SHARED CONDITIONAL TRANSFORMATIONS OF ARITHMETIC DIRECTIONS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic interpretability seeks to explain model behavior, in part by identifying feature directions in internal representations and testing their effects through interventions. Yet a central question is whether shared structural principles can explain and predict how representations change across contexts. In controlled addition, we find that arithmetic directions exhibit condition-dependent variation described by low-dimensional orthogonal transformations shared across values at certain layers. These transformations predict target-condition directions for values excluded from fitting. At the reporting layer of each of four language models, hidden-state interventions using directions predicted by these shared transformations achieve exact complete target-answer match rates of 71.13%–99.43%, compared with 29.66%–41.68% using untransformed source directions. In three models, these effects approach those of directions estimated directly under the target condition. The shared-relation pattern also recurs after independent refitting in a tens-column setting. We further characterize a more constrained structure in terms of direction-level approximate conditional equivariance: powers of a shared orthogonal operator approximate the transformations associated with cyclic condition shifts. Predictions under this description induce expected output changes at multiple tested layers. Together, these findings suggest that interpretable structure can lie not only in feature directions themselves, but also in the shared, predictable relations organizing their variation across contexts.

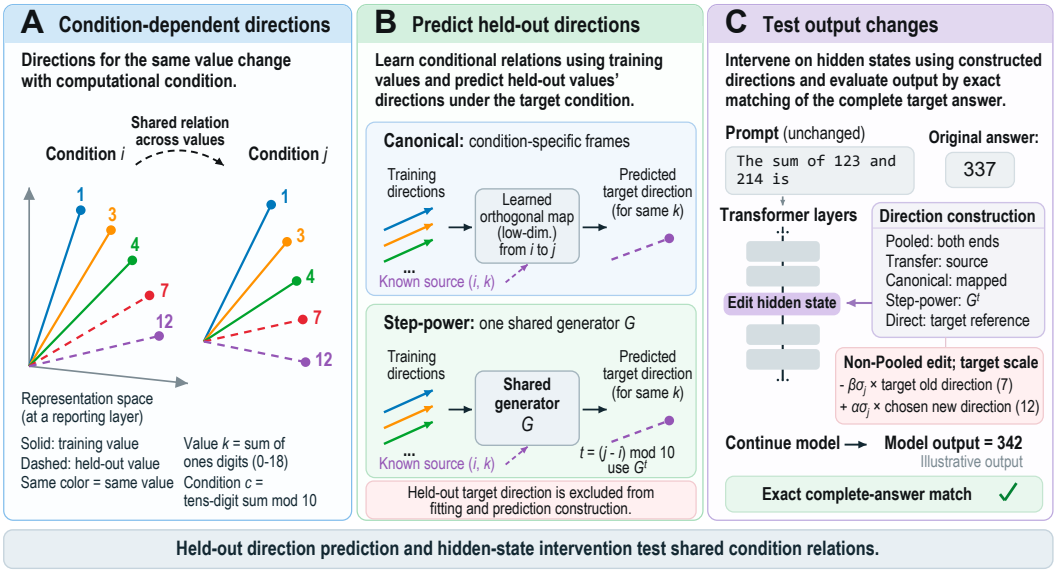


Figure 1: **Testing shared condition relations through held-out prediction and intervention. A:** Condition-dependent value directions (schematic). **B:** Canonical and Step-power learn conditional relations using only training values to predict held-out target-condition directions. **C:** Interventions leave the prompt unchanged and are evaluated by exact matching of the complete target answer.

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.

1 INTRODUCTION

A central goal of mechanistic interpretability is to understand how neural networks produce outputs through internal representations and computations. One approach is to find feature directions associated with particular information or behavior and use interventions to test their effects on output. Intuitively, a feature direction can be understood as an arrow in the model's internal activation space that characterizes an activation change associated with a type of information or behavior. Prior work has found that modifying intermediate activations along certain feature directions can change a model's output content or behavioral tendencies (Turner et al., 2023; Zou et al., 2023; Arditi et al., 2024).

However, a feature direction identified in one input context may not transfer directly to another. Prior work has found that the output changes produced by adjusting hidden states along the same feature direction can vary with the input and prompt (Tan et al., 2024); directions extracted for the same behavior may themselves vary with prompt format (Braun et al., 2025). For example, function vectors (Todd et al., 2024) extracted for the same task can be nearly orthogonal between open-ended and multiple-choice formats (Opielka et al., 2026). We ask whether this context dependence has shared structure: when the information under study takes multiple values, do the corresponding directions change according to a common rule as one moves between contexts? Can a rule learned from some values predict directions for the remaining values?

In this paper, we find shared, predictable structure in how arithmetic feature directions change across computational conditions in controlled addition. For example, $103+204 = 307$ and $123+214 = 337$ share the ones-digit sum 7, while their tens-digit sums define conditions 0 and 3, respectively. Under such changes in condition, the directions representing 7 and other ones-digit sums approximately follow a shared transformation rule at some layers. Relations learned from some values predict target-condition directions for held-out values, and interventions using these predictions can guide the model to produce the expected outputs.

To characterize these relations, *Canonical* fits low-dimensional orthogonal transformations between conditions using directions for a subset of values. It predicts a held-out value's target-condition direction from its source-condition direction and uses the predicted new-value direction in a hidden-state intervention (Figure 1). We further test direction-level *approximate conditional equivariance*: can the same cyclic condition displacement approximately correspond to the same direction transformation across starting conditions? To this end, *Step-power* fits a shared orthogonal operator in a common low-dimensional space using relations between training-value directions in adjacent conditions, and uses its powers to predict held-out directions.

We evaluate intervention effects by the exact-match rate for the complete target answer and geometric correspondence by cosine similarity between predicted directions and held-out target-condition directions, with additional cyclic-consistency diagnostics for Step-power. At the exploratory reporting layers of four models, interventions using Canonical predictions produce the expected outputs more often than those using untransformed source directions, and the predicted directions also have higher cosine similarity to the held-out target directions. Exact target-answer match rates are 96.76%–99.43% in Llama-3-8B, Phi-4, and OLMo-3-7B, close to the reference using directions matched to the current computational condition; GLM-Z1-9B yields 71.13%, with a corresponding reference of 73.42%. Step-power also supports held-out direction prediction and expected output changes at multiple tested layers, with its applicability varying across models, layers, and held-out values. Canonical layer scans further delineate the layer range of its output effects; after independent refitting in a second setting with tens-column values and hundreds-column conditions, interventions using predicted directions again produce the expected output effects.

2 RELATED WORK

2.1 FEATURE DIRECTIONS AND CONDITION DEPENDENCE

Work on the linear representation hypothesis formalizes concept directions and their links to probing and intervention (Park et al., 2024); activation steering tests their output effects (Rimsky et al., 2024; Arditi et al., 2024). Features can also be irreducibly multidimensional: circular weekday and month representations support modular-arithmetic behavior under subspace interventions (Engels et al.,

2025). Across contexts, the effects of a fixed direction can vary with the input or prompt (Tan et al., 2024), and extracted behavior or task directions can themselves vary across prompt formats (Braun et al., 2025; Opielka et al., 2026).

Other approaches adapt interventions to input semantics or current activations (Wang et al., 2025; Li et al., 2026), transport activation distributions (Rodríguez et al., 2025; Abdullaev et al., 2026), or edit nonlinear kernel coordinates (Raval et al., 2026). We ask whether explicitly defined computational conditions induce a shared, learnable relation among the directions for different values.

2.2 Shared Geometry Across Conditions

Structured relations can persist across changes of coordinates or subspaces, as studied in categorical geometry (Park et al., 2025), representation comparison (Williams et al., 2021), and shared geometry across tasks and contexts (Hu et al., 2026a;b). Geometry-informed interventions have also been tested directly: manifold steering compares activation paths through their output trajectories (Wurgaft et al., 2026; Oozeer et al., 2026), while work on distributional representations examines local-probe interpolation and geometry-informed edits (Sarfati et al., 2026).

Numerical representations support vector edits of addition outputs (Zhu et al., 2025) and successor-head predictions (Gould et al., 2024). Studies of modular addition analyze Fourier-based computation and alternative algorithms (Nanda et al., 2023; Zhong et al., 2023); pretrained-model studies examine Fourier and helical numerical structure (Zhou et al., 2024; Kantamneni & Tegmark, 2025). Shared sinusoidal number representations have also been studied through cross-model comparisons, numeric-probe generalization, and probe-derived activation edits (Štefánik et al., 2026). In models trained on multiplication, prior work analyzes partial-product caching and Minkowski-sum and Fourier geometry (Bai et al., 2025).

Related arithmetic studies examine addition and carry algorithms (Quirke & Barez, 2024; Kruthoff, 2024), information flow and last-token computation (Stolfo et al., 2023; Mamidanna et al., 2025), and heuristic computation (Nikankin et al., 2025). Lindsey et al. (2025) also ask whether addition-related features form equivariant families with a coherent geometric structure.

Here, within a fixed model and layer, we learn transformations shared across values between known computational conditions, predict held-out new-value directions, and test their output effects, further examining a shared generator for cyclic condition shifts. Multi-digit addition provides a controlled setting in which values and conditions can be varied separately and each intervention has a defined complete target answer.

3 Methods

3.1 Conditional Value Directions

For addends a, b, let a_0, b_0 denote the ones digits and a_1, b_1 the tens digits. We define the arithmetic value k and computational condition c as

$$k = a_0 + b_0 \in \{0, \ldots, 18\}, \qquad c = (a_1 + b_1) \bmod 10 \in \{0, \ldots, 9\}. \tag{1}$$

The condition c excludes the carry from adding the ones digits.

For a fixed model and selected layer, let $h(a, b) \in \mathbb{R}^d$ denote the residual state at the output of the selected Transformer block at the final prompt token, represented as a row vector. Let $\mathcal{D}_{c,k}$ be the set of problems used to estimate the direction for condition–value pair (c, k). The centroid of their hidden states is $\mu_{c,k} = \frac{1}{|\mathcal{D}_{c,k}|} \sum_{(a,b)\in\mathcal{D}_{c,k}} h(a, b)$.

In each fold, let $\mathcal{K}_{\mathrm{tr}}$ and $\mathcal{K}_{\mathrm{te}}$ denote the training and held-out value sets, respectively; the partition rule is given in Section 4. For each condition c, we use the equally weighted mean of the training-value centroids as the center and normalize the displacement of each value's centroid from this center to obtain its condition-dependent direction:

$$\bar{\mu}_{c,\mathrm{tr}} = \frac{1}{|\mathcal{K}_{\mathrm{tr}}|} \sum_{v \in \mathcal{K}_{\mathrm{tr}}} \mu_{c,v}, \qquad u_{c,k} = \mathrm{unit}(\mu_{c,k} - \bar{\mu}_{c,\mathrm{tr}}), \tag{2}$$

Here, $\mathrm{unit}(v) = v/\|v\|_2$ for nonzero v. Training and held-out value directions use the same training-value center within each condition. In every fold, the centers are recomputed and the subsequent subspaces and transformations are fitted using only the training values; fold indices are omitted.

3.2 Learning Conditional Transformations

For each condition c, we arrange the directions for the same training values in the same order as rows of $U_{c,\mathrm{tr}} \in \mathbb{R}^{m \times d}$, where $m = |\mathcal{K}_{\mathrm{tr}}|$. Its leading $r = 12$ right singular vectors define a condition-specific subspace basis $B_c \in \mathbb{R}^{d \times r}$, giving low-dimensional coordinates $X_{c,\mathrm{tr}} = U_{c,\mathrm{tr}} B_c$.

Generalized Procrustes analysis (Gower, 1975) fits an orthogonal frame $Q_c \in \mathbb{R}^{r \times r}$ for each condition so that the transformed coordinates of the same training values across conditions align with a shared template $Z \in \mathbb{R}^{m \times r}$:

$$\min_{Z, \{Q_c\}} \sum_{c=0}^{9} \|X_{c,\mathrm{tr}} Q_c - Z\|_F^2, \qquad Q_c^\top Q_c = I_r. \tag{3}$$

Each condition's frame is shared across training values, yielding the relation $R_{i \to j} = Q_i Q_j^\top$ from source condition i to target condition j. For $k \in \mathcal{K}_{\mathrm{te}}$, Canonical predicts the target direction from the known source direction $u_{i,k}$; the held-out target direction $u_{j,k}$ enters neither fitting nor prediction construction:

$$\widehat{u}^{\mathrm{Canonical}}_{j \leftarrow i,k} = \mathrm{unit}\left[(u_{i,k} B_i) Q_i Q_j^\top B_j^\top\right]. \tag{4}$$

This prediction projects the source direction into the source condition's training subspace, transforms it through the shared relation, reconstructs it in the target subspace, and normalizes it. Orthogonality permits rotations and reflections in the low-dimensional coordinates; the full map $B_i R_{i \to j} B_j^\top$ is low rank.

Transfer directly reuses the same held-out source direction $u_{i,k}$. All computational conditions participate in fitting, so the test concerns cross-value generalization between known conditions. We evaluate geometric correspondence by cosine similarity with $u_{j,k}$ and expected output effects by hidden-state intervention (Section 3.4). Fitting and geometric evaluation details appear in Appendix A.8.

These conditional relations may vary with the source–target condition pair. In common coordinates, the next section tests whether the same cyclic condition displacement can be described by one transformation across starting conditions.

3.3 Shared-Generator Model for Condition Shifts

Retaining the condition-specific centers determined from training values in each fold, we stack the training direction matrices from all ten conditions. Their leading $r = 12$ right singular vectors define a common basis $B \in \mathbb{R}^{d \times r}$, with $Y_{c,\mathrm{tr}} = U_{c,\mathrm{tr}} B$. Replacing the condition-specific subspaces of Section 3.2 with this common subspace introduces an additional approximation.

In common coordinates, we fit a shared orthogonal step operator using the relations between training-value directions in adjacent conditions:

$$\widehat{G} = \operatorname*{arg\,min}_{G^\top G = I_r} \sum_{c=0}^{9} \left\| Y_{c,\mathrm{tr}} G - Y_{(c+1) \bmod 10,\mathrm{tr}} \right\|_F^2. \tag{5}$$

For source and target conditions $i \neq j$ and a held-out value k, *Step-power* predicts the target direction using a power of this operator:

$$\widehat{u}^{\mathrm{Step}}_{j \leftarrow i,k} = \mathrm{unit}\left[(u_{i,k} B) \widehat{G}^{\,t} B^\top\right], \qquad t = (j - i) \bmod 10. \tag{6}$$

The held-out protocol follows Section 3.2. Within each model, layer, and fold, the same $\widehat{G}$ is used across all starting conditions, displacements, and held-out values.

Equivariant representations relate transformations of represented variables to corresponding transformations in representation space (Higgins et al., 2018). We call this direction-level hypothesis *approximate conditional equivariance*: within the common subspace, the same cyclic condition displacement across starting conditions and values can be approximately described by the same power of the shared operator. Fitting includes the wraparound relation $9 \to 0$ but does not impose $\widehat{G}^{10} = I_r$; we measure ten-step closure error after fitting to assess the operator's cyclic consistency. Details of common-subspace coverage, operator fitting, and closure diagnostics are given in Appendix B.

Two comparisons use the same common basis and training-value centers. *Common-Identity* replaces only $\widehat{G}^t$ in the prediction formula with I_r, testing the role of a condition transformation. *Common-GPA* applies the GPA objective of Section 3.2 in the same common coordinates $Y_{c,\mathrm{tr}}$, fitting condition-specific frames Q_c^{com} and replacing $\widehat{G}^t$ with $Q_i^{\mathrm{com}}(Q_j^{\mathrm{com}})^\top$ to test the effect of the shared-generator constraint. With the common basis fixed and counting only the transformation components, the shared generator has 66 continuous degrees of freedom, compared with 594 for Common-GPA after removing the global orthogonal gauge.

3.4 Hidden-State Interventions

We use the direction constructions in Table 1 to test whether different representational interpretations support expected output changes under hidden-state intervention.

For a problem under target condition j, let $s = a + b$. Changing the value from k_{old} to k_{new} defines the target answer $s_{\mathrm{target}} = s + k_{\mathrm{new}} - k_{\mathrm{old}}$.

Keeping the input unchanged, we intervene once at the extraction layer and token position. For all constructions other than Pooled, the intervention takes the form

$$h' = h - \beta\sigma_j u_{j,k_{\mathrm{old}}} + \alpha\sigma_j d_{j\leftarrow i,k_{\mathrm{new}}}. \tag{7}$$

Here, σ_j is a target-condition scale estimated from training values; α, β control the addition and removal magnitudes. Table 1 specifies the new-value direction $d_{j\leftarrow i,k}$.

Table 1: Five direction constructions and their experimental roles.

Construction	**New direction added** $d_{j\leftarrow i,k}$	**Experimental role**
Pooled	p_k	Condition-averaged baseline
Transfer	$u_{i,k}$	Untransformed source reuse
Canonical	$\widehat{u}^{\mathrm{Canonical}}_{j\leftarrow i,k}$	Held-out target prediction
Projected	$\mathrm{unit}[(u_{j,k}B_j)B_j^\top]$	Subspace reference
Direct	$u_{j,k}$	Target-condition reference

Pooled uses p_k for both removal and addition, while its scale and coefficients remain target-condition dependent. All other constructions remove the old-value direction obtained directly under the target condition, including for held-out old values. Pooled, Projected, and Direct use target new-value information; Projected projects only the addition term.

Common-Identity, Step-power, and Common-GPA use the same intervention form with the addition directions defined in Section 3.3. For these constructions and Canonical, the held-out prediction test concerns the addition term.

Intervention coefficients are calibrated separately for each construction on training-value edits and fixed for evaluation. Evaluation follows Section 4; scale, calibration, and decoding details appear in Appendix A.9.

4 EXPERIMENTAL SETUP

Models and task. We study Llama-3-8B (Llama Team, 2024), Phi-4 (Abdin et al., 2024), OLMo-3-7B (Team Olmo et al., 2025), and GLM-Z1-9B (Zeng et al., 2024; Z.ai, 2025), abbreviated as Llama, Phi-4, OLMo, and GLM. Llama, Phi-4, and OLMo encode each integer from 0 to 999 as a single token, whereas GLM does so for only a subset of this range. Addends range from 1–999, with both original and target answers restricted to 100–999. We use the fixed prompt `The sum of {a} and {b} is` and the intervention position defined in Section 3.4.

Data splits and evaluation. The 19 values are partitioned into five folds by $k \bmod 5$. Testing covers all 54 directed edits between distinct values within the same held-out fold. Fitting, calibration, independent checks, and testing use disjoint ordered addend pairs within each experimental package. For each model, constructions share preset calibration budgets and matched test problems and edits within each comparison.

In the primary setting, geometric evaluation comprises 1,710 held-out route–value comparisons per model, layer, and construction. At each layer, the main direction and shared-generator comparisons each use 13,500 input–edit pairs per construction for every model. Source-dependent constructions evaluate each pair from nine source conditions; reused route records are not independent samples.

The behavioral metric is the exact complete target-answer match rate: the first parsed signed integer must equal the target answer, including any carry change. Parsing failures and problems answered incorrectly before intervention remain in the denominator. Fold results are aggregated by actual record counts.

Layer ranges and reporting layers. The Canonical scan covers Llama L17–L28, Phi-4 L20–L34, OLMo L21–L30, and GLM L24–L39, for 53 model–layer settings (Figure 2); reporting layers in Table 2 are selected post hoc by each model's peak Canonical match rate. Common-coordinate geometric diagnostics cover 47 model–layer settings, with Phi-4 spanning L26–L34. Shared-generator interventions cover five layers per model, for 20 settings; Table 3 uses the same exploratory reporting layers as Table 2.

Tens-column extension. We also study $T = a_1 + b_1 \in \{0, \ldots, 18\}$ under $H = (a_2 + b_2) \bmod 10 \in \{4, 5, 6, 7, 8\}$, where a_2, b_2 are the hundreds digits and the incoming ones-to-tens carry is zero. The target is $s_{\text{target}} = s + 10(T_{\text{new}} - T_{\text{old}})$. Directions and Canonical relations are independently refitted using five folds defined by $T \bmod 5$ and rank-12 subspaces.

Full experimental details are provided in Appendix A for the primary and tens-column settings, and Appendix B for common-coordinate geometry and shared-generator interventions.

5 RESULTS

5.1 SHARED RELATIONS SUPPORT HELD-OUT PREDICTION AND OUTPUT CHANGES

At each model's exploratory reporting layer, interventions using Canonical-predicted held-out directions yield higher exact complete target-answer match rates than Transfer and Pooled (Table 2). Canonical approaches Direct in Llama, Phi-4, and OLMo.

Table 2: Exact complete target-answer match rates (%) at each model's post-hoc peak-Canonical layer, using matched target problems.

Model	Layer	Pooled	Transfer	Canonical	Direct (ref.)
Llama-3-8B	L21	25.07	33.27	**99.43**	99.52
Phi-4	L26	30.78	40.23	**98.05**	97.70
OLMo-3-7B	L25	27.70	41.68	**96.76**	97.47
GLM-Z1-9B	L33	18.71	29.66	**71.13**	73.42

At the four reporting layers, Canonical's mean cosine to held-out target directions is 0.8964–0.9619, exceeding Transfer by 0.0534–0.2037 (Table A2). Relations learned from training values predict held-out new-value directions from known source directions. The contrast with Transfer shows that weaker effects of unchanged source directions can coexist with predictable condition dependence. Together with held-out geometric correspondence, the near-Direct output effects in three models support a shared relation with both cross-value predictive content and behavioral relevance.

In GLM, the match rates for Canonical and Direct are 71.13% and 73.42%, respectively, differing by only 2.29 percentage points. Lower intervention success is therefore also present with directions matched to the current computational condition, rather than being specific to cross-condition prediction under this intervention setup. One possible interpretation is that the arithmetic information and interactions relevant to GLM's expected outputs are not fully captured by the current direction-based representation.

5.2 A Shared Generator Supports Output Changes Across Multiple Layers

Across all 20 model–layer settings, interventions using Step-power predictions produce complete target answers more often than Common-Identity, supporting an output effect beyond common-subspace projection alone. Match rates are 90.83%–99.21% within Llama L19–L22, Phi-4 L27–L28, and OLMo L23–L25. Table 3 summarizes four reporting layers.

Table 3: Exact complete target-answer match rates (%) at the exploratory reporting layers in Table 2, using the five-layer runs. Each construction uses 13,500 input–edit pairs and 121,500 route-level records per layer on matched problems and edits.

Model	Layer	Common-Identity	Step-power	Common-GPA
Llama-3-8B	L21	32.62	99.21	99.10
Phi-4	L26	38.89	90.96	95.29
OLMo-3-7B	L25	41.04	93.67	95.95
GLM-Z1-9B	L33	28.63	68.25	68.73

The comparison with Common-GPA tests whether powers of one operator can retain the output effects of condition-specific frames in the same common basis. At Llama L21, Step-power and Common-GPA yield similar match rates (99.21% and 99.10%) with 66 and 594 continuous transformation degrees of freedom, respectively, illustrating that a more compact description can preserve the observed output effects at this layer.

The cost of this constraint varies across layers and held-out values. Relative to Common-GPA, Step-power is lower by 7.27 percentage points at Llama L23 and 4.33 at Phi-4 L26; the latter difference is concentrated in a subset of held-out values. At OLMo L27, both constructions yield about 8.4%, while Direct yields 98.01%, so this decline is not specific to the single-operator constraint (Appendix B.5).

At the four reporting layers, mean cosines between Step-power predictions and held-out target directions are 0.8534–0.9366, exceeding Common-Identity in every model and remaining within 0.030 of Common-GPA (Table B7). Fold-averaged normalized ten-step closure errors are 0.1065–0.2933, compared with normalized single-step distances from identity of 0.7526–1.0434, supporting approximate cyclic consistency of nontrivial transformations. Closure is a post-fit diagnostic: fitting includes $9 \to 0$ without imposing $\widehat{G}^{10} = I_r$. Held-out generator predictions have lower residuals for the arithmetic condition ordering than for refitted shuffled-label controls ($p_{\mathrm{MC}} = 0.001$ at each diagnostic reporting layer; Appendices B.3–B.4). These results support a direction-level description tied to the cyclic ordering of conditions: within the supported layer and value ranges, powers of a shared transformation predict held-out directions and retain the corresponding output effects.

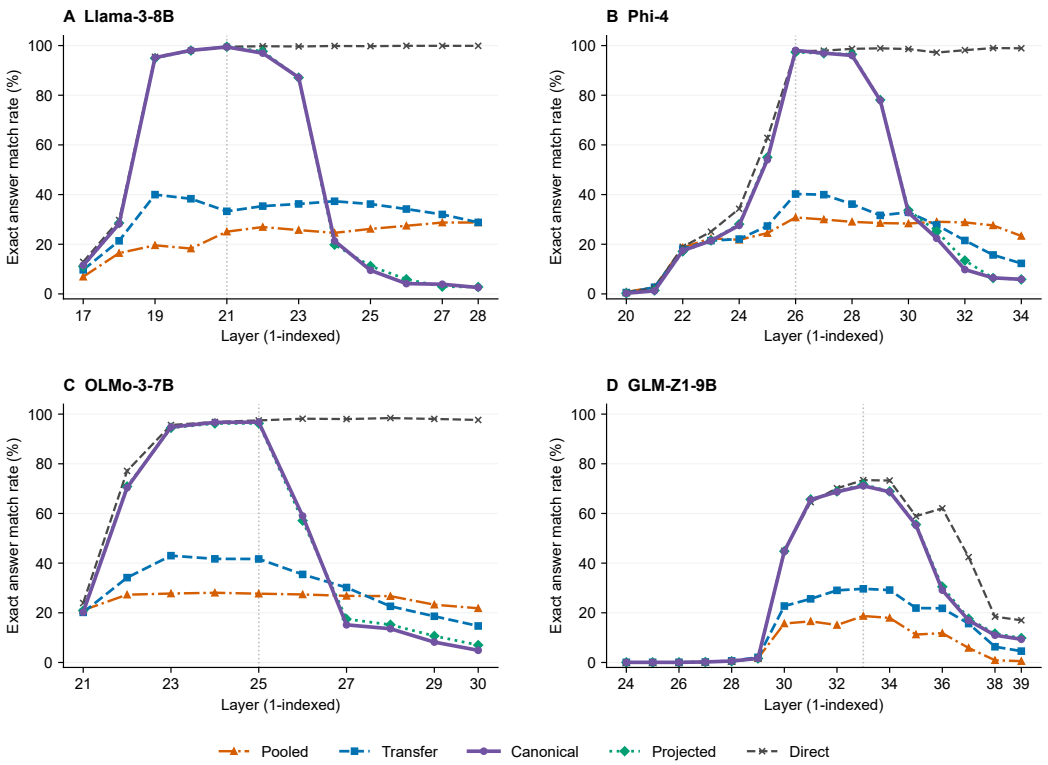


Figure 2: Exact complete target-answer match rates (%) across 53 model–layer settings, aggregated over five held-out folds. Projected and Direct provide subspace and target-direction references. Dashed lines mark the exploratory reporting layers in Table 2; constructions use matched target problems.

5.3 Generalization Across Values Has a Model-Specific Layer Range

Canonical interventions yield complete target-answer match rates of 94.78%–99.43% within Llama L19–L22, Phi-4 L26–L28, and OLMo L23–L25; GLM peaks at L33 (Figure 2). All four models show declines at deeper scanned layers.

At some deeper layers of Llama, Phi-4, and OLMo, Canonical and Projected decline while Direct remains effective. Thus, effective condition-matched directions remain available even where the learned low-dimensional description does not reliably recover their output effects for held-out values. Projected restricts the known target new-value direction to the training subspace in the addition term; its decline, together with reduced coverage and high match rates after all-value fitting, points to a possible training-subspace limitation. The all-value comparison changes multiple fitting and calibration components and does not isolate coverage (Appendix A.7). GLM shows declines in the reference interventions as well.

Geometric proximity alone does not guarantee the expected output effect: at Llama L17, the mean cosine between Canonical predictions and held-out target directions is 0.8699, yet Canonical interventions produce the complete target answer in only 11.52% of test records. The layer range supporting the shared description must therefore be established through both held-out geometric prediction and direct behavioral evaluation.

5.4 The Shared-Relation Pattern Extends to Tens-Column Values

After independent refitting in the tens-value/hundreds-condition setting, interventions using Canonical-predicted held-out directions yield higher exact complete target-answer match rates than Transfer in all four models (Table 4). Llama, Phi-4, and OLMo approach Direct; GLM remains

below Direct (71.14% versus 82.07%). Recurrence after changing both the represented column and the column defining the condition extends the shared-relation pattern beyond the ones-value/tens-condition pairing.

Table 4: Exact complete target-answer match rates (%) for independently refitted tens-value/hundreds-condition directions, using matched problems and edits at exploratory layers.

Model	Layer	Transfer	Canonical	Direct (ref.)
Llama-3-8B	L22	39.37	94.42	95.19
Phi-4	L27	37.67	95.46	96.56
OLMo-3-7B	L25	37.43	96.69	98.03
GLM-Z1-9B	L33	41.75	71.14	82.07

6 Discussion and Limitations

6.1 Shared Relations as Interpretable Structure

Our findings suggest that interpretable representational structure need not depend on feature directions remaining fixed across conditions. At some layers in controlled addition, directions for the same arithmetic value vary with condition, while directions for different values change according to shared, predictable patterns. The object of interpretation can therefore extend from a single direction to a family of condition-dependent directions and their relations. Context dependence thus provides a way to understand how representations are organized: how the same information is expressed systematically under different conditions.

6.2 From Shared Relations to a Shared Generator

Canonical and Step-power characterize this organization at two levels: shared relations across values, and a common rule across cyclic condition displacements. The latter imposes a stronger constraint by linking multiple conditional relations through powers of one operator within each model, layer, and fold. Where supported, approximate conditional equivariance offers a compact direction-level account of how conditions organize a family of value directions.

6.3 Connecting Structure to Model Behavior

Held-out direction prediction tests whether shared relations generalize beyond training values; interventions test whether predicted directions induce the expected complete target answers. Together, these tests go beyond identifying individually effective directions, connecting a predictive account of representational organization to model behavior. Layer-wise evaluation delineates where both forms of evidence support this account.

Scope and future directions. Our evidence comes from controlled addition in the tested models, layers, and numerical tokenization settings, with generalization assessed across values between known conditions. Whether similar shared relations support direction prediction and behavioral interventions under other numerical tokenization schemes, including digit-level tokenization, and in natural-language semantics, factual knowledge, or open-ended reasoning remains to be tested.

7 Conclusion

In controlled addition, shared conditional relations predict held-out arithmetic directions that support expected output changes under intervention at appropriate layers. Shared-generator results further support direction-level approximate conditional equivariance at some layers. These findings suggest that interpretable structure need not require fixed feature directions: it can also lie in the predictable relations organizing their variation across contexts.

Reproducibility Statement

Sections 3 and 4 specify the direction constructions, held-out protocols, and intervention procedures. Appendix A provides sampling budgets, model identifiers, numerical settings, calibration, and output scoring; Appendix B details the shared-generator analyses and layer-wise results. The supplementary material includes experiment code, frozen arithmetic inputs, and reported-result summaries.

Ethics Statement

This work studies internal representations in pretrained language models through controlled arithmetic experiments using programmatically generated inputs. It does not involve human participants or personal data. We do not identify specific ethical concerns arising from this study.

AI Use Statement

Generative AI tools were used to assist with manuscript drafting and language editing, LaTeX formatting, literature organization, coding assistance for method implementation and debugging, and experimental prioritization. The authors take full responsibility for the final content of the paper.

References

Marah I Abdin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. doi: 10.48550/arXiv.2412.08905. URL https://arxiv.org/abs/2412.08905.

Laziz U. Abdullaev, Noelle Y. L. Wong, Ryan T. Z. Lee, Shiqi Jiang, Khoi N. M. Nguyen, and Tan M. Nguyen. Concept heterogeneity-aware representation steering. *arXiv preprint arXiv:2603.02237*, 2026. doi: 10.48550/arXiv.2603.02237. URL https://arxiv.org/abs/2603.02237. ICML 2026.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Xiaoyan Bai, Itamar Pres, Yuntian Deng, Chenhao Tan, Stuart M. Shieber, Fernanda B. Viégas, Martin Wattenberg, and Andrew Lee. Why can't transformers learn multiplication? reverse-engineering reveals long-range dependency pitfalls. *arXiv preprint arXiv:2510.00184*, 2025. doi: 10.48550/arXiv.2510.00184. URL https://arxiv.org/abs/2510.00184.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. URL https://openreview.net/forum?id=JZiKuvIK1t.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024. URL https://openreview.net/forum?id=kvcbV8KQsi.

John C Gower. Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. doi: 10.1007/BF02291478. URL https://doi.org/10.1007/BF02291478.

Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. URL https://arxiv.org/abs/1812.02230.

Zhimin Hu, Lanhao Niu, and Sashank Varma. The representational geometry of number. *arXiv preprint arXiv:2602.06843*, 2026a. URL https://arxiv.org/abs/2602.06843.

Zhimin Hu, Lanhao Niu, and Sashank Varma. Language models represent and transform concepts with shared geometry. *arXiv preprint arXiv:2607.04525*, 2026b. URL https://arxiv.org/abs/2607.04525.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition. *arXiv preprint arXiv:2502.00873*, 2025. URL https://arxiv.org/abs/2502.00873.

Jorrit Kruthoff. Carrying over algorithm in transformers. *arXiv preprint arXiv:2401.07993*, 2024. doi: 10.48550/arXiv.2401.07993. URL https://arxiv.org/abs/2401.07993.

Jiaqian Li, Yanshu Li, and Kuan-Hao Huang. Steering vector fields for context-aware inference-time control in large language models. *arXiv preprint arXiv:2602.01654*, 2026. URL https://arxiv.org/abs/2602.01654.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Llama Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Siddarth Mamidanna, Daking Rai, Ziyu Yao, and Yilun Zhou. All for one: LLMs solve mental math at the last token with information transferred from other tokens. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30747–30760. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1565. URL https://aclanthology.org/2025.emnlp-main.1565/.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2025. URL https://openreview.net/forum?id=O9YTt26r2P.

Narmeen Oozeer, Shivam Raval, Philip Quirke, Manikandan Ravikiran, Jeff M. Phillips, Shriyash Upadhyay, and Amirali Abdullah. Riemannian-manifold steering: Geometry-aware generative autoencoders for label-free steering. *arXiv preprint arXiv:2605.24942*, 2026. doi: 10.48550/arXiv.2605.24942. URL https://arxiv.org/abs/2605.24942.

Gustaw Opielka, Hannes Rosenbusch, and Claire E Stevenson. Causality $\neq$ invariance: Function and concept vectors in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=LmLmhb6GEL.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585. URL https://doi.org/10.2202/1544-6115.1585.

Philip Quirke and Fazl Barez. Understanding addition in transformers. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024. URL https://openreview.net/forum?id=rIx1YXVWZb.

Shivam Raval, Hae Jin Song, Linlin Wu, Abir Harrasse, Jeff M. Phillips, Fazl Barez, and Amirali Abdullah. Curveball steering: The right direction to steer isn't always linear. *arXiv preprint arXiv:2603.09313*, 2026. doi: 10.48550/arXiv.2603.09313. URL https://arxiv.org/abs/2603.09313.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. URL https://doi.org/10.18653/v1/2024.acl-long.828.

Pau Rodríguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2025. URL https://openreview.net/forum?id=l2zFn6TIQi.

Raphaël Sarfati, Eric Bigelow, Daniel Wurgaft, Siddharth Boppana, Jack Merullo, Atticus Geiger, Owen Lewis, Tom McGrath, and Ekdeep Singh Lubana. The shape of beliefs: Geometry, dynamics, and interventions along representation manifolds of language models' posteriors. *arXiv preprint arXiv:2602.02315*, 2026. doi: 10.48550/arXiv.2602.02315. URL https://arxiv.org/abs/2602.02315.

Peter H Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451. URL https://doi.org/10.1007/BF02289451.

Michal Štefánik, Timothee Mickus, Marek Kadlčík, Bertram Højer, Michal Spiegel, Raúl Vázquez, Aman Sinha, Josef Kuchař, Philipp Mondorf, and Pontus Stenetorp. Language models learn universal representations of numbers and here's why you should care. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30663–30681. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1415. URL https://aclanthology.org/2026.acl-long.1415/.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7035–7052. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.435. URL https://aclanthology.org/2023.emnlp-main.435/.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/fb3ad59a84799bfb8d700e56d19c231b-Abstract-Conference.html.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. doi: 10.48550/arXiv.2512.13961. URL https://arxiv.org/abs/2512.13961.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2025. URL https://openreview.net/forum?id=8WQ7VTfPTl.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/252a3dbaeb32e7690242ad3b556e626b-Abstract.html.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric J. Bigelow, Raphaël Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah D. Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior. *arXiv preprint arXiv:2605.05115*, 2026. doi: 10.48550/arXiv.2605.05115. URL https://arxiv.org/abs/2605.05115.

Z.ai. GLM-Z1-9B-0414 model card, 2025. URL https://huggingface.co/zai-org/GLM-Z1-9B-0414. Accessed September 24, 2026.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. doi: 10.48550/arXiv.2406.12793. URL https://arxiv.org/abs/2406.12793.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/56cbfbf49937a0873d451343ddc8c57d-Abstract-Conference.html.

Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. Pre-trained large language models use Fourier features to compute addition. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/2cc8dc30e52798b27d37b795cc153310-Abstract-Conference.html.

Fangwei Zhu, Damai Dai, and Zhifang Sui. Language models encode the value of numbers linearly. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 693–709. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.47/.

Andy Zou, Long Phan, Sarah Li Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song,

Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A Technical Configuration and Supplementary Results

A.1 Sample Configuration and Independent Calibration Checks

Table A1: Sample budgets and aggregate evaluation sizes. Each condition–value centroid is estimated from 200 fitting problems, giving 38,000 fitting problems across 10 conditions and 19 values. Formal output evaluation covers all 54 held-out directed value edits under 10 target conditions, yielding 13,500 input–edit pairs for every model. Canonical and Transfer each evaluate every input–edit pair from nine source conditions; these route-level evaluations are not treated as independent samples.

Model	Fit / (c, k)	Fit total	Test / target–edit	Formal test total	Route evals / layer
Llama-3-8B	200	**38,000**	25	**13,500**	121,500
Phi-4	200	**38,000**	25	**13,500**	121,500
OLMo-3-7B	200	**38,000**	25	**13,500**	121,500
GLM-Z1-9B	200	**38,000**	25	**13,500**	121,500

The formal test totals in Table A1 count input–edit pairs: an input can recur under different edits. For each target condition–edit pair, we use the first 25 problems in the frozen sample order, shared across constructions. These totals correspond to 13,224 distinct ordered (a, b) inputs for every model.

Main calibration uses 20 problems per target condition and old value for Llama and 5 for the other models; independent checks use 10 for all models, with inputs reused across the corresponding new values. Each branch uses 48 calibration edits for Llama and 24 for the other models, and 24 independent-check edits for every model.

For the original 47-layer scan, Pooled comes from a separate experiment using shared directions at both endpoints. It uses frozen samples that are byte-for-byte identical to those of the corresponding main experiment, with the same task configuration, while re-extracting directions and independently calibrating coefficients. Both endpoints use the shared dictionary; all five constructions use saved predictions on the matched test problems specified in the table. Comparisons use the same target problems, but direction extraction and scale calibration belong to their respective runs.

The four sample roles within each experiment package have zero overlap in ordered (a, b) pairs; (a, b) and (b, a) are treated as different inputs. Main calibration and independent checks use nonoverlapping sets of training edits, each covering the training values of that branch. We search all 16 combinations of $\alpha, \beta \in \{0.5, 1, 1.5, 2\}$. Main calibration selects by the number of exact answer matches; exact ties are resolved by taking the first combination ordered by α, then β. Each construction selects its coefficients independently.

Independent checks use L17, L22, and L28 in Llama; L20, L22, and L25 in the Phi-4 extension, and L26, L30, and L34 in its original scan; L21, L25, and L30 in OLMo; and L24, L31, and L39 in GLM. These additional samples assess the stability of scale selection as a training-value diagnostic. For a two-fold cross-fitted diagnostic reference, coefficients selected on each half of the independent-check questions are evaluated on the other half, without changing the formal coefficients. These checks do not replace the formal held-out test and are not used to change formal coefficients, reselect reporting layers, or add samples.

Shared numerical settings are seed 42, rank 12, GPA reference condition 0, a maximum of 200 iterations, and relative-change tolerance 10^{-9}. The exact prompt is `The sum of {a} and {b} is ` , including the trailing space, with no chat template or added special tokens and with left padding. Model weights use BF16; intervention terms are computed in FP32 and written back in BF16. Inference uses SDPA, greedy decoding, a single beam, at most two new tokens, and the model's EOS configuration.

A.2 Aggregation, Reuse, and Paired Comparisons

The five folds do not contain exactly the same number of held-out values. We aggregate folds by their actual numbers of test records rather than taking an unweighted average of five match rates. At each layer, Canonical and Transfer cover all sources, targets, edits, and associated test problems; source-independent constructions are evaluated once per target problem. Because each target has the same number of nine sources, equal reuse of these predictions across routes does not change their match rates. Paired differences between constructions are differences between success indicators on the same problems and edits. Differences between all-value and held-out branches are also computed problem by problem over the 54 shared edits.

Reusing an input across constructions, sources, layers, or branches does not increase the number of independent samples. Main-text tables and curves describe fixed fits over the current scan ranges. We do not construct ordinary confidence intervals that treat all inference records as independent Bernoulli samples, or treat reporting-layer peaks as independently validated layer selections. Cross-fitting in the independent calibration check examines coefficient selection only; it does not replace value holdout or uncertainty analysis based on refitting.

A.3 Geometry at Reporting Layers and Across Scanned Layers

Table A2 reports direction cosines at the main-text reporting layers, and Table A3 summarizes all scanned layers of each model.

Table A2: Held-out direction cosines at the main-text reporting layers. Each model and construction includes 1,710 direction comparisons; mean differences use unrounded values.

Model	**Layer**	**Transfer**	**Canonical**	**Canonical − Transfer**
Llama-3-8B	L21	0.7910	0.9619	+0.1709
Phi-4	L26	0.8430	0.8964	+0.0534
OLMo-3-7B	L25	0.7534	0.9082	+0.1548
GLM-Z1-9B	L33	0.7097	0.9134	+0.2037

Table A3: Mean held-out direction cosines across all scanned layers of each model. Each layer has the same number of comparisons, so layers are equally weighted. Scan ranges differ across models. This table describes each model's scanned results and should not be conflated with the single-layer results in Table A2.

Model	**Layers scanned**	**Transfer**	**Canonical**	**Canonical − Transfer**
Llama-3-8B	12	0.7269	0.7475	+0.0207
Phi-4	15	0.7536	0.7542	+0.0006
OLMo-3-7B	10	0.7069	0.7289	+0.0220
GLM-Z1-9B	16	0.7144	0.7902	+0.0759

A substantial geometric difference at reporting layers can coexist with a smaller average difference across the full scan. Across Phi-4 L20–L34, the mean difference is only 0.0006. At L25, Canonical has a lower mean cosine than Transfer (0.9300 versus 0.9433), yet its interventions yield more exact target answers (54.05% versus 27.33%). Functional relevance must still be evaluated through complete answers; we do not directly pair a geometric difference averaged across layers with a functional difference at one peak layer.

A.4 Definition of the Coverage Diagnostic

Figure A2 uses the squared projection norm of a unit direction, $\|u_{j,k}B_j\|_2^2$. For each model and layer, we average the 190 target directions defined by 10 target conditions and 19 values. A projection repeated across source routes or constructions is counted only once. The held-out scheme uses the

training center and basis of the corresponding fold for each value; the all-value scheme uses the all-value center and basis.

Let $P_j = B_j B_j^\top$, $\rho = \|uP_j\|_2 > 0$, and $\widetilde{u} = uP_j/\rho$. Any unit prediction w in that subspace satisfies

$$\cos(w, u) = \rho \cos(w, \widetilde{u}) \leq \rho. \tag{8}$$

This is a geometric bound on direction cosine, not an upper bound on exact answer match rate. Differences in match rates among Direct, Projected, and Canonical can diagnose functional changes across direction constructions, but do not constitute an exact additive decomposition of the errors.

A.5 Run Configuration

The attention backend is explicitly set to SDPA for all four models. The model identifiers are listed below; tokenizers were loaded from the corresponding model directories. The original four-model main runs used PyTorch 2.6.0+cu124 and Transformers 5.5.4.

Model	Checkpoint identifier
Llama	`meta-llama/Meta-Llama-3-8B`
Phi-4	`microsoft/phi-4`
OLMo	`allenai/Olmo-3-1025-7B`
GLM	`zai-org/GLM-Z1-9B-0414`

Two additional frame-and-template updates follow the main GPA iterations. Normalization uses a norm floor of 10^{-8}. The unmodified baseline describes original answering ability and targets the original answer $a + b$, which differs from the target after the hidden-state intervention.

Table A4 summarizes saved baseline outputs in the primary setting. Each run counts each unique ordered input (a, b) in the reported formal test set once, without duplication across edits, source routes, or layers. Outputs use the prompt and greedy decoding protocol in Appendix A.1; the first parsed signed integer is scored against $a + b$, with parsing failures counted as incorrect. The three Phi-4 batches use the same test inputs and are reported separately; agreement on these baseline outputs does not verify checkpoint identity.

Table A4: Unmodified original-answer accuracy in the primary setting. The first four rows use the original non-Pooled main runs; the last two report the separate Phi-4 Pooled and extension batches. Counts refer to unique ordered inputs, and denominators are not pooled across runs.

Model	Run	Correct / unique inputs	Accuracy (%)
Llama-3-8B	Original main	13,224 / 13,224	100.000
Phi-4	Original main	13,224 / 13,224	100.000
OLMo-3-7B	Original main	13,177 / 13,224	99.645
GLM-Z1-9B	Original main	12,551 / 13,224	94.911
Phi-4	Pooled L26–L34	13,224 / 13,224	100.000
Phi-4	L20–L25 extension	13,224 / 13,224	100.000

A.6 Output-Category Breakdown

Figure A1 breaks down intervention outcomes at the reporting layers by output category, distinguishing complete target-answer matches from outputs with only the target ones digit correct.

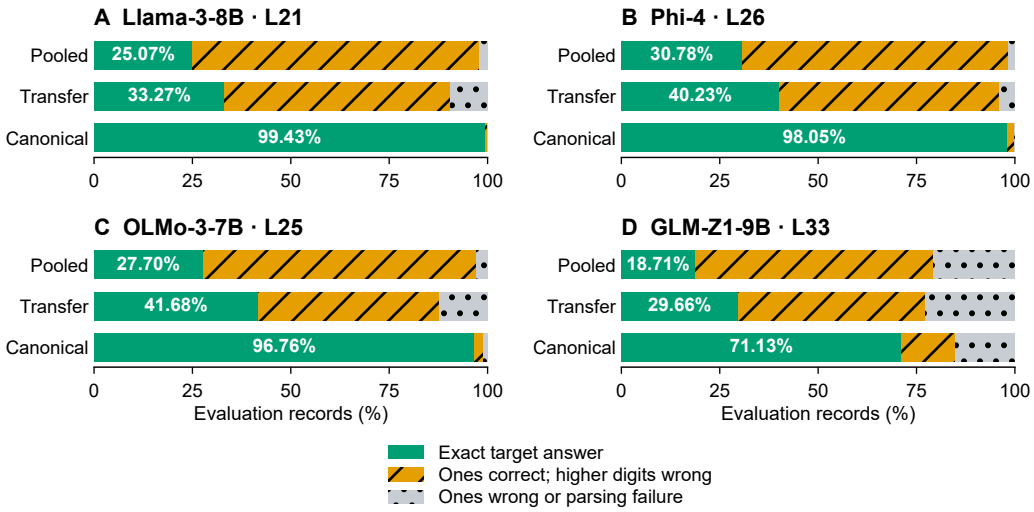


Figure A1: Output-category breakdown at the reporting layers. Each bar partitions evaluation records into exact target answers, correct target ones digits with incorrect higher digits, and incorrect ones digits or parsing failures. Numbers indicate exact target-answer match rates.

A.7 Layer-Wise Coverage and All-Value Diagnostics

This section expands the deeper-layer diagnostics summarized in the main text. Coverage and its geometric bound are defined in Appendix A.4; the all-value comparison is a fitting diagnostic rather than a held-out prediction result.

Condition-matched directions remain effective at deeper layers of Llama, Phi-4, and OLMo. At Llama L24, Direct yields 99.83%, whereas Projected and Canonical yield 19.80% and 21.35%, respectively. At L28, Direct remains at 99.89%, compared with 2.79% and 2.57% for the latter two constructions. Direct also remains at 98.90% in Phi-4 L34 and 97.65% in OLMo L30. The concurrent decline of Projected and Canonical, while Direct retains its output effects, identifies the training subspace as a possible limitation to examine.

Keeping the rank fixed, we include all values in centering, subspace estimation, and frame fitting, and evaluate the same 54 edits on the same problems as the held-out branch. The output panels of Figure A2 show that Canonical with all-value fitting remains at 99.62%–99.93% in Llama L23–L28, 98.35%–99.86% in Phi-4 L26–L34, and 95.72%–98.94% in OLMo L23–L30. At these measured layers, including the values in fitting recovers high output match rates.

The geometric diagnostics are consistent with a training-subspace limitation. Coverage energy for Llama held-out target directions decreases from 93.17% at L21 to 40.45% at L24, whereas coverage of the corresponding all-value dictionary is 99.37% and 98.83%. Reduced coverage of held-out directions, declining output match rates with Projected, and high match rates after all-value fitting jointly suggest that a subspace determined by a limited set of training values can constrain generalization at deeper layers. The all-value comparison jointly changes centering, subspaces, shared frames, scales, and coefficient calibration, so it does not isolate the contribution of coverage.

GLM exhibits a different decline. Its Canonical exact answer match rate decreases from 71.13% at L33 to 9.34% at L39; Direct and all-value Canonical at L39 also fall to 16.99% and 16.88%, respectively. In GLM, the deeper-layer decline thus extends to condition-matched directions and all-value fitting. Together, the scans and diagnostics show that generalization across values has model- and layer-specific limits.

A.8 Direction Fitting and Geometric Evaluation

Each condition–value centroid averages 200 sampled problems, whose specific addends may differ within that combination (Table A1). Within each value fold, we recompute the training center of

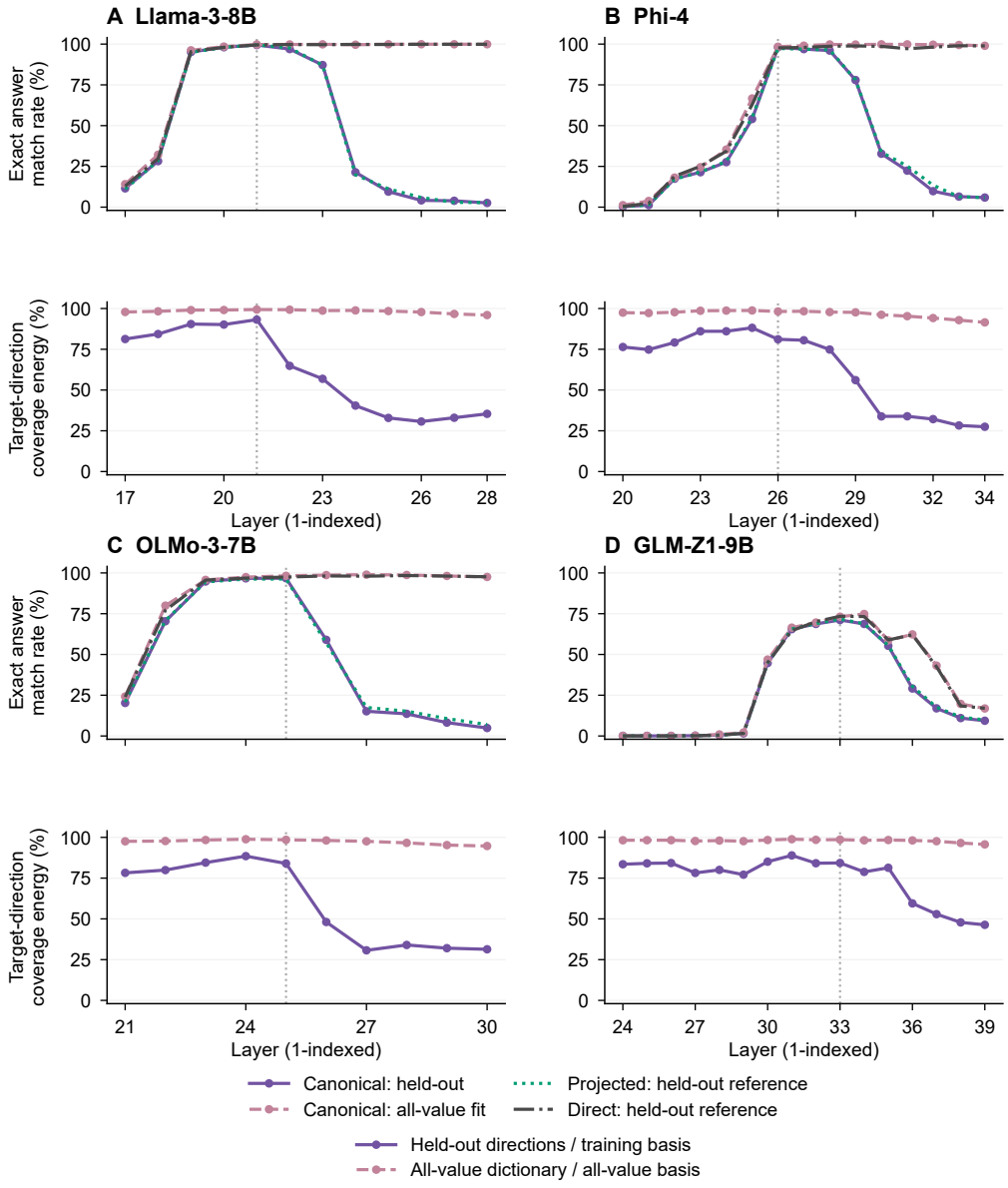


Figure A2: Layer-wise output effects and subspace coverage under held-out and all-value fitting. Within each model group, top: Canonical on held-out values, Canonical with all-value fitting, and Projected and Direct references from the held-out branch. The two Canonical variants are paired problem by problem over the 54 shared edits. Bottom: Mean coverage energy $\|u_{j,k}B_j\|_2^2$ of target directions in the corresponding subspaces. Held-out curves use each fold's training basis and held-out target directions; all-value curves use an all-value basis and the dictionary included in fitting. Each layer averages 190 distinct target condition–value directions. Vertical dashed lines mark the reporting layers in Table 2.

every condition, center both training and held-out centroids relative to it, and estimate the subspace bases and shared frames using training directions only. The source direction of a held-out value remains available for prediction; its target direction is excluded from fitting the center, subspace, and shared transformation.

This objective brings the aligned coordinates of the same training value across conditions close together. Generalized Procrustes analysis alternates between updating the condition frames and the common template: with the template fixed, we fit an orthogonal transformation for each condition, then update the template by averaging the transformed coordinates. All training values within a condition jointly determine one frame for subsequent held-out direction prediction.

The rank, reference condition, stopping tolerance, and maximum iteration count are specified in Appendix A.1; the final frame-and-template updates and numerical normalization convention are described in Appendix A.5.

Geometric evaluation compares each prediction with the separately estimated held-out direction in the target condition. The Transfer and Canonical metrics are, respectively,

$$\cos(u_{i,k}, u_{j,k}), \qquad \cos(\widehat{u}^{\text{Canonical}}_{j\leftarrow i,k}, u_{j,k}). \tag{9}$$

These cosines quantify correspondence with the target direction.

A.9 Intervention Scaling, Pooled Directions, and Output Evaluation

A.9.1 Scale and Pooled dictionary

The scale σ_j is the mean distance of training-value centroids from the training center in the current fold. Omitting the fold index,

$$\sigma_j = \frac{1}{|\mathcal{K}_{\text{tr}}|} \sum_{v \in \mathcal{K}_{\text{tr}}} \|\mu_{j,v} - \bar{\mu}_{j,\text{tr}}\|_2. \tag{10}$$

Pooled averages centered centroid displacements across the ten conditions before normalization:

$$p_k = \text{unit}\left[\frac{1}{10}\sum_{c=0}^{9}(\mu_{c,k} - \bar{\mu}_{c,\text{tr}})\right]. \tag{11}$$

Both old-value removal and new-value addition use this condition-averaged dictionary:

$$h' = h - \beta\sigma_j p_{k_{\text{old}}} + \alpha\sigma_j p_{k_{\text{new}}}. \tag{12}$$

Within a fold, p_k does not vary with the target condition; σ_j and the independently calibrated α, β are still determined for each target condition. Thus, the dictionary is shared across conditions, while the intervention scale and coefficients remain condition dependent. The Pooled run and its frozen-sample correspondence with the other constructions are specified in Appendix A.1.

Transfer and Canonical use the same source-condition direction for the new value; Canonical estimates its frames and subspaces from training values only. Pooled's shared dictionary includes both old- and new-value information from the target condition. Projected and Direct also use target new-value information. All constructions other than Pooled remove an old-value direction matched to the target condition, including for held-out old values during evaluation. The held-out prediction test concerns the new-value direction and its effect on outputs. Projected projects only the addition term.

A.9.2 Calibration and output scoring

For each model, layer, held-out fold, target condition, and construction, α, β are selected independently on training-value edits. The selected coefficients are shared across source conditions and

test edits. Both the old and new values in the formal test belong to the held-out set of that fold. The model resumes generation from the modified state using greedy decoding for at most two new tokens, and we extract the first signed integer from its output. A record is successful only if that integer equals s_{target} exactly. Parsing failures and problems that the unmodified model answered incorrectly remain in the denominator. Exact answer scoring includes any valid carry caused by the target change.

The calibration grid, tie-breaking, and independent-check protocol are specified in Appendix A.1; record aggregation, source reuse, and paired comparisons follow Appendix A.2.

A.10 Tens-Column Extension: Geometry and Diagnostic Scope

The second setting uses the same Canonical fitting and intervention logic as the primary experiment, with the represented value T and condition H defined in Section 4. The reported layers in Table 4 are exploratory choices for this setting. All five conditions $H = 4, \ldots, 8$ are known during fitting; values held out by $T \bmod 5$ do not contribute to the rank-12 bases or Procrustes maps. A held-out source-condition direction remains available to the prediction, and the target-condition old-value direction remains available for removal. Direct uses the target-condition new-value direction as its reference.

Sampling and evaluation budgets. We use $H \in \{4, 5, 6, 7, 8\}$ to retain sufficient samples in each condition–value cell while separating fitting, calibration, independent checks, and testing. Each model and layer has 6,750 input–edit pairs drawn from 6,490 distinct ordered addend pairs. Five target conditions with four other source conditions give 20 directed condition routes. Canonical and Transfer each contribute 27,000 route-level records; Direct is evaluated once per input–edit pair, with a denominator of 6,750. The constructions use the same target problems and edits. Intervention and output-scoring procedures follow Section 4.

Geometric correspondence. Table A5 reports mean cosine similarity to the held-out target-condition direction at the same layers as the behavioral comparison. In Llama, Phi-4, and OLMo, Canonical has both higher cosine and higher exact-answer match rate than Transfer. In GLM, Canonical has a higher answer match rate despite a slightly lower mean cosine.

Table A5: Mean cosine similarity to held-out target-condition directions for the tens-column extension at the exploratory layers in Table 4.

Model	Layer	Transfer	Canonical
Llama-3-8B	L22	0.763	0.958
Phi-4	L27	0.778	0.959
OLMo-3-7B	L25	0.725	0.926
GLM-Z1-9B	L33	0.923	0.916

Subspace and layer diagnostics. At GLM L33, the Projected target-direction reference matches 72.49% of complete target answers, compared with 71.14% for Canonical and 82.07% for Direct. Restricting the known target direction to the training subspace thus reduces its output effect to a rate close to Canonical. In Llama, the all-value Canonical branch reaches 98.33% at L28, but the held-out branch matches only 16.05% there; its reported held-out peak is 94.42% at L22. These observations show that a layer with strong all-value fitting need not support held-out value generalization. The fitting components changed by the all-value comparison are described in Appendix A.7.

B Common-Coordinate Geometry and Generator Interventions

This appendix supplements the shared-generator model with geometric diagnostics in common coordinates and an intervention evaluation at five layers per model. The geometric scan uses the same four models and five-fold value split as the main text, retaining the original 47-layer subset of the

expanded 53-layer main scan: Llama L17–L28, Phi-4 L26–L34, OLMo L21–L30, and GLM L24–L39. This gives 235 layer–fold fits. Phi-4 L20–L25 is not included in these geometric diagnostics. Quantities in this scan are computed from cached directions. Reporting layers remain those in Table 2 and are not reselected using the geometric results or the additional interventions.

Table B1 compares held-out direction predictions at the reporting layers, and Figure B1 shows the shared step operator's spectrum. Figure B2 compares geometry with the original Canonical, Direct, and Transfer intervention results across the 47-layer range. Common-Identity, Step-power, and Common-GPA interventions at five layers per model are reported in Appendix B.5, with further geometric and route diagnostics in Appendix B.6.

B.1 Gauge Dependence of Local Coordinates and a Common Basis

Suppose each condition-specific basis is independently changed to $B'_c = B_c O_c$, where O_c is orthogonal. Taking $Q'_c = O_c^\top Q_c$ gives

$$\begin{aligned} R'_{i\to j} &= O_i^\top R_{i\to j} O_j, \\ B'_i R'_{i\to j} B'^\top_j &= B_i R_{i\to j} B_j^\top, \\ R'_{i\to j} R'_{j\to t} &= O_i^\top R_{i\to t} O_t = R'_{i\to t}. \end{aligned} \tag{13}$$

The ambient map is therefore unchanged, and the composition identity guaranteed by the frame construction transforms covariantly. However, O_i and O_j may differ, so the first line is generally not a similarity transformation, and the spectrum of $R_{i\to j}$ lacks the same invariance. Merely changing the signs of SVD columns demonstrates this: all 235 sign-flip probes across folds change the mean distance of local-map eigenvalue angles from the nearest 36-degree root grid, while the maximum absolute change in ambient direction predictions is numerically zero. This conclusion concerns spectral interpretations of operators expressed in independently chosen endpoint coordinates; it does not invalidate the original ambient direction predictions.

To compare spectra and operators in common coordinates, we stack the training unit directions from the ten conditions within each fold, take the leading $r = 12$ right singular vectors as a common basis B, and define $Y_{c,\mathrm{tr}} = U_{c,\mathrm{tr}} B$. We retain each condition-specific center determined by the training values of that fold. Held-out values do not enter fitting of the centers, common basis, or transformations. The stacked matrix contains 150 training directions in each of the first four folds and 160 in the last fold. These directions are not independent samples.

Replacing the separate condition-specific subspaces with one common subspace is an additional representational assumption and therefore requires reporting coverage; it is not merely a change of local coordinates. Once that common subspace is fixed, a shared basis change $B' = BO$ transforms operators as $R' = O^\top R O$, preserving their spectra. Common coordinates are determined directly from the training directions, without first using GPA alignment to remove the conditional differences being tested.

B.2 Shared Step Fitting and Held-Out Prediction

Let $G \in O(r)$ be the shared orthogonal step operator, fitted using the training directions of all adjacent conditions:

$$\widehat{G} = \operatorname*{arg\,min}_{G^\top G = I_r} \sum_{c=0}^{9} \left\| Y_{c,\mathrm{tr}} G - Y_{(c+1) \bmod 10,\mathrm{tr}} \right\|_F^2 . \tag{14}$$

The objective is solved by SVD for orthogonal Procrustes (Schönemann, 1966). Fitting includes the $9 \to 0$ relation but does not impose $G^{10} = I_r$. For source condition i, target condition j, and held-out value k, prediction uses

$$\widehat{u}^{\,\mathrm{step}}_{j\leftarrow i,k} = \mathrm{unit}\left[(u_{i,k} B) \widehat{G}^{\,t} B^\top \right], \qquad t = (j - i) \bmod 10. \tag{15}$$

When closure is not exact, exponents modulo 10 are not fully equivalent to integer exponents without the modulo operation, so we also report ten-step closure error. Under exact closure with frames constructed from a single generator, the convention consistent with the main-text row vectors is $Q_c = G^{-c}$, giving $Q_i Q_j^\top = G^{j-i}$.

All comparison models are fitted in the same common basis. Common-basis GPA uses the main-text common-template objective; common-basis pairwise fits an orthogonal map independently for each condition pair, with the reverse map defined as its transpose; the identity reference uses I_r. Every construction is evaluated on the same 90 directed condition pairs and 19 values held out in their respective folds, giving 1,710 cosines of full ambient directions. Direction metrics are weighted by the five folds' held-out value counts, $4/4/4/4/3$; operator and spectral metrics are averaged equally across folds. Original Transfer reuses the ambient source direction directly and is reported separately because its representational construction differs.

Table B1: Subspace coverage and held-out direction prediction at the original reporting layers. Coverage is the squared projection norm of a unit target direction; all other values are cosines of full directions, not answer match rates. Local coverage uses the separate condition-specific bases of the main text; common coverage uses the single basis of this appendix. Identity, step powers, common GPA, and common pairwise share the same common basis; original Transfer is reported separately.

Model	Layer	Local coverage	Common coverage	Identity	Step powers	Common GPA	Common pairwise	Original Transfer
Llama	21	0.9317	0.9080	0.7573	0.9366	0.9502	0.9502	0.7910
Phi-4	26	0.8113	0.7990	0.7496	0.8830	0.8903	0.8903	0.8430
OLMo	25	0.8397	0.7921	0.6875	0.8534	0.8829	0.8829	0.7534
GLM	33	0.8429	0.7934	0.6310	0.8684	0.8839	0.8840	0.7097

Coverage with the common basis is 0.0123–0.0495 lower than with the local bases, supporting a shared low-dimensional approximation of the sampled directions at these layers without establishing equality of the condition subspaces. With the common basis fixed, $\dim O(12) = 66$: the step operator has 66 continuous transformation degrees of freedom; the ten common-GPA frames have $(10-1)66 = 594$ after removing the global orthogonal gauge; and the 45 independent pairwise maps, with transposed reverses, have $45 \times 66 = 2970$. These counts exclude parameters for the common basis, condition centers, and direction estimation. Discrete determinant branches do not change the dimension.

The cosine decrease from common GPA to step powers is 0.0073–0.0296, indicating that far fewer transformation degrees of freedom retain most of the held-out prediction accuracy. The difference from original Transfer also involves the choice of representation and cannot be attributed entirely to cyclic structure. Rank 12 is inherited from the main experiment; optimality and stability across ranks have not been tested.

B.3 Nonidentity, Closure, and Observed Spectral Deviations

Define $D_1 = \|G - I_r\|_F/\sqrt{r}$ and $D_{10} = \|G^{10} - I_r\|_F/\sqrt{r}$ to measure distance from identity and ten-step closure error, respectively. In Table B2, D_1 ranges from 0.7526–1.0434, so small closure errors at the reporting layers cannot be explained by generators close to identity. Relative to the identity map, the absolute reductions in held-out normalized squared residual are 0.3745, 0.2982, 0.3684, and 0.5295, respectively; these are not percentage improvements. Within each fold, the residual is computed as $E = \sum_{i\neq j,k} \|x_{i,k}G^t - x_{j,k}\|_2^2 / \sum_{i\neq j,k} \|x_{j,k}\|_2^2$, with $x_{c,k} = u_{c,k}B$, then weighted by the held-out value count.

Let $e^{i\theta_\ell}$ be an eigenvalue of G, with signed angular deviation from the nearest tenth root of unity $\delta_\ell = \mathrm{Arg}(e^{10i\theta_\ell})/10$. Then

$$D_{10}^2 = \frac{1}{r}\sum_{\ell=1}^{r} 2\left[1 - \cos(10\delta_\ell)\right]. \tag{16}$$

An equivalent angle inferred from closure error summarizes aggregate deviation and does not bound the deviation of every eigenvalue. Applying this relation across folds also requires averaging D_{10}^2 first. We therefore report errors for individual eigenvalues directly and do not claim that every spectral angle lies within 2° of a tenth root.

Most spectral points are close to 0°, while nontrivial components often lie near $\pm 72^\circ$ and 180°. Phi-4 fold2 lacks a block near $\pm 72^\circ$; its nearest-root pattern has only order 2. OLMo fold0 has an angular

Table B2: Nonidentity and spectral deviations of the shared step operator at reporting layers. D_1 and D_{10} are means over five folds. Angular deviation is the distance of each eigenvalue from the nearest tenth root of unity. The mean, RMS, and maximum retain all 60 eigenvalues across the five folds, including multiplicities. These 60 spectral points are not independent replications.

Model	D_1	D_{10}	Mean angle (°)	Angle RMS (°)	Max. angle (°)	Count above 2°
Llama	0.9476	0.2933	0.713	1.844	6.403	8/60
Phi-4	0.9191	0.1264	0.312	1.057	5.504	2/60
OLMo	0.7526	0.1753	0.520	2.180	11.855	2/60
GLM	1.0434	0.1065	0.316	0.630	2.007	2/60

deviation of about $11.85°$. All five OLMo folds and three GLM folds have determinant -1, so we use the general term orthogonal transformations. Nearest-root patterns are descriptive; operators are not rounded to exact group representations. Fitting already includes all adjacent relations, including $9 \to 0$, so the spectra and closure do not independently validate a held-out wraparound relation.

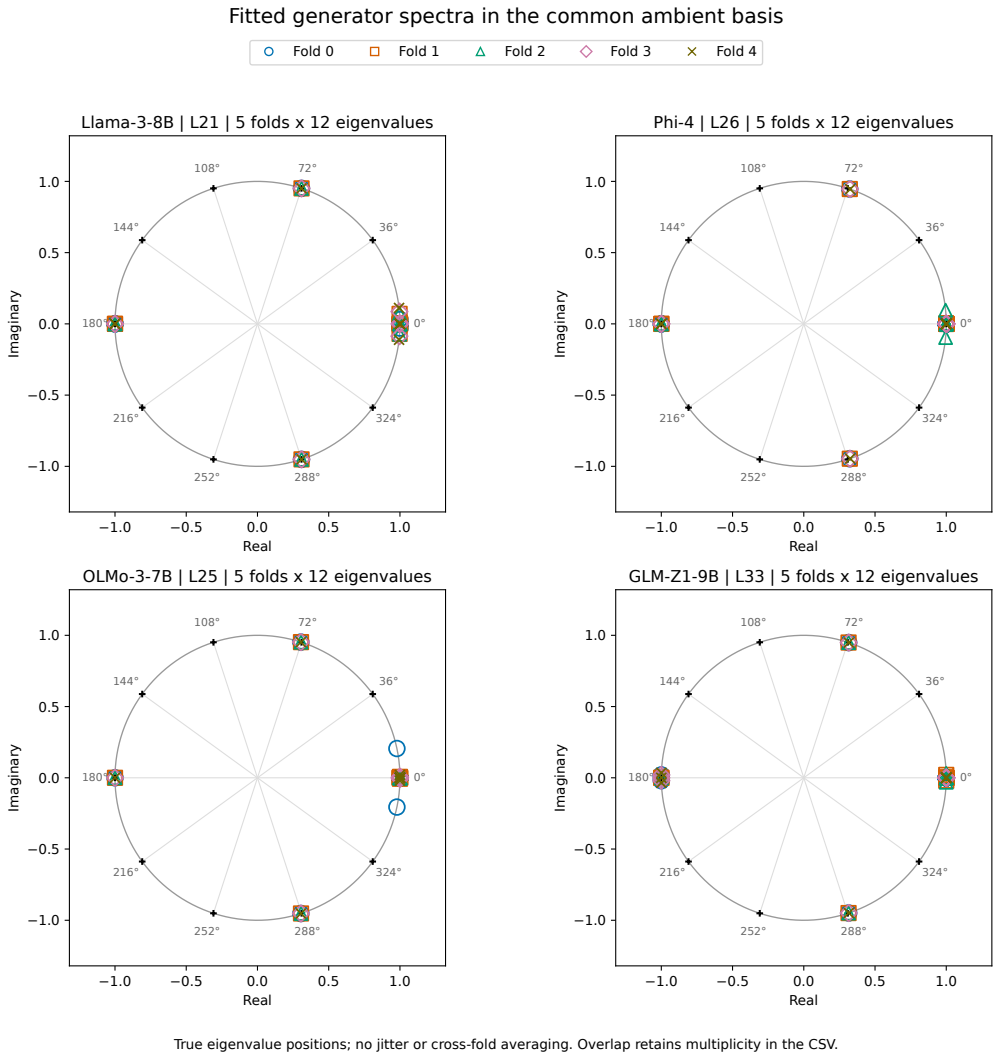


Figure B1: Complex-plane spectra of the shared step operators at reporting layers. Each panel contains 12 eigenvalues from each of five folds, distinguished by color and marker; the ten markers on the circle indicate tenth roots of unity. Positions are not jittered, matrices are not averaged across folds, and ± 1 and all multiplicities are retained, so some points overlap. Nontrivial components often occur near $\pm 72^\circ$ and 180°, with variation across folds and outlying angular deviations (Table B2).

Table B3: Distinguishing the spectra of condition-pair maps from powers of the generator. The first four metric columns average over all 90 directed condition pairs, all eigenvalues, and five folds, in degrees. The grid order specific to displacement t is $10/\gcd(10, t)$. The final column is the mean normalized Frobenius distance between step powers and common-GPA maps.

Model	GPA: 36-degree grid	GPA: displacement grid	Pairwise: 36-degree grid	Pairwise: displacement grid	Powers vs. GPA distance
Llama	4.315	7.304	4.217	7.174	0.6061
Phi-4	2.548	3.899	2.607	3.962	0.3395
OLMo	4.787	6.931	4.762	7.004	0.6259
GLM	5.107	7.978	4.995	8.114	0.6053

Condition-pair spectra of common GPA and common pairwise deviate more from the root grids than the generators themselves (Table B3). The final column is $\mathrm{mean}_{f,i\neq j}\,\|G_f^{(j-i) \bmod 10} - R^{\mathrm{GPA}}_{f,i\to j}\|_F/\sqrt{12}$, which remains 0.3395–0.6259. Thus, similar predictions on the sampled held-out directions coexist with appreciable operator differences over the full 12-dimensional common subspace: Step-power provides a compact predictor, while the common-GPA maps retain additional operator variation.

B.4 Comparisons Across Layers and Statistical Conventions

Table B4: Descriptive comparisons of extrema within the scanned layer ranges. Original Canonical answer match rate, step-power direction cosine, and closure error each use their own extremum. These extrema are not used to redefine reporting layers or select effective windows post hoc.

Model	Scan	Highest original Canonical match	Highest step-power cosine	Lowest closure error
Llama	17–28	21	21	19
Phi-4	26–34	26	27	27
OLMo	21–30	25	23	23
GLM	24–39	33	31	31

Within the scanned ranges, descriptive Spearman correlations between step-power cosine and original Canonical match rate are 0.853, 0.983, 0.818, and 0.474, respectively; their peaks do not generally coincide (Table B4). Adjacent layers are correlated, and reporting layers were already selected post hoc using the original intervention results. We therefore do not interpret these layer correlations as independent significance evidence or as evidence for generator effects on outputs. The complete curves also show weak cyclic structure at some early layers, declining coverage at deeper layers, and differences between models.

The auxiliary condition-label permutation tests use 999 random permutations with seed 20260918, sharing each permutation across folds and layers. For the two common-basis condition-pair constructions, the statistic is the difference between mean normalized squared Frobenius distances for different-displacement and same-displacement groups. The step model is refitted under each permutation using training values only and evaluated through the weighted held-out residual. Statistics are first aggregated across folds; one-sided values are then computed as $p_{\mathrm{MC}} = (1 + b)/(999 + 1)$ (Phipson & Smyth, 2010). Benjamini–Hochberg adjustment (Benjamini & Hochberg, 1995) covers all 141 values from the two condition-pair tests and the step-prediction test across all 47 layers. At each of the four reporting layers, all three tests have 0 more-extreme permutations out of 999, giving $p_{\mathrm{MC}} = 0.001$ and adjusted values of approximately 0.0013. This is the Monte Carlo resolution limit and must not be reported as $p < 0.001$. The comparison establishes structure relative to shuffled labels; it does not replace prediction comparisons within a common basis or functional validation.

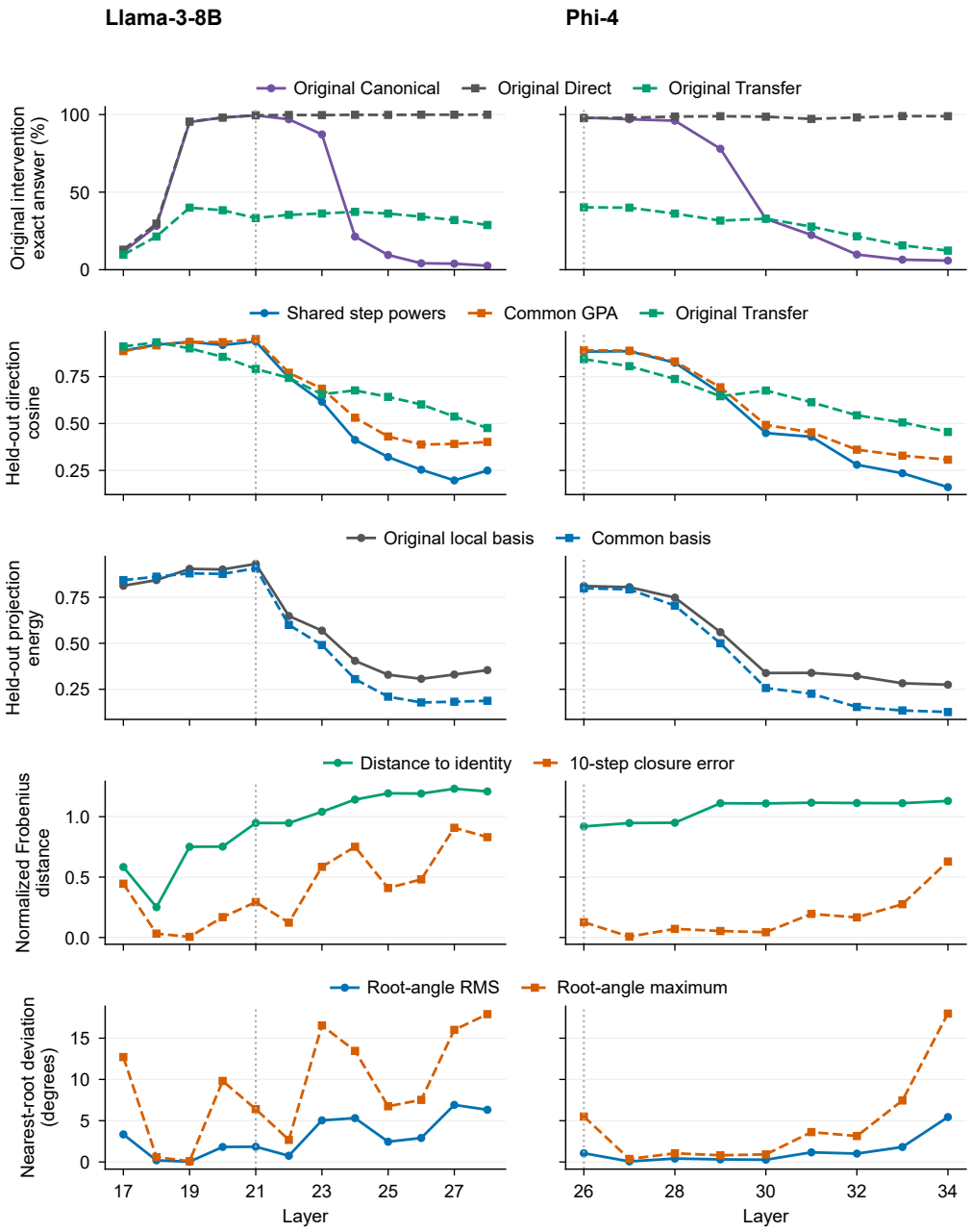


Figure B2: Geometry and original intervention results across this appendix's 47 scanned layers. Columns correspond to models. The top row shows exact answer match rates for the original Canonical, Direct, and Transfer interventions. The remaining rows show held-out direction cosines; subspace coverage energy; shared-generator distance from identity and ten-step closure error; and RMS and maximum spectral deviations from the nearest root, following this appendix's protocol. Vertical dashed lines mark the original reporting layers. The Phi-4 range is L26–L34.

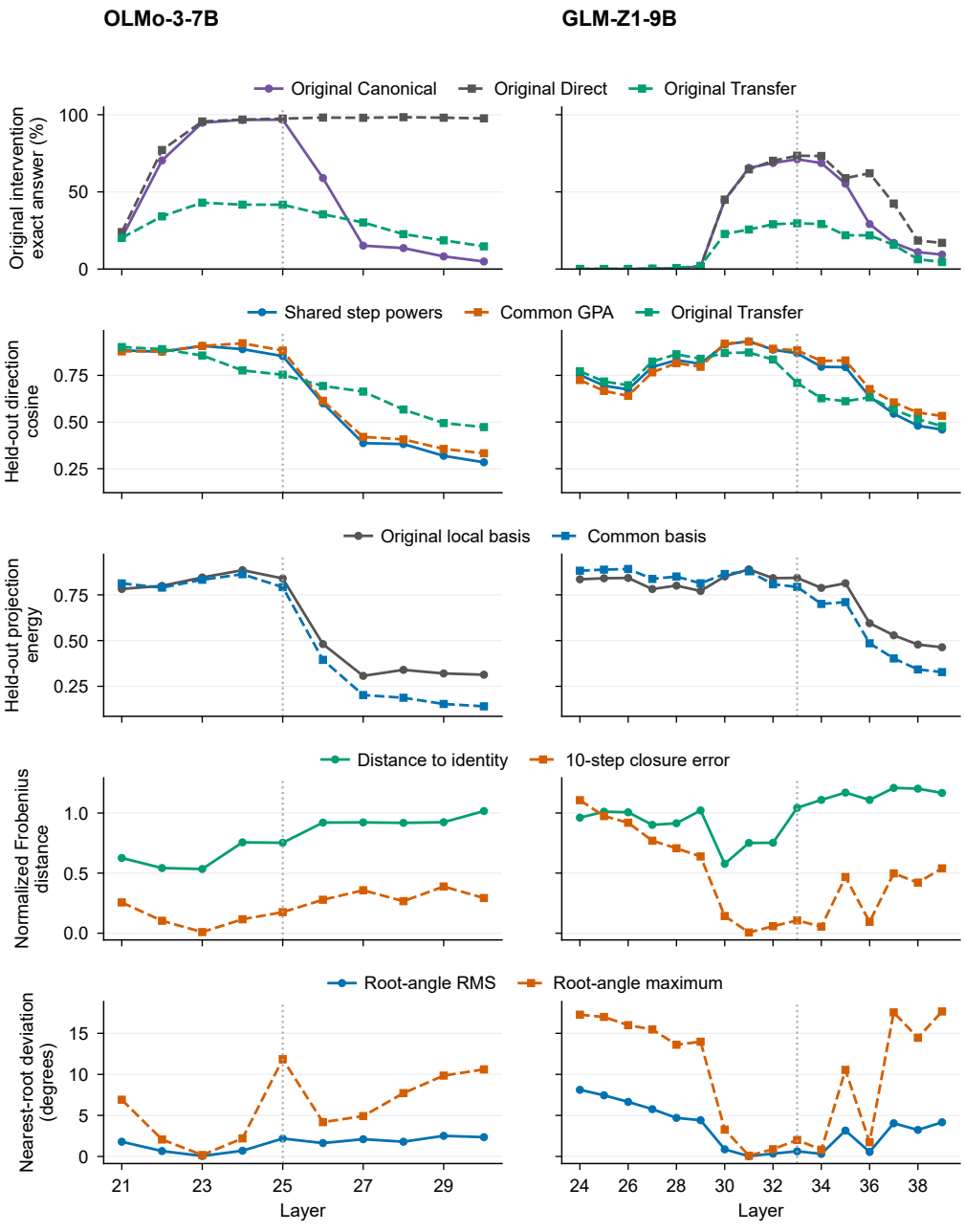


Figure B2: Geometry and original intervention results across the 47-layer scan (continued): OLMo-3-7B and GLM-Z1-9B. Rows, line styles, and vertical reporting-layer markers follow the preceding page.

B.5 Five-Layer Interventions with Common-Basis Predictions

We evaluate Common-Identity, Step-power, and Common-GPA at Llama L19–L23, Phi-4 L24–L28, OLMo L23–L27, and GLM L31–L35, using the constructions in Section 3.3 and the intervention interface in Section 3.4. The compact main-text summary uses Llama L21, Phi-4 L26, OLMo L25, and GLM L33, matching the exploratory reporting layers selected post hoc by peak Canonical match rate in the original scans (Table 2). These layers are not reselected by the five-layer results. The generator is fitted separately for every model, layer, and value fold. This evaluation is distinct from the 47-layer geometric scan.

Frozen geometry and intervention inputs. Within each value fold, the three constructions use the same training-value centers, common rank-12 basis, and source directions. The common basis, shared step operator, and common-GPA frames are estimated from training values and frozen before intervention calibration. Held-out target new-value directions do not enter their fitting or the predicted addition term. All three constructions retain the target-condition old-value removal direction and training-value scale σ_j defined in Appendix A.9. The operator predicts the direction added to the residual state; it is not applied to the entire hidden state. Step fitting observes all ten conditions, including $9 \to 0$, without imposing $\widehat{G}^{10} = I_r$.

Calibration and independent checks. For each model, layer, fold, target condition, and construction, we search the 16 combinations $\alpha, \beta \in \{0.5, 1, 1.5, 2\}$. Coefficients are selected by the number of complete target-answer matches on training-value calibration edits; exact ties use the first pair ordered by α, then β. The selected pair is shared across all nine source conditions and formal test edits for that fold and target. Each model uses 24 calibration edits per fold and 5 problems per target condition and old value, together with 24 independent-check edits and 10 problems per target condition and old value, reusing inputs across the corresponding new values. Checks evaluate the selected coefficients on separate training edits and do not reselect them. Fitting, calibration, independent checks, and formal testing use disjoint ordered addend pairs, with reuse within a role as in Appendix A.1. Formal test results do not change the geometry, coefficient grid, selected coefficients, or reporting layer. The three common-basis constructions therefore use equal preset calibration budgets, not identical selected intervention magnitudes.

Formal test and output scoring. Every distinct old–new value pair within a held-out fold is tested. The fold sizes $4, 4, 4, 4, 3$ give $12, 12, 12, 12, 6$ directed edits, respectively, for 54 edits in total; their change magnitudes are 5, 10, or 15. Each edit is evaluated under ten target conditions and nine distinct source conditions, covering all 90 routes. Each model uses 25 problems per target–edit pair. All three constructions use the same frozen problems, edits, and routes within a model. Generation uses the prompt and decoding settings in Appendix A.1: no chat template, one hidden-state intervention, and greedy generation of at most two new tokens. Following Appendix A.9, success requires the first parsed signed integer to equal $s_{\text{target}} = s + k_{\text{new}} - k_{\text{old}}$ exactly. Parsing failures and problems answered incorrectly before the intervention remain in the denominator.

For each model and layer, the shared-generator comparison evaluates 54 directed held-out value edits under 10 target conditions with 25 problems per edit, giving 13,500 input–edit pairs. Each is evaluated from nine distinct source conditions, yielding 121,500 route-level records. There are 13,224 distinct ordered addend pairs. The first four value folds contain 27,000 route records each and the final fold contains 13,500; aggregate rates use these record counts.

Aggregation and paired comparisons. Exact-match rates use actual test-record counts, with fold weights $2/9, 2/9, 2/9, 2/9, 1/9$. Geometry instead averages over the 1,710 unique route–value predictions, with fold weights $4/19, 4/19, 4/19, 4/19, 3/19$; operator and spectral diagnostics retain the equal-fold convention of the geometric scan. Paired behavioral differences compare success indicators for the same problem, edit, and route. All folds, values, and routes remain in the main aggregation. Repeated inputs, shared fitted directions, and source reuse do not create independent observations. These are descriptive comparisons of fixed fits, without confidence intervals or significance tests that treat route records as independent Bernoulli samples.

Evaluation runs and reference predictions. The shared-generator intervention results reported here use the five-layer evaluation runs. Common-Identity, Step-power, and Common-GPA were

Table B5: Formal evaluation sizes for the five-layer common-basis interventions. Counts are per layer and construction. Input–edit pairs can reuse an input under different edits; each pair is evaluated from nine source conditions. Route records are not independent samples.

Model	Layers	Problems / target–edit	Input–edit pairs	Route records	Distinct inputs
Llama	L19–L23	25	13,500	121,500	13,224
Phi-4	L24–L28	25	13,500	121,500	13,224
OLMo	L23–L27	25	13,500	121,500	13,224
GLM	L31–L35	25	13,500	121,500	13,224

evaluated in these runs; Canonical, Transfer, and Direct use frozen reference predictions matched by problem, edit, fold, and condition route. The Llama comparison uses the same formal test set as the main direction comparison, with 25 problems per target-condition edit. Run configurations are recorded with the corresponding experimental artifacts.

Direct's source-independent output is expanded over nine source routes for logical pairing. The common-basis constructions are newly evaluated on NVIDIA A100-PCIE-40GB GPUs; the frozen references retain their historical execution settings. The within-common-basis comparison uses the same problems and edits, while comparisons with original Canonical also involve its condition-specific subspaces and historical predictions. The Llama direction fits are unchanged; its three common-basis constructions use the calibration budget above, which is smaller than Llama's main-comparison budget. The frozen references retain the coefficients selected for the main comparison (Appendix A.1). Matching evaluation records does not imply bitwise equivalence across execution environments.

Complete layer-wise results. Table B6 gives all 20 model–layer settings. Step-power has a higher complete target-answer match rate than Common-Identity in each setting. The relative output effects of Step-power and Common-GPA vary across models and layers.

Table B6: Exact complete target-answer match rates (%) across the five-layer shared-generator evaluation. Id: Common-Identity; Step: Step-power; GPA: Common-GPA; Can: Canonical; Tr: Transfer; Dir: Direct. The first three constructions use the common-basis evaluation runs; the last three are matched frozen references. Each row and construction contains 121,500 route-level records. Llama uses the same formal test set as the main direction comparison.

Model	Layer	Common-Identity	Step-power	Common-GPA	Canonical	Transfer	Direct
Llama-3-8B	L19	41.37	92.48	91.89	95.15	39.98	95.44
Llama-3-8B	L20	41.86	95.62	96.05	98.08	38.31	98.03
Llama-3-8B	L21	32.62	99.21	99.10	99.43	33.27	99.52
Llama-3-8B	L22	15.04	97.69	97.04	96.95	35.36	99.73
Llama-3-8B	L23	8.06	77.41	84.69	87.16	36.23	99.64
Phi-4	L24	21.12	26.28	27.90	27.61	22.08	34.28
Phi-4	L25	27.95	54.06	54.34	54.05	27.33	62.83
Phi-4	L26	38.89	90.96	95.29	98.05	40.23	97.70
Phi-4	L27	37.70	98.64	97.26	96.94	39.96	97.94
Phi-4	L28	28.74	98.55	98.07	96.03	36.16	98.68
OLMo-3-7B	L23	42.99	93.52	93.50	94.78	43.01	95.61
OLMo-3-7B	L24	42.80	90.83	96.23	96.68	41.72	96.84
OLMo-3-7B	L25	41.04	93.67	95.95	96.76	41.68	97.47
OLMo-3-7B	L26	7.79	55.02	52.89	58.97	35.46	98.16
OLMo-3-7B	L27	0.75	8.45	8.44	15.15	30.18	98.01
GLM-Z1-9B	L31	26.99	63.54	63.23	65.53	25.61	64.59
GLM-Z1-9B	L32	28.79	63.24	63.85	68.71	29.04	70.13
GLM-Z1-9B	L33	28.63	68.25	68.73	71.13	29.66	73.42
GLM-Z1-9B	L34	25.80	63.51	65.07	68.71	29.16	73.26
GLM-Z1-9B	L35	19.63	47.72	50.93	55.29	21.91	58.90

Detailed fold, displacement, and route results. The following diagnostics use the same five-layer evaluations as Table B6; all folds and routes remain in the aggregate rates. At Llama L21, Step-power produces the complete target answer in 120,538 of 121,500 records (99.21%), compared with 32.62% for Common-Identity and 99.10% for Common-GPA. Relative to Common-Identity, it has a higher match rate on all 90 routes. Relative to Common-GPA, 38 routes have higher rates, 4 equal rates, and 48 lower rates. At Llama L23, the paired Step-power–Common-GPA differences for f0–f4 are -6.44, -4.03, -4.69, $+0.40$, and -35.94 percentage points. The final fold has 13,500 records, half the count of each other fold.

Value-fold variation at Phi-4 L26. The Step-power–Common-GPA difference at Phi-4 L26 is concentrated in fold f2, which holds out values $2, 7, 12, 17$. Step-power yields 70.35% exact target-answer matching in this fold and is 21.49 percentage points below Common-GPA. Its paired differences in the other four folds are $+0.57$, $+0.79$, $+0.25$, and $+0.76$ percentage points for f0, f1, f3, and f4, respectively. Across the full evaluation, the difference is -4.33 percentage points. At the neighboring layers L27 and L28, Step-power instead yields 98.64% and 98.55%, respectively, with paired differences of $+1.38$ and $+0.48$ points relative to Common-GPA.

At Phi-4 L26, Step-power produces the complete target answer in 110,515 of 121,500 records. Its paired difference from Common-GPA is negative at each of the nine displacements, with differences of about -6.5 to -6.9 percentage points at displacements $2, 3, 7, 8$. There are 13 routes with higher match rates, 3 equal, and 74 lower. At OLMo L25, the Step-power–Common-GPA difference is negative in all five folds (-3.37 to -1.58 percentage points). At GLM L33, the fold differences range from -1.69 to $+1.17$ points. These comparisons describe layer- and value-dependent limits of the fitted constructions. Further geometry and route diagnostics are reported in Appendix B.6.

B.6 Geometric and Route Diagnostics for Shared-Generator Predictions

Evaluation scope. These diagnostics use the five-layer evaluations described in Appendix B.5. The full six-construction behavioral comparison is in Table B6. The compact generator summary uses OLMo L25, matching Table 2; OLMo L23 is retained as an additional diagnostic case.

Direction geometry. Table B7 reports the four summary-layer geometric comparisons and adds Phi-4 L27 and OLMo L23. The common-basis predictions use the corresponding frozen training fits. At Phi-4 L26 and L27, Step-power cosines are similar (0.883 and 0.885), whereas complete target-answer match rates are 90.96% and 98.64%. Geometric correspondence and output effects therefore require separate evaluation. The cyclic-closure and spectral measurements remain those of the original 47-layer scan in Appendix B.3.

Table B7: Mean cosine similarity to held-out target directions for common-basis constructions at the four reporting layers in Table 2, followed by Phi-4 L27 and OLMo L23 as additional diagnostic layers. Each entry averages 1,710 unique route–value predictions. The corresponding five-layer behavioral results are in Table B6. These geometric measurements do not extend the 47-layer cyclic-closure scan.

Model	Layer	Common-Identity	Step-power	Common-GPA
Llama	L21	0.7573	0.9366	0.9502
Phi-4	L26	0.7496	0.8830	0.8903
OLMo	L25	0.6875	0.8534	0.8829
GLM	L33	0.6310	0.8684	0.8839
Phi-4	L27	0.7072	0.8853	0.8880
OLMo	L23	0.8048	0.9076	0.9074

Paired differences and route comparisons. Table B8 reports paired differences from the unrounded records in the five-layer runs. At OLMo L23 and GLM L33, Step-power has a higher match rate than Common-Identity and Transfer on all 90 routes, and a lower rate than Direct on all 90. Relative to Common-GPA, the net difference is 26 successful route records in OLMo and -587 in GLM. The route comparisons and net differences describe fixed fits and reuse the evaluation inputs.

Table B8: Paired differences at OLMo L23 and GLM L33, defined as Step-power minus comparator. Differences are in percentage points. Higher/equal/lower counts compare aggregate exact-match rates over 90 directed condition routes, not individual outcomes or independent statistical tests.

Comparator	OLMo Δ (pp)	OLMo routes higher/equal/lower	GLM Δ (pp)	GLM routes higher/equal/lower
Common-Identity	$+50.526$	90/0/0	$+39.613$	90/0/0
Transfer	$+50.509$	90/0/0	$+38.590$	90/0/0
Common-GPA	$+0.021$	47/7/36	-0.483	33/2/55
Canonical	-1.263	7/0/83	-2.884	19/0/71
Direct	-2.090	0/0/90	-5.174	0/0/90

Displacement averages and route heterogeneity. Table B9 separates variation across displacement averages, directed routes, and held-out folds. Step-power displacement ranges span 0.541 percentage points in OLMo and 1.667 in GLM, compared with spans of 54.711 and 43.207 for Common-Identity and 51.052 and 38.541 for Transfer. Similar displacement averages coexist with route variation: GLM's lowest Step-power route is $9 \to 3$ at 50.593%, and its highest is $6 \to 9$ at 81.556%.

Table B9: Minimum–maximum Step-power exact complete-answer match rates (%) under three groupings. These descriptive ranges are not uncertainty intervals.

Model/layer	Displacement means	Route-level rates	Fold-level rates
OLMo L23	93.222–93.763	90.593–96.074	91.878–96.370
GLM L33	67.548–69.215	50.593–81.556	62.507–70.722