while its size varies across models, depths, and modules.

The module-level results show a similar pattern. Q, K, O, and FFN have positive model-level means for both measures in all 21 models. V shows a weaker pattern, with positive means in 19 of the 21 models for each measure. These counts are based on model-level module averages. The intervals in Figure 8 and Table 7 show the variation across the 15 depth-by-module cells within each model.

We also repeat the analysis after removing each model family in turn. Across all 13 such settings, the mean excess-kurtosis difference remains between 2.195 and 2.472, while the Top-1% ℓ_2-energy

difference remains between 0.0403 and 0.0432. Both measures stay positive for every remaining model in every setting. The positive pattern persists after removing any one model family.

Table 7: Model-level matched-null differences across 21 models. Each entry reports the mean Δ [95% CI]. For each model, the interval is computed by resampling its 15 depth-by-module cells. The last column shows how many of the 15 cells have a positive difference in excess kurtosis and Top-1% ℓ_2-energy, respectively. The final row instead resamples the 21 model means, with counts taken over models. Model names and ordering match the intervention figures; DS-R1 denotes DeepSeek-R1-Distill-Qwen.

Model	Excess kurtosis Δ [95% CI]	Top-1% ℓ_2-energy Δ [95% CI]
DS-R1-1.5B	1.159 [0.516, 1.870]	0.0356 [0.0167, 0.0565]
DS-R1-7B	1.502 [0.772, 2.300]	0.0308 [0.0173, 0.0453]
Falcon3-3B	2.048 [0.694, 4.183]	0.0373 [0.0212, 0.0555]
Falcon3-7B	1.453 [0.776, 2.262]	0.0378 [0.0229, 0.0566]
Granite-3.0-2B	2.540 [1.401, 3.888]	0.0527 [0.0331, 0.0735]
Granite-3.3-8B	5.391 [1.714, 10.192]	0.0647 [0.0353, 0.0996]
Llama-3.2-1B	3.375 [1.173, 6.308]	0.0540 [0.0299, 0.0796]
Llama-3.2-3B	3.107 [1.372, 5.190]	0.0533 [0.0323, 0.0747]
Llama-3.1-8B	4.624 [1.678, 8.645]	0.0603 [0.0324, 0.0908]
Mistral-7B	3.263 [1.224, 5.771]	0.0498 [0.0269, 0.0741]
OLMo-2-7B	0.630 [0.284, 1.054]	0.0205 [0.0104, 0.0324]
Qwen1.5-14B	2.858 [0.594, 6.522]	0.0261 [0.0138, 0.0412]
Qwen2.5-0.5B	1.488 [0.787, 2.316]	0.0393 [0.0250, 0.0550]
Qwen2.5-1.5B	2.678 [1.230, 4.815]	0.0570 [0.0351, 0.0818]
Qwen2.5-3B	1.905 [0.391, 4.551]	0.0282 [0.0141, 0.0467]
Qwen2.5-7B	1.950 [0.665, 3.504]	0.0376 [0.0172, 0.0629]
Qwen2.5-14B	2.534 [1.128, 4.283]	0.0460 [0.0265, 0.0668]
Qwen3-1.7B	2.043 [0.966, 3.283]	0.0462 [0.0260, 0.0687]
SOLAR-10.7B	2.261 [0.857, 4.252]	0.0377 [0.0201, 0.0579]
StableLM-2-1.6B	1.693 [0.965, 2.624]	0.0413 [0.0283, 0.0542]
Yi-1.5-9B	1.128 [0.561, 1.735]	0.0278 [0.0146, 0.0428]
Mean	**2.363 [1.911, 2.869]**	**0.0421 [0.0371, 0.0472]**

Table 8: Matched-null differences across all 21 models and by module. Each estimate is the unweighted mean of the 21 model-level values. The two intervals are obtained by resampling models and model families, respectively. The two p-values are one-sided exact sign-test and Wilcoxon-test results computed across models.

Scope	Metric	Mean Δ	95% CI (model)	95% CI (family)	Positive models	Sign p	Wilcoxon p
All modules	Excess kurtosis	2.363	[1.911, 2.869]	[1.858, 2.958]	21/21	4.768×10^{-7}	4.768×10^{-7}
All modules	Top-1% ℓ_2-energy	0.0421	[0.0371, 0.0472]	[0.0363, 0.0478]	21/21	4.768×10^{-7}	4.768×10^{-7}
Q	Excess kurtosis	1.577	[0.621, 2.823]	[0.427, 2.986]	21/21	4.768×10^{-7}	4.768×10^{-7}
Q	Top-1% ℓ_2-energy	0.0264	[0.0178, 0.0373]	[0.0148, 0.0380]	21/21	4.768×10^{-7}	4.768×10^{-7}
K	Excess kurtosis	2.447	[1.510, 3.664]	[1.502, 3.668]	21/21	4.768×10^{-7}	4.768×10^{-7}
K	Top-1% ℓ_2-energy	0.0443	[0.0330, 0.0573]	[0.0326, 0.0561]	21/21	4.768×10^{-7}	4.768×10^{-7}
V	Excess kurtosis	0.313	[0.167, 0.504]	[0.171, 0.545]	19/21	1.106×10^{-4}	2.384×10^{-6}
V	Top-1% ℓ_2-energy	0.0092	[0.0054, 0.0139]	[0.0055, 0.0146]	19/21	1.106×10^{-4}	3.338×10^{-6}
O	Excess kurtosis	2.043	[1.595, 2.529]	[1.642, 2.564]	21/21	4.768×10^{-7}	4.768×10^{-7}
O	Top-1% ℓ_2-energy	0.0571	[0.0466, 0.0682]	[0.0474, 0.0699]	21/21	4.768×10^{-7}	4.768×10^{-7}
FFN	Excess kurtosis	5.437	[3.945, 7.103]	[3.756, 7.505]	21/21	4.768×10^{-7}	4.768×10^{-7}
FFN	Top-1% ℓ_2-energy	0.0734	[0.0613, 0.0858]	[0.0580, 0.0885]	21/21	4.768×10^{-7}	4.768×10^{-7}

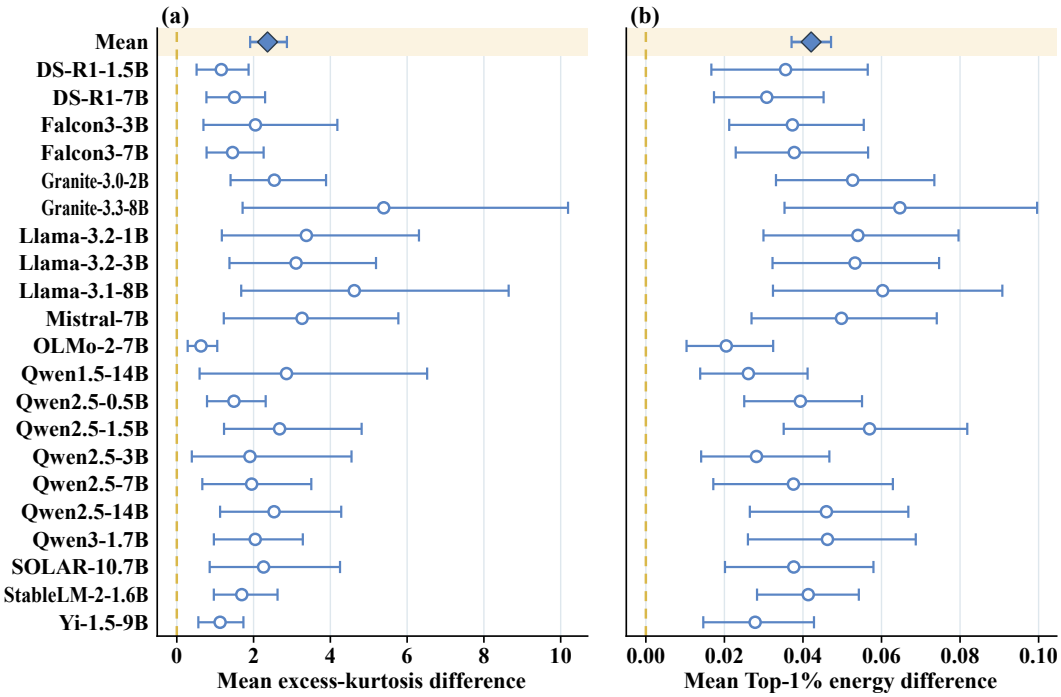


Figure 8: Matched-null differences for the 21 models. Circles are model means with cell-bootstrap 95% intervals over the 15 depth-by-module cells; the diamond is the mean over models with a model-bootstrap 95% interval; the gold dashed line marks zero. Both differences are positive for every model.

Table 9: Matched-null differences by model family. Family values are descriptive unweighted means of the model-level estimates in Table 7.

Family	N	Models	Excess kurtosis Δ	Top-1% ℓ_2-energy Δ
DS-R1	2	DS-R1-1.5B; DS-R1-7B	1.331	0.0332
Falcon3	2	Falcon3-3B; Falcon3-7B	1.751	0.0376
Granite	2	Granite-3.0-2B; Granite-3.3-8B	3.965	0.0587
Llama-3.1	1	Llama-3.1-8B	4.624	0.0603
Llama-3.2	2	Llama-3.2-1B; Llama-3.2-3B	3.241	0.0536
Mistral	1	Mistral-7B	3.263	0.0498
OLMo-2	1	OLMo-2-7B	0.630	0.0205
Qwen1.5	1	Qwen1.5-14B	2.858	0.0261
Qwen2.5	5	Qwen2.5-0.5B; Qwen2.5-1.5B; Qwen2.5-3B; Qwen2.5-7B; Qwen2.5-14B	2.111	0.0416
Qwen3	1	Qwen3-1.7B	2.043	0.0462
SOLAR	1	SOLAR-10.7B	2.261	0.0377
StableLM	1	StableLM-2-1.6B	1.693	0.0413
Yi-1.5	1	Yi-1.5-9B	1.128	0.0278

Table 10: Matched-null differences after removing one family at a time. Every row removes one complete family and recomputes the unweighted model mean over the 12 families that remain. Positive models counts the retained models with a positive difference, which is the same for both metrics in every row.

Omitted family	N	Excess kurtosis Δ	Top-1% ℓ_2-energy Δ	Positive models
DS-R1	19	2.472	0.0430	19/19
Falcon3	19	2.428	0.0426	19/19
Granite	19	2.195	0.0403	19/19
Llama-3.1	20	2.250	0.0412	20/20
Llama-3.2	19	2.271	0.0409	19/19
Mistral	20	2.318	0.0417	20/20
OLMo-2	20	2.450	0.0432	20/20
Qwen1.5	20	2.339	0.0429	20/20
Qwen2.5	16	2.442	0.0422	16/16
Qwen3	20	2.379	0.0419	20/20
SOLAR	20	2.368	0.0423	20/20
StableLM	20	2.397	0.0421	20/20
Yi-1.5	20	2.425	0.0428	20/20

B.2 Fixed functional inspection

We fix the selection rule before examining any token. We consider five models, seven scopes (Q, K, V, O, FFN, Q+K+V+O, and Q+K+V+O+FFN), and directions with high, median, and low cohesion, giving 105 single-direction cases. For each case, we inspect both score tails.

Figure 9 shows two forms of structure in Mistral-7B. One Q direction separates a group of field-related tokens from a group of complete-related tokens. In the centroid examples, one group shows a clear terrain-versus-grade pattern, while other groups contain many token fragments and are harder to interpret. These examples show that some vocabulary tails have clear and readable structure, while others are less easy to describe. They therefore support local interpretability without assuming that every parameter or cluster has a single clean meaning.

C Intervention Details and Cases

C.1 Aggregate intervention setup and denominators

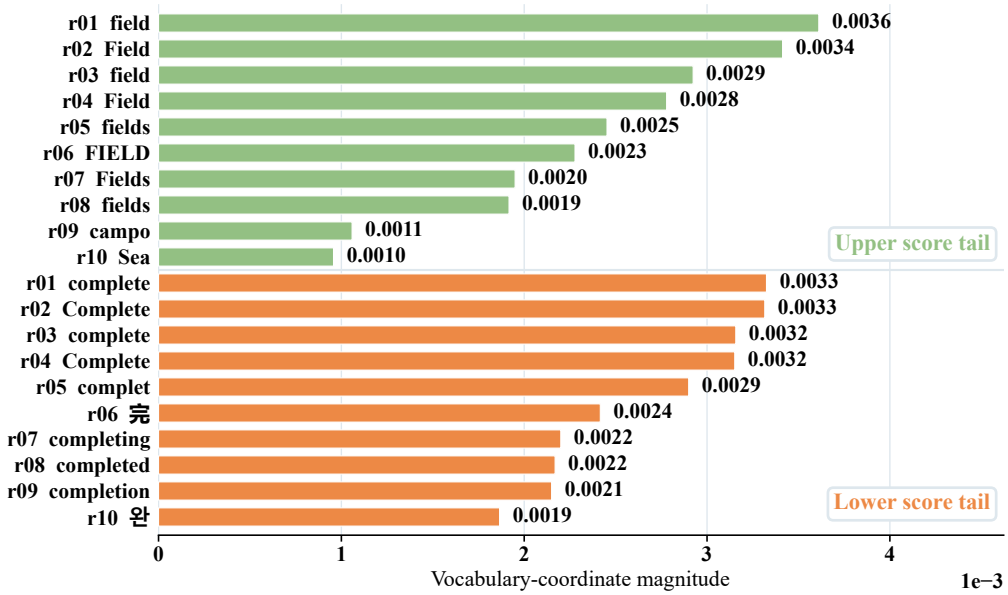
The aggregate results in Section 4.2 cover 21 models, ten concepts, six scopes, three operations, and eight intervention fractions. Each model contributes $10 \times 6 \times 3 \times 8 = 1{,}440$ concept-level cells, giving 30,240 cells in total and 3,024 aggregate rows for each measure. The full-Jacobian/vector–Jacobian-product directions, eligible layers, intervention strengths, generation-time edits, and model set are kept fixed throughout the experiments.

Intervention Operators, Rankings, and Strengths. For a selected parameter column w_i, let u_c be the unit-normalized full-Jacobian/vector–Jacobian-product direction for concept c, let u_t denote the paired target direction, and let α be the fixed strength for the corresponding model, operation, and scope. Ablation applies $w_i \leftarrow w_i - \alpha u_c(u_c^\top w_i)$, while steering applies $w_i \leftarrow w_i + \alpha\|w_i\|_2 u_t$. Swapping uses separate source and target rankings: source columns receive the ablation update for u_c, while target columns receive the steering update for u_t.

The edits remain active during next-token scoring, full-answer scoring, and the complete 16-token continuation, after which all edited columns are restored exactly. For each concept, scope, and operation, columns follow one fixed full-Jacobian ranking. The eight intervention fractions are nested prefixes of this ranking, with no filtering based on decoded outputs.

At the reference fraction of 0.2%, α is selected separately for each model, operation, and scope from $0.25, 0.5, 1, 2, 4, 8$. We choose the value with the highest mean of the first-token and full-answer probability-direction indicators across the ten concepts, breaking ties in favor of the smaller strength. This value is then used for all reported fractions.

(a) Mistral-7B | Q | w1394 | c8



(b) Mistral-7B, six same-model centroids

Q | c8 | w1394

Upper score tail			Lower score tail		
r1	field	0.073	r1	Complete	0.058
r2	Field	0.067	r2	complete	0.057
r3	field	0.061	r3	complet	0.054
r4	Field	0.056	r4	Complete	0.053

FFN | c20 | w10769

Upper score tail			Lower score tail		
r1	terrain	0.046	r1	Grade	0.022
r2	slope	0.021	r2	grade	0.020
r3	landscape	0.019	r3	grades	0.019
r4	sl	0.019	r4	Maur	0.019

FFN | c22 | w9551

Upper score tail			Lower score tail		
r1	&	0.027	r1	wie	0.022
r2	display	0.019	r2	/******/	0.018
r3	entertainment	0.016	r3	members	0.016
r4	Choice	0.016	r4	likewise	0.016

O | c18 | w2188

Upper score tail			Lower score tail		
r1	uses	0.021	r1	->	0.022
r2	hist	0.020	r2	swer	0.021
r3	rag	0.019	r3	>	0.020
r4	groups	0.019	r4	:	0.020

K | c7 | w602

Upper score tail			Lower score tail		
r1	e	0.042	r1	op	0.020
r2	Jun	0.035	r2	op	0.020
r3	Q	0.034	r3	Op	0.019
r4	DO	0.030	r4	opacity	0.018

V | c14 | w2291

Upper score tail			Lower score tail		
r1	Distribution	0.018	r1	Rest	0.020
r2	Wr	0.017	r2	rest	0.019
r3	DES	0.017	r3	re	0.017
r4	Q	0.017	r4	doc	0.016

Figure 9: Single directions and same-model centroids expose distinct vocabulary-tail groups whose readability varies. **(a)** One fixed Mistral-7B query-projection direction: the upper score tail is dominated by variants of the word field and the lower score tail by variants of the word complete, with coordinate magnitudes and within-tail ordering retained. **(b)** Six fixed centroids from the same model, including both concept-like and fragment-heavy groups, with four tokens from each score tail per centroid. These selected cases illustrate organization, not population readability.

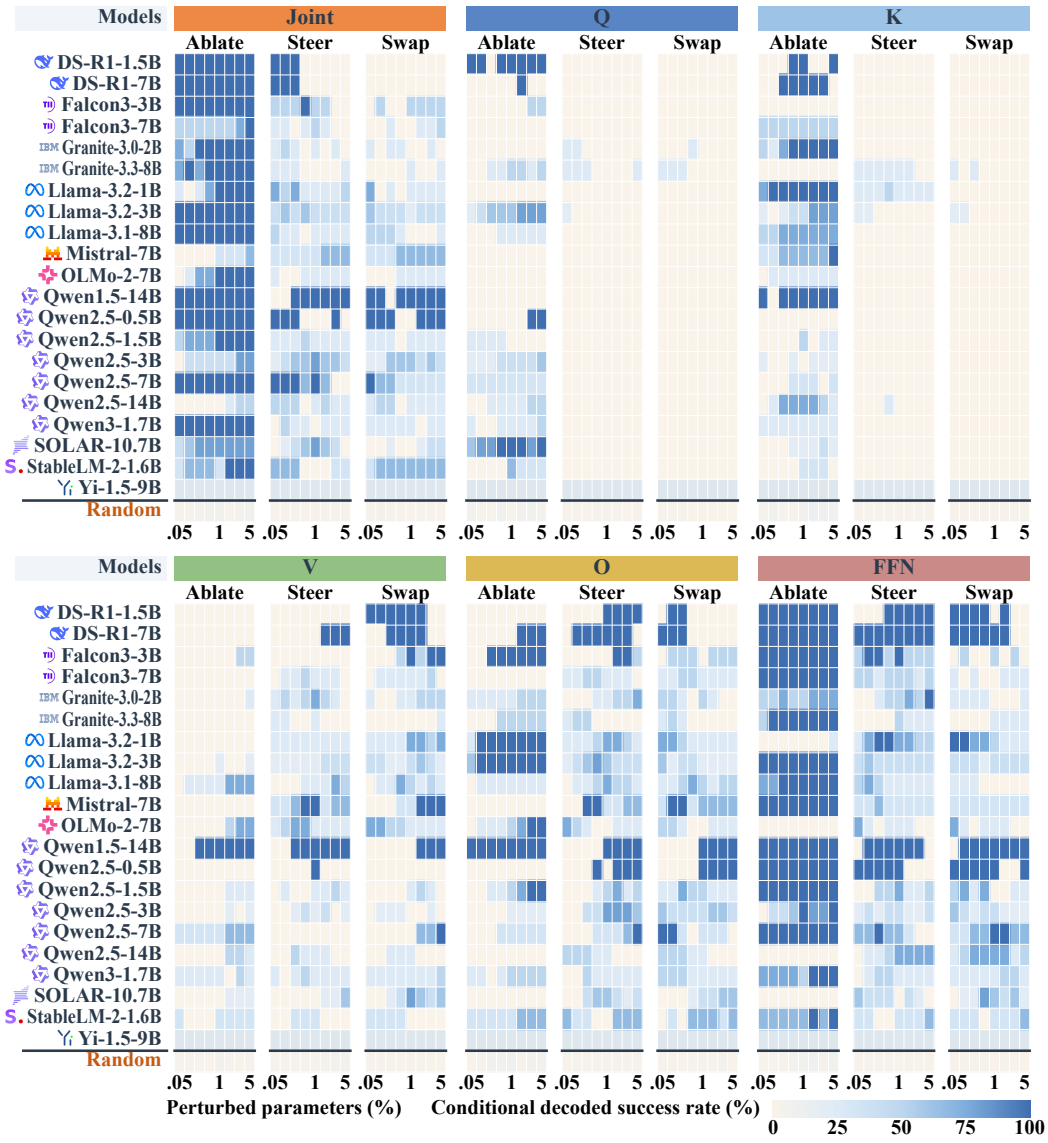


Figure 10: Conditional decoded success rates for all 21 models. Layout as in Figure 4; the bottom row is the matched random baseline averaged over defined model rows; grey cells have a zero denominator.

The conditional decoded success rate is evaluated only for concept cells whose clean continuation contains the original answer. Among the 3,024 aggregate rows, 2,880 have a nonzero denominator and 144 are undefined.

C.2 Complete conditional decoded success-rate grid

Figure 10 extends Figure 4 to all 21 models with the same scopes, operations, and fractions.

C.3 Future-token probability direction

Figure 11 reports the probability-based measure separately from the conditional decoded success rate. A concept cell counts as successful when at least one future answer token moves in the required direction at one or more checked positions. Since all ten concepts contribute to every aggregate row, all 3,024 rows are defined.

At Joint scope, the highest pooled rates are $209/210$ for ablation at 1%, 3%, and 5%; $210/210$ for steering from 0.05% through 0.2%; and $198/210$ for swapping at 0.05% and 0.2%.

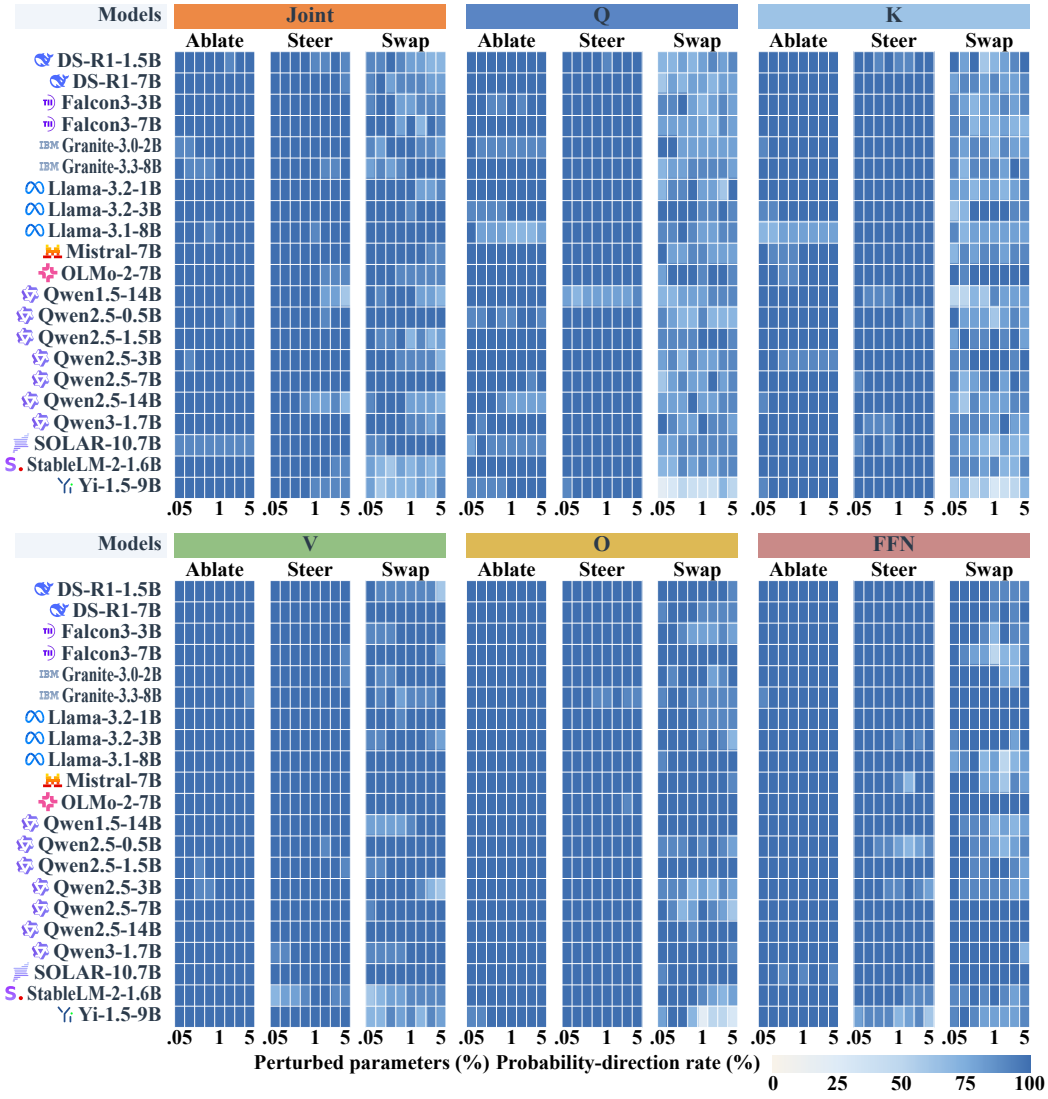


Figure 11: Future-token probability-direction events for the 21 models. Layout and scale as in Figure 10. A concept counts when at least one future answer token moves in the specified direction at any checked position; all 3,024 rows are defined. This event is not a decoded-success measure and carries no multiplicity correction.

This measure checks several future positions and answer tokens without correcting for the number of checks. The high rates show that this broad event is frequent under the fixed protocol.

Figure 11 shows high rates across many models and module scopes, especially for steering at small fractions, while differences across operations and scopes remain visible. The near-saturation of some settings is consistent with this event being relatively easy to satisfy.

C.4 Matched random-column and same-norm control

Each of the 3,024 primary model–scope–operation–fraction rows has a matched random-control row over the same concepts. The control keeps the checkpoint, prompt, module scope, eligible matrices and layers, operation, intervention fraction, strength, clean-generation denominator, future-token positions, and evaluation rules unchanged.

Within each target matrix, the selected columns are replaced by the same number of distinct non-primary columns. These columns are sampled from within $\pm 5\%$ of the original weight-column norm rank, using the nearest available rank when necessary. For each target, we then sample an independent random direction and rescale the update so that both the columnwise and total Frobenius norms match the paired primary intervention. A cell is discarded if the relative norm mismatch is larger than 2.5%.

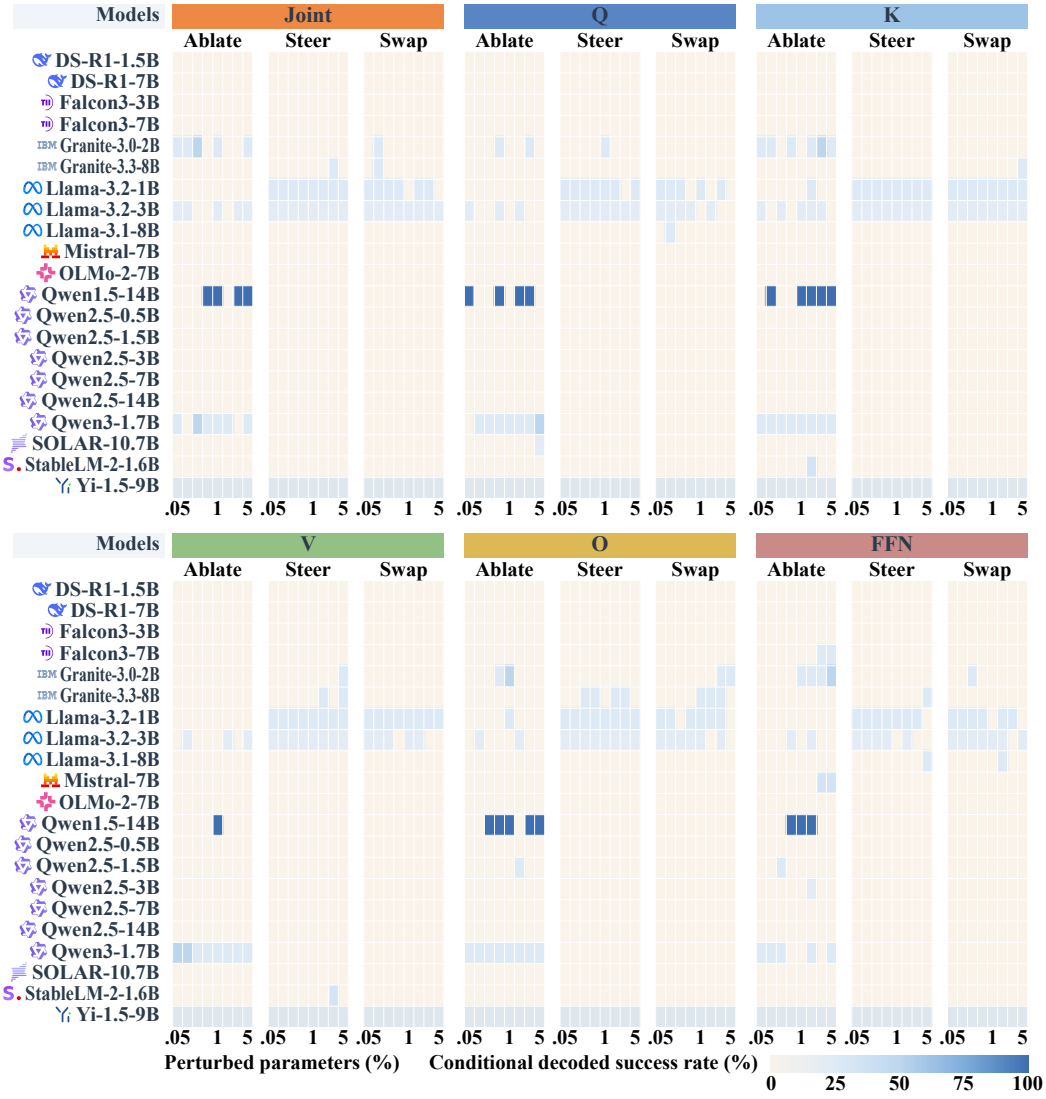


Figure 12: Matched random-column and same-norm control for the 21 models. Each control cell keeps the primary cell's model, concept, scope, operation, fraction, strength, eligible layers, denominator, and evaluation rule, but uses norm-matched non-primary columns and isotropic directions rescaled to the primary update norms. Grey cells have a zero denominator. The conditional decoded success rate averages 3.13% against 30.82% for the primary selection; this single draw is a reference control, not a null distribution.

One matched random draw is generated for every primary cell using a fixed seed. Figures 4 and 10 show the model-averaged control as their bottom row; Figure 12 shows the full control matrix. Averages include only model rows with a defined paired conditional decoded denominator; undefined cases are shown in grey.

Across all 144 matched scope–operation–fraction settings, the mean conditional decoded success rate is 30.82% for the primary intervention and 3.13% for the control. At Joint scope, the corresponding averages are 49.49% and 3.27%. The overall difference of 27.69 percentage points is based on this single matched random draw and does not measure uncertainty across repeated random controls.

Figure 12 shows that the matched control has low rates across all six scopes and does not reproduce the broad high-rate regions of the primary intervention in Figure 10. A small number of nonzero control cells remain, so the comparison does not imply that every random intervention has no effect.

Table 11: IBM Granite-3.3-8B: decoded intervention cases across all six scopes.

Scope / edit	Prompt excerpt	Clean	Edited
Q · Ablate 2%, α=0.5 Replay	Discounted cash flow is used for valuation in this field. …	finance …	programming …
K · Swap 2%, α=2 Replay	… An isotope has a characteristic number of protons but a different number of neutrons here. …	chemistry …	Physics …
V · Steer 2%, α=0.5 Replay	… The Dolomites rise in the northeastern part of this country. …	Italy …	France …
O · Steer 0.2%, α=8 Factual	Norbert Wiener's expertise is	in the field of mathematics and engineering, particularly in the areas of control systems	evident in the harmony and rhythm of the music, and the
FFN · Ablate 2%, α=0.25 Replay	Choirs coordinate pitch, phrasing, and vocal blend in this discipline. …	music …	mathematics …
Joint · Swap 2%, α=0.5 Replay	… Reaction heat and enthalpy are topics in this discipline. …	chemistry …	Physics …

C.5 Qualitative intervention cases

Tables 11–16 present 36 selected cases across six models, with one case for each of Q, K, V, O, FFN, and Joint. We select models with complete module coverage and cases with readable continuations that satisfy the recorded decoded-success criterion.

The tables combine the archived fixed-2% replay run with successful cells from the factual intervention grid in Figure 10. Each row retains its original operation, perturbed fraction, and strength α, followed by the prompt and clean and edited continuation excerpts. For ablation, the edited text omits the source answer; steering and swapping cases show the paired target. Ellipses indicate omitted text. Figure 5 highlights four factual cases in the main text.

Table 12: ∞ Llama-3.2-3B: decoded intervention cases across all six scopes.

Scope / edit	Prompt excerpt	Clean	Edited
Q · Ablate 2%, α=0.5 Replay	Molecular spectroscopy identifies substances in this science. …	Chemistry. …	Physics. …
K · Ablate 0.2%, α=2 Factual	Shawn Fanning's profession is a	computer programmer and entrepreneur. He is best known for co-founding Napster,	professional wrestler, actor, and stuntman. He is best known for his role
V · Swap 0.05%, α=8 Factual	William Herschel's expertise is	in the field of astronomy, particularly in the study of the universe's structure and	in the field of biology, genetics, and evolutionary biology. He is a renowned
O · Swap 0.05%, α=4 Factual	William Herschel's expertise is	in the field of astronomy, particularly in the study of the universe's structure and	in genetics and biology, and he is a renowned biologist. He is a cell
FFN · Swap 2%, α=0.5 Replay	Fugue, symphony, concerto, and sonata are forms in this discipline. …	music. …	mathematics. …
Joint · Ablate 0.05%, α=8 Factual	William Herschel's expertise is	in the field of astronomy, particularly in the study of the universe's structure and	in the field of chemistry, particularly in the area of organic chemistry. His research

Table 13: ∞ Llama-3.1-8B: decoded intervention cases across all six scopes.

Scope / edit	Prompt excerpt	Clean	Edited
Q · Ablate 0.5%, α=1 Factual	The domain of work of Henri Becquerel is	a very important part of the history of physics. He was a French physicist who	a good example of the way in which the scientific community has been able to recognize
K · Ablate 0.05%, α=4 Factual	William Herschel's expertise is	in the field of astronomy, particularly in the study of the solar system and the	in the field of mathematics, specifically in the area of algebraic geometry. His
V · Swap 0.5%, α=2 Factual	Norbert Wiener's expertise is	in the field of mathematics, particularly in the areas of harmonic analysis, stochastic processes	in the field of music, and he is a renowned composer and pianist.
O · Swap 2%, α=1 Replay	… Arithmetic operations on fractions are elementary work in this subject. …	mathematics. (Fractions are a fundamental concept in mathematics.) (Source: Stanford	music …
FFN · Steer 0.1%, α=0.5 Factual	William Herschel's expertise is	in the field of astronomy, particularly in the study of the solar system and the	in biology and genetics, with a focus on the biology of aging and the genetics
Joint · Swap 2%, α=0.25 Replay	… The euro-using republic between Spain and Germany is this country. …	France. (The country between Spain and Germany is France, and the country that	Italy. (Italy is the country between Spain and Germany, and it uses the

Table 14: Mistral-7B: decoded intervention cases across all six scopes.

Scope / edit	Prompt excerpt	Clean	Edited
Q · Swap 2%, α=4 Replay	… A spectrometer can identify a substance from the wavelengths it absorbs in this science. …	chemistry …	Physics …
K · Ablate 0.5%, α=2 Factual	Norbert Wiener's expertise is	in the field of mathematics, but he is also known for his work in	in the field of international law, international relations, and international organizations. He is
V · Steer 0.05%, α=4 Factual	Norbert Wiener's expertise is	in the field of mathematics, but he is also known for his work in	in the field of music, and he is a composer of some distinction. He
O · Swap 2%, α=2 Replay	… Ecological succession describes how communities change after disturbance in this field. …	biology	astronomy
FFN · Ablate 2%, α=0.5 Replay	… This euro-using country lies between Austria and the Mediterranean Sea. …	Italy …	France …
Joint · Swap 2%, α=0.5 Replay	… Testimony is given under oath and may be challenged by cross-examination here. …	Law	Medicine …

Table 15: Qwen2.5-3B: decoded intervention cases across all six scopes.

Scope / edit	Prompt excerpt	Clean	Edited
Q · Ablate 0.05%, α=8 Factual	Grand Mosque of Paris, which is located in	the 10th arrondissement of Paris, France, is a mosque	the 12th arrondissement, is a mosque for Muslims. It
K · Ablate 0.05%, α=0.25 Factual	Grand Mosque of Paris, which is located in	the 10th arrondissement of Paris, France, is a mosque	the 10th arrondissement, is a mosque that was built in
V · Steer 0.2%, α=2 Factual	Limoncello, that was created in	the south of Italy, is a clear, yellow liqueur made from the zest	the south of France, is a clear, yellow liqueur made from the zest
O · Steer 0.2%, α=0.5 Factual	Limoncello, that was created in	the south of Italy, is a clear, yellow liqueur made from the zest	the south of France, is a clear and very light yellow liqueur made from
FFN · Swap 0.05%, α=4 Factual	William Herschel's expertise is	in the field of astronomy. He discovered Uranus and made significant contributions to the	in the field of biology, specifically in the area of bioinformatics. He has
Joint · Steer 0.05%, α=0.25 Factual	Limoncello, that was created in	the south of Italy, is a clear, yellow liqueur made from the zest	the south of France, is a clear yellow liqueur made from the zest of

Table 16: S. StableLM-2-1.6B: decoded intervention cases across all six scopes.

Scope / edit	Prompt excerpt	Clean	Edited
Q · Steer 2%, α=0.5 Replay	… The canals and gondolas of Venice attract visitors to this country. …	Italy …	France …
K · Ablate 1%, α=0.25 Factual	Grand Mosque of Paris, which is located in	the heart of Paris, France, is one of the most famous and largest mosques	the heart of Paris, is one of the most famous and iconic mosques in the
V · Swap 2%, α=8 Replay	… Diversification reduces concentration risk in this discipline. …	Diversification reduces concentration risk in finance. …	Diversification reduces concentration risk in the field of programming. …
O · Ablate 2%, α=4 Replay	A researcher examining DNA inheritance works in this branch of science. …	The researcher examines DNA inheritance in the field of biology. …	A researcher examining DNA inheritance works in the field of astronomy. …
FFN · Swap 2%, α=4 Replay	… A white coat, stethoscope, and examination room are associated with this profession. …	Medical Doctor (MD) …	Law Enforcement Officer …
Joint · Ablate 2%, α=2 Replay	A hospital resident training under senior physicians works in this profession. …	A hospital resident training under senior physicians works in the profession of medicine. …	A resident in a specific field of study, such as law or a related field

D ADDITIONAL DV-LENS COMPUTATION DETAILS

This section records the contractions, the reusable operators, and the input-conditioned scores used by the reported experiments.

Concept-Space Contractions. The parameter signal requires only the Jacobian contraction $\bar{J}_{\ell,m}^{\top} s_c$, rather than the full Jacobian $\bar{J}_{\ell,m}$. For each prompt x, this contraction can be computed directly as

$$J_{\ell,m\to L}(x)^{\top} s_c = \nabla_{h_{\ell,m}(x)} \left[s_c^{\top} z(x) \right], \tag{18}$$

so that

$$d_{\ell,m,c} = \frac{1}{|\mathcal{D}_J|} \sum_{x\in\mathcal{D}_J} \nabla_{h_{\ell,m}(x)} \left[s_c^{\top} z(x) \right]. \tag{19}$$

A vector–Jacobian product therefore produces the required $d_{\ell,m,c}$ without forming $J_{\ell,m\to L}(x)$, and several concepts can be evaluated together.

When concepts from the same output space are queried repeatedly, we further reuse these computations across concepts. Let $\mathcal{S} = \mathrm{span}\{s_1, \ldots, s_C\}$ and let $Q \in \mathbb{R}^{r\times d_{\text{model}}}$ contain an orthonormal basis of $\mathcal{S}$ in its rows. We precompute $B_{\ell,m} = Q\bar{J}_{\ell,m}$. Since $Q^{\top}Qs_c = s_c$ for every $s_c \in \mathcal{S}$,

$$d_{\ell,m,c} = \bar{J}_{\ell,m}^{\top} s_c = B_{\ell,m}^{\top} Q s_c. \tag{20}$$

Hence $B_{\ell,m}$ is computed once for the concept space, after which each concept signal is recovered by one matrix multiplication, without a further backward pass through the model.

Input-Conditioned Intervention Scores. The parameter signal captures the association between a concept and learned parameters over the Jacobian-estimation prompts. For a particular intervention, input-dependent activations identify the parameters participating in the relevant behavior.

For concept c, define $\ell_c(x) = \log \sum_{v\in\mathcal{T}_c} p(v \mid x)$ and $\widehat{d}_{\ell,m,c} = d_{\ell,m,c}/(\|d_{\ell,m,c}\|_2 + \varepsilon)$. At position t, we measure the sensitivity of the concept probability to this signal as

$$q_{\ell,m,c,t}(x) = \left(\nabla_{h_{\ell,m,t}} \ell_c(x) \right)^{\top} \widehat{d}_{\ell,m,c}. \tag{21}$$

Combining this sensitivity with the corresponding input-dependent activation gives

$$\alpha_{\ell,m,i,c}(x) = \frac{1}{|\mathcal{P}(x)|} \sum_{t\in\mathcal{P}(x)} q_{\ell,m,c,t}(x) f_{\ell,m,i,t}(x). \tag{22}$$

For removing an existing concept, we use $M^{\text{remove}}_{\ell,m,i,c} = \alpha_{\ell,m,i,c}\beta_{\ell,m,i,c}$; for adding a target concept, we use $M^{\text{add}}_{\ell,m,i,c} = \alpha_{\ell,m,i,c}\|w_{\ell,m,i}\|_2$. We aggregate these quantities over calibration prompts and rank the eligible parameter columns within the specified module scope.

Shared Module-Level Readout Operator. All parameters in the same weight matrix share the same mapping from the module output to the final vocabulary. We therefore define the module-level readout operator $A_a = W_U\bar{J}_a$. For an individual parameter $w_{a,i}$, its DV-LENS readout is simply $\bar{r}_{a,i} = A_a w_{a,i}$. Since $W_a = [w_{a,1}, \ldots, w_{a,n_a}]$, the readouts of all parameters can be computed jointly as

$$R_a = A_a W_a = W_U \bar{J}_a W_a. \tag{23}$$

The mapping A_a is therefore computed once for each module and shared across all of its parameter columns, after which the full DV-LENS readout matrix follows from one matrix multiplication. The columns of R_a are then normalized and used directly for clustering.

E THEORY OF THE DV-COMPLEXITY SCORE

The DV-COMPLEXITY score is the normalized parameter energy removed by replacing each column with its role centroid. We first establish this identity, then relate the removed energy to the first-order and finite cross-entropy costs of the replacement.

E.1 SETTING

Role models and score. Let $\mathcal{A}$ be a finite, nonempty collection of analysed matrices $W_a = [w_{a,1}, \dots, w_{a,n_a}] \in \mathbb{R}^{d_a \times n_a}$, with $n_a \geq 1$ and $m_a := \|W_a\|_F^2 > 0$. Fix an integer $k \geq 1$ and a labelling $c_{a,i} \in \{1, \dots, k\}$. All cluster sums below range over nonempty clusters. With $C_{a,c} := \{i : c_{a,i} = c\}$ and $n_{a,c} := |C_{a,c}|$, define

$$\bar{w}_{a,c} = \frac{1}{n_{a,c}} \sum_{i \in C_{a,c}} w_{a,i}, \qquad \bar{W}_a[:, i] = \bar{w}_{a,c_{a,i}}, \qquad \Delta_a = W_a - \bar{W}_a. \tag{24}$$

Thus each $\bar{W}_a$ has at most k distinct columns. Write $W = (W_a)_a$, $\bar{W} = (\bar{W}_a)_a$, $\Delta = W - \bar{W}$, and use $\langle U, V \rangle = \sum_a \langle U_a, V_a \rangle_F$ and $\|U\|^2 = \sum_a \|U_a\|_F^2$ for tuples of matrices. The scores in Equations 12–13 are

$$S_a = \frac{\|\Delta_a\|_F^2}{m_a}, \qquad \bar{S} = \frac{1}{|\mathcal{A}|} \sum_a S_a, \qquad \mathcal{E}_{\text{role}} := \|\Delta\|^2 = \sum_a S_a m_a. \tag{25}$$

Prediction loss and regularity. Let $z_W(x) \in \mathbb{R}^{|\mathcal{V}|}$ be the logits with all unanalysed parameters held fixed, and let $p_W(\cdot \mid x) = \mathrm{softmax}(z_W(x))$. Define

$$\ell_y(z) = \mathrm{LSE}(z) - z_y, \qquad \mathrm{LSE}(z) = \log \sum_v e^{z_v}, \qquad L(W) = \mathbb{E}_{(x,y)} \ell_y(z_W(x)). \tag{26}$$

Assume finite logits almost surely, finite $L(W)$ and $L(\bar{W})$, and differentiability at W, with differentiation commuting with the expectation defining L. All expectations below are finite. The labelling and the centroids are fixed at the trained W when taking directional derivatives. Set

$$J_W(x) = D_W z_W(x), \qquad \delta z(x) = z_{\bar{W}}(x) - z_W(x) = d(x) + r(x), \qquad d(x) = -J_W(x)\Delta. \tag{27}$$

Here r is the exact remainder of the finite displacement. For $\omega \geq 0$, define

$$\rho = \nabla_W L(W) + \omega W, \qquad \varepsilon = \|\rho\|, \qquad \varepsilon_a = \|\rho_a\|_F. \tag{28}$$

Thus $\rho = 0$ means stationarity of $L(W) + \frac{\omega}{2}\|W\|^2$ on the analysed parameters. Write $DL(W)[V] = \langle \nabla_W L(W), V \rangle$ and define

$$\kappa_1 = \mathbb{E}\langle p_W - e_y, d \rangle, \qquad R = \mathbb{E}\langle p_W - e_y, r \rangle, \tag{29}$$

$$\bar{D} = \mathbb{E}_x \mathrm{KL}(p_W(\cdot \mid x) \| p_{\bar{W}}(\cdot \mid x)), \qquad \rho_R = 2\mathbb{E}_x \|r(x)\|_\infty, \tag{30}$$

$$\Lambda_1 = \mathbb{E}_x \|J_W(x)\|_{F \to \infty}. \tag{31}$$

The operator norm uses the tuple norm above.

E.2 ENERGY REMOVED BY ROLE COLLAPSE

Lemma E.1 (Orthogonal decomposition induced by a labelling)**.** *For every fixed labelling,*

$$\langle \bar{W}_a, \Delta_a \rangle_F = 0, \qquad m_a = \|\bar{W}_a\|_F^2 + \|\Delta_a\|_F^2, \qquad S_a = 1 - \frac{\|\bar{W}_a\|_F^2}{m_a} \in [0, 1]. \tag{32}$$

These identities imply

$$\langle W, \Delta \rangle = \|\Delta\|^2 = \sum_a S_a m_a \leq \|W\|^2, \tag{33}$$

$$|\mathcal{A}| \bar{S} \min_a m_a \leq \|\Delta\|^2 \leq |\mathcal{A}| \bar{S} \max_a m_a. \tag{34}$$

Proof. The residuals sum to zero in each cluster. Hence

$$\langle \bar{W}_a, \Delta_a \rangle_F = \sum_c \left\langle \bar{w}_{a,c}, \sum_{i \in C_{a,c}} (w_{a,i} - \bar{w}_{a,c}) \right\rangle = 0,$$

$$
\begin{aligned}
m_a &= \left\|\bar{W}_a + \Delta_a\right\|_F^2 = \left\|\bar{W}_a\right\|_F^2 + \left\|\Delta_a\right\|_F^2, \\
S_a &= \frac{\left\|\Delta_a\right\|_F^2}{m_a} = 1 - \frac{\left\|\bar{W}_a\right\|_F^2}{m_a} \in [0,1], \\
\langle W, \Delta \rangle &= \sum_a \langle \bar{W}_a + \Delta_a, \Delta_a \rangle_F = \sum_a \left\|\Delta_a\right\|_F^2 = \sum_a S_a m_a \le \|W\|^2, \\
\left(\min_a m_a\right) \sum_a S_a &\le \sum_a S_a m_a \le \left(\max_a m_a\right) \sum_a S_a. \qquad \square
\end{aligned}
$$

E.3 Cross-Entropy under a Logit Change

Lemma E.2 (Softmax cross-entropy identity and bounds). *For finite logits z, z' and $p = \mathrm{softmax}(z)$, $q = \mathrm{softmax}(z')$,*

$$\ell_y(z') - \ell_y(z) = \langle p - e_y, z' - z \rangle + \mathrm{KL}(p\|q), \tag{35}$$

$$0 \le \mathrm{KL}(p\|q) \le 2\left\|z' - z\right\|_\infty, \qquad |\ell_y(z') - \ell_y(z)| \le 2\left\|z' - z\right\|_\infty. \tag{36}$$

Proof. Put $h = z' - z$. The softmax log-probabilities and nonnegativity of Kullback–Leibler (KL) divergence give

$$
\begin{aligned}
\mathrm{KL}(p\|q) &= \sum_v p_v (z_v - z'_v) + \mathrm{LSE}(z') - \mathrm{LSE}(z), \\
\ell_y(z') - \ell_y(z) &= \mathrm{KL}(p\|q) + \langle p, h \rangle - h_y = \mathrm{KL}(p\|q) + \langle p - e_y, h \rangle, \\
0 \le \mathrm{KL}(p\|q) &\le \mathrm{KL}(p\|q) + \mathrm{KL}(q\|p) = \langle p - q, z - z' \rangle \le 2\|h\|_\infty, \\
\ell_y(z') - \ell_y(z) &= \int_0^1 \langle \mathrm{softmax}(z + th) - e_y, h \rangle \, dt, \\
|\ell_y(z') - \ell_y(z)| &\le \int_0^1 \|\mathrm{softmax}(z + th) - e_y\|_1 \, \|h\|_\infty \, dt \le 2\|h\|_\infty. \qquad \square
\end{aligned}
$$

E.4 Exact Loss Decomposition

Theorem E.3 (Exact role-collapse decomposition). *Under the setting above, for every labelling and $\omega \ge 0$,*

$$L(\bar{W}) - L(W) = \kappa_1 + \bar{D} + R, \qquad \kappa_1 = \omega\|\Delta\|^2 - \langle \rho, \Delta \rangle, \tag{37}$$

$$|\kappa_1 - \omega\|\Delta\|^2| \le \varepsilon\|\Delta\|, \qquad |R| \le \rho_R. \tag{38}$$

At stationarity, $\kappa_1 = \omega \sum_a S_a m_a$.

Proof. Using the chain rule under the stated expectation condition and Lemma E.2,

$$
\begin{aligned}
\nabla_W L(W) &= \mathbb{E}\big[J_W(x)^\top (p_W(x) - e_y)\big], \\
L(\bar{W}) - L(W) &= \mathbb{E}\langle p_W - e_y, \delta z \rangle + \bar{D} \\
&= \mathbb{E}\langle p_W - e_y, d \rangle + \mathbb{E}\langle p_W - e_y, r \rangle + \bar{D} = \kappa_1 + R + \bar{D}, \\
\kappa_1 &= -\mathbb{E}\langle p_W - e_y, J_W \Delta \rangle = -\langle \nabla_W L(W), \Delta \rangle \\
&= -\langle \rho - \omega W, \Delta \rangle = \omega\|\Delta\|^2 - \langle \rho, \Delta \rangle, \\
|\kappa_1 - \omega\|\Delta\|^2| &= |\langle \rho, \Delta \rangle| \le \varepsilon\|\Delta\|, \\
|R| &\le \mathbb{E}\big[\|p_W - e_y\|_1 \|r\|_\infty\big] \le 2\mathbb{E}_x\|r\|_\infty = \rho_R, \\
\rho = 0 \quad &\Longrightarrow \quad \kappa_1 = \omega\|\Delta\|^2 = \omega \sum_a S_a m_a. \qquad \square
\end{aligned}
$$

Corollary E.4 (Local collapse cost at stationarity). *If $\omega > 0$, $\rho = 0$, and $\Delta \ne 0$, the unit collapse direction $u = -\Delta/\|\Delta\|$ satisfies*

$$DL(W)[u] = \omega\|\Delta\| = \omega\left(\sum_a S_a m_a\right)^{1/2} > 0. \tag{39}$$

For all Δ, the unnormalized derivative is $DL(W)[-\Delta] = \omega \|\Delta\|^2$, which vanishes exactly when $\Delta = 0$.

Proof. By Theorem E.3,

$$DL(W)[-\Delta] = -\langle \nabla_W L(W), \Delta \rangle = \kappa_1 = \omega \|\Delta\|^2,$$
$$DL(W)[u] = \frac{DL(W)[-\Delta]}{\|\Delta\|} = \omega \|\Delta\| \quad (\Delta \neq 0),$$
$$DL(W)[-\Delta] = 0 \iff \Delta = 0.$$

□

E.5 The Score as a Ratio of First-Order Costs

Proposition E.5 (Matrix-level collapse-to-deletion ratio)**.** *Fix a and $\omega > 0$. Interpret $-\Delta_a$ and $-W_a$ as tuple directions supported only on block a, and write $N_a = DL(W)[-\Delta_a]$, $D_a = DL(W)[-W_a]$. These derivatives satisfy*

$$N_a = \omega \|\Delta_a\|_F^2 - \langle \rho_a, \Delta_a \rangle_F, \qquad D_a = \omega m_a - \langle \rho_a, W_a \rangle_F. \tag{40}$$

If $\rho_a = 0$, $N_a/D_a = S_a$; if $\rho = 0$, consequently

$$\bar{S} = \frac{1}{|\mathcal{A}|} \sum_a \frac{N_a}{D_a}. \tag{41}$$

More generally, if $\varepsilon_a < \omega \|W_a\|_F$, then $D_a > 0$ and

$$\left| \frac{N_a}{D_a} - S_a \right| \leq \frac{2\varepsilon_a \sqrt{S_a}}{\omega \|W_a\|_F - \varepsilon_a}. \tag{42}$$

Proof. All products below are on block a. For the residual bound, assume $\varepsilon_a < \omega \|W_a\|_F$. Orthogonality gives

$$N_a = -\langle \rho_a - \omega W_a, \Delta_a \rangle_F = \omega \|\Delta_a\|_F^2 - \langle \rho_a, \Delta_a \rangle_F,$$
$$D_a = -\langle \rho_a - \omega W_a, W_a \rangle_F = \omega m_a - \langle \rho_a, W_a \rangle_F,$$
$$\rho_a = 0 \implies \frac{N_a}{D_a} = \frac{\omega \|\Delta_a\|_F^2}{\omega m_a} = S_a,$$
$$D_a \geq \|W_a\|_F (\omega \|W_a\|_F - \varepsilon_a) > 0,$$
$$\frac{N_a}{D_a} - S_a = \frac{S_a \langle \rho_a, W_a \rangle_F - \langle \rho_a, \Delta_a \rangle_F}{D_a},$$
$$\left| \frac{N_a}{D_a} - S_a \right| \leq \frac{\varepsilon_a (S_a \|W_a\|_F + \|\Delta_a\|_F)}{D_a}$$
$$\leq \frac{\varepsilon_a (S_a + \sqrt{S_a})}{\omega \|W_a\|_F - \varepsilon_a} \leq \frac{2\varepsilon_a \sqrt{S_a}}{\omega \|W_a\|_F - \varepsilon_a}.$$

□

E.6 Finite Role-Collapse Cost

Theorem E.6 (Score certificate for role-collapse cost)**.** *For every fixed labelling and $\omega \geq 0$,*

$$L(\bar{W}) - L(W) \geq \omega \sum_a S_a m_a + \bar{D} - \varepsilon \|\Delta\| - \rho_R, \tag{43}$$

$$L(\bar{W}) - L(W) \geq \omega |\mathcal{A}| \bar{S} \min_a m_a + \bar{D} - \varepsilon \|\Delta\| - \rho_R. \tag{44}$$

A strictly positive loss increase is certified when $\varepsilon \|\Delta\| + \rho_R < \omega \sum_a S_a m_a + \bar{D}$.

Proof. The exact decomposition and the residual bounds give

$$L(\bar{W}) - L(W) = \omega \|\Delta\|^2 - \langle \rho, \Delta \rangle + \bar{D} + R$$

$$\geq \omega \|\Delta\|^2 + \bar{D} - \varepsilon \|\Delta\| - \rho_R$$
$$= \omega \sum_a S_a m_a + \bar{D} - \varepsilon \|\Delta\| - \rho_R$$
$$\geq \omega |\mathcal{A}| \bar{S} \min_a m_a + \bar{D} - \varepsilon \|\Delta\| - \rho_R. \qquad \square$$

Corollary E.7 (Stationarity with zero logit remainder)**.** *If $\rho = 0$ and $r(x) = 0$ almost surely, then*

$$L(\bar{W}) - L(W) = \omega \sum_a S_a m_a + \bar{D} \geq \omega |\mathcal{A}| \bar{S} \min_a m_a. \tag{45}$$

Proof. Substituting $R = 0$ and $\rho = 0$ into the exact decomposition yields

$$L(\bar{W}) - L(W) = \omega \|\Delta\|^2 + \bar{D} = \omega \sum_a S_a m_a + \bar{D}$$
$$\geq \omega \sum_a S_a m_a \geq \omega |\mathcal{A}| \bar{S} \min_a m_a. \qquad \square$$

E.7 Loss Advantage Requires Role-External Energy

With unanalysed parameters fixed, let

$$L_k^\star = \inf_{W' :\, \#\,\mathrm{cols}_{\mathrm{distinct}}(W'_a) \leq k \;\forall a} L(W'). \tag{46}$$

Theorem E.8 (Necessary energy for a loss advantage)**.** *Fix a labelling. If $L(W) \leq L_k^\star - \delta$ with $\delta > \rho_R$, then $\Lambda_1 > 0$ and*

$$\|\Delta\| = \left(\sum_a S_a m_a \right)^{1/2} \geq \frac{\delta - \rho_R}{2\Lambda_1}, \tag{47}$$

$$\bar{S} \geq \frac{(\delta - \rho_R)^2}{4|\mathcal{A}| \max_a m_a \, \Lambda_1^2}. \tag{48}$$

Proof. Feasibility of the role model and Lemma E.2 give

$$L(\bar{W}) \geq L_k^\star \geq L(W) + \delta,$$
$$\delta \leq L(\bar{W}) - L(W) \leq \mathbb{E}\big|\ell_y(z_{\bar{W}}) - \ell_y(z_W)\big|$$
$$\leq 2\mathbb{E}_x \|\delta z(x)\|_\infty$$
$$\leq 2\mathbb{E}_x \|J_W(x)\Delta\|_\infty + 2\mathbb{E}_x \|r(x)\|_\infty$$
$$\leq 2\Lambda_1 \|\Delta\| + \rho_R,$$
$$0 < \delta - \rho_R \leq 2\Lambda_1 \|\Delta\| \quad \Longrightarrow \quad \Lambda_1 > 0, \quad \|\Delta\| \geq \frac{\delta - \rho_R}{2\Lambda_1},$$
$$|\mathcal{A}| \bar{S} \max_a m_a \geq \|\Delta\|^2 \geq \frac{(\delta - \rho_R)^2}{4\Lambda_1^2}. \qquad \square$$

E.8 What the DV-Lens Labelling Determines

The following propositions characterize the normalized-readout clustering objective and the role model's readouts.

Proposition E.9 (Normalized-readout clustering objective)**.** *Fix the trained readout map $A_a = W_U \bar{J}_a$ and set $r_{a,i} = A_a w_{a,i}$ and $\widetilde{r}_{a,i} = r_{a,i}/(\|r_{a,i}\|_2 + \eta)$, where $\eta > 0$ is the numerical stabilizer denoted by ε in the main text. For any labelling c, let $\widetilde{\mu}_{a,c}$ be the mean of the normalized readouts in that cluster and define*

$$J_a(c) = \sum_i \|\widetilde{r}_{a,i} - \widetilde{\mu}_{a,c_i}\|_2^2. \tag{49}$$

An ideal global minimizer $c_a^\star$ satisfies

$$J_a(c_a^\star) = \min_{c:\ \text{at most } k \text{ labels}} J_a(c) \le J_a(c_a^{\mathrm{DVL}}). \tag{50}$$

Lloyd assignment and centroid updates each weakly decrease the normalized-readout objective.

Proof. Suppress a, and put $F(c,\mu) = \sum_i \|\widetilde{r}_i - \mu_{c_i}\|_2^2$. With μ^t the current cluster means, the centroid identity and nearest-centre assignment imply

$$\begin{aligned}
F(c,\mu) &= J(c) + \sum_{j:n_j>0} n_j \,\|\widetilde{\mu}_j - \mu_j\|_2^2\,, \qquad J(c) = \min_\mu F(c,\mu),\\
\min_c J(c) &= \min_\mu \sum_i \min_j \|\widetilde{r}_i - \mu_j\|_2^2\,,\\
F(c^{t+1},\mu^t) &\le F(c^t,\mu^t),\\
J(c^{t+1}) = F(c^{t+1},\mu^{t+1}) &\le F(c^{t+1},\mu^t) \le F(c^t,\mu^t) = J(c^t).
\end{aligned}$$

□

Proposition E.10 (Readouts under a fixed trained map)**.** *For any labelling and the fixed linear map A_a, the role centroid reproduces the cluster mean of the unnormalized readouts: $A_a\bar{w}_{a,c} = \bar{r}_{a,c} := n_{a,c}^{-1}\sum_{i\in C_{a,c}} r_{a,i}$. The removed readout matrix is $A_a\Delta_a$ and its squared norm is the within-cluster readout dispersion. If A_a has full column rank,*

$$S_a \le \frac{\sum_i \left\|r_{a,i} - \bar{r}_{a,c_{a,i}}\right\|_2^2}{\sigma_{\min}(A_a)^2 m_a}. \tag{51}$$

Proof. Linearity and the smallest singular value give

$$\begin{aligned}
A_a\bar{w}_{a,c} &= \frac{1}{n_{a,c}}\sum_{i\in C_{a,c}} A_a w_{a,i} = \bar{r}_{a,c},\\
A_a\bar{W}_a - A_aW_a &= -A_a\Delta_a,\\
\|A_a\Delta_a\|_F^2 &= \sum_i \left\|r_{a,i} - \bar{r}_{a,c_{a,i}}\right\|_2^2\\
&\ge \sigma_{\min}(A_a)^2 \sum_i \|\Delta_{a,i}\|_2^2 = \sigma_{\min}(A_a)^2 m_a S_a.
\end{aligned}$$

□

Proposition E.11 (Score bounds and permutation reference)**.** *For a matrix with $n = n_a$ columns and $m = m_a > 0$, define*

$$\begin{gathered}
\bar{w}^{(1)} = \frac{1}{n}\sum_i w_i, \qquad m^{\mathrm{c}} = \sum_i \left\|w_i - \bar{w}^{(1)}\right\|_2^2, \qquad S^{(1)} = m^{\mathrm{c}}/m,\\
S^{\mathrm{opt}}(k) = \min_{c:\ \text{at most } k \text{ labels}} S_a(c).
\end{gathered} \tag{52}$$

For any labelling with $k_c \le k$ nonempty clusters, $S^{\mathrm{opt}}(k) \le S_a(c) \le S^{(1)}$. A raw-weight k-means run satisfies $S^{\mathrm{opt}}(k) \le S^{\mathrm{run}}$. For $n > 1$ and a uniform random permutation π of the labels,

$$\mathbb{E}_\pi S_a(c\circ\pi) = \frac{n-k_c}{n-1} S^{(1)}. \tag{53}$$

For $n = 1$, all these scores are zero.

Proof. Write $\mathrm{SSE}(c) = mS_a(c)$ and $v_i = w_i - \bar{w}^{(1)}$, so that $\sum_i v_i = 0$. For a permuted cluster C_j^π of fixed size n_j, each column is included with probability n_j/n and each distinct pair with probability $n_j(n_j-1)/(n(n-1))$. Therefore, for $n > 1$,

$$\begin{aligned}
mS^{\mathrm{opt}}(k) \le \mathrm{SSE}(c) &\le \sum_i \left\|w_i - \bar{w}^{(1)}\right\|_2^2 = m^{\mathrm{c}},\\
\sum_{i\ne l}\langle v_i, v_l\rangle &= -\sum_i \|v_i\|_2^2 = -m^{\mathrm{c}},
\end{aligned}$$

$$\mathbb{E}_\pi \left\| \sum_{i \in C_j^\pi} v_i \right\|_2^2 = \frac{n_j}{n} m^{\mathrm{c}} - \frac{n_j(n_j-1)}{n(n-1)} m^{\mathrm{c}} = \frac{n_j(n-n_j)}{n(n-1)} m^{\mathrm{c}},$$

$$\begin{aligned} \mathbb{E}_\pi \mathrm{SSE}(c \circ \pi) &= m^{\mathrm{c}} - \sum_{j=1}^{k_c} \frac{1}{n_j} \mathbb{E}_\pi \left\| \sum_{i \in C_j^\pi} v_i \right\|_2^2 \\ &= m^{\mathrm{c}} \left(1 - \frac{\sum_j (n-n_j)}{n(n-1)} \right) = \frac{n-k_c}{n-1} m^{\mathrm{c}}, \\ \mathbb{E}_\pi S_a(c \circ \pi) &= \frac{n-k_c}{n-1} S^{(1)}. \end{aligned}$$

□

F LIMITATIONS

This work studies parameter organization from the perspective of downstream vocabulary behavior, which provides only one view of the internal structure of language models. Learned representations are shaped by many interacting factors, including architecture, optimization, training data, and scale, and the relationships observed from a particular analytical perspective may not fully capture this broader complexity. More generally, connecting internal representational structure to high-level model behavior remains an open problem, especially across different model families and stages of training. Future work could explore how the proposed perspective interacts with complementary views of representation, training dynamics, and model scaling, toward a more unified understanding of how internal structure relates to emergent capabilities.

G MORE DISCUSSIONS

Q1. What does a DV-LENS readout describe for one parameter column?

The readout is the prompt-set average of the local logit response to a unit change in that column's activation coefficient (Equation 6). For an individual prompt, the realized first-order contribution is $c_{\ell,m,i}(x) = f_{\ell,m,i}(x) r_{\ell,m,i}(x)$, as defined in Section 3.1. Thus the signed vocabulary tail of a column identifies a response direction; the coefficient determines how that direction participates in a particular computation.

Q2. Does the readout predict actual changes to the model output?

The finite-difference experiment directly compares the predicted local response with a small parameter perturbation. Across 720 cases from nine models, pooled coordinate-orientation agreement is 0.980, while exact two-tail Top-32 overlap is 0.407 and Recall@128 is 0.613 (Table 2). This distinction matters: the readout is more accurate about response direction and the broader affected vocabulary region than about the exact rank of every token. The ablation, steering, and swapping study then tests decoded behavior across 21 models (Section 4.2).

Q3. Could the observed structure be a generic effect of random parameter directions?

The matched-null analysis compares concept-mean directions with Gaussian directions under the same model, depth, and module cells. Both excess kurtosis and top-1% energy exceed the matched null on average in all 21 models, and the positive pattern remains after each model family is removed in turn (Appendix B.1). These controls support the presence of structured vocabulary responses, while the finite-difference and intervention tests separately examine whether the responses track functional changes.

Q4. Do all modules provide equally precise vocabulary readouts?

No. In Table 2, orientation agreement ranges from 0.947 for K to 0.997 for FFN; exact Top-32 overlap ranges from 0.209 for K to 0.565 for FFN. The method uses the same local-response definition for Q, K, V, O, and FFN, but the empirical coordinate precision differs across these modules. The module-level values are therefore more informative than a single pooled score when interpreting a particular projection.