FROM NUMERICAL RULES TO ANSWER TOKENS

Anonymous authors
Paper under double-blind review

ABSTRACT

The effect of an intervention on a language model's hidden state can depend on the token used to report an answer. We test this dependence on parity and divisibility by three, exchanging the answer labels while keeping the number and rule fixed. Class directions fitted under the two mappings yield a symmetric component that keeps its sign under the exchange and an antisymmetric component that reverses it. In Qwen2.5-7B, natural projection removal of the symmetric component at layer 22 lowers the correct-answer logit margin by about three, while removing the antisymmetric component changes it by less than 0.1. At the final layer, their order of influence reverses. This change holds under norm-matched edits, across three calibration splits, and for divisibility under a second answer vocabulary. An exact analysis of the final readout explains the final-layer effect. The intermediate symmetric direction also supports classification across answer vocabularies: a classifier fixed on one vocabulary recovers all 100 held-out parity labels under another mapping where the model's own candidate answers are only 50% accurate.

1 INTRODUCTION

On a parity test with reversed answer labels, Qwen2.5 gives the same candidate answer for all 100 test numbers. Its accuracy is 50%, yet a linear classifier fixed after calibration on a different answer vocabulary recovers every label from layer 22. The labels remain readable under an output convention that the model fails to follow. We use answer exchange to study how this convention affects the directions that control the model's prediction (Figure 1).

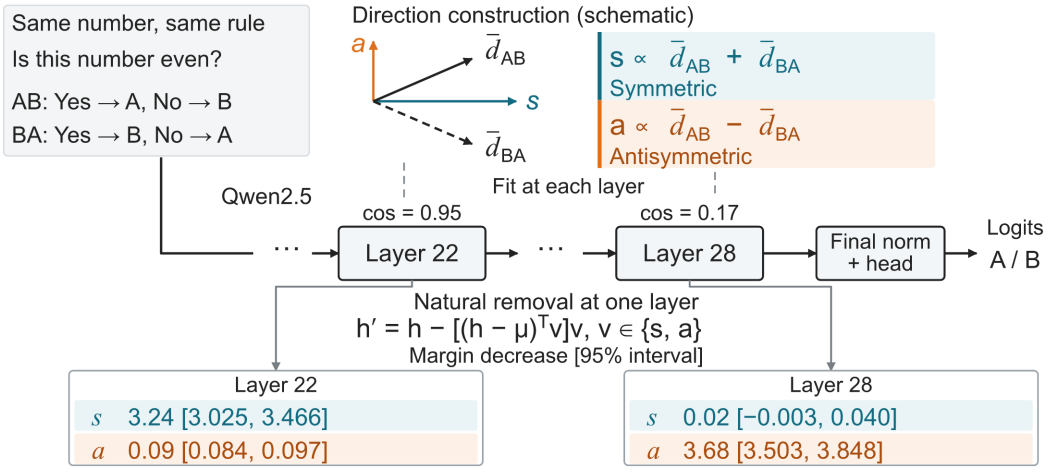


Figure 1: Answer exchange separates a class direction into two components with different intervention effects. AB and BA are evaluated separately in Qwen2.5-7B. The schematic illustrates the per-layer sum and difference construction (Section 3); cosine values are measured. Each edit removes one component at one layer. Numbers give the mean correct-answer margin decrease and 95% bootstrap interval on 100 test numbers, paired across mappings.

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.

Exchanging answer labels changes the required token while preserving the rule's truth value. For example, 12 is even whether the prompt assigns A or B to yes. This gives a controlled way to examine a class direction, the difference between mean representations of positive and negative examples. Class means and probes recover truth-related information, and interventions along concept directions can change predictions (Burns et al., 2023; Marks & Tegmark, 2024; Park et al., 2024). Near the output, an intervention can also act through alignment with the output head. Its interpretation therefore depends on where and along which direction the state is edited (Hase et al., 2023; Makelov et al., 2024; Zhang & Nanda, 2024). We ask how much of the effect changes when the required answer labels are exchanged.

Numerical rules give exact labels and let us hold the input fixed across mappings. Earlier studies established that embeddings and language models contain structured numerical information (Spithourakis & Riedel, 2018; Wallace et al., 2019). Mechanistic work has since identified arithmetic circuits, representations, and learned heuristics (Nanda et al., 2023; Hanna et al., 2023; Stolfo et al., 2023; Nikankin et al., 2025). Studies of multiple-choice questions also locate answer-symbol selection before the final layers (Wiegreffe et al., 2025) and identify attention heads that increase the correct answer label's logit (Lieberum et al., 2023). We examine the binary outcome at the answer position: how does its representation change when the same decision must be expressed by a different token?

We fit the two class directions on the same calibration numbers. Each compares class means after subtracting states from a baseline prompt asking whether the number is an integer. The sum and difference of the normalized directions give symmetric and antisymmetric components. The former keeps its sign when the mappings are exchanged; the latter reverses it. We then edit the same target states along each component, using both natural projection removal and perturbations of matched size. This comparison measures their relative effects without changing the numerical task.

The two components exchange their relative influence between the intermediate and final layers of Qwen2.5-7B. At layer 22, natural symmetric removal lowers the correct-answer margin by 3.24 for parity and 2.92 for divisibility, compared with 0.09 and 0.05 for antisymmetric removal. At the final layer 28, the antisymmetric component has the greater effect. Both mappings are accurate in these conditions. The ordering holds across three calibration splits and, for divisibility, a second answer vocabulary. The final-layer effect follows the components' projections onto the answer readout; the layer-22 effect is measured before the remaining network computes the output. Together with the frozen classifier, this gives two complementary measurements: which label information transfers across mappings, and which directions affect the answer under each mapping.

2 Related work

Numerical representations can support several different computational strategies. Early studies measured whether embeddings and language models preserve magnitude and predict numerical expressions (Spithourakis & Riedel, 2018; Wallace et al., 2019). Interpretable modular-arithmetic models provide detailed examples of learned algorithmic structure (Nanda et al., 2023). Studies of pretrained models instead identify greater-than circuits, arithmetic mediation pathways, and collections of numerical heuristics (Hanna et al., 2023; Stolfo et al., 2023; Nikankin et al., 2025). Our experiments concern the representation of a rule's binary outcome at the answer position. Holding the rule fixed while exchanging its answer labels complements studies that identify the operations used to compute the rule. Recent work reports shared relational structure across numerical task subspaces (Hu et al., 2026), category-boundary geometry that can differ from explicit category behavior (Cacioli, 2026), and class-related residual updates with causal effects on numerical answers (Zhang et al., 2026). We examine which part of such an effect depends on the required answer mapping.

Linear structure provides a useful description of information in language-model representations. Spatial and temporal variables admit structured readouts (Gurnee & Tegmark, 2024), while the linear representation hypothesis relates concept directions to interventions and measurement geometry (Park et al., 2024). Unsupervised consistency objectives and supervised class means can recover truth-related information (Burns et al., 2023; Marks & Tegmark, 2024). Bürger et al. (2024) distinguish general and polarity-sensitive truth directions under negation. Representation engineering also uses concept directions to influence model behavior (Zou et al., 2023). We start from the class-

mean construction and vary the answer mapping while keeping the truth label fixed. The resulting components are then compared through interventions.

Output labels already play a central role in analyses of in-context learning. Demonstrations convey the label space, input distribution, and formatting conventions (Min et al., 2022). Larger models can also learn input-label mappings that oppose familiar semantic associations (Wei et al., 2023). Information-flow analyses show that label-word representations aggregate information used in prediction (Wang et al., 2023). Task vectors and function vectors capture transferable task information in hidden activations (Hendel et al., 2023; Todd et al., 2024). We examine explicit rule prompts without demonstrations and use paired mapping reversal to compare the geometry and causal effect of two components of a class direction.

Intervention methods require controls that distinguish information availability from the role of that information. Probe controls address confounds in decodability (Hewitt & Liang, 2019; Belinkov, 2022). Nullspace projection and linear concept erasure formalize particular removal objectives (Ravfogel et al., 2020; Belrose et al., 2023b). Causal abstraction and alignment methods connect distributed representations to interpretable variables (Geiger et al., 2021; 2024). Activation-patching studies show that the subspace, perturbation, and evaluation metric can change the resulting explanation (Makelov et al., 2024; Zhang & Nanda, 2024). We specify these choices separately: answer-mapping contrasts define the directions, natural projection defines one removal operation, and within-layer norm matching defines a distinct intervention comparison. Intermediate predictions from the tuned lens provide a related view of changing readouts across layers (Belrose et al., 2023a). Here we compare layerwise interventions with an exact calculation of the final head.

3 Comparing directions under answer exchange

3.1 Tasks, representations, and calibration

We study binary numerical rules with deterministic labels. The primary tasks are parity and divisibility by three on integers 1 to 200. Under mapping AB, true answers require A and false answers require B, while mapping BA reverses this assignment. The input number and rule remain fixed. For each mapping, a baseline prompt asks whether the number is an integer using the same answer assignment. This baseline holds the numerical input and output convention constant while changing the requested property.

The representation is the residual state at the final prompt token. Write $h^m_{it\ell} \in \mathbb{R}^d$ for number i, task t, layer ℓ, and mapping m, and $b^m_{i\ell}$ for the corresponding baseline state. Layer ℓ denotes the output of decoder block ℓ. The final-layer state is measured before the last normalization. The task update is

$$u^m_{it\ell} = h^m_{it\ell} - b^m_{i\ell}. \tag{1}$$

We retain both constituent states so that the contribution of baseline subtraction can be measured directly.

Calibration and evaluation use disjoint numbers. A joint stratification by parity and divisibility assigns 100 numbers to each split. Within a task, both mappings use the calibration numbers answered correctly under both mappings. If this set contains fewer than two numbers in either class, both mappings use the full calibration split and the fallback is recorded. All 100 test numbers enter the main evaluation, including initially incorrect answers. No test label enters the edit operation.

The class direction is a difference of mean updates on this common calibration set C:

$$d^m_{t\ell} = \mathbb{E}_C[u^m_{it\ell} \mid y_{it} = 1] - \mathbb{E}_C[u^m_{it\ell} \mid y_{it} = 0]. \tag{2}$$

Subtracting a shared calibration mean update leaves this difference unchanged. The direction also satisfies $d^m = d^m_h - d^m_b$, where d^m_h and d^m_b are the task-state and baseline-state class differences. This identity allows a direct check of whether subtraction substantially changes the class geometry.

3.2 SYMMETRIC AND ANTISYMMETRIC DIRECTIONS

Normalizing the two class directions removes length imbalance from the mapping contrast. Suppressing task and layer indices, let $\bar{d}^m = d^m/\|d^m\|$. We define

$$s_0 = \tfrac{1}{2}(\bar{d}^{\mathrm{AB}} + \bar{d}^{\mathrm{BA}}), \qquad a_0 = \tfrac{1}{2}(\bar{d}^{\mathrm{AB}} - \bar{d}^{\mathrm{BA}}), \qquad s = \frac{s_0}{\|s_0\|}, \quad a = \frac{a_0}{\|a_0\|}. \tag{3}$$

The symmetric component retains its sign when the mappings are exchanged, and the antisymmetric component reverses sign. These are operational definitions under answer exchange. Their interpretation follows from the intervention results.

The normalized construction gives orthogonal components whenever both are nonzero. Indeed, $s_0^\top a_0 = (\|\bar{d}^{\mathrm{AB}}\|^2 - \|\bar{d}^{\mathrm{BA}}\|^2)/4 = 0$. Without normalization, the same inner product equals $(\|d^{\mathrm{AB}}\|^2 - \|d^{\mathrm{BA}}\|^2)/4$, so the raw contrasts need not be orthogonal. We evaluate both constructions and record degenerate directions explicitly. The normalization makes within-layer comparisons easier to interpret without assuming that either component identifies a pure computational factor.

3.3 NATURAL REMOVAL AND MATCHED DIRECTIONAL PERTURBATIONS

Natural projection removal deletes the component already present along a candidate unit direction v. Let μ^m be the mean task state over the full calibration split. For a test state h,

$$\beta_v(h) = (h - \mu^m)^\top v, \qquad h'_{\mathrm{natural}} = h - \beta_v(h)v. \tag{4}$$

The centered projection onto v is then zero. Its removal size is determined by the state itself, so different directions can receive different perturbation norms.

Matched directional perturbations compare directions at equal size within a target condition. We first measure $r(h)$, the actual edit norm from naturally removing the target's native class direction $\bar{d}^m$. Each candidate is then edited as

$$h'_{\mathrm{matched}} = h - \mathrm{sgn}_+(\beta_v(h))\, r(h) v, \tag{5}$$

where $\mathrm{sgn}_+(0) = 1$. This operation uses the candidate projection's sign and the native removal's size. It can exceed the candidate's natural projection. We therefore report the two operations separately, together with actual edit norms and the frequency of crossing the centered projection's zero point.

Controls connect intervention effects to specific alternatives. Every layer includes the unchanged prediction, a no-operation hook, native-axis removal, and eight isotropic random directions with matched edit norms. We also transfer each task's symmetric and antisymmetric directions to the other task. These comparisons test whether the effect is shared across rules or selective for its source rule. Intermediate edits are propagated through the remaining network, while final-layer edits use the actual final normalization and output head.

3.4 MEASURING THE FUNCTIONAL CONTRAST

The main response measure is the decrease in semantic logit margin. Candidate logits are always ordered by their meaning under the current mapping: positive, then negative. For label $y_i \in \{0, 1\}$, define

$$M_i = (2y_i - 1)(z_{i,+} - z_{i,-}), \qquad H_{i,v} = M_i - M'_{i,v}. \tag{6}$$

Positive H denotes damage to the correct answer's margin. Accuracy, balanced accuracy (mean class recall), and prediction flips supplement this continuous measure.

The layer comparison is paired by input number and mapping. For a given task and calibration split, we average $H_{i,a} - H_{i,s}$ over the two mappings at each layer. We then compute

$$I_i = \tfrac{1}{2} \sum_m \left[(H^m_{i,a} - H^m_{i,s})_{\ell_f} - (H^m_{i,a} - H^m_{i,s})_{\ell_m} \right]. \tag{7}$$

A positive mean indicates a shift toward the antisymmetric component's relative effect. We report both mappings separately and give 95% percentile intervals from 20,000 paired bootstrap resamples of test numbers. These intervals are conditional on the fitted calibration directions. Different calibration splits are reported separately, since their test sets overlap.

Table 1: Held-out rule accuracy (%) under forward and reversed answer mappings. Each entry uses 100 integers from split 1. A dagger marks a mapping that gives the same semantic answer for every held-out input.

Model	Labels	Rule	Forward	Reversed
Qwen2.5	AB	Parity	100	94
Qwen2.5	AB	Divisibility by 3	100	99
Qwen2.5	CD	Parity	100	50†
Qwen2.5	CD	Divisibility by 3	100	98
Gemma 2	AB	Parity	86	50†
Gemma 2	AB	Divisibility by 3	63	32†

4 RESULTS

4.1 THE COMPARISON BEGINS WITH MAPPING CAPABILITY

Qwen2.5-7B-Instruct is the primary model, with Gemma-2-9B-it as a second architecture and Qwen3-4B for answer-interface analysis (Qwen et al., 2024; Gemma Team et al., 2024; Yang et al., 2025). All models use single-precision inference without quantization. The mapping experiment uses the stated prompts as raw text. Qwen layers 14, 22, and 28 and Gemma layers 21, 33, and 42 were fixed before the confirmation run. We refit Qwen AB directions under three calibration splits and add a CD/DC vocabulary whose final instruction always lists C before D. Appendix B provides the exact prompts and splits.

Both Qwen AB tasks and Qwen CD divisibility support accurate decisions under both mappings. Table 1 also retains conditions in which answer exchange breaks behavior. In particular, reversed CD parity and the reversed Gemma mappings produce a constant positive answer. These conditions remain in the analysis as boundaries on interpretation. A nonzero intervention contrast alone cannot establish successful rule execution under both mappings.

4.2 NATURAL REMOVAL PRESERVES THE CHANGE IN FUNCTIONAL ROLE

The symmetric and antisymmetric components exchange their relative importance between Qwen's intermediate and final layers. Figure 2 compares their effects on the same test numbers. Under natural projection removal, the layer-22 symmetric component decreases the mean margin by 3.244 for parity and 2.924 for divisibility. The antisymmetric decreases are 0.090 and 0.048. At layer 28, the antisymmetric decreases are 3.683 and 3.439, while the symmetric averages are close to zero. The two mappings separately follow the same ordering at both layers.

The change also affects discrete predictions. For divisibility, natural symmetric removal at layer 22 reduces both mappings to 68% accuracy, from 100% and 99%, while antisymmetric removal preserves their baseline accuracies. At layer 28, symmetric removal retains 100% and 99%, while antisymmetric removal reduces accuracy to 35% and 68%. Margin and accuracy therefore support the same ordering, while revealing different sensitivity to the initial decision margin.

This result survives the controls that directly address the contrast construction. Length balancing makes the two directions orthogonal, with measured absolute inner products below 10^{-15}. Natural removal produces no centered-projection crossings, while matched perturbations often cross zero for small candidate projections. Despite this difference, both operations preserve the intermediate-to-final change. The natural-removal layer interaction is 6.818 for parity and 6.344 for divisibility in the first split. The result thus extends beyond a single large perturbation or an imbalance between the two original class-direction norms.

The small final symmetric effect occurs despite a substantial edit. For parity, the mean natural symmetric edit norm is 50.64, compared with 41.63 for the antisymmetric edit on the same target states. Removing more of the residual state can therefore have less effect on the answer. The final-readout analysis in Section 4.4 explains this difference through the directions' output projections.

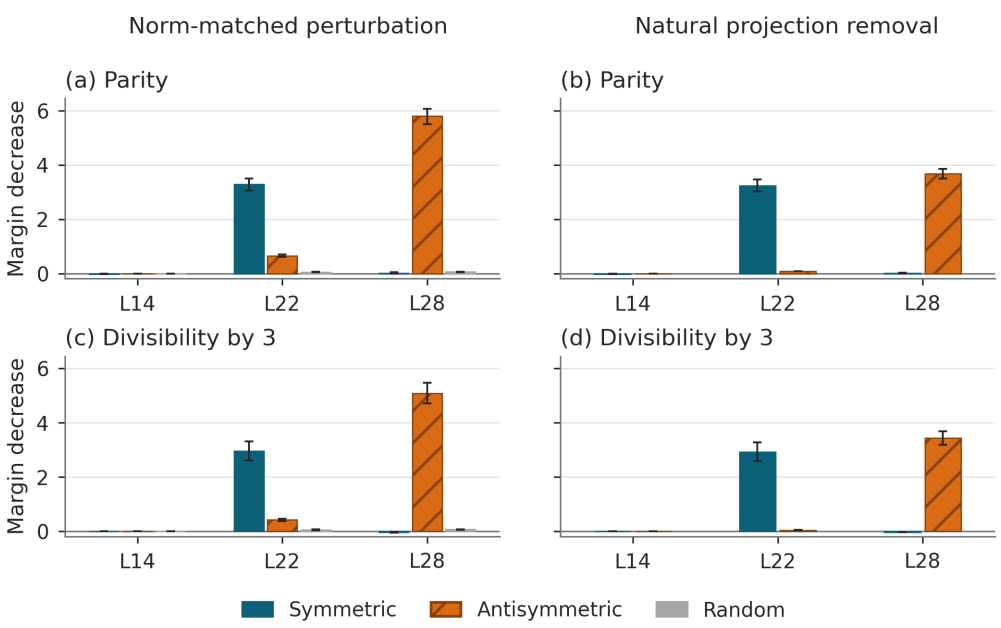


Figure 2: The two components exchange their relative influence between layers 22 and 28 under both interventions. Bars show the mean semantic margin decrease in Qwen2.5 (AB/BA, split 1), paired over answer mappings, with 95% paired-number bootstrap intervals. The random control averages eight directions with matched edit norms. Natural removal uses each direction's own projection, so the two columns differ in intervention size.

The shared effect across rules clarifies what the intermediate direction represents. At layer 22, the symmetric parity and divisibility directions formed from unnormalized class contrasts have cosine similarity 0.947. Their norm-matched transfer between rules is effective, whereas the advantage of using a task's own direction changes sign across mappings. The complete raw and normalized transfer comparisons appear in Appendix G. This pattern supports a shared decision-related component. Across the three AB calibration splits at layers 22 and 28, the task-state and update class directions have cosine similarity above 0.995, and the baseline class-difference norm is below 9.8% of the task-state class-difference norm. Baseline subtraction therefore makes a small change to the geometry underlying the main comparison.

4.3 Calibration and vocabulary controls preserve the effect

Refitting the directions yields the same functional ordering across all three Qwen AB splits. Table 2 reports natural-removal interactions, and Figure 3 shows their paired uncertainty. The parity estimates range from 6.668 to 6.993, and divisibility ranges from 6.344 to 6.604. The matched-perturbation comparison has the same sign for each split. Because both directions are compared on the same target states, this replication tests the fitted geometry without changing the evaluation rule.

Changing the answer vocabulary also preserves the result for divisibility. Under CD/DC, the final instruction lists the same two candidate letters in the same order for both mappings. Both mappings remain accurate, and the natural-removal layer interaction is 3.914. Each mapping separately gives a positive interaction. This condition changes the answer tokens and removes the moving final-option order of AB/BA, while retaining the numerical task. It therefore provides a direct check beyond the original answer vocabulary.

The second architecture gives a useful boundary on the layer interpretation. Gemma has positive middle-to-final interaction estimates, but the antisymmetric component already dominates at the selected intermediate layer, and the reversed mappings produce constant answers. Its result is therefore different from Qwen's change in ordering. Likewise, CD parity retains a positive interaction despite

Table 2: Natural-removal layer interaction for Qwen2.5. Positive values indicate a shift in sensitivity toward the antisymmetric direction from layer 22 to layer 28. Intervals pair both mappings and both layers by integer and use 20,000 bootstrap resamples.

Split	Labels	Rule	Interaction	95% interval
1	AB	Parity	6.82	[6.46, 7.17]
1	AB	Divisibility by 3	6.34	[5.76, 6.95]
2	AB	Parity	6.67	[6.24, 7.06]
2	AB	Divisibility by 3	6.41	[5.79, 7.05]
3	AB	Parity	6.99	[6.67, 7.31]
3	AB	Divisibility by 3	6.60	[6.00, 7.23]
1	CD	Divisibility by 3	3.91	[3.54, 4.30]



Figure 3: Natural-removal layer interactions are positive across the Qwen conditions with accurate answers under both mappings. Points and 95% paired-number intervals show Equation 7; each split refits the class directions. Gray rows retain conditions in which one mapping gives the same answer to every input. Qwen compares layers 22 and 28; Gemma compares layers 33 and 42.

failure of its reversed mapping. Reporting these cases alongside the capable conditions separates a measured intervention effect from a broader claim about successful numerical decisions.

Label information can remain available when an answer mapping fails. In an exploratory analysis of the first Qwen split, we freeze the symmetric direction and a class-midpoint threshold fitted using AB/BA calibration labels. Applied to CD/DC task states at layer 22, this classifier attains 100%/100% accuracy for parity and 100%/99% for divisibility. Reversed CD parity therefore retains recoverable label information despite its constant native answer and 50% candidate-pair accuracy. The source threshold uses no target calibration labels or test labels. Appendix H gives the protocol, the layer-28 results, and an integer-baseline control.

4.4 The final readout explains the output-sensitive effect

The final-layer decision has an exact linear sign representation before root mean square normalization (RMSNorm) (Zhang & Sennrich, 2019). Let g denote its gain, w_+, w_- the two output-weight rows, and $q = g \odot (w_+ - w_-)$. For the bias-free heads used here,

$$\operatorname{sgn}(z_+ - z_-) = \operatorname{sgn}(q^\top h). \tag{8}$$

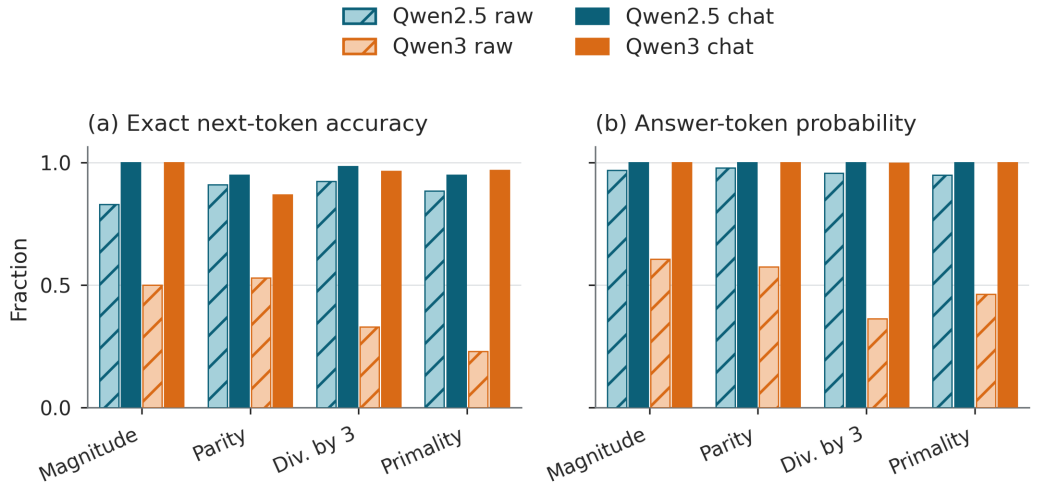


Figure 4: Chat rendering improves answer accuracy and increases probability on the audited answer tokens. Light bars use the raw prompt and dark bars the chat template. (a) Exact next-token accuracy: the greedy token must be an audited answer form for the true class. (b) Mean probability of the audited answer-token set, using the full-vocabulary denominator and deduplicated token IDs. Each value uses all 200 numbers.

The normalization denominator is positive. Gemma's shared monotone logit softcap also preserves this sign. The identity concerns the decision, while the normalized or softcapped margin can have a different magnitude.

Equation 8 makes the effect of natural removal explicit. Editing along v changes the numerator by $-\beta_v(h)q^\top v$. In the two-dimensional span of the orthogonal symmetric and antisymmetric directions, the antisymmetric direction accounts for 99.66% of the readout projection energy for parity and 99.94% for divisibility in Qwen AB. For CD divisibility, the proportion is 97.05%. These are proportions within the two-direction span, rather than proportions of the entire readout. They explain why a geometrically substantial symmetric component can have little final decision effect. In particular, when the two normalized class directions have equal and opposite projections onto a fixed answer readout, their symmetric sum has zero readout projection. Appendix C derives this relation. The final preference is therefore consistent with balanced answer encoding. The intermediate result shows a different functional ordering before the remaining network produces that output.

Direct numerical checks agree with the sign analysis. Across 42,000 final-layer test-condition evaluations, the saved readout sign matches the actual candidate prediction. In the separate four-rule Yes/No experiment, readout-orthogonal edits produce no flips on 400 test cases. For parity and divisibility, restricting the class-axis edit to its readout-parallel component preserves the full edit's predictions on all 100 cases per task, using approximately 1.13% and 0.81% of its squared norm. The final-layer effect is consequently interpretable in terms of a specific output dependency. Intermediate effects require propagation through the remaining network and provide a different functional comparison.

4.5 Answer interfaces determine the behavioral comparison

An answer interface can change which decision a candidate-pair score measures. We compare raw prompts and each model's chat template on magnitude, parity, divisibility, and primality. For each state, we evaluate Yes/No with and without leading spaces, the most probable token over the full vocabulary, and the total probability of the distinct audited answer tokens. Figure 4 reports these quantities without selecting an interface by its accuracy.

Qwen3's raw-prompt failures largely disappear under its chat interface. Exact next-token accuracy changes from 0.50, 0.53, 0.33, and 0.23 to 1.00, 0.87, 0.965, and 0.97 across the four tasks. The

audited answer-token probability also rises toward one. Qwen2.5 illustrates a subtler distinction: under chat rendering, leading-space answer tokens receive very little probability even when their pairwise ranking is accurate. For primality, that pair scores 0.995 while the full-vocabulary next-token accuracy is 0.95. A representation analysis should therefore establish both task capability and the meaning of its answer score.

5 Discussion

Answer exchange turns a semantic distinction into a controlled representation comparison. Truth-direction studies establish that class means and probes can recover information relevant to truth (Burns et al., 2023; Marks & Tegmark, 2024). Label-word and task-vector studies show that output conventions and task information organize hidden states (Wang et al., 2023; Hendel et al., 2023; Todd et al., 2024). Our results connect these views at the answer position. In capable Qwen conditions, a symmetric decision-related component has the stronger intermediate effect, while an antisymmetric component aligned with the answer readout has the stronger final effect. The same truth labels support different functional interpretations at these two sites.

The useful methodological consequence is to evaluate a direction together with its output mapping and intervention. Decodability, projection removal, norm-matched perturbation, and causal abstraction impose different requirements (Belinkov, 2022; Belrose et al., 2023b; Geiger et al., 2021; 2024). Our experiments show why these choices need to remain explicit: large matched edits can cross a component's natural projection, and a low-energy readout-parallel edit can reproduce a final-layer decision change. Mapping controls add a complementary test to the subspace and metric choices emphasized in activation-patching work (Makelov et al., 2024; Zhang & Nanda, 2024).

The frozen classifier gives this distinction a concrete diagnostic use. A failed answer mapping can preserve label information that a classifier calibrated on another vocabulary recovers. This observation connects truth-related decodability (Burns et al., 2023; Marks & Tegmark, 2024) to the answer-interface problem: failure to report a decision does not establish loss of its information. Evaluating a fixed classifier across answer mappings can identify conditions in which labels remain recoverable despite poor answer accuracy. In the reversed parity condition, the labels remain recoverable even when the native candidate answers are constant.

The strongest supported finding concerns the representation of an answer at a fixed token position. Identifying the arithmetic procedure itself requires interventions on the tokens and circuits that compute the rule, as in mechanistic arithmetic analyses (Hanna et al., 2023; Stolfo et al., 2023; Nikankin et al., 2025). The weaker reversed mappings also establish a concrete limit to architectural generalization. Model-dependent truth-direction consistency has also been observed across logical transformations (Bao et al., 2025). Our failed mappings show why a paired causal comparison should establish behavioral capability in both conditions. We retain these conditions and the wider numerical and sentiment results in the appendix. Within this scope, the combination of paired mappings, calibrated directions, natural removal, and exact final readout yields an interpretable comparison that a single class-direction intervention leaves unresolved.

6 Conclusion

Changing answer labels provides a direct test of what numerical decision directions control. In capable Qwen conditions, symmetric and antisymmetric components have different relative effects at intermediate and final layers. The distinction survives calibration changes, natural projection removal, and a second vocabulary for divisibility. Frozen source-vocabulary classification further separates label availability from the model's reported answer. The final readout and answer-interface results explain why output dependence must accompany geometric and causal measurements. These controls make the interpretation of class-related directions more specific and their limits observable.

AI USE STATEMENT

Generative AI tools supported mathematical formalization, experimental code checking, and manuscript polishing. The authors developed the theoretical framework and experimental design, interpreted the results, and verified the citations, equations, code, figures, insights, claims, and prose. The authors take responsibility for all content.

REPRODUCIBILITY STATEMENT

The complete reproduction code and the data underlying all reported figures and tables are provided in the Supplementary Material. The program computes these results from public inputs in one Google Colab cell. Appendix B specifies the prompts, models, data splits, and calibration choices. Appendix C provides the mathematical derivations. Appendices E and G retain the broader experiments and complete mapping conditions, and Appendix I describes the saved outputs and their organization.

REFERENCES

Yuntai Bao, Xuhong Zhang, Tianyu Du, Xinkui Zhao, Zhengwen Feng, Hao Peng, and Jianwei Yin. Probing the Geometry of Truth: Consistency and Generalization of Truth Directions in LLMs Across Logical Transformations and Question Answering Tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 682–700. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.38. URL https://aclanthology.org/2025.findings-acl.38/.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens. arXiv preprint arXiv:2303.08112, 2023a. URL https://arxiv.org/abs/2303.08112.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063, 2023b. doi: 10.52202/075280-2884. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is Universal: Robust Detection of Lies in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f9f54762cbb4fe4dbffdd4f792c31221-Abstract-Conference.html.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Jon-Paul Cacioli. Categorical perception in large language model hidden states: Structural warping at digit-count boundaries. *arXiv preprint arXiv:2603.28258*, 2026. URL https://arxiv.org/abs/2603.28258.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.02997.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume

236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving Open Language Models at a Practical Size. arXiv preprint arXiv:2408.00118, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin,

Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi (Jack) Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon,

Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu (Sid) Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

Wes Gurnee and Max Tegmark. Language Models Represent Space and Time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0a6059857ae5c82ea9726ee9282a7145-Abstract-Conference.html.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.00586.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2301.04213.

Roee Hendel, Mor Geva, and Amir Globerson. In-Context Learning Creates Task Vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. URL https://arxiv.org/abs/2310.15916.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019. URL https://arxiv.org/abs/1909.03368.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. arXiv preprint arXiv:2310.06825, 2023. URL https://arxiv.org/abs/2310.06825.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does Circuit Analysis Interpretability Scale? Evidence from Multiple Choice Capabilities in Chinchilla. arXiv preprint arXiv:2307.09458, 2023. URL https://arxiv.org/abs/2307.09458.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/70b8505ac79e3e131756f793cd80eb8d-Abstract-Conference.html.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://arxiv.org/abs/2202.12837.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2301.05217.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic Without Algorithms: Language Models Solve Math With a Bag of Heuristics. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8c5f30296296d2ae402ebbd09aaa9c12-Abstract-Conference.html.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. arXiv preprint arXiv:2412.15115, 2024. URL https://arxiv.org/abs/2412.15115.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://arxiv.org/abs/2004.07667.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642. Association for Computational Linguistics, 2013. URL https://aclanthology.org/D13-1170/.

Georgios P. Spithourakis and Sebastian Riedel. Numeracy for Language Models: Evaluating and Improving their Ability to Predict Numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. doi: 10.18653/v1/P18-1196. URL https://arxiv.org/abs/1805.08154.

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*, 2023. URL https://arxiv.org/abs/2305.15054.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function Vectors in Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019. URL https://arxiv.org/abs/1909.07940.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2305.14160.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. arXiv preprint arXiv:2303.03846, 2023. URL https://arxiv.org/abs/2303.03846.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, Assemble, Ace: Understanding How LMs Answer Multiple Choice Questions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c248154176c08147e82c0b30961604f7-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1910.07467.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Long Zhang, Dai-jun Lin, and Wei-neng Chen. The geometric price of discrete logic: Context-driven manifold dynamics of number representations. *arXiv preprint arXiv:2603.23577*, 2026. URL https://arxiv.org/abs/2603.23577.

Andy Zou, Long Phan, Sarah 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 Appendix guide

The appendix provides the protocols, derivations, and complete condition tables behind the main comparisons. Section B specifies inputs, answer mappings, calibration, and evaluation. Section C derives the mapping contrasts and final-readout identities. Section D explains the additional geometric measurements and their interpretation. Section E reports the broader numerical and sentiment experiments. Section F specifies the answer-interface audit. Section G contains every mapping intervention condition, including unsuccessful mappings and cross-rule controls. Section H tests frozen classification across answer vocabularies. Section I describes the accompanying code and machine-readable results.

B Experimental protocol

B.1 Models and representation sites

The experiments use publicly released transformer language models (Vaswani et al., 2017). Qwen2.5 is the primary model for mapping interventions. Gemma 2 supplies a second architecture, and Qwen3 supplies a contrasting answer-interface condition. The broader numerical experiment also includes Llama 3.1 and Mistral (Qwen et al., 2024; Gemma Team et al., 2024; Yang et al., 2025; Grattafiori et al., 2024; Jiang et al., 2023). Table 3 identifies the exact checkpoints. These are inference experiments; no model weights are trained or modified.

Table 3: Model checkpoints and decoder depth. The representation at the final layer precedes the final normalization.

Model	Checkpoint identifier	Blocks
Qwen2.5	`Qwen/Qwen2.5-7B-Instruct`	28
Gemma 2	`google/gemma-2-9b-it`	42
Qwen3	`Qwen/Qwen3-4B`	36
Llama 3.1	`meta-llama/Llama-3.1-8B-Instruct`	32
Mistral	`mistralai/Mistral-7B-Instruct-v0.3`	32

All states are captured at the last non-padding prompt token. Layer numbers count decoder blocks from one, so Qwen layer 28 is the last decoder-block output. At an intermediate layer, an edit replaces only this token's residual state, after which the remaining blocks execute normally. At the final layer, the edited state passes through the model's final normalization and output head. The same candidate-token ordering is used for ordinary inference, hook inference, and final-head replay.

The returned measurements use single-precision model inference and full-precision post-processing of saved arrays. Evaluation disables stochastic training behavior and does not sample generated continuations. Batch adaptation changes execution size without changing input IDs, edit directions, or random-direction seeds. Unchanged-state replay and zero-edit hooks check the equivalence of the inference paths. The largest fresh zero-edit logit discrepancy is 2.67×10^{-5}, with no prediction changes.

B.2 Numerical inputs and exact prompts

All primary numerical inputs are the integers from 1 through 200. The four elementary rules are $n > 100$, $n \bmod 2 = 0$, $n \bmod 3 = 0$, and primality. Their positive-class counts are 100, 100, 66, and 46, respectively. Primality excludes 1. The two compositional rules are the conjunction (AND) and exclusive disjunction (XOR) of even parity and $n > 100$. Labels follow these definitions directly.

The standard question strings are fixed across numbers. They are "Is this number greater than 100?", "Is this number even?", "Is this number divisible by 3?", and "Is this number prime?". The conjunction question is "Is this number even and greater than 100?". The exclusive-disjunction question is "Is exactly one of these statements true: this number is even; this number is greater than 100?". This last question defines the operation in ordinary language as well as naming its two predicates.

The AB/BA intervention prompts explicitly assign answer letters to truth values. For a number n, question Q, positive-answer letter P, and negative-answer letter N, the full prompt is:

> `Consider the number` $\langle n \rangle$`.` $\langle Q \rangle$ `If yes, answer` $\langle P \rangle$`; if no, answer` $\langle N \rangle$`. Answer only` $\langle P \rangle$ `or` $\langle N \rangle$`:`

Mapping AB sets $(P, N) = (\mathrm{A}, \mathrm{B})$; BA sets $(P, N) = (\mathrm{B}, \mathrm{A})$. The corresponding baseline replaces Q with "Is this number an integer?". Baseline inputs are always integers, but their hidden states can still differ by number and hence by the target rule's classes.

The CD/DC prompts keep the last candidate list fixed. They use the same prefix and assignment sentences, with $(P, N) = (\mathrm{C}, \mathrm{D})$ or (D, C), but always end with "Answer only C or D:". This condition jointly changes the vocabulary and the final candidate-order convention. It tests whether the result persists under that joint change; it does not separately estimate the two factors.

The broader Yes/No experiment uses a different baseline prompt. Its task prompt is "Consider the number $\langle n \rangle$. $\langle Q \rangle$ Answer only Yes or No:", while its baseline is "Consider the number $\langle n \rangle$. Repeat the number only:". The alternative wording uses "Focus on the integer $\langle n \rangle$." with "Does it exceed 100?", "Does it have even parity?", "Is it divisible by 3?", or "Is it a prime number?", followed by "Reply with Yes or No only:". Its baseline ends with "Write the same integer only:". The English-number condition replaces the standard prefix with "Consider the number written as ⟨number in words⟩." and retains the standard question and ending.

The mapping experiment supplies these strings directly to the tokenizer. Chat rendering is confined to the separately reported interface diagnostic. Candidate-pair scores use the designated single-token answer forms, ordered as semantic positive and negative under each mapping. The diagnostic records token IDs and single-token validity rather than assuming that a displayed answer string has the same tokenization across models.

B.3 Calibration and planned condition coverage

The mapping experiment shares a joint stratified split across tasks and mappings. The stratification label is $2\,\mathbf{1}[n \bmod 2 = 0] + \mathbf{1}[n \bmod 3 = 0]$. A stratified 50/50 split produces 100 calibration and 100 evaluation numbers. The three split seeds are 42, 43, and 44, denoted splits 1, 2, and 3 in the paper. All layers for a given split use the same numbers.

Class-direction fitting uses a shared correctness filter within each rule. On the calibration split, a number is retained if both mappings answer it correctly. If either class then has fewer than two members, both mappings instead use the full calibration split. The task-state center μ^m and mean update use all calibration numbers regardless of this filter. Thus the two mappings use identical class membership for their contrasts and independently computed centers for their target states. The retained indices, class counts, and fallback mode accompany every fitted direction.

The intervention panel contains 13 combinations of model, vocabulary, split, and layer. Qwen AB split 1 uses layers 14, 22, and 28; AB splits 2 and 3 use layers 22 and 28; Qwen CD split 1 uses layers 14, 22, and 28; and Gemma AB split 1 uses layers 21, 33, and 42. Each combination evaluates four targets: two rules under two mappings. Each target has 21 conditions: baseline, zero edit, native-axis removal, eight random directions, eight raw or normalized cross-rule contrast directions, and two natural removals of its own normalized contrasts. This gives 1,092 target conditions. Reused identical baseline or intervention outputs are not treated as independent replications.

The wider numerical study uses per-rule stratified splits. It fits on 100 numbers and evaluates the other 100, with a primary seed of 42, additional geometric split repeats for all four Qwen rules, and intervention repeats for parity and divisibility. These splits differ from the joint mapping split. All-input capability summaries use 200 numbers and are explicitly distinguished from held-out intervention accuracy. For range transfer, directions fitted on 1 through 200 are evaluated on the 800 integers from 201 through 1,000.

B.4 Endpoints and uncertainty

Semantic prediction is $\mathbf{1}[z_+ > z_-]$. In a tie, the decision is the negative class. Ordinary accuracy averages agreement with the deterministic rule label; balanced accuracy averages the two class re-

calls. A prediction flip compares the edited decision with the original decision, so it can either repair an error or introduce one. The signed margin in Equation 6 gives larger values for stronger support for the correct class.

The main intervals preserve the paired structure of the experiment. Each bootstrap draw samples 100 test-number indices with replacement and carries their two mappings, two candidate directions, and two layers together. We use 20,000 draws and take the 2.5th and 97.5th percentiles of the resulting mean interaction. This procedure conditions on the fitted directions and describes variation across the evaluated numbers. We report calibration splits separately because overlapping test sets cannot be pooled as independent replications.

Supplementary summaries retain initially correct and initially incorrect subsets. The all-test result remains primary. A common-correct comparison uses numbers answered correctly by both mappings before intervention; its conditioning changes the population being described. For random controls, we average eight fixed directions within each number before forming the main comparison. The machine-readable output also preserves all individual directions and class-specific outcomes.

C Mapping contrasts and final-readout derivations

C.1 Orthogonality and answer exchange

The sum and difference have a simple behavior under exchanging the two mappings. Let x and y be their unit class directions. Exchanging x and y preserves $(x + y)/2$ and negates $(x - y)/2$. Their inner product is

$$\left(\frac{x+y}{2}\right)^{\top}\left(\frac{x-y}{2}\right) = \frac{\|x\|^2 - \|y\|^2 + y^{\top}x - x^{\top}y}{4} = 0. \tag{9}$$

Their squared norms are $(1 + x^{\top}y)/2$ and $(1 - x^{\top}y)/2$. The two components therefore describe different amounts of the two-mapping geometry even after they are separately normalized for intervention.

Degenerate components have a direct geometric meaning. If $x = y$, the antisymmetric component vanishes; if $x = -y$, the symmetric component vanishes. A zero original class difference cannot be unit-normalized in the first place. Implementations must record these cases instead of assigning an arbitrary direction. The evaluated main contrasts are nondegenerate. With unnormalized original directions, the same calculation yields $(\|d^{\mathrm{AB}}\|^2 - \|d^{\mathrm{BA}}\|^2)/4$, which explains why the raw components can be correlated.

Baseline subtraction admits a separate exact check. With a common calibration set and class labels,

$$d_u^m = \mathbb{E}[h^m - b^m \mid y = 1] - \mathbb{E}[h^m - b^m \mid y = 0] \tag{10}$$

$$= d_h^m - d_b^m. \tag{11}$$

Subtracting any common mean update before computing the two class means also cancels from this difference. The identity does not imply $d_b^m = 0$, even though the integer-baseline answer is constant. We consequently retain all three directions and measure their norms and cosines. For the principal Qwen AB layers, the task-state and update class directions align closely. At layer 14 the normalized mapping components are less aligned, so the same approximation should not be extended to that earlier site.

The contrast definitions specify an experimental symmetry, not a unique decomposition of internal computation. A symmetric component can reflect information shared by the two mappings, common prompt structure, or a representation of their binary decision. An antisymmetric component can reflect the required answer token or another quantity that changes with the mapping. The interventions test their relative function under the stated prompts; the final-head calculation supplies a more specific interpretation at the output.

C.2 Class and mapping effects before normalization

A four-cell mean decomposition clarifies what normalization preserves. Let $\tau = 2y-1$ be the signed task label and $m \in \{-1, +1\}$ the signed answer mapping, with $m = +1$ for AB and $m = -1$ for

BA. Using the same calibration population and filtering rule in all four cells, write their task-state means as

$$\bar{h}_{\tau,m} = \mu + \tau U + mV + \tau mW, \qquad (\mu, U, V, W) = \tfrac{1}{4}\sum_{\tau,m}(1, \tau, m, \tau m)\bar{h}_{\tau,m}. \tag{12}$$

Here U is the task-label main effect, V is the mapping main effect, and W is their interaction. This identity represents four conditional means; it does not assume an exact decomposition of each example's state. With a fixed ordering of the answer tokens, their required signed encoding is τm.

For these task-state means, the class contrasts recover the effects directly. They satisfy $d_+ = 2(U + W)$ and $d_- = 2(U - W)$, so their half-sum and half-difference are $2U$ and $2W$. Let $\lambda_\pm = 1/\|d_\pm\|$. Normalizing the class contrasts before taking their half-sum and half-difference instead gives

$$s_0 = (\lambda_+ + \lambda_-)U + (\lambda_+ - \lambda_-)W, \tag{13}$$

$$a_0 = (\lambda_+ - \lambda_-)U + (\lambda_+ + \lambda_-)W. \tag{14}$$

Unequal class-contrast norms thus generally recombine the main effect and interaction. The resulting orthogonality describes the operational directions; it does not establish independent semantic factors.

The choice of population matters for this identity. The class-direction fit uses a common-correct calibration subset when available, whereas natural removal uses a full-calibration task-state center. These quantities cannot jointly define the four-cell decomposition above unless they are recomputed on the same population. A task-update decomposition is also valid, but its effects subtract the corresponding baseline effects. We keep these descriptive identities separate from claims about the variables that the model causally uses.

C.3 Natural removal and crossing

Natural removal sets one centered coordinate to zero. For unit v and $\beta = (h - \mu)^\top v$,

$$(h - \beta v - \mu)^\top v = \beta - \beta\|v\|^2 = 0. \tag{15}$$

Its squared edit norm is β^2. Removing orthogonal symmetric and antisymmetric coordinates therefore tests two defined projections, but their natural edit sizes can differ substantially.

Norm matching changes this operation when the reference size differs from the candidate projection. Equation 5 leaves the centered coordinate

$$\beta' = \beta - \mathrm{sgn}_+(\beta)r(h). \tag{16}$$

For nonzero β, crossing occurs precisely when $r(h) > |\beta|$ in exact arithmetic. A matched edit is natural removal only at equality. We store the coordinates before and after editing, planned and realized norms, and crossing indicators. Across 10,400 evaluated natural edits, no centered coordinate crosses zero beyond the numerical tolerance. Large crossing rates in matched conditions are reported in Section G.

C.4 The final decision identity

Root mean square normalization scales every coordinate by an effective gain and a shared positive denominator (Zhang & Sennrich, 2019). For hidden dimension d, gain vector g, and $\varepsilon > 0$, write

$$\mathrm{RMSNorm}(h) = \frac{g \odot h}{\sigma(h)}, \qquad \sigma(h) = \sqrt{\|h\|^2/d + \varepsilon}. \tag{17}$$

For Gemma, the effective gain is one plus the stored normalization weight. The output heads in this analysis have no additive bias.

Proposition 1. *For a bias-free output head following RMSNorm, let $q = g \odot (w_+ - w_-)$. If any subsequent scalar transformation of logits is shared and strictly increasing, then the binary candidate decision is $\mathbf{1}[q^\top h > 0]$.*

Proof. Before any scalar transformation, the logit difference is

$$w_+^\top \operatorname{RMSNorm}(h) - w_-^\top \operatorname{RMSNorm}(h) = \frac{[g \odot (w_+ - w_-)]^\top h}{\sigma(h)}.$$

The denominator is positive, so this difference has the sign of $q^\top h$. A shared strictly increasing scalar transformation preserves the ordering of the two logits, including equality. Gemma's logit transformation $c\tanh(z/c)$, with $c > 0$, satisfies this condition. The specified tie rule then gives the stated decision. □

The proposition applies to signs rather than to a linear model of all margins. The RMSNorm denominator changes after editing, and softcapping changes logit differences nonlinearly. Thus $q^\top e = 0$ guarantees no change in the binary decision for an edit e, but need not preserve the margin magnitude. Similarly, the numerator change $-\beta_v(h)q^\top v$ explains the direction of a natural edit's influence on the decision without equating it to the reported semantic margin decrease.

The same identity decomposes any final-layer edit by its output effect. Let $\hat{q} = q/\|q\|$ and write

$$e_\parallel = (e^\top \hat{q})\hat{q}, \qquad e_\perp = e - e_\parallel. \tag{18}$$

Then $q^\top(h - e) = q^\top(h - e_\parallel)$ and $q^\top(h - e_\perp) = q^\top h$. The full edit and its parallel part therefore produce identical candidate decisions. This concerns the same edit decomposed into two pieces; projecting the state independently onto a newly normalized direction would be a different operation. Numerical checks report the actually evaluated comparisons separately from this algebraic statement.

C.5 Readout energy within the two-direction span

The reported readout share is conditional on the symmetric and antisymmetric span. Since s and a are orthonormal, the squared norm of the readout projected onto their span is $(q^\top s)^2 + (q^\top a)^2$. We report

$$\eta_a = \frac{(q^\top a)^2}{(q^\top s)^2 + (q^\top a)^2}. \tag{19}$$

The denominator excludes any component of q perpendicular to both directions. Consequently, a value near one states how this two-direction projection is allocated; it does not state that a contains nearly all of q.

The saved natural-edit quantities suffice to recover these scalar projections. Each record retains its natural coefficient β_i and the signed state change $\Delta h_i = h_i' - h_i = -\beta_i v$ through the scalar $q^\top \Delta h_i$. We recover the magnitude of $q^\top v$ from the zero-intercept relation $q^\top \Delta h_i = -\beta_i (q^\top v)$, using all available test coefficients. The fitted residuals are below 10^{-5}. Applying Equation 19 gives antisymmetric shares of 99.66% and 99.94% for Qwen AB parity and divisibility, and 97.05% for CD divisibility. The two mappings reverse the candidate ordering, which negates q but preserves these squared projections.

Balanced readout contrasts give a direct explanation for a small symmetric output effect. Fix the AB readout q, and let $x = \bar{d}^{\mathrm{AB}}$, $y = \bar{d}^{\mathrm{BA}}$, $c = x^\top y$, $p = q^\top x$, and $r = -q^\top y$. The negative sign aligns the reversed mapping's class contrast with its semantic answer ordering. For $|c| < 1$,

$$q^\top s = \frac{p - r}{\sqrt{2(1 + c)}}, \qquad q^\top a = \frac{p + r}{\sqrt{2(1 - c)}}. \tag{20}$$

If $p = r \neq 0$, the symmetric readout projection is zero and the antisymmetric projection is nonzero, even when the symmetric geometric component is large. This algebra explains a possible final-layer preference. The intermediate-layer intervention result establishes the different functional ordering before the remaining network computes the final readout.

D Additional geometric measurements

D.1 Centered updates and baseline-orthogonal components

The wider geometric analysis separates a centered task update into radial and orthogonal components relative to each baseline state. To distinguish this construction from the main mapping directions s

and a, write

$$\delta_i = h_i - b_i, \quad \bar{\delta} = \mathbb{E}_{\text{cal}}[\delta_i], \quad \xi_i = \delta_i - \bar{\delta}, \quad e_i = \frac{b_i}{\|b_i\|}, \quad o_i = \xi_i - (e_i^\top \xi_i) e_i, \quad \nu_i = \frac{o_i}{\|o_i\|}. \tag{21}$$

The mean shift uses the calibration split. Zero baseline states or orthogonal components make their normalized directions undefined and are excluded from the relevant pairwise metric with counts recorded.

Two matrices summarize different geometric relations. We define $U_{ij} = \nu_i^\top \nu_j$ and $C_{ij} = e_i^\top \nu_j$. The first compares orthogonal-update directions across inputs. The second compares a baseline direction with another input's orthogonal update; it need not be symmetric. Within-class and cross-class averages exclude diagonal pairs and retain both directed orders. The separation statistic is $\Delta U = U_{\text{same}} - U_{\text{cross}}$. These are class-conditional angular summaries.

D.2 Exact reconstruction includes the mean shift

The radial and orthogonal components reconstruct the shifted task state. Let $p_i = e_i^\top \xi_i$, $\rho_i = \|o_i\|$, and $z_i = h_i - \bar{\delta}$. Then

$$z_i = (\|b_i\| + p_i) e_i + \rho_i \nu_i. \tag{22}$$

For nonzero z_i, define $\alpha_i = (\|b_i\| + p_i)/\|z_i\|$ and $\kappa_i = \rho_i/\|z_i\|$. Their squared sum is one, and the exact cosine expansion is

$$\hat{z}_i^\top \hat{z}_j = \alpha_i \alpha_j e_i^\top e_j + \alpha_i \kappa_j e_i^\top \nu_j \tag{23}$$

$$+ \kappa_i \alpha_j \nu_i^\top e_j + \kappa_i \kappa_j \nu_i^\top \nu_j. \tag{24}$$

The radial coefficient can be negative. An angle representation therefore uses both the radial and orthogonal coordinates instead of assuming an acute angle.

The unshifted task state requires the remaining mean-shift terms. Since $h_i = z_i + \bar{\delta}$,

$$h_i^\top h_j = z_i^\top z_j + z_i^\top \bar{\delta} + \bar{\delta}^\top z_j + \|\bar{\delta}\|^2. \tag{25}$$

Its cosine divides this expression by $\|h_i\|\|h_j\|$. Dropping the mean-shift terms instead computes the geometry of z_i. The numerical analysis retains both quantities and their reconstruction residuals.

D.3 Class-mean identities for angular separation

The pairwise separation statistic can be written in terms of class means. For class size n_c and mean $m_c = n_c^{-1} \sum_{i:y_i=c} \nu_i$, unit-length directions give

$$U_{\text{cross}} = m_0^\top m_1, \qquad U_{\text{within},c} = \frac{n_c \|m_c\|^2 - 1}{n_c - 1}, \quad n_c > 1. \tag{26}$$

The within-class expression follows by expanding $\|\sum_i \nu_i\|^2$ and subtracting the n_c self-inner-products. The pooled within-class mean weights these expressions by $n_c(n_c - 1)$, the number of ordered non-self pairs. It differs from a macro average when class sizes differ.

The macro version makes the relation to class-mean separation explicit:

$$\Delta U_{\text{macro}} = \frac{1}{2}\|m_0 - m_1\|^2 - \frac{1}{2} \sum_{c=0}^{1} \frac{1 - \|m_c\|^2}{n_c - 1}. \tag{27}$$

Thus a large pairwise contrast can follow from separated class means. Counting the many pairs does not create additional independent inputs. The geometric uncertainty calculation resamples or permutes input labels at the number level, retaining the fitted calibration transform.

Centering also constrains class vectors estimated on the same centering set. If $\sum_i \xi_i = 0$, then $n_0 \bar{\xi}_0 + n_1 \bar{\xi}_1 = 0$. The two class means are therefore oppositely directed whenever both are nonzero. This identity applies when the class-mean set equals the centering set. A subsequent correctness filter need not preserve it. We use the identity as an algebraic check, without counting the two resulting class-vector alignments as independent evidence.

D.4 Distance, translation, and normalization controls

Angular change does not by itself establish a change in Euclidean distances. A shared translation $h_i = b_i + t$ preserves every pairwise Euclidean distance but can change cosines computed relative to the origin. A shared orthogonal transform $h_i = Rb_i$ preserves both Euclidean distances and baseline pairwise cosines, while the update $(R - I)b_i$ can still have class-related directions. For example, two opposite groups on a circle remain opposite after a common rotation; their nonzero orthogonal updates can also point in opposite directions. Class organization in U consequently has no general implication of non-isometric computation.

The supplementary measurements test distance and angular changes separately. They compare baseline and task squared distances, raw cosine changes, the cosine change produced by the shared mean translation, and residual change beyond that translation. An orthogonal Procrustes transform and translation are fitted only on calibration states and evaluated on test states. A separate control matches within-class and cross-class pairs in ten bins of baseline cosine. These quantities assess particular geometric explanations; they do not identify a numerical algorithm.

RMSNorm also distinguishes two useful metrics. Writing $G = \mathrm{diag}(g)$, its output cosine is the input cosine under G^2:

$$\cos(\mathrm{RMSNorm}(x), \mathrm{RMSNorm}(y)) = \frac{x^\top G^2 y}{\sqrt{x^\top G^2 x}\sqrt{y^\top G^2 y}}. \tag{28}$$

By contrast, for nonzero gains the normalized output z satisfies

$$z^\top G^{-2} z = \frac{\|x\|^2}{\|x\|^2/d + \varepsilon}. \tag{29}$$

This is constant at d only when $\varepsilon = 0$ and $x \neq 0$. With a zero gain, the inverse metric is restricted to nonzero-gain coordinates. The robustness analysis recomputes the full decomposition after gain scaling; it does not substitute one metric into a formula derived for another.

E Broader numerical and sentiment experiments

E.1 Intervention definitions

The broader experiments retain class-vector edits as well as the class-axis edits used in the main comparison. A class vector is $v_c = \mathbb{E}[\xi_i \mid y_i = c]$, fitted on correctly answered calibration inputs if both classes have at least two examples, otherwise on the full calibration split. For a current task state h_i, define $P_i = I - \widehat{h}_i \widehat{h}_i^\top$. The class-vector edit is $h_i' = h_i - \gamma P_i v_c$. Its projection is perpendicular to the current task state, whereas o_i in Section D is perpendicular to the baseline state. These are distinct operations.

The choice of class determines whether an edit uses test-label information. Oracle class removal uses $c = y_i$ and serves as a descriptive upper-information intervention. Predicted class removal uses the original candidate prediction. A class swap subtracts $\gamma P_i(v_c - v_{1-c})$. The class-axis operation instead removes $\gamma[(h_i - \mu)^\top \hat{d}]\hat{d}$, with $d = v_1 - v_0$ and no test label. Replacing v_c with the class mean of o_i gives an additional orthogonal-update comparator, labeled an innovation direction in the supplementary tables. Oracle edits are reported as such and do not support the main claim of a test-label-free operation.

Random and shuffled-label directions have different control objectives. A random-axis projection removes the state projection onto an isotropic random direction. A shuffled-label axis is fitted after permuting calibration labels. The matched-random control instead matches each example's actual native-axis edit norm. Merely matching the lengths of the stored direction vectors would not match these projection sizes, because natural projection uses a unit direction.

The dose parameter scales the final edit vector. At $\gamma = 0$ the edit is zero, and at $\gamma = 1$ a class-axis edit removes its centered projection. Values above one can cross that center. The dose tables report the executed values rather than interpreting every dose as a complete removal. On initially correct inputs, oracle and predicted class selection coincide algebraically; their agreement on that subset is not an independent test of label information.

E.2 ADDITIONAL TASK AND RANGE COVERAGE

The numerical tables retain the complete task families that motivated the mapping comparison. They report five-model capability and geometry, prompt wording, numbers written in English, compositional rules, layerwise effects, dose dependence, calibration changes, and range transfer. These measurements establish the breadth and variability of class-related structure. They use their own stated answer interface and cannot substitute for the paired-mapping confirmation in the main text.

The sentiment experiment uses the binary Stanford Sentiment Treebank (SST-2) task (Socher et al., 2013). We sample 400 examples per class from its training split for calibration and 200 per class from its validation split for evaluation, without replacement and with selection seed 42. The task prompt is "Review: ⟨text⟩ / Question: Is the sentiment of this review positive? / Answer only Yes or No:". The baseline replaces the question with "Does this review contain English text?". The saved selection indices fix the sample independently of model predictions. This is a supplementary test of class-related updates outside numbers, with no exchanged-mapping experiment on sentiment.

Range transfer keeps the fitted numerical directions fixed. It evaluates parity, divisibility, and primality on 201 through 1,000, including initially incorrect inputs. Accuracy, class recalls, and the intervention changes therefore describe extrapolation of both model behavior and a fitted direction. The range-transfer results should be read alongside baseline capability, since an edit's effect on a weak initial decision has a different interpretation from its effect on a correct one.

Table 4: Yes/No rule accuracy across five models. Each rule uses integers 1–200 with 100 training and 100 held-out examples. All accuracy columns are percentages; Pred. yes is the percentage of held-out answers classified as Yes. Bal. denotes balanced accuracy and Maj. denotes the held-out majority-class baseline.

Model	Rule	Layer	All	Train	Test	Bal.	Maj.	Pred. yes
Qwen2.5-7B	Magnitude	28	83.0	83.0	83.0	83.0	50.0	33.0
Qwen2.5-7B	Parity	28	91.0	91.0	91.0	91.0	50.0	41.0
Qwen2.5-7B	Divisibility	28	92.5	90.0	95.0	92.4	67.0	28.0
Qwen2.5-7B	Primality	28	88.5	85.0	92.0	82.6	77.0	15.0
Gemma-2-9B	Magnitude	42	100.0	100.0	100.0	100.0	50.0	50.0
Gemma-2-9B	Parity	42	96.0	95.0	97.0	97.0	50.0	47.0
Gemma-2-9B	Divisibility	42	92.0	92.0	92.0	94.0	67.0	41.0
Gemma-2-9B	Primality	42	97.5	98.0	97.0	93.5	77.0	20.0
Qwen3-4B†	Magnitude	36	50.0	50.0	50.0	50.0	50.0	100.0
Qwen3-4B†	Parity	36	53.5	51.0	56.0	56.0	50.0	94.0
Qwen3-4B†	Divisibility	36	33.0	33.0	33.0	50.0	67.0	100.0
Qwen3-4B†	Primality	36	23.0	23.0	23.0	50.0	77.0	100.0
Llama-3.1-8B	Magnitude	32	94.5	95.0	94.0	94.0	50.0	44.0
Llama-3.1-8B	Parity	32	77.0	74.0	80.0	80.0	50.0	30.0
Llama-3.1-8B	Divisibility	32	71.5	72.0	71.0	56.1	67.0	4.0
Llama-3.1-8B†	Primality	32	79.0	78.0	80.0	56.5	77.0	3.0
Mistral-7B	Magnitude	32	98.5	99.0	98.0	98.0	50.0	52.0
Mistral-7B†	Parity	32	51.0	50.0	52.0	52.0	50.0	98.0
Mistral-7B†	Divisibility	32	33.0	33.0	33.0	50.0	67.0	100.0
Mistral-7B†	Primality	32	24.5	24.0	25.0	51.3	77.0	98.0

† Class-direction fitting uses all training examples because fewer than two correctly answered training examples occur in at least one class. Other rows use correctly answered training examples for class directions; shared centering uses all training examples. The displayed model names identify the same instruction-tuned checkpoints used throughout the corresponding experiments.

Table 5: Final-residual innovation geometry on all held-out examples. Counts are class 0/class 1. U_{same} and U_{cross} average innovation cosines over ordered, distinct within-class and cross-class pairs; $\Delta U = U_{\text{same}} - U_{\text{cross}}$. Intervals are the stored example-bootstrap 95% intervals conditional on the fitted training shift. n/a denotes an undefined cross-class quantity when one class is absent.

Model	Rule	n_0/n_1	U_{same}	U_{cross}	ΔU	95% interval
Qwen2.5-7B	Magnitude	50/50	0.344	-0.251	0.595	[0.553, 0.671]
Qwen2.5-7B	Parity	50/50	0.151	-0.086	0.237	[0.212, 0.312]
Qwen2.5-7B	Divisibility	67/33	0.104	0.003	0.101	[0.084, 0.167]
Qwen2.5-7B	Primality	77/23	0.060	-0.023	0.083	[0.056, 0.155]
Gemma-2-9B	Magnitude	50/50	0.645	-0.613	1.259	[1.216, 1.307]
Gemma-2-9B	Parity	50/50	0.242	-0.211	0.453	[0.392, 0.556]
Gemma-2-9B	Divisibility	67/33	0.201	-0.138	0.338	[0.296, 0.424]
Gemma-2-9B	Primality	77/23	0.103	-0.118	0.221	[0.180, 0.303]
Qwen3-4B	Magnitude	50/50	0.399	-0.363	0.762	[0.729, 0.824]
Qwen3-4B	Parity	50/50	0.190	-0.173	0.363	[0.308, 0.464]
Qwen3-4B	Divisibility	67/33	0.090	0.010	0.080	[0.066, 0.148]
Qwen3-4B	Primality	77/23	0.052	-0.055	0.108	[0.091, 0.160]
Llama-3.1-8B	Magnitude	50/50	0.258	-0.139	0.397	[0.384, 0.442]
Llama-3.1-8B	Parity	50/50	0.106	0.005	0.101	[0.096, 0.150]
Llama-3.1-8B	Divisibility	67/33	0.076	0.062	0.014	[0.014, 0.057]
Llama-3.1-8B	Primality	77/23	0.067	0.042	0.026	[0.019, 0.067]
Mistral-7B	Magnitude	50/50	0.474	-0.369	0.844	[0.798, 0.902]
Mistral-7B	Parity	50/50	0.088	0.038	0.050	[0.031, 0.127]
Mistral-7B	Divisibility	67/33	0.102	0.096	0.005	[-0.001, 0.064]
Mistral-7B	Primality	77/23	0.081	0.052	0.029	[0.017, 0.078]

Table 6: Final-residual innovation geometry restricted to examples answered correctly before intervention. Counts are class 0/class 1. U_{same} and U_{cross} average innovation cosines over ordered, distinct within-class and cross-class pairs; $\Delta U = U_{\text{same}} - U_{\text{cross}}$. Intervals are the stored example-bootstrap 95% intervals conditional on the fitted training shift. n/a denotes an undefined cross-class quantity when one class is absent.

Model	Rule	n_0/n_1	U_{same}	U_{cross}	ΔU	95% interval
Qwen2.5-7B	Magnitude	50/33	0.349	-0.260	0.609	[0.567, 0.680]
Qwen2.5-7B	Parity	50/41	0.160	-0.095	0.256	[0.228, 0.335]
Qwen2.5-7B	Divisibility	67/28	0.103	-0.004	0.108	[0.086, 0.178]
Qwen2.5-7B	Primality	77/15	0.053	-0.029	0.082	[0.040, 0.159]
Gemma-2-9B	Magnitude	50/50	0.645	-0.613	1.259	[1.216, 1.307]
Gemma-2-9B	Parity	50/47	0.263	-0.237	0.501	[0.446, 0.603]
Gemma-2-9B	Divisibility	59/33	0.238	-0.172	0.409	[0.378, 0.481]
Gemma-2-9B	Primality	77/20	0.100	-0.123	0.223	[0.182, 0.306]
Qwen3-4B	Magnitude	0/50	0.507	n/a	n/a	n/a
Qwen3-4B	Parity	6/50	0.176	-0.155	0.331	[0.252, 0.460]
Qwen3-4B	Divisibility	0/33	0.141	n/a	n/a	n/a
Qwen3-4B	Primality	0/23	0.268	n/a	n/a	n/a
Llama-3.1-8B	Magnitude	50/44	0.260	-0.142	0.402	[0.384, 0.449]
Llama-3.1-8B	Parity	50/30	0.120	-0.018	0.138	[0.124, 0.203]
Llama-3.1-8B	Divisibility	67/4	0.073	0.025	0.048	[0.016, 0.127]
Llama-3.1-8B	Primality	77/3	0.064	0.026	0.038	[0.003, 0.104]
Mistral-7B	Magnitude	48/50	0.487	-0.382	0.869	[0.834, 0.921]
Mistral-7B	Parity	2/50	0.081	0.032	0.049	[0.001, 0.152]
Mistral-7B	Divisibility	0/33	0.097	n/a	n/a	n/a
Mistral-7B	Primality	2/23	0.163	0.100	0.063	[0.021, 0.194]

Table 7: Raw and final-normalization-gain-weighted innovation separation. The first pair of columns uses all held-out examples; the second uses the correctly answered subset. All entries report ΔU. Gain weighting changes the coordinate metric used for the geometric calculation.

Model	Rule	All: raw	All: gain	Correct: raw	Correct: gain
Qwen2.5-7B	Magnitude	0.595	0.624	0.609	0.648
Qwen2.5-7B	Parity	0.237	0.262	0.256	0.278
Qwen2.5-7B	Divisibility	0.101	0.105	0.108	0.113
Qwen2.5-7B	Primality	0.083	0.084	0.082	0.084
Gemma-2-9B	Magnitude	1.259	1.270	1.259	1.270
Gemma-2-9B	Parity	0.453	0.473	0.501	0.521
Gemma-2-9B	Divisibility	0.338	0.359	0.409	0.433
Gemma-2-9B	Primality	0.221	0.239	0.223	0.245
Qwen3-4B	Magnitude	0.762	0.767	n/a	n/a
Qwen3-4B	Parity	0.363	0.356	0.331	0.326
Qwen3-4B	Divisibility	0.080	0.077	n/a	n/a
Qwen3-4B	Primality	0.108	0.107	n/a	n/a
Llama-3.1-8B	Magnitude	0.397	0.396	0.402	0.402
Llama-3.1-8B	Parity	0.101	0.101	0.138	0.139
Llama-3.1-8B	Divisibility	0.014	0.014	0.048	0.048
Llama-3.1-8B	Primality	0.026	0.025	0.038	0.036
Mistral-7B	Magnitude	0.844	0.839	0.869	0.864
Mistral-7B	Parity	0.050	0.051	0.049	0.047
Mistral-7B	Divisibility	0.005	0.006	n/a	n/a
Mistral-7B	Primality	0.029	0.028	0.063	0.055

Table 8: Qwen2.5-7B coupling and cosine distortion. C is the stored directed coupling average. D is baseline cosine minus task-conditioned cosine. D_{adj} subtracts the cosine change caused by the shared translation alone; s/c indicate within-class/cross-class pairs. These are signed geometric differences, not accuracy changes.

Rule	Subset	C_s	C_c	D_s	D_c	$D_{\mathrm{adj},s}$	$D_{\mathrm{adj},c}$
Magnitude	All	0.126	0.181	-0.025	-0.025	-0.027	-0.027
Magnitude	Correct	0.131	0.183	-0.027	-0.025	-0.029	-0.027
Parity	All	0.154	0.169	-0.030	-0.027	-0.032	-0.029
Parity	Correct	0.154	0.169	-0.030	-0.027	-0.032	-0.029
Divisibility	All	0.156	0.155	-0.027	-0.023	-0.034	-0.030
Divisibility	Correct	0.156	0.155	-0.027	-0.023	-0.034	-0.030
Primality	All	0.158	0.172	-0.027	-0.028	-0.034	-0.035
Primality	Correct	0.158	0.167	-0.027	-0.026	-0.033	-0.033

Table 9: Qwen2.5-7B prompt variants. Primary prompts ask the rule directly, reworded prompts use alternate wording, and number-word prompts spell out the same integer. Accuracy is reported over all 200 inputs, the 100 training inputs, and the 100 held-out inputs separately. Geometry uses all held-out inputs.

Prompt	Rule	All (%)	Train (%)	Test (%)	Bal. (%)	ΔU
Primary	Magnitude	83.0	83.0	83.0	83.0	0.595
Primary	Parity	91.0	91.0	91.0	91.0	0.237
Primary	Divisibility	92.5	90.0	95.0	92.4	0.101
Primary	Primality	88.5	85.0	92.0	82.6	0.083
Reworded	Magnitude	100.0	100.0	100.0	100.0	0.825
Reworded	Parity	84.0	81.0	87.0	87.0	0.172
Reworded	Divisibility	99.5	100.0	99.0	98.5	0.333
Reworded	Primality	77.5	77.0	78.0	52.2	0.137
Number words	Magnitude	85.0	85.0	85.0	85.0	0.549

Table 9 (continued)

Prompt	Rule	All (%)	Train (%)	Test (%)	Bal. (%)	ΔU
Number words	Parity	50.5	50.0	51.0	51.0	0.071
Number words	Divisibility	68.5	67.0	70.0	54.5	-0.004
Number words	Primality	77.0	77.0	77.0	50.0	0.000

Table 10: Correct-answer geometry under the alternate Qwen2.5-7B prompts. The class counts show when conditional geometry excludes a class. Entries use the raw residual metric and the same pair aggregation as Table 6.

Prompt	Rule	n_0/n_1	U_{same}	U_{cross}	ΔU	95% interval
Reworded	Magnitude	50/50	0.446	-0.378	0.825	[0.768, 0.899]
Reworded	Parity	37/50	0.149	-0.033	0.182	[0.163, 0.249]
Reworded	Divisibility	67/32	0.202	-0.135	0.337	[0.296, 0.405]
Reworded	Primality	77/1	0.076	0.013	0.063	n/a
Number words	Magnitude	50/35	0.402	-0.175	0.576	[0.546, 0.636]
Number words	Parity	50/1	0.164	-0.026	0.190	n/a
Number words	Divisibility	67/3	0.151	0.110	0.041	[-0.027, 0.152]
Number words	Primality	77/0	0.120	n/a	n/a	n/a

Table 11: Qwen2.5-7B intervention doses at the final residual. Each cell gives accuracy (%) / mean correct-answer logit margin, conditional on an originally correct answer ($n = 91$ for parity and $n = 95$ for divisibility). The multiplier γ scales the specified edit. Class (true) and class (pred.) choose a class vector using the true label or the original prediction. Class axis uses a training-fitted axis without the test label. Class swap subtracts the selected class vector and adds the other class vector.

Rule	Edit	$\gamma = 0$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$
Parity	Unedited	100.0 / 1.364	100.0 / 1.364	100.0 / 1.364	100.0 / 1.364
Parity	Class (true)	100.0 / 1.364	63.7 / 0.745	54.9 / 0.126	47.3 / -0.489
Parity	Class (pred.)	100.0 / 1.364	63.7 / 0.745	54.9 / 0.126	47.3 / -0.489
Parity	Class axis	100.0 / 1.364	82.4 / 0.978	59.3 / 0.591	54.9 / 0.203
Parity	Random projection	100.0 / 1.364	100.0 / 1.364	100.0 / 1.364	100.0 / 1.364
Parity	Class swap	100.0 / 1.364	54.9 / 0.126	12.1 / -1.100	0.0 / -2.298
Parity	Shuffled axis	100.0 / 1.364	100.0 / 1.360	100.0 / 1.357	100.0 / 1.353
Divisibility	Unedited	100.0 / 1.333	100.0 / 1.333	100.0 / 1.333	100.0 / 1.333
Divisibility	Class (true)	100.0 / 1.333	71.6 / 0.922	70.5 / 0.512	69.5 / 0.104
Divisibility	Class (pred.)	100.0 / 1.333	71.6 / 0.922	70.5 / 0.512	69.5 / 0.104
Divisibility	Class axis	100.0 / 1.333	86.3 / 1.103	70.5 / 0.873	70.5 / 0.644
Divisibility	Random projection	100.0 / 1.333	100.0 / 1.333	100.0 / 1.334	100.0 / 1.334
Divisibility	Class swap	100.0 / 1.333	67.4 / 0.364	22.1 / -0.599	0.0 / -1.544
Divisibility	Shuffled axis	100.0 / 1.333	98.9 / 1.318	98.9 / 1.302	98.9 / 1.286

Table 12: Qwen2.5-7B interventions at $\gamma = 1$. All denotes all 100 held-out inputs; correct denotes the originally correct subset. Margin and edit norm use the correct subset. Norm is the mean Euclidean magnitude of the actual residual change. The original random-projection and shuffled-axis controls have different edit magnitudes from class edits; equal-energy evidence is reported separately.

Rule	Edit	All (%)	Correct (%)	Margin	Edit norm
Parity	Unedited	91.0	100.0	1.364	0.000
Parity	Class (true)	50.0	54.9	0.126	32.343
Parity	Class (pred.)	59.0	54.9	0.126	32.343
Parity	Class axis	54.0	59.3	0.591	20.733
Parity	Random projection	91.0	100.0	1.364	0.294

Table 12 (continued)

Rule	Edit	All (%)	Correct (%)	Margin	Edit norm
Parity	Class swap	11.0	12.1	-1.100	64.682
Parity	Shuffled axis	91.0	100.0	1.357	3.449
Divisibility	Unedited	95.0	100.0	1.333	0.000
Divisibility	Class (true)	67.0	70.5	0.512	23.129
Divisibility	Class (pred.)	72.0	70.5	0.512	23.129
Divisibility	Class axis	67.0	70.5	0.873	13.999
Divisibility	Random projection	95.0	100.0	1.334	0.281
Divisibility	Class swap	21.0	22.1	-0.599	55.706
Divisibility	Shuffled axis	95.0	98.9	1.302	4.087

Table 13: Additional final-residual controls for Qwen2.5-7B at $\gamma = 1$. Matched random reports the mean over eight random directions, each using the class-axis edit amplitude for every example. Parallel and orthogonal split the same class-axis edit relative to the answer-readout direction; their edit vectors sum to the original axis edit. All and correct use the same denominators as Table 12.

Rule	Edit	Subset	Acc. (%)	Margin	Edit norm
Parity	Innovation (true)	All	50.0	-0.001	33.531
Parity	Innovation (true)	Correct	54.9	0.149	33.307
Parity	Innovation axis	All	52.0	0.470	19.193
Parity	Innovation axis	Correct	57.1	0.574	20.179
Parity	Matched random	All	89.1	1.137	19.741
Parity	Matched random	Correct	97.7	1.275	20.733
Parity	Readout axis	All	50.0	-0.012	3.467
Parity	Readout axis	Correct	54.9	0.066	3.649
Parity	Axis edit: parallel	All	54.0	0.476	2.094
Parity	Axis edit: parallel	Correct	59.3	0.581	2.199
Parity	Axis edit: orthogonal	All	91.0	1.232	19.629
Parity	Axis edit: orthogonal	Correct	100.0	1.375	20.616
Divisibility	Innovation (true)	All	67.0	0.425	27.089
Divisibility	Innovation (true)	Correct	70.5	0.531	26.238
Divisibility	Innovation axis	All	67.0	0.765	13.877
Divisibility	Innovation axis	Correct	70.5	0.841	13.902
Divisibility	Matched random	All	91.4	1.195	14.117
Divisibility	Matched random	Correct	95.9	1.271	13.999
Divisibility	Readout axis	All	67.0	0.366	2.378
Divisibility	Readout axis	Correct	70.5	0.446	2.374
Divisibility	Axis edit: parallel	All	67.0	0.784	1.270
Divisibility	Axis edit: parallel	Correct	70.5	0.865	1.259
Divisibility	Axis edit: orthogonal	All	95.0	1.265	14.060
Divisibility	Axis edit: orthogonal	Correct	100.0	1.341	13.942

Table 14: Layerwise Qwen2.5-7B parity geometry and class-vector deletion. Layer 0 is the input to the first decoder block; layer $\ell \geq 1$ is the output of block ℓ. Geometry and edited performance use the 91 originally correct test examples. The deletion uses the true-label class vector and $\gamma = 1$. Layer-0 innovation geometry is undefined because the stored innovation is zero.

Layer	U_{same}	U_{cross}	ΔU	Acc. (%)	Margin
0	n/a	n/a	n/a	100.0	1.364
1	0.051	-0.042	0.093	100.0	1.359
2	0.067	-0.043	0.110	100.0	1.371
3	0.056	-0.030	0.085	100.0	1.371
4	0.052	-0.022	0.074	100.0	1.365
5	0.053	-0.028	0.081	100.0	1.365
6	0.050	-0.027	0.077	100.0	1.371

Table 14 (continued)

Layer	U_{same}	U_{cross}	ΔU	Acc. (%)	Margin
7	0.046	-0.028	0.074	100.0	1.369
8	0.050	-0.031	0.081	100.0	1.357
9	0.049	-0.029	0.078	100.0	1.339
10	0.050	-0.032	0.083	100.0	1.385
11	0.047	-0.026	0.072	100.0	1.404
12	0.049	-0.025	0.074	100.0	1.404
13	0.058	-0.031	0.089	100.0	1.416
14	0.063	-0.035	0.098	98.9	1.344
15	0.069	-0.037	0.106	100.0	1.420
16	0.070	-0.036	0.105	98.9	1.344
17	0.078	-0.040	0.118	100.0	1.393
18	0.094	-0.055	0.149	100.0	1.378
19	0.115	-0.075	0.190	89.0	1.119
20	0.177	-0.138	0.315	57.1	0.539
21	0.198	-0.162	0.360	54.9	0.175
22	0.204	-0.167	0.371	54.9	0.228
23	0.221	-0.181	0.402	54.9	0.233
24	0.210	-0.164	0.374	54.9	0.151
25	0.193	-0.142	0.334	54.9	0.113
26	0.174	-0.104	0.277	54.9	0.160
27	0.171	-0.097	0.269	54.9	0.163
28	0.160	-0.095	0.256	54.9	0.126

Table 15: Descriptive correlations between cross-class innovation cosine and edited performance across decoder layers 1–28. First differences compare adjacent layers. Neighboring layers are dependent, so these coefficients are descriptive associations without independent-layer significance tests.

Outcome	Comparison	n	Pearson	Spearman
Accuracy	Levels	28	0.933	0.825
Accuracy	First differences	27	0.752	0.351
Margin	Levels	28	0.917	0.694
Margin	First differences	27	0.731	0.114

Table 16: Qwen2.5-7B sentiment control on balanced samples from the Stanford Sentiment Treebank. Training uses 800 examples and evaluation uses 400 held-out examples. Accuracy and Pred. positive are percentages; the latter is the proportion classified as positive. The majority baseline is 50% in both splits.

Split	n	Acc. (%)	Bal. (%)	Pred. positive	Margin
Training	800	85.0	85.0	35.5	5.486
Held out	400	90.5	90.5	42.0	7.565

Table 17: Sentiment innovation geometry for all held-out examples and the originally correct subset. The same raw and gain-weighted definitions as the numerical study are used.

Subset	Metric	n_0/n_1	U_{same}	U_{cross}	ΔU	95% interval
All	Raw	200/200	0.680	-0.514	1.193	[1.103, 1.283]
All	Gain weighted	200/200	0.661	-0.482	1.143	[1.057, 1.227]
Correct	Raw	197/165	0.806	-0.617	1.423	[1.380, 1.465]
Correct	Gain weighted	197/165	0.778	-0.582	1.361	[1.316, 1.405]

Table 18: Sentiment intervention doses on the 362 originally correct test examples. Each cell gives accuracy (%) / mean correct-answer logit margin. Edits use the same definitions as Table 11.

Edit	$\gamma = 0$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$
Unedited	100.0 / 8.712	100.0 / 8.712	100.0 / 8.712	100.0 / 8.712
Class (true)	100.0 / 8.712	95.3 / 5.675	79.6 / 2.639	49.2 / -0.257
Class (pred.)	100.0 / 8.712	95.3 / 5.675	79.6 / 2.639	49.2 / -0.257
Class axis	100.0 / 8.712	94.5 / 5.238	81.5 / 1.584	17.1 / -1.995
Random projection	100.0 / 8.712	100.0 / 8.719	100.0 / 8.726	100.0 / 8.732
Class swap	100.0 / 8.712	84.8 / 2.633	5.8 / -2.919	0.0 / -7.350
Shuffled axis	100.0 / 8.712	99.4 / 8.036	97.8 / 7.290	97.0 / 6.514

Table 19: Sentiment edits at $\gamma = 1$. All denotes 400 held-out examples and correct denotes the 362 originally correct examples. Margin and edit norm are conditional on the originally correct subset.

Edit	All (%)	Correct (%)	Margin	Edit norm
Unedited	90.5	100.0	8.712	0.000
Class (true)	72.0	79.6	2.639	113.874
Class (pred.)	79.0	79.6	2.639	113.874
Class axis	74.2	81.5	1.584	130.960
Random projection	90.5	100.0	8.726	4.466
Class swap	5.2	5.8	-2.919	224.026
Shuffled axis	88.8	97.8	7.290	90.366

Table 20: Qwen2.5-7B compositional controls. AND is the conjunction of even parity and magnitude greater than 100. XOR is true when exactly one of those predicates holds. Test accuracy and the majority baseline always use all 100 held-out inputs, even in rows whose geometry is conditioned on correctness. XOR predicts Yes for every held-out input, so its correct subset contains only one class.

Rule	Geometry	n_0/n_1	Test (%)	Maj. (%)	U_{same}	U_{cross}	ΔU
AND	All	75/25	96.0	75.0	0.082	-0.045	0.127
AND	Correct	75/21	96.0	75.0	0.072	-0.045	0.117
XOR	All	50/50	50.0	50.0	0.039	0.037	0.003
XOR	Correct	0/50	50.0	50.0	0.038	n/a	n/a

Table 21: Qwen2.5-7B extrapolation to integers 201–1000. All performance columns use the 800 new integers only. Geometry is shown for all new inputs and their originally correct subset. The training shift and class directions use the original 1–200 training split; no new-range examples are fitted.

Rule	Geometry	n_0/n_1	Test (%)	Bal. (%)	Maj. (%)	U_{cross}	ΔU
Parity	All	400/400	79.0	79.0	50.0	0.384	0.113
Parity	Correct	400/232	79.0	79.0	50.0	0.337	0.190
Divisibility	All	533/267	79.0	68.5	66.6	0.462	0.046
Divisibility	Correct	533/99	79.0	68.5	66.6	0.430	0.081
Primality	All	678/122	90.1	67.6	84.8	0.446	0.002
Primality	Correct	678/43	90.1	67.6	84.8	0.424	0.019

Table 22: Extrapolation interventions at $\gamma = 1$. All accuracy uses the 800 integers 201–1000; n_c gives the size of the originally correct subset. Correct accuracy, margin and edit norm use that subset. All directions are fitted in the original training range.

Rule	Edit	All (%)	n_c	Correct (%)	Margin	Edit norm
Parity	Unedited	79.0	632	100.0	1.370	0.000
Parity	Class (true)	48.8	632	61.7	0.132	31.990
Parity	Class (pred.)	67.1	632	61.7	0.132	31.990
Parity	Class axis	50.2	632	63.6	0.650	19.119
Parity	Random projection	79.0	632	100.0	1.370	0.346
Parity	Class swap	2.1	632	2.7	-1.109	64.719
Parity	Shuffled axis	80.4	632	99.8	1.353	5.495
Divisibility	Unedited	79.0	632	100.0	1.123	0.000
Divisibility	Class (true)	64.0	632	81.0	0.391	19.604
Divisibility	Class (pred.)	83.4	632	81.0	0.391	19.604
Divisibility	Class axis	66.6	632	84.3	0.948	8.227
Divisibility	Random projection	79.0	632	100.0	1.124	0.433
Divisibility	Class swap	3.1	632	4.0	-0.861	55.704
Divisibility	Shuffled axis	81.6	632	99.8	1.047	7.369
Primality	Unedited	90.1	721	100.0	1.798	0.000
Primality	Class (true)	83.1	721	92.2	1.206	18.042
Primality	Class (pred.)	91.1	721	92.2	1.206	18.042
Primality	Class axis	84.8	721	94.0	1.824	10.908
Primality	Random projection	90.1	721	100.0	1.799	0.457
Primality	Class swap	21.4	721	23.7	-0.587	79.791
Primality	Shuffled axis	85.8	721	95.1	1.891	11.437

Table 23: Qwen2.5-7B training/test split repeats for the Yes/No numerical protocol. Each row fits its own training shift and directions and evaluates its corresponding 100 held-out integers. The two separation columns use all held-out examples and the originally correct subset. These repeats change the split, not model training.

Rule	Split seed	Test (%)	ΔU: all	ΔU: correct	95% interval: all
Magnitude	42	83.0	0.595	0.609	[0.553, 0.671]
Magnitude	43	78.0	0.549	0.549	[0.510, 0.619]
Magnitude	44	82.0	0.549	0.635	[0.504, 0.624]
Parity	42	91.0	0.237	0.256	[0.212, 0.312]
Parity	43	94.0	0.231	0.247	[0.204, 0.302]
Parity	44	94.0	0.225	0.236	[0.205, 0.290]
Divisibility	42	95.0	0.101	0.108	[0.084, 0.167]
Divisibility	43	94.0	0.117	0.127	[0.096, 0.185]
Divisibility	44	93.0	0.098	0.101	[0.090, 0.143]
Primality	42	92.0	0.083	0.082	[0.056, 0.155]
Primality	43	89.0	0.118	0.120	[0.082, 0.194]
Primality	44	89.0	0.095	0.094	[0.054, 0.182]

Table 24: Final-residual interventions across five models at $\gamma = 1$. Each cell gives accuracy (%) on all held-out inputs / the originally correct subset. Class denotes true-label class-vector deletion and Axis denotes label-free class-axis deletion. These controls use their original edit amplitudes. Large changes in weak-baseline settings do not establish successful rule computation.

Model	Rule	None	Class	Axis	Random	Swap
Qwen2.5-7B	Parity	91.0/100.0	50.0/54.9	54.0/59.3	91.0/100.0	11.0/12.1
Qwen2.5-7B	Divisibility	95.0/100.0	67.0/70.5	67.0/70.5	95.0/100.0	21.0/22.1
Gemma-2-9B	Parity	97.0/100.0	67.0/69.1	61.0/59.8	97.0/100.0	0.0/0.0

Table 24 (continued)

Model	Rule	None	Class	Axis	Random	Swap
Gemma-2-9B	Divisibility	92.0/100.0	40.0/43.5	41.0/44.6	92.0/100.0	0.0/0.0
Qwen3-4B	Parity	56.0/100.0	49.0/87.5	50.0/89.3	56.0/100.0	38.0/67.9
Qwen3-4B	Divisibility	33.0/100.0	33.0/100.0	33.0/100.0	33.0/100.0	33.0/100.0
Llama-3.1-8B	Parity	80.0/100.0	48.0/60.0	51.0/63.7	80.0/100.0	4.0/5.0
Llama-3.1-8B	Divisibility	71.0/100.0	67.0/94.4	67.0/94.4	71.0/100.0	31.0/43.7
Mistral-7B	Parity	52.0/100.0	51.0/98.1	50.0/96.2	52.0/100.0	45.0/86.5
Mistral-7B	Divisibility	33.0/100.0	33.0/100.0	33.0/100.0	33.0/100.0	33.0/100.0

Table 25: Qwen2.5-7B alignment between Arabic-numeral and number-word formats. The entries are cosines between format-specific class mean vectors. Class update uses the centered residual update, and innovation uses its component orthogonal to the baseline representation. These are geometric comparisons of fitted class vectors, not cross-format intervention tests.

Rule	Vector	Class 0	Class 1
Magnitude	Class update	0.381	0.300
Magnitude	Innovation	0.417	0.418
Parity	Class update	0.551	0.512
Parity	Innovation	0.605	0.465
Divisibility	Class update	0.625	0.581
Divisibility	Innovation	0.609	0.378
Primality	Class update	0.307	0.176
Primality	Innovation	0.489	0.410
AND	Class update	0.506	0.429
AND	Innovation	0.549	0.540

Table 26: Qwen2.5-7B innovation separation across all four numerical rules and every stored layer. Each cell gives ΔU on all 100 held-out inputs / the originally correct subset. The layer convention is the same as Table 14. Only parity has the full layerwise intervention sweep.

Layer	Magnitude	Parity	Divisibility	Primality
0	n/a/n/a	n/a/n/a	n/a/n/a	n/a/n/a
1	0.902/0.889	0.083/0.093	-0.015/-0.015	0.005/-0.011
2	0.943/0.909	0.108/0.110	-0.011/-0.010	0.032/0.022
3	0.883/0.845	0.087/0.085	-0.012/-0.011	0.028/0.018
4	0.876/0.836	0.077/0.074	-0.012/-0.010	0.029/0.018
5	0.749/0.713	0.087/0.081	-0.009/-0.004	0.036/0.025
6	0.730/0.693	0.082/0.077	-0.007/0.000	0.035/0.023
7	0.741/0.699	0.076/0.074	-0.004/0.006	0.040/0.025
8	0.679/0.634	0.079/0.081	-0.002/0.010	0.046/0.031
9	0.660/0.617	0.074/0.078	0.002/0.014	0.044/0.029
10	0.594/0.560	0.079/0.083	0.017/0.028	0.046/0.033
11	0.577/0.541	0.069/0.072	0.020/0.031	0.043/0.030
12	0.552/0.523	0.072/0.074	0.023/0.035	0.050/0.038
13	0.532/0.510	0.088/0.089	0.025/0.038	0.061/0.052
14	0.504/0.490	0.097/0.098	0.027/0.038	0.075/0.066
15	0.489/0.490	0.104/0.106	0.029/0.040	0.080/0.071
16	0.463/0.467	0.103/0.105	0.030/0.038	0.083/0.075
17	0.466/0.466	0.117/0.118	0.031/0.039	0.095/0.089
18	0.475/0.485	0.148/0.149	0.064/0.071	0.114/0.109
19	0.521/0.559	0.192/0.190	0.082/0.088	0.122/0.122
20	0.595/0.638	0.307/0.315	0.118/0.126	0.147/0.146
21	0.629/0.670	0.348/0.360	0.138/0.146	0.162/0.159
22	0.647/0.682	0.357/0.371	0.138/0.146	0.165/0.161
23	0.709/0.738	0.382/0.402	0.148/0.155	0.151/0.147

Table 26 (continued)

Layer	Magnitude	Parity	Divisibility	Primality
24	0.717/0.733	0.351/0.374	0.150/0.157	0.127/0.122
25	0.726/0.735	0.312/0.334	0.136/0.142	0.113/0.106
26	0.671/0.666	0.252/0.277	0.112/0.117	0.091/0.085
27	0.651/0.648	0.244/0.269	0.109/0.115	0.089/0.083
28	0.595/0.609	0.237/0.256	0.101/0.108	0.083/0.082

Table 27: Qwen2.5-7B final-residual intervention repeats for the additional training/test splits, at $\gamma = 1$. All accuracy uses the complete test split; correct accuracy, margin and edit norm use the originally correct subset. The split-42 results appear in Table 12.

Rule	Seed	Edit	All (%)	Correct (%)	Margin	Edit norm
Parity	43	Unedited	94.0	100.0	1.285	0.000
Parity	43	Class (true)	50.0	53.2	0.012	31.286
Parity	43	Class (pred.)	56.0	53.2	0.012	31.286
Parity	43	Class axis	50.0	53.2	0.436	21.492
Parity	43	Random projection	94.0	100.0	1.285	0.293
Parity	43	Class swap	6.0	6.4	-1.250	62.416
Parity	43	Shuffled axis	84.0	89.4	1.047	9.053
Parity	44	Unedited	94.0	100.0	1.301	0.000
Parity	44	Class (true)	49.0	52.1	0.023	31.541
Parity	44	Class (pred.)	55.0	52.1	0.023	31.541
Parity	44	Class axis	50.0	53.2	0.461	21.269
Parity	44	Random projection	94.0	100.0	1.298	0.592
Parity	44	Class swap	7.0	7.4	-1.244	62.987
Parity	44	Shuffled axis	90.0	95.7	1.179	5.985
Divisibility	43	Unedited	94.0	100.0	1.331	0.000
Divisibility	43	Class (true)	67.0	71.3	0.469	22.158
Divisibility	43	Class (pred.)	73.0	71.3	0.469	22.158
Divisibility	43	Class axis	67.0	71.3	0.803	14.112
Divisibility	43	Random projection	94.0	100.0	1.331	0.312
Divisibility	43	Class swap	15.0	16.0	-0.727	53.630
Divisibility	43	Shuffled axis	79.0	84.0	1.198	12.016
Divisibility	44	Unedited	93.0	100.0	1.374	0.000
Divisibility	44	Class (true)	67.0	72.0	0.585	21.960
Divisibility	44	Class (pred.)	74.0	72.0	0.585	21.960
Divisibility	44	Class axis	67.0	72.0	0.896	14.105
Divisibility	44	Random projection	93.0	100.0	1.374	0.574
Divisibility	44	Class swap	27.0	29.0	-0.510	53.775
Divisibility	44	Shuffled axis	93.0	100.0	1.372	2.113

F ANSWER-INTERFACE DIAGNOSTIC

The diagnostic compares the stated raw prompts with each tokenizer's chat rendering. Chat prompts contain one user message and request the generation prefix. Qwen3 explicitly disables its optional thinking mode for this comparison. The diagnostic is performed on Qwen2.5 and Qwen3 for all four elementary numerical tasks, using all 200 numbers. It does not alter the raw-text mapping intervention protocol.

Answer forms are checked against each model's tokenizer before scoring. The audit includes Yes/No forms with and without a leading space and their declared capitalization variants. A form is eligible for next-token scoring only if it encodes as one token. Distinct strings that map to the same token ID contribute probability once. The full-vocabulary greedy token is considered correct only if it belongs to the audited single-token forms for the true class; any other token is counted as incorrect.

Pairwise accuracy and answer probability measure different properties. A positive-answer token can outrank a negative-answer token even when both have negligible mass under the full vocabu-

lary. We therefore preserve both the pair ranking and the sum of audited answer-token probabilities. Qwen2.5's leading-space chat condition illustrates this distinction, while Qwen3's raw/chat comparison shows that rendering can change the actual next-token answer. These are next-token measurements, with no claim about the eventual correctness of a multi-token generated response.

G COMPLETE MAPPING-INTERVENTION RESULTS

The following tables report all intervention conditions for every layer, label vocabulary, and calibration split. Each target uses 100 held-out integers. Source identifies the rule used to estimate the direction; it can differ from the target rule. Symmetric and antisymmetric directions use normalized constituent class contrasts. Raw directions use the corresponding unnormalized contrasts.

Accuracy is reported as a percentage. Harm is the baseline signed margin minus the edited signed margin, and $\|e\|$ is the mean realized edit norm. Crossing is the percentage of integers whose centered projection coefficient changes sign. Matched denotes a norm-matched directional perturbation; Natural denotes unscaled projection removal. Native removes the original class axis. Identity runs the intervention path with a zero edit.

Random mean (8) averages each reported metric across eight independently evaluated random directions. The accompanying machine-readable table retains every individual random direction and every source rule. Conditions where one mapping gives a constant answer are included in full.

Table 28: Qwen2.5, AB/BA, split 1, layer 14: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	100.0	-0.012	0.348	0.0
Control	Random mean (8)	Matched	100.0	0.000	0.348	98.0
Parity	Raw symmetric	Matched	100.0	-0.012	0.348	43.0
Parity	Raw antisymmetric	Matched	100.0	-0.009	0.348	86.0
Parity	Symmetric	Matched	100.0	-0.012	0.348	45.0
Parity	Antisymmetric	Matched	100.0	-0.001	0.348	86.0
Div. by 3	Raw symmetric	Matched	100.0	0.002	0.348	49.0
Div. by 3	Raw antisymmetric	Matched	100.0	-0.002	0.348	87.0
Div. by 3	Symmetric	Matched	100.0	0.003	0.348	49.0
Div. by 3	Antisymmetric	Matched	100.0	-0.003	0.348	85.0
Parity	Symmetric	Natural	100.0	-0.011	0.348	0.0
Parity	Antisymmetric	Natural	100.0	-0.002	0.111	0.0
Target: Parity, mapping BA						
Control	Baseline	None	94.0	0.000	0.000	0.0
Control	Identity	None	94.0	0.000	0.000	0.0
Control	Native axis	Native	94.0	-0.004	0.267	0.0
Control	Random mean (8)	Matched	93.9	0.000	0.267	96.4
Parity	Raw symmetric	Matched	94.0	-0.004	0.267	60.0
Parity	Raw antisymmetric	Matched	94.0	-0.001	0.267	87.0
Parity	Symmetric	Matched	94.0	-0.004	0.267	61.0
Parity	Antisymmetric	Matched	94.0	0.001	0.267	88.0
Div. by 3	Raw symmetric	Matched	94.0	0.000	0.267	59.0
Div. by 3	Raw antisymmetric	Matched	94.0	0.002	0.267	79.0
Div. by 3	Symmetric	Matched	94.0	0.000	0.267	59.0
Div. by 3	Antisymmetric	Matched	94.0	0.002	0.267	79.0
Parity	Symmetric	Natural	94.0	-0.004	0.264	0.0
Parity	Antisymmetric	Natural	94.0	0.000	0.087	0.0
Target: Divisibility by 3, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	100.0	0.003	0.369	0.0
Control	Random mean (8)	Matched	100.0	0.000	0.369	99.0
Parity	Raw symmetric	Matched	100.0	-0.008	0.369	81.0
Parity	Raw antisymmetric	Matched	100.0	0.000	0.369	88.0
Parity	Symmetric	Matched	100.0	-0.008	0.369	81.0
Parity	Antisymmetric	Matched	100.0	0.001	0.369	89.0
Div. by 3	Raw symmetric	Matched	100.0	0.004	0.369	39.0
Div. by 3	Raw antisymmetric	Matched	100.0	-0.007	0.369	86.0
Div. by 3	Symmetric	Matched	100.0	0.004	0.369	39.0
Div. by 3	Antisymmetric	Matched	100.0	0.000	0.369	86.0
Div. by 3	Symmetric	Natural	100.0	0.004	0.376	0.0
Div. by 3	Antisymmetric	Natural	100.0	-0.001	0.101	0.0
Target: Divisibility by 3, mapping BA						
Control	Baseline	None	99.0	0.000	0.000	0.0
Control	Identity	None	99.0	0.000	0.000	0.0
Control	Native axis	Native	99.0	0.002	0.386	0.0
Control	Random mean (8)	Matched	99.0	0.000	0.386	98.5
Parity	Raw symmetric	Matched	99.0	-0.011	0.386	87.0
Parity	Raw antisymmetric	Matched	99.0	0.006	0.386	92.0
Parity	Symmetric	Matched	99.0	-0.011	0.386	87.0
Parity	Antisymmetric	Matched	99.0	0.002	0.386	95.0
Div. by 3	Raw symmetric	Matched	99.0	0.002	0.386	82.0
Div. by 3	Raw antisymmetric	Matched	99.0	0.003	0.386	90.0
Div. by 3	Symmetric	Matched	99.0	0.002	0.386	82.0
Div. by 3	Antisymmetric	Matched	99.0	0.003	0.386	90.0
Div. by 3	Symmetric	Natural	99.0	0.002	0.360	0.0
Div. by 3	Antisymmetric	Natural	99.0	0.001	0.161	0.0

Table 29: Qwen2.5, AB/BA, split 1, layer 22: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	89.0	3.803	15.996	0.0
Control	Random mean (8)	Matched	100.0	0.061	15.996	100.0
Parity	Raw symmetric	Matched	90.0	3.793	15.996	82.0
Parity	Raw antisymmetric	Matched	100.0	1.883	15.996	100.0
Parity	Symmetric	Matched	90.0	3.789	15.996	83.0
Parity	Antisymmetric	Matched	100.0	0.515	15.996	100.0
Div. by 3	Raw symmetric	Matched	98.0	3.475	15.996	96.0
Div. by 3	Raw antisymmetric	Matched	100.0	2.228	15.996	100.0
Div. by 3	Symmetric	Matched	98.0	3.468	15.996	96.0
Div. by 3	Antisymmetric	Matched	100.0	0.473	15.996	100.0
Parity	Symmetric	Natural	91.0	3.750	15.851	0.0
Parity	Antisymmetric	Natural	100.0	0.060	2.203	0.0
Target: Parity, mapping BA						
Control	Baseline	None	94.0	0.000	0.000	0.0
Control	Identity	None	94.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	2.903	13.831	0.0
Control	Random mean (8)	Matched	92.8	0.053	13.831	100.0
Parity	Raw symmetric	Matched	50.0	2.771	13.831	86.0
Parity	Raw antisymmetric	Matched	88.0	0.317	13.831	100.0
Parity	Symmetric	Matched	50.0	2.783	13.831	85.0
Parity	Antisymmetric	Matched	80.0	0.807	13.831	100.0
Div. by 3	Raw symmetric	Matched	50.0	2.808	13.831	92.0
Div. by 3	Raw antisymmetric	Matched	75.0	0.956	13.831	100.0
Div. by 3	Symmetric	Matched	50.0	2.815	13.831	92.0
Div. by 3	Antisymmetric	Matched	83.0	0.500	13.831	100.0
Parity	Symmetric	Natural	50.0	2.737	13.641	0.0
Parity	Antisymmetric	Natural	91.0	0.121	2.337	0.0
Target: Divisibility by 3, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	68.0	3.213	15.404	0.0
Control	Random mean (8)	Matched	100.0	0.053	15.404	100.0
Parity	Raw symmetric	Matched	68.0	3.402	15.404	97.0
Parity	Raw antisymmetric	Matched	100.0	1.665	15.404	99.0
Parity	Symmetric	Matched	68.0	3.399	15.404	97.0
Parity	Antisymmetric	Matched	100.0	0.507	15.404	99.0
Div. by 3	Raw symmetric	Matched	68.0	3.184	15.404	97.0
Div. by 3	Raw antisymmetric	Matched	100.0	2.001	15.404	100.0
Div. by 3	Symmetric	Matched	68.0	3.179	15.404	97.0
Div. by 3	Antisymmetric	Matched	100.0	0.430	15.404	100.0
Div. by 3	Symmetric	Natural	68.0	3.150	15.275	0.0
Div. by 3	Antisymmetric	Natural	100.0	0.047	1.992	0.0
Target: Divisibility by 3, mapping BA						
Control	Baseline	None	99.0	0.000	0.000	0.0
Control	Identity	None	99.0	0.000	0.000	0.0
Control	Native axis	Native	68.0	2.771	12.972	0.0
Control	Random mean (8)	Matched	99.0	0.058	12.972	99.5
Parity	Raw symmetric	Matched	68.0	2.550	12.972	96.0
Parity	Raw antisymmetric	Matched	99.0	0.339	12.972	97.0
Parity	Symmetric	Matched	68.0	2.561	12.972	95.0
Parity	Antisymmetric	Matched	99.0	0.555	12.972	97.0
Div. by 3	Raw symmetric	Matched	68.0	2.722	12.972	95.0
Div. by 3	Raw antisymmetric	Matched	96.0	0.903	12.972	99.0
Div. by 3	Symmetric	Matched	68.0	2.729	12.972	90.0
Div. by 3	Antisymmetric	Matched	99.0	0.415	12.972	99.0
Div. by 3	Symmetric	Natural	68.0	2.699	12.864	0.0
Div. by 3	Antisymmetric	Natural	99.0	0.049	1.686	0.0

Table 30: Qwen2.5, AB/BA, split 1, layer 28: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	4.242	70.523	0.0
Control	Random mean (8)	Matched	100.0	0.183	70.523	100.0
Parity	Raw symmetric	Matched	100.0	0.755	70.523	97.0
Parity	Raw antisymmetric	Matched	1.0	6.194	70.523	97.0
Parity	Symmetric	Matched	100.0	0.294	70.523	97.0
Parity	Antisymmetric	Matched	1.0	6.194	70.523	97.0
Div. by 3	Raw symmetric	Matched	100.0	0.606	70.523	100.0
Div. by 3	Raw antisymmetric	Matched	1.0	5.911	70.523	97.0
Div. by 3	Symmetric	Matched	100.0	-0.175	70.523	100.0
Div. by 3	Antisymmetric	Matched	1.0	6.003	70.523	97.0
Parity	Symmetric	Natural	100.0	0.226	55.063	0.0
Parity	Antisymmetric	Natural	54.0	3.881	44.188	0.0
Target: Parity, mapping BA						
Control	Baseline	None	94.0	0.000	0.000	0.0
Control	Identity	None	94.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	3.286	60.516	0.0
Control	Random mean (8)	Matched	93.4	-0.047	60.516	100.0
Parity	Raw symmetric	Matched	96.0	-0.642	60.516	100.0
Parity	Raw antisymmetric	Matched	23.0	5.391	60.516	100.0
Parity	Symmetric	Matched	96.0	-0.245	60.516	100.0
Parity	Antisymmetric	Matched	23.0	5.396	60.516	99.0
Div. by 3	Raw symmetric	Matched	96.0	-0.541	60.516	99.0
Div. by 3	Raw antisymmetric	Matched	29.0	5.149	60.516	99.0
Div. by 3	Symmetric	Matched	92.0	0.138	60.516	99.0
Div. by 3	Antisymmetric	Matched	26.0	5.234	60.516	99.0
Parity	Symmetric	Natural	95.0	-0.190	46.217	0.0
Parity	Antisymmetric	Natural	50.0	3.486	39.067	0.0
Target: Divisibility by 3, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	38.0	3.776	68.470	0.0
Control	Random mean (8)	Matched	100.0	0.193	68.470	100.0
Parity	Raw symmetric	Matched	100.0	0.731	68.470	97.0
Parity	Raw antisymmetric	Matched	1.0	5.919	68.470	97.0
Parity	Symmetric	Matched	100.0	0.287	68.470	99.0
Parity	Antisymmetric	Matched	1.0	5.912	68.470	97.0
Div. by 3	Raw symmetric	Matched	100.0	0.575	68.470	100.0
Div. by 3	Raw antisymmetric	Matched	1.0	5.646	68.470	97.0
Div. by 3	Symmetric	Matched	100.0	-0.174	68.470	100.0
Div. by 3	Antisymmetric	Matched	1.0	5.731	68.470	97.0
Div. by 3	Symmetric	Natural	100.0	-0.137	50.761	0.0
Div. by 3	Antisymmetric	Natural	35.0	3.854	46.019	0.0
Target: Divisibility by 3, mapping BA						
Control	Baseline	None	99.0	0.000	0.000	0.0
Control	Identity	None	99.0	0.000	0.000	0.0
Control	Native axis	Native	68.0	3.093	52.698	0.0
Control	Random mean (8)	Matched	99.1	-0.054	52.698	100.0
Parity	Raw symmetric	Matched	100.0	-0.555	52.698	97.0
Parity	Raw antisymmetric	Matched	67.0	4.575	52.698	100.0
Parity	Symmetric	Matched	99.0	-0.220	52.698	97.0
Parity	Antisymmetric	Matched	67.0	4.581	52.698	99.0
Div. by 3	Raw symmetric	Matched	100.0	-0.461	52.698	100.0
Div. by 3	Raw antisymmetric	Matched	68.0	4.369	52.698	100.0
Div. by 3	Symmetric	Matched	99.0	0.112	52.698	100.0
Div. by 3	Antisymmetric	Matched	68.0	4.443	52.698	98.0
Div. by 3	Symmetric	Natural	99.0	0.079	38.517	0.0
Div. by 3	Antisymmetric	Natural	68.0	3.023	35.968	0.0

Table 31: Qwen2.5, AB/BA, split 2, layer 22: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping AB						
Control	Baseline	None	99.0	0.000	0.000	0.0
Control	Identity	None	99.0	0.000	0.000	0.0
Control	Native axis	Native	87.0	3.736	16.016	0.0
Control	Random mean (8)	Matched	99.0	0.062	16.016	100.0
Parity	Raw symmetric	Matched	87.0	3.728	16.016	77.0
Parity	Raw antisymmetric	Matched	99.0	1.861	16.016	100.0
Parity	Symmetric	Matched	87.0	3.723	16.016	80.0
Parity	Antisymmetric	Matched	99.0	0.491	16.016	100.0
Div. by 3	Raw symmetric	Matched	97.0	3.419	16.016	98.0
Div. by 3	Raw antisymmetric	Matched	99.0	2.215	16.016	100.0
Div. by 3	Symmetric	Matched	97.0	3.412	16.016	98.0
Div. by 3	Antisymmetric	Matched	99.0	0.466	16.016	100.0
Parity	Symmetric	Natural	88.0	3.691	15.890	0.0
Parity	Antisymmetric	Natural	99.0	0.052	2.103	0.0
Target: Parity, mapping BA						
Control	Baseline	None	90.0	0.000	0.000	0.0
Control	Identity	None	90.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	2.881	13.864	0.0
Control	Random mean (8)	Matched	90.0	0.050	13.864	100.0
Parity	Raw symmetric	Matched	50.0	2.748	13.864	91.0
Parity	Raw antisymmetric	Matched	86.0	0.323	13.864	100.0
Parity	Symmetric	Matched	50.0	2.760	13.864	90.0
Parity	Antisymmetric	Matched	78.0	0.828	13.864	100.0
Div. by 3	Raw symmetric	Matched	50.0	2.790	13.864	95.0
Div. by 3	Raw antisymmetric	Matched	73.0	0.956	13.864	100.0
Div. by 3	Symmetric	Matched	50.0	2.797	13.864	95.0
Div. by 3	Antisymmetric	Matched	83.0	0.520	13.864	100.0
Parity	Symmetric	Natural	50.0	2.711	13.661	0.0
Parity	Antisymmetric	Natural	90.0	0.128	2.400	0.0
Target: Divisibility by 3, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	67.0	3.282	15.369	0.0
Control	Random mean (8)	Matched	100.0	0.055	15.369	99.4
Parity	Raw symmetric	Matched	67.0	3.468	15.369	97.0
Parity	Raw antisymmetric	Matched	100.0	1.727	15.369	98.0
Parity	Symmetric	Matched	67.0	3.465	15.369	97.0
Parity	Antisymmetric	Matched	100.0	0.517	15.369	97.0
Div. by 3	Raw symmetric	Matched	67.0	3.256	15.369	96.0
Div. by 3	Raw antisymmetric	Matched	99.0	2.069	15.369	99.0
Div. by 3	Symmetric	Matched	67.0	3.251	15.369	97.0
Div. by 3	Antisymmetric	Matched	100.0	0.417	15.369	99.0
Div. by 3	Symmetric	Natural	67.0	3.223	15.242	0.0
Div. by 3	Antisymmetric	Natural	100.0	0.044	1.976	0.0
Target: Divisibility by 3, mapping BA						
Control	Baseline	None	99.0	0.000	0.000	0.0
Control	Identity	None	99.0	0.000	0.000	0.0
Control	Native axis	Native	67.0	2.812	12.995	0.0
Control	Random mean (8)	Matched	99.0	0.053	12.995	98.9
Parity	Raw symmetric	Matched	67.0	2.587	12.995	96.0
Parity	Raw antisymmetric	Matched	99.0	0.367	12.995	97.0
Parity	Symmetric	Matched	67.0	2.598	12.995	96.0
Parity	Antisymmetric	Matched	99.0	0.549	12.995	97.0
Div. by 3	Raw symmetric	Matched	67.0	2.760	12.995	88.0
Div. by 3	Raw antisymmetric	Matched	95.0	0.929	12.995	97.0
Div. by 3	Symmetric	Matched	67.0	2.768	12.995	87.0
Div. by 3	Antisymmetric	Matched	99.0	0.441	12.995	98.0
Div. by 3	Symmetric	Natural	67.0	2.737	12.885	0.0
Div. by 3	Antisymmetric	Natural	99.0	0.052	1.704	0.0

Table 32: Qwen2.5, AB/BA, split 2, layer 28: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping AB						
Control	Baseline	None	99.0	0.000	0.000	0.0
Control	Identity	None	99.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	4.166	70.476	0.0
Control	Random mean (8)	Matched	99.0	0.178	70.476	100.0
Parity	Raw symmetric	Matched	99.0	0.735	70.476	95.0
Parity	Raw antisymmetric	Matched	3.0	6.067	70.476	98.0
Parity	Symmetric	Matched	99.0	0.268	70.476	95.0
Parity	Antisymmetric	Matched	3.0	6.070	70.476	98.0
Div. by 3	Raw symmetric	Matched	99.0	0.578	70.476	100.0
Div. by 3	Raw antisymmetric	Matched	4.0	5.806	70.476	98.0
Div. by 3	Symmetric	Matched	99.0	-0.184	70.476	100.0
Div. by 3	Antisymmetric	Matched	3.0	5.896	70.476	98.0
Parity	Symmetric	Natural	99.0	0.209	55.890	0.0
Parity	Antisymmetric	Natural	54.0	3.701	43.459	0.0
Target: Parity, mapping BA						
Control	Baseline	None	90.0	0.000	0.000	0.0
Control	Identity	None	90.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	3.285	60.592	0.0
Control	Random mean (8)	Matched	90.4	-0.048	60.592	100.0
Parity	Raw symmetric	Matched	93.0	-0.626	60.592	100.0
Parity	Raw antisymmetric	Matched	20.0	5.325	60.592	100.0
Parity	Symmetric	Matched	92.0	-0.218	60.592	100.0
Parity	Antisymmetric	Matched	20.0	5.333	60.592	100.0
Div. by 3	Raw symmetric	Matched	93.0	-0.518	60.592	100.0
Div. by 3	Raw antisymmetric	Matched	25.0	5.100	60.592	100.0
Div. by 3	Symmetric	Matched	90.0	0.150	60.592	100.0
Div. by 3	Antisymmetric	Matched	23.0	5.183	60.592	100.0
Parity	Symmetric	Natural	90.0	-0.170	46.263	0.0
Parity	Antisymmetric	Natural	50.0	3.453	39.130	0.0
Target: Divisibility by 3, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	45.0	3.794	68.831	0.0
Control	Random mean (8)	Matched	100.0	0.196	68.831	100.0
Parity	Raw symmetric	Matched	100.0	0.714	68.831	99.0
Parity	Raw antisymmetric	Matched	0.0	5.924	68.831	98.0
Parity	Symmetric	Matched	100.0	0.257	68.831	99.0
Parity	Antisymmetric	Matched	0.0	5.920	68.831	98.0
Div. by 3	Raw symmetric	Matched	100.0	0.559	68.831	100.0
Div. by 3	Raw antisymmetric	Matched	0.0	5.668	68.831	97.0
Div. by 3	Symmetric	Matched	100.0	-0.190	68.831	100.0
Div. by 3	Antisymmetric	Matched	0.0	5.752	68.831	98.0
Div. by 3	Symmetric	Natural	100.0	-0.149	50.930	0.0
Div. by 3	Antisymmetric	Natural	41.0	3.874	46.306	0.0
Target: Divisibility by 3, mapping BA						
Control	Baseline	None	99.0	0.000	0.000	0.0
Control	Identity	None	99.0	0.000	0.000	0.0
Control	Native axis	Native	67.0	3.133	53.031	0.0
Control	Random mean (8)	Matched	99.0	-0.054	53.031	100.0
Parity	Raw symmetric	Matched	99.0	-0.557	53.031	96.0
Parity	Raw antisymmetric	Matched	66.0	4.598	53.031	98.0
Parity	Symmetric	Matched	99.0	-0.207	53.031	96.0
Parity	Antisymmetric	Matched	66.0	4.607	53.031	97.0
Div. by 3	Raw symmetric	Matched	99.0	-0.452	53.031	100.0
Div. by 3	Raw antisymmetric	Matched	67.0	4.403	53.031	98.0
Div. by 3	Symmetric	Matched	99.0	0.123	53.031	100.0
Div. by 3	Antisymmetric	Matched	67.0	4.478	53.031	97.0
Div. by 3	Symmetric	Natural	99.0	0.087	38.839	0.0
Div. by 3	Antisymmetric	Natural	67.0	3.030	36.126	0.0

Table 33: Qwen2.5, AB/BA, split 3, layer 22: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	87.0	3.913	16.349	0.0
Control	Random mean (8)	Matched	100.0	0.064	16.349	100.0
Parity	Raw symmetric	Matched	88.0	3.900	16.349	86.0
Parity	Raw antisymmetric	Matched	100.0	1.936	16.349	100.0
Parity	Symmetric	Matched	88.0	3.895	16.349	87.0
Parity	Antisymmetric	Matched	100.0	0.552	16.349	100.0
Div. by 3	Raw symmetric	Matched	98.0	3.573	16.349	96.0
Div. by 3	Raw antisymmetric	Matched	100.0	2.276	16.349	100.0
Div. by 3	Symmetric	Matched	98.0	3.566	16.349	97.0
Div. by 3	Antisymmetric	Matched	100.0	0.472	16.349	100.0
Parity	Symmetric	Natural	91.0	3.850	16.181	0.0
Parity	Antisymmetric	Natural	100.0	0.068	2.355	0.0
Target: Parity, mapping BA						
Control	Baseline	None	95.0	0.000	0.000	0.0
Control	Identity	None	95.0	0.000	0.000	0.0
Control	Native axis	Native	51.0	2.968	14.103	0.0
Control	Random mean (8)	Matched	94.0	0.056	14.103	100.0
Parity	Raw symmetric	Matched	51.0	2.839	14.103	85.0
Parity	Raw antisymmetric	Matched	87.0	0.344	14.103	100.0
Parity	Symmetric	Matched	51.0	2.851	14.103	83.0
Parity	Antisymmetric	Matched	84.0	0.789	14.103	100.0
Div. by 3	Raw symmetric	Matched	51.0	2.870	14.103	89.0
Div. by 3	Raw antisymmetric	Matched	76.0	0.977	14.103	100.0
Div. by 3	Symmetric	Matched	51.0	2.877	14.103	89.0
Div. by 3	Antisymmetric	Matched	85.0	0.508	14.103	100.0
Parity	Symmetric	Natural	51.0	2.810	13.932	0.0
Parity	Antisymmetric	Natural	93.0	0.111	2.224	0.0
Target: Divisibility by 3, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	67.0	3.359	15.873	0.0
Control	Random mean (8)	Matched	100.0	0.065	15.873	98.4
Parity	Raw symmetric	Matched	67.0	3.564	15.873	95.0
Parity	Raw antisymmetric	Matched	100.0	1.759	15.873	96.0
Parity	Symmetric	Matched	67.0	3.560	15.873	96.0
Parity	Antisymmetric	Matched	100.0	0.554	15.873	96.0
Div. by 3	Raw symmetric	Matched	67.0	3.333	15.873	95.0
Div. by 3	Raw antisymmetric	Matched	100.0	2.088	15.873	99.0
Div. by 3	Symmetric	Matched	67.0	3.328	15.873	96.0
Div. by 3	Antisymmetric	Matched	100.0	0.430	15.873	98.0
Div. by 3	Symmetric	Natural	67.0	3.299	15.743	0.0
Div. by 3	Antisymmetric	Natural	100.0	0.045	2.043	0.0
Target: Divisibility by 3, mapping BA						
Control	Baseline	None	99.0	0.000	0.000	0.0
Control	Identity	None	99.0	0.000	0.000	0.0
Control	Native axis	Native	67.0	2.886	13.435	0.0
Control	Random mean (8)	Matched	99.0	0.065	13.435	98.4
Parity	Raw symmetric	Matched	67.0	2.668	13.435	94.0
Parity	Raw antisymmetric	Matched	99.0	0.373	13.435	98.0
Parity	Symmetric	Matched	67.0	2.678	13.435	94.0
Parity	Antisymmetric	Matched	99.0	0.529	13.435	97.0
Div. by 3	Raw symmetric	Matched	67.0	2.837	13.435	91.0
Div. by 3	Raw antisymmetric	Matched	95.0	0.949	13.435	98.0
Div. by 3	Symmetric	Matched	67.0	2.844	13.435	89.0
Div. by 3	Antisymmetric	Matched	99.0	0.434	13.435	98.0
Div. by 3	Symmetric	Natural	67.0	2.812	13.322	0.0
Div. by 3	Antisymmetric	Natural	99.0	0.050	1.753	0.0

Table 34: Qwen2.5, AB/BA, split 3, layer 28: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	49.0	4.323	71.963	0.0
Control	Random mean (8)	Matched	100.0	0.190	71.963	100.0
Parity	Raw symmetric	Matched	100.0	0.758	71.963	100.0
Parity	Raw antisymmetric	Matched	0.0	6.362	71.963	98.0
Parity	Symmetric	Matched	100.0	0.299	71.963	100.0
Parity	Antisymmetric	Matched	0.0	6.361	71.963	98.0
Div. by 3	Raw symmetric	Matched	100.0	0.635	71.963	100.0
Div. by 3	Raw antisymmetric	Matched	0.0	6.086	71.963	98.0
Div. by 3	Symmetric	Matched	100.0	-0.179	71.963	100.0
Div. by 3	Antisymmetric	Matched	0.0	6.178	71.963	98.0
Parity	Symmetric	Natural	100.0	0.228	55.650	0.0
Parity	Antisymmetric	Natural	53.0	4.019	45.632	0.0
Target: Parity, mapping BA						
Control	Baseline	None	95.0	0.000	0.000	0.0
Control	Identity	None	95.0	0.000	0.000	0.0
Control	Native axis	Native	51.0	3.317	61.154	0.0
Control	Random mean (8)	Matched	94.8	-0.051	61.154	100.0
Parity	Raw symmetric	Matched	99.0	-0.639	61.154	100.0
Parity	Raw antisymmetric	Matched	20.0	5.455	61.154	100.0
Parity	Symmetric	Matched	98.0	-0.247	61.154	100.0
Parity	Antisymmetric	Matched	20.0	5.460	61.154	99.0
Div. by 3	Raw symmetric	Matched	98.0	-0.561	61.154	100.0
Div. by 3	Raw antisymmetric	Matched	28.0	5.222	61.154	99.0
Div. by 3	Symmetric	Matched	94.0	0.140	61.154	99.0
Div. by 3	Antisymmetric	Matched	25.0	5.307	61.154	99.0
Parity	Symmetric	Natural	96.0	-0.192	46.756	0.0
Parity	Antisymmetric	Natural	51.0	3.522	39.417	0.0
Target: Divisibility by 3, mapping AB						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	31.0	3.960	70.880	0.0
Control	Random mean (8)	Matched	100.0	0.207	70.880	100.0
Parity	Raw symmetric	Matched	100.0	0.742	70.880	99.0
Parity	Raw antisymmetric	Matched	0.0	6.133	70.880	98.0
Parity	Symmetric	Matched	100.0	0.291	70.880	100.0
Parity	Antisymmetric	Matched	0.0	6.125	70.880	98.0
Div. by 3	Raw symmetric	Matched	100.0	0.612	70.880	100.0
Div. by 3	Raw antisymmetric	Matched	0.0	5.864	70.880	98.0
Div. by 3	Symmetric	Matched	100.0	-0.183	70.880	100.0
Div. by 3	Antisymmetric	Matched	0.0	5.950	70.880	99.0
Div. by 3	Symmetric	Natural	100.0	-0.144	52.460	0.0
Div. by 3	Antisymmetric	Natural	29.0	4.007	47.674	0.0
Target: Divisibility by 3, mapping BA						
Control	Baseline	None	99.0	0.000	0.000	0.0
Control	Identity	None	99.0	0.000	0.000	0.0
Control	Native axis	Native	67.0	3.252	54.464	0.0
Control	Random mean (8)	Matched	99.1	-0.058	54.464	99.6
Parity	Raw symmetric	Matched	100.0	-0.573	54.464	98.0
Parity	Raw antisymmetric	Matched	66.0	4.760	54.464	99.0
Parity	Symmetric	Matched	99.0	-0.232	54.464	98.0
Parity	Antisymmetric	Matched	66.0	4.766	54.464	99.0
Div. by 3	Raw symmetric	Matched	100.0	-0.493	54.464	99.0
Div. by 3	Raw antisymmetric	Matched	67.0	4.557	54.464	99.0
Div. by 3	Symmetric	Matched	99.0	0.116	54.464	99.0
Div. by 3	Antisymmetric	Matched	67.0	4.633	54.464	99.0
Div. by 3	Symmetric	Natural	99.0	0.082	40.177	0.0
Div. by 3	Antisymmetric	Natural	67.0	3.123	36.789	0.0

Table 35: Qwen2.5, CD/DC, split 1, layer 14: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping CD						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	100.0	-0.008	0.184	0.0
Control	Random mean (8)	Matched	100.0	0.000	0.184	94.7
Parity	Raw symmetric	Matched	100.0	-0.007	0.184	64.0
Parity	Raw antisymmetric	Matched	100.0	-0.002	0.184	58.0
Parity	Symmetric	Matched	100.0	-0.007	0.184	64.0
Parity	Antisymmetric	Matched	100.0	-0.002	0.184	59.0
Div. by 3	Raw symmetric	Matched	100.0	0.001	0.184	41.0
Div. by 3	Raw antisymmetric	Matched	100.0	-0.001	0.184	50.0
Div. by 3	Symmetric	Matched	100.0	0.001	0.184	41.0
Div. by 3	Antisymmetric	Matched	100.0	-0.001	0.184	53.0
Parity	Symmetric	Natural	100.0	-0.007	0.174	0.0
Parity	Antisymmetric	Natural	100.0	-0.001	0.153	0.0
Target: Parity, mapping DC						
Control	Baseline	None	50.0	0.000	0.000	0.0
Control	Identity	None	50.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	0.003	0.179	0.0
Control	Random mean (8)	Matched	50.0	0.000	0.179	93.9
Parity	Raw symmetric	Matched	50.0	0.003	0.179	45.0
Parity	Raw antisymmetric	Matched	50.0	-0.001	0.179	78.0
Parity	Symmetric	Matched	50.0	0.003	0.179	45.0
Parity	Antisymmetric	Matched	50.0	-0.001	0.179	78.0
Div. by 3	Raw symmetric	Matched	50.0	0.000	0.179	41.0
Div. by 3	Raw antisymmetric	Matched	50.0	-0.001	0.179	45.0
Div. by 3	Symmetric	Matched	50.0	0.000	0.179	41.0
Div. by 3	Antisymmetric	Matched	50.0	-0.001	0.179	46.0
Parity	Symmetric	Natural	50.0	0.003	0.180	0.0
Parity	Antisymmetric	Natural	50.0	0.000	0.084	0.0
Target: Divisibility by 3, mapping CD						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	100.0	0.000	0.278	0.0
Control	Random mean (8)	Matched	100.0	0.000	0.278	97.1
Parity	Raw symmetric	Matched	100.0	-0.007	0.278	81.0
Parity	Raw antisymmetric	Matched	100.0	-0.001	0.278	91.0
Parity	Symmetric	Matched	100.0	-0.007	0.278	81.0
Parity	Antisymmetric	Matched	100.0	-0.001	0.278	91.0
Div. by 3	Raw symmetric	Matched	100.0	0.001	0.278	10.0
Div. by 3	Raw antisymmetric	Matched	100.0	0.000	0.278	83.0
Div. by 3	Symmetric	Matched	100.0	0.001	0.278	10.0
Div. by 3	Antisymmetric	Matched	100.0	0.000	0.278	83.0
Div. by 3	Symmetric	Natural	100.0	0.001	0.315	0.0
Div. by 3	Antisymmetric	Natural	100.0	-0.001	0.091	0.0
Target: Divisibility by 3, mapping DC						
Control	Baseline	None	98.0	0.000	0.000	0.0
Control	Identity	None	98.0	0.000	0.000	0.0
Control	Native axis	Native	98.0	-0.001	0.370	0.0
Control	Random mean (8)	Matched	98.0	0.001	0.370	98.0
Parity	Raw symmetric	Matched	98.0	-0.001	0.370	86.0
Parity	Raw antisymmetric	Matched	98.0	-0.001	0.370	95.0
Parity	Symmetric	Matched	98.0	-0.001	0.370	86.0
Parity	Antisymmetric	Matched	98.0	-0.001	0.370	96.0
Div. by 3	Raw symmetric	Matched	98.0	-0.001	0.370	87.0
Div. by 3	Raw antisymmetric	Matched	98.0	0.000	0.370	95.0
Div. by 3	Symmetric	Matched	98.0	-0.001	0.370	87.0
Div. by 3	Antisymmetric	Matched	98.0	0.000	0.370	94.0
Div. by 3	Symmetric	Natural	98.0	-0.001	0.333	0.0
Div. by 3	Antisymmetric	Natural	98.0	0.000	0.180	0.0

Table 36: Qwen2.5, CD/DC, split 1, layer 22: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping CD						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	79.0	2.770	11.389	0.0
Control	Random mean (8)	Matched	100.0	0.033	11.389	100.0
Parity	Raw symmetric	Matched	66.0	2.874	11.389	100.0
Parity	Raw antisymmetric	Matched	100.0	0.872	11.389	100.0
Parity	Symmetric	Matched	66.0	2.872	11.389	100.0
Parity	Antisymmetric	Matched	100.0	-0.260	11.389	100.0
Div. by 3	Raw symmetric	Matched	84.0	2.642	11.389	90.0
Div. by 3	Raw antisymmetric	Matched	98.0	2.006	11.389	100.0
Div. by 3	Symmetric	Matched	84.0	2.635	11.389	90.0
Div. by 3	Antisymmetric	Matched	100.0	-0.209	11.389	100.0
Parity	Symmetric	Natural	73.0	2.823	11.221	0.0
Parity	Antisymmetric	Natural	100.0	-0.047	1.950	0.0
Target: Parity, mapping DC						
Control	Baseline	None	50.0	0.000	0.000	0.0
Control	Identity	None	50.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	1.020	12.406	0.0
Control	Random mean (8)	Matched	52.8	-0.068	12.406	100.0
Parity	Raw symmetric	Matched	50.0	0.932	12.406	93.0
Parity	Raw antisymmetric	Matched	50.0	0.803	12.406	100.0
Parity	Symmetric	Matched	50.0	0.927	12.406	94.0
Parity	Antisymmetric	Matched	50.0	0.607	12.406	100.0
Div. by 3	Raw symmetric	Matched	50.0	0.800	12.406	92.0
Div. by 3	Raw antisymmetric	Matched	50.0	0.846	12.406	100.0
Div. by 3	Symmetric	Matched	50.0	0.788	12.406	94.0
Div. by 3	Antisymmetric	Matched	50.0	0.415	12.406	100.0
Parity	Symmetric	Natural	50.0	0.911	12.203	0.0
Parity	Antisymmetric	Natural	50.0	0.112	2.246	0.0
Target: Divisibility by 3, mapping CD						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	68.0	2.172	9.665	0.0
Control	Random mean (8)	Matched	100.0	0.034	9.665	99.9
Parity	Raw symmetric	Matched	68.0	2.340	9.665	95.0
Parity	Raw antisymmetric	Matched	100.0	0.618	9.665	99.0
Parity	Symmetric	Matched	68.0	2.339	9.665	95.0
Parity	Antisymmetric	Matched	100.0	-0.180	9.665	99.0
Div. by 3	Raw symmetric	Matched	68.0	2.259	9.665	88.0
Div. by 3	Raw antisymmetric	Matched	91.0	1.786	9.665	98.0
Div. by 3	Symmetric	Matched	68.0	2.252	9.665	85.0
Div. by 3	Antisymmetric	Matched	100.0	-0.212	9.665	99.0
Div. by 3	Symmetric	Natural	68.0	2.224	9.542	0.0
Div. by 3	Antisymmetric	Natural	100.0	-0.037	1.563	0.0
Target: Divisibility by 3, mapping DC						
Control	Baseline	None	98.0	0.000	0.000	0.0
Control	Identity	None	98.0	0.000	0.000	0.0
Control	Native axis	Native	64.0	1.464	13.567	0.0
Control	Random mean (8)	Matched	98.0	-0.048	13.567	100.0
Parity	Raw symmetric	Matched	90.0	1.306	13.567	98.0
Parity	Raw antisymmetric	Matched	98.0	0.637	13.567	99.0
Parity	Symmetric	Matched	90.0	1.303	13.567	98.0
Parity	Antisymmetric	Matched	98.0	0.298	13.567	99.0
Div. by 3	Raw symmetric	Matched	80.0	1.412	13.567	96.0
Div. by 3	Raw antisymmetric	Matched	90.0	1.336	13.567	100.0
Div. by 3	Symmetric	Matched	82.0	1.398	13.567	97.0
Div. by 3	Antisymmetric	Matched	98.0	0.488	13.567	100.0
Div. by 3	Symmetric	Natural	84.0	1.376	13.374	0.0
Div. by 3	Antisymmetric	Natural	98.0	0.081	2.289	0.0

Table 37: Qwen2.5, CD/DC, split 1, layer 28: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping CD						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	49.0	3.157	51.192	0.0
Control	Random mean (8)	Matched	100.0	0.059	51.192	100.0
Parity	Raw symmetric	Matched	99.0	1.367	51.192	100.0
Parity	Raw antisymmetric	Matched	0.0	4.415	51.192	99.0
Parity	Symmetric	Matched	100.0	0.988	51.192	100.0
Parity	Antisymmetric	Matched	0.0	4.300	51.192	100.0
Div. by 3	Raw symmetric	Matched	100.0	0.802	51.192	100.0
Div. by 3	Raw antisymmetric	Matched	0.0	4.291	51.192	99.0
Div. by 3	Symmetric	Matched	100.0	0.741	51.192	100.0
Div. by 3	Antisymmetric	Matched	0.0	4.272	51.192	99.0
Parity	Symmetric	Natural	100.0	0.817	42.462	0.0
Parity	Antisymmetric	Natural	80.0	2.402	28.618	0.0
Target: Parity, mapping DC						
Control	Baseline	None	50.0	0.000	0.000	0.0
Control	Identity	None	50.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	1.101	37.616	0.0
Control	Random mean (8)	Matched	51.6	0.008	37.616	100.0
Parity	Raw symmetric	Matched	90.0	-1.041	37.616	100.0
Parity	Raw antisymmetric	Matched	12.0	3.301	37.616	100.0
Parity	Symmetric	Matched	81.0	-0.758	37.616	100.0
Parity	Antisymmetric	Matched	12.0	3.217	37.616	100.0
Div. by 3	Raw symmetric	Matched	71.0	-0.598	37.616	100.0
Div. by 3	Raw antisymmetric	Matched	12.0	3.203	37.616	100.0
Div. by 3	Symmetric	Matched	67.0	-0.554	37.616	100.0
Div. by 3	Antisymmetric	Matched	12.0	3.189	37.616	100.0
Parity	Symmetric	Natural	72.0	-0.642	31.864	0.0
Parity	Antisymmetric	Natural	50.0	1.710	19.995	0.0
Target: Divisibility by 3, mapping CD						
Control	Baseline	None	100.0	0.000	0.000	0.0
Control	Identity	None	100.0	0.000	0.000	0.0
Control	Native axis	Native	44.0	2.630	44.106	0.0
Control	Random mean (8)	Matched	100.0	0.039	44.106	100.0
Parity	Raw symmetric	Matched	100.0	1.154	44.106	100.0
Parity	Raw antisymmetric	Matched	2.0	3.728	44.106	97.0
Parity	Symmetric	Matched	100.0	0.834	44.106	100.0
Parity	Antisymmetric	Matched	2.0	3.630	44.106	97.0
Div. by 3	Raw symmetric	Matched	100.0	0.674	44.106	99.0
Div. by 3	Raw antisymmetric	Matched	2.0	3.621	44.106	97.0
Div. by 3	Symmetric	Matched	100.0	0.623	44.106	99.0
Div. by 3	Antisymmetric	Matched	2.0	3.605	44.106	97.0
Div. by 3	Symmetric	Natural	100.0	0.501	35.551	0.0
Div. by 3	Antisymmetric	Natural	96.0	2.135	26.141	0.0
Target: Divisibility by 3, mapping DC						
Control	Baseline	None	98.0	0.000	0.000	0.0
Control	Identity	None	98.0	0.000	0.000	0.0
Control	Native axis	Native	29.0	1.609	42.698	0.0
Control	Random mean (8)	Matched	98.0	0.018	42.698	100.0
Parity	Raw symmetric	Matched	99.0	-1.167	42.698	100.0
Parity	Raw antisymmetric	Matched	1.0	3.712	42.698	100.0
Parity	Symmetric	Matched	99.0	-0.848	42.698	100.0
Parity	Antisymmetric	Matched	1.0	3.619	42.698	100.0
Div. by 3	Raw symmetric	Matched	99.0	-0.681	42.698	100.0
Div. by 3	Raw antisymmetric	Matched	1.0	3.609	42.698	100.0
Div. by 3	Symmetric	Matched	99.0	-0.631	42.698	100.0
Div. by 3	Antisymmetric	Matched	1.0	3.594	42.698	100.0
Div. by 3	Symmetric	Natural	98.0	-0.510	34.428	0.0
Div. by 3	Antisymmetric	Natural	1.0	2.128	25.255	0.0

Table 38: Gemma 2, AB/BA, split 1, layer 21: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping AB						
Control	Baseline	None	86.0	0.000	0.000	0.0
Control	Identity	None	86.0	0.000	0.000	0.0
Control	Native axis	Native	86.0	0.018	2.242	0.0
Control	Random mean (8)	Matched	86.0	-0.001	2.242	99.0
Parity	Raw symmetric	Matched	86.0	0.021	2.242	78.0
Parity	Raw antisymmetric	Matched	86.0	-0.012	2.242	93.0
Parity	Symmetric	Matched	86.0	0.021	2.242	78.0
Parity	Antisymmetric	Matched	86.0	-0.010	2.242	93.0
Div. by 3	Raw symmetric	Matched	86.0	0.018	2.242	59.0
Div. by 3	Raw antisymmetric	Matched	86.0	0.004	2.242	97.0
Div. by 3	Symmetric	Matched	86.0	0.018	2.242	59.0
Div. by 3	Antisymmetric	Matched	86.0	0.004	2.242	97.0
Parity	Symmetric	Natural	86.0	0.020	2.161	0.0
Parity	Antisymmetric	Natural	86.0	-0.003	0.754	0.0
Target: Parity, mapping BA						
Control	Baseline	None	50.0	0.000	0.000	0.0
Control	Identity	None	50.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	0.012	2.265	0.0
Control	Random mean (8)	Matched	50.0	0.003	2.265	98.1
Parity	Raw symmetric	Matched	50.0	0.013	2.265	53.0
Parity	Raw antisymmetric	Matched	50.0	0.000	2.265	92.0
Parity	Symmetric	Matched	50.0	0.013	2.265	53.0
Parity	Antisymmetric	Matched	50.0	0.000	2.265	92.0
Div. by 3	Raw symmetric	Matched	50.0	0.007	2.265	53.0
Div. by 3	Raw antisymmetric	Matched	50.0	0.000	2.265	97.0
Div. by 3	Symmetric	Matched	50.0	0.007	2.265	53.0
Div. by 3	Antisymmetric	Matched	50.0	0.000	2.265	96.0
Parity	Symmetric	Natural	50.0	0.013	2.245	0.0
Parity	Antisymmetric	Natural	50.0	0.000	0.560	0.0
Target: Divisibility by 3, mapping AB						
Control	Baseline	None	63.0	0.000	0.000	0.0
Control	Identity	None	63.0	0.000	0.000	0.0
Control	Native axis	Native	61.0	0.039	2.958	0.0
Control	Random mean (8)	Matched	64.1	-0.005	2.958	98.8
Parity	Raw symmetric	Matched	61.0	0.049	2.958	75.0
Parity	Raw antisymmetric	Matched	65.0	-0.019	2.958	96.0
Parity	Symmetric	Matched	61.0	0.049	2.958	75.0
Parity	Antisymmetric	Matched	65.0	-0.020	2.958	98.0
Div. by 3	Raw symmetric	Matched	61.0	0.038	2.958	83.0
Div. by 3	Raw antisymmetric	Matched	63.0	0.007	2.958	95.0
Div. by 3	Symmetric	Matched	61.0	0.038	2.958	84.0
Div. by 3	Antisymmetric	Matched	63.0	0.006	2.958	95.0
Div. by 3	Symmetric	Natural	62.0	0.037	2.891	0.0
Div. by 3	Antisymmetric	Natural	63.0	0.001	0.660	0.0
Target: Divisibility by 3, mapping BA						
Control	Baseline	None	32.0	0.000	0.000	0.0
Control	Identity	None	32.0	0.000	0.000	0.0
Control	Native axis	Native	32.0	0.011	2.869	0.0
Control	Random mean (8)	Matched	32.0	0.000	2.869	97.9
Parity	Raw symmetric	Matched	32.0	0.030	2.869	71.0
Parity	Raw antisymmetric	Matched	32.0	0.001	2.869	88.0
Parity	Symmetric	Matched	32.0	0.030	2.869	71.0
Parity	Antisymmetric	Matched	32.0	-0.001	2.869	92.0
Div. by 3	Raw symmetric	Matched	32.0	0.012	2.869	78.0
Div. by 3	Raw antisymmetric	Matched	32.0	-0.008	2.869	93.0
Div. by 3	Symmetric	Matched	32.0	0.012	2.869	78.0
Div. by 3	Antisymmetric	Matched	32.0	-0.007	2.869	94.0
Div. by 3	Symmetric	Natural	32.0	0.012	2.823	0.0
Div. by 3	Antisymmetric	Natural	32.0	-0.001	0.642	0.0

Table 39: Gemma 2, AB/BA, split 1, layer 33: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping AB						
Control	Baseline	None	86.0	0.000	0.000	0.0
Control	Identity	None	86.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	1.485	46.784	0.0
Control	Random mean (8)	Matched	85.6	0.040	46.784	100.0
Parity	Raw symmetric	Matched	50.0	1.039	46.784	93.0
Parity	Raw antisymmetric	Matched	50.0	1.559	46.784	93.0
Parity	Symmetric	Matched	73.0	0.666	46.784	94.0
Parity	Antisymmetric	Matched	50.0	1.419	46.784	96.0
Div. by 3	Raw symmetric	Matched	52.0	0.930	46.784	94.0
Div. by 3	Raw antisymmetric	Matched	50.0	1.539	46.784	94.0
Div. by 3	Symmetric	Matched	77.0	0.626	46.784	94.0
Div. by 3	Antisymmetric	Matched	50.0	1.422	46.784	96.0
Parity	Symmetric	Natural	78.0	0.489	34.150	0.0
Parity	Antisymmetric	Natural	50.0	0.932	32.054	0.0
Target: Parity, mapping BA						
Control	Baseline	None	50.0	0.000	0.000	0.0
Control	Identity	None	50.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	0.860	25.886	0.0
Control	Random mean (8)	Matched	50.0	0.020	25.886	100.0
Parity	Raw symmetric	Matched	50.0	-0.045	25.886	99.0
Parity	Raw antisymmetric	Matched	50.0	0.876	25.886	97.0
Parity	Symmetric	Matched	50.0	0.232	25.886	99.0
Parity	Antisymmetric	Matched	50.0	0.978	25.886	98.0
Div. by 3	Raw symmetric	Matched	50.0	0.047	25.886	98.0
Div. by 3	Raw antisymmetric	Matched	50.0	0.868	25.886	97.0
Div. by 3	Symmetric	Matched	50.0	0.259	25.886	96.0
Div. by 3	Antisymmetric	Matched	50.0	0.952	25.886	97.0
Parity	Symmetric	Natural	50.0	0.163	18.288	0.0
Parity	Antisymmetric	Natural	50.0	0.685	18.338	0.0
Target: Divisibility by 3, mapping AB						
Control	Baseline	None	63.0	0.000	0.000	0.0
Control	Identity	None	63.0	0.000	0.000	0.0
Control	Native axis	Native	32.0	1.052	33.154	0.0
Control	Random mean (8)	Matched	62.5	0.026	33.154	100.0
Parity	Raw symmetric	Matched	32.0	0.731	33.154	96.0
Parity	Raw antisymmetric	Matched	32.0	1.096	33.154	92.0
Parity	Symmetric	Matched	40.0	0.460	33.154	96.0
Parity	Antisymmetric	Matched	32.0	0.998	33.154	96.0
Div. by 3	Raw symmetric	Matched	33.0	0.671	33.154	93.0
Div. by 3	Raw antisymmetric	Matched	32.0	1.099	33.154	91.0
Div. by 3	Symmetric	Matched	43.0	0.449	33.154	94.0
Div. by 3	Antisymmetric	Matched	32.0	1.014	33.154	95.0
Div. by 3	Symmetric	Natural	48.0	0.318	23.189	0.0
Div. by 3	Antisymmetric	Natural	32.0	0.714	23.859	0.0
Target: Divisibility by 3, mapping BA						
Control	Baseline	None	32.0	0.000	0.000	0.0
Control	Identity	None	32.0	0.000	0.000	0.0
Control	Native axis	Native	32.0	0.749	22.525	0.0
Control	Random mean (8)	Matched	32.0	0.020	22.525	100.0
Parity	Raw symmetric	Matched	32.0	-0.061	22.525	99.0
Parity	Raw antisymmetric	Matched	32.0	0.767	22.525	99.0
Parity	Symmetric	Matched	32.0	0.184	22.525	100.0
Parity	Antisymmetric	Matched	32.0	0.854	22.525	98.0
Div. by 3	Raw symmetric	Matched	32.0	0.047	22.525	99.0
Div. by 3	Raw antisymmetric	Matched	32.0	0.741	22.525	98.0
Div. by 3	Symmetric	Matched	32.0	0.233	22.525	97.0
Div. by 3	Antisymmetric	Matched	32.0	0.818	22.525	98.0
Div. by 3	Symmetric	Natural	32.0	0.164	15.770	0.0
Div. by 3	Antisymmetric	Natural	32.0	0.577	16.117	0.0

Table 40: Gemma 2, AB/BA, split 1, layer 42: all intervention conditions.

Source	Direction	Operation	Acc. (%)	Harm	$\|e\|$	Crossing (%)
Target: Parity, mapping AB						
Control	Baseline	None	86.0	0.000	0.000	0.0
Control	Identity	None	86.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	1.553	63.165	0.0
Control	Random mean (8)	Matched	85.2	0.061	63.165	100.0
Parity	Raw symmetric	Matched	69.0	0.538	63.165	92.0
Parity	Raw antisymmetric	Matched	50.0	2.147	63.165	94.0
Parity	Symmetric	Matched	86.0	0.079	63.165	94.0
Parity	Antisymmetric	Matched	50.0	2.200	63.165	96.0
Div. by 3	Raw symmetric	Matched	77.0	0.386	63.165	94.0
Div. by 3	Raw antisymmetric	Matched	50.0	2.184	63.165	95.0
Div. by 3	Symmetric	Matched	87.0	0.000	63.165	94.0
Div. by 3	Antisymmetric	Matched	50.0	2.233	63.165	96.0
Parity	Symmetric	Natural	86.0	0.061	48.101	0.0
Parity	Antisymmetric	Natural	50.0	1.450	41.127	0.0
Target: Parity, mapping BA						
Control	Baseline	None	50.0	0.000	0.000	0.0
Control	Identity	None	50.0	0.000	0.000	0.0
Control	Native axis	Native	50.0	1.019	39.759	0.0
Control	Random mean (8)	Matched	50.0	0.019	39.759	99.6
Parity	Raw symmetric	Matched	50.0	-0.296	39.759	98.0
Parity	Raw antisymmetric	Matched	50.0	1.431	39.759	99.0
Parity	Symmetric	Matched	50.0	0.016	39.759	97.0
Parity	Antisymmetric	Matched	50.0	1.477	39.759	99.0
Div. by 3	Raw symmetric	Matched	50.0	-0.164	39.759	98.0
Div. by 3	Raw antisymmetric	Matched	50.0	1.467	39.759	99.0
Div. by 3	Symmetric	Matched	50.0	0.100	39.759	98.0
Div. by 3	Antisymmetric	Matched	50.0	1.513	39.759	99.0
Parity	Symmetric	Natural	50.0	0.012	29.465	0.0
Parity	Antisymmetric	Natural	50.0	0.997	26.739	0.0
Target: Divisibility by 3, mapping AB						
Control	Baseline	None	63.0	0.000	0.000	0.0
Control	Identity	None	63.0	0.000	0.000	0.0
Control	Native axis	Native	32.0	1.081	43.896	0.0
Control	Random mean (8)	Matched	62.4	0.035	43.896	99.0
Parity	Raw symmetric	Matched	39.0	0.380	43.896	96.0
Parity	Raw antisymmetric	Matched	31.0	1.540	43.896	92.0
Parity	Symmetric	Matched	63.0	0.050	43.896	97.0
Parity	Antisymmetric	Matched	29.0	1.570	43.896	94.0
Div. by 3	Raw symmetric	Matched	47.0	0.268	43.896	94.0
Div. by 3	Raw antisymmetric	Matched	29.0	1.566	43.896	93.0
Div. by 3	Symmetric	Matched	65.0	-0.011	43.896	96.0
Div. by 3	Antisymmetric	Matched	28.0	1.598	43.896	96.0
Div. by 3	Symmetric	Natural	64.0	-0.007	31.793	0.0
Div. by 3	Antisymmetric	Natural	32.0	1.104	30.488	0.0
Target: Divisibility by 3, mapping BA						
Control	Baseline	None	32.0	0.000	0.000	0.0
Control	Identity	None	32.0	0.000	0.000	0.0
Control	Native axis	Native	32.0	0.891	32.014	0.0
Control	Random mean (8)	Matched	32.0	0.022	32.014	98.5
Parity	Raw symmetric	Matched	32.0	-0.244	32.014	97.0
Parity	Raw antisymmetric	Matched	32.0	1.161	32.014	97.0
Parity	Symmetric	Matched	32.0	0.008	32.014	97.0
Parity	Antisymmetric	Matched	32.0	1.198	32.014	94.0
Div. by 3	Raw symmetric	Matched	32.0	-0.140	32.014	96.0
Div. by 3	Raw antisymmetric	Matched	32.0	1.187	32.014	97.0
Div. by 3	Symmetric	Matched	32.0	0.075	32.014	93.0
Div. by 3	Antisymmetric	Matched	32.0	1.227	32.014	95.0
Div. by 3	Symmetric	Natural	32.0	0.055	23.026	0.0
Div. by 3	Antisymmetric	Natural	32.0	0.849	22.350	0.0

H FROZEN CLASSIFICATION ACROSS ANSWER VOCABULARIES

This exploratory analysis asks whether a change in answer vocabulary preserves information accessible to a fixed classifier. We use the first Qwen calibration split and the symmetric update direction from AB/BA, with the original common-correct fitting rule. Its sign, weights, and threshold receive no CD/DC labels. The same 100 calibration numbers and 100 disjoint test numbers are used under both vocabularies. The parity test split has 50 positives and 50 negatives; divisibility has 32 positives and 68 negatives. We report this single split descriptively.

The source threshold is the midpoint of the two source class means. For the frozen symmetric direction s, let $f_i^m = s^\top h_i^m$ and

$$c_y = \tfrac{1}{2} \sum_{m \in \{\mathrm{AB},\mathrm{BA}\}} \mathbb{E}_{\mathrm{cal}}[f_i^m \mid y_i = y], \qquad \theta = \tfrac{1}{2}(c_0 + c_1). \tag{30}$$

Both classes and both source mappings receive equal weight in this threshold. All source calibration numbers enter its fit. The prediction is $\mathbf{1}[\omega(f_i^m - \theta) > 0]$, with $\omega = \mathrm{sgn}_+(c_1 - c_0)$ determined on the source calibration set. We apply this rule unchanged to each target mapping's task states.

An additional protocol tests sensitivity to mapping-dependent offsets. For each source and target mapping separately, we subtract its mean calibration score before computing predictions. The threshold is then fitted on the centered source scores, using the same source labels and class weighting. Target centering uses calibration features without labels and never uses test features. This protocol therefore permits unlabeled target calibration; the unchanged source-threshold protocol does not require it. Table 41 reports both accuracies and the integer-baseline area under the receiver operating characteristic curve (AUC).

Table 41: Frozen source classification on 100 held-out numbers per task. Native accuracy uses the model's candidate-pair decision. Source accuracy uses the AB/BA direction and threshold unchanged; centered accuracy additionally uses each target mapping's unlabeled calibration mean. Accuracies are percentages. Baseline AUC uses the original sign of the fixed symmetric direction on integer-baseline states, with no baseline-specific reorientation.

Layer	Task	Mapping	Native	Source	Centered	Baseline AUC
22	Parity	CD	100	100	100	0.514
22	Parity	DC	50	100	100	0.500
22	Divisibility	CD	100	100	100	0.501
22	Divisibility	DC	98	99	100	0.540
28	Parity	CD	100	100	100	0.755
28	Parity	DC	50	96	100	0.595
28	Divisibility	CD	100	100	100	0.541
28	Divisibility	DC	98	100	100	0.536

The integer-baseline control tests whether the same direction separates the numerical labels under the integer-baseline prompt. Its score uses the original source symmetric direction applied to baseline states, without reorienting it on baseline labels. We compute AUC by pairwise positive-negative comparisons with half credit for ties and independently check it through average ranks. At layer 22, baseline AUC ranges from 0.500 to 0.540, while task-state and task-update AUC are 1 for both tasks and target mappings. At layer 28, parity baseline AUC rises to 0.755 under CD, so near-chance baseline separation is specific to the intermediate comparison. An AUC of 1 establishes complete class ordering on this test set; fixed-threshold accuracy remains a separate measurement.

The result distinguishes recoverable labels from the native answer. Reversed CD parity has 50% native candidate-pair accuracy, yet its layer-22 states support 100% frozen-source classification accuracy. The native answer and the classifier use different readouts of the same state. These measurements establish information recovery under the changed vocabulary; testing whether a transferred direction controls the native answer requires a new intervention.

I REPRODUCIBILITY AND RESULT ORGANIZATION

The accompanying program runs in one Google Colab cell. It downloads the public checkpoints and SST-2 data, constructs the numerical tasks, captures representations, fits calibration directions, applies interventions, and computes the reported figures and tables. Execution and storage settings appear at the top. The program reuses an existing persistent-drive mount, resumes completed experimental units, adapts batch size to available accelerator memory, and reports global progress. Resumption uses the declared experimental settings, saved completion status, and required outputs.

The exported results preserve the information needed to audit each scientific comparison. Input records contain numbers or sampled dataset indices, exact prompts, labels, and token IDs. Calibration records contain split indices, class counts, direction vectors, centers, and fallback decisions. Intervention records contain original and edited logits, edit norms, projection coefficients, and final-readout scalar checks where applicable. Summaries distinguish all-test, initially correct, and initially incorrect inputs. Numeric arrays preserve individual random directions even when a printed table averages them.

Paper figures and tables are generated from the per-example measurements within the same program. These computations include semantic predictions, margin decreases, paired intervals, mapping capability, readout-span shares, and frozen source classification. The exported tables and machine-readable summaries retain the individual random controls and outcomes omitted from the compact printed tables. This organization allows the main contrasts and the wider experiment families to be checked from one implementation.