AMPLIFICATION WITHOUT REPAIR: CAUSAL LIMITS OF ACTIVATION STEERING IN ABSTRACT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering modifies hidden representations to control language-model behavior. Current approaches select interpretable features or learn directions from demonstrations, seeking improvements in reasoning. How these interventions translate into correction on new examples remains unclear. We investigate this transfer through a mechanistic study of text-serialized ARC-AGI-1, ARC-AGI-2, and symbolic Bongard-LOGO. Feature audits combine input controls, activation localization, and random-controlled interventions; support-adaptation experiments examine gradient alignment, decision boundaries, and support aggregation. Steering improves output validity and produces concept-specific held-out margin gains, while decision repair remains sparse. Fit/check gradient alignment differs between initially correct and incorrect concepts across development and fresh-combination confirmation. Aggregating support gradients amplifies both groups' responses, strengthening correct and incorrect preferences. These findings connect correction to directional transfer and boundary crossing, motivating interventions that target held-out errors and evaluations that track repairs and harms alongside margin gains.

1 INTRODUCTION

Activation steering modifies a language model's behavior by intervening on its hidden representations. Existing methods control behavioral attributes, elicit truthful answers, and transfer task information across inputs (Turner et al., 2023; Li et al., 2023; Hendel et al., 2023; Todd et al., 2024). Sparse autoencoders extend this approach by exposing interpretable features for targeted intervention (Huben et al., 2024; Bricken et al., 2023; Templeton et al., 2024). Recent work reports reasoning improvements through strategy-specific and conditional steering (Fang et al., 2026; Valentino et al., 2026). These raise a question about transfer: *when does a direction extracted from demonstrations correct a model's mistakes on new examples?*

Answering this question requires connecting internal measurements to task outcomes. Research on probing, causal intervention, and SAE evaluation motivates testing feature interpretations through controlled behavioral effects (Belinkov, 2022; Geiger et al., 2021; Zhang & Nanda, 2024; Karvonen et al., 2025). Abstract reasoning provides a demanding setting for such tests: models must infer a rule from demonstrations and apply it to an unfamiliar instance. We study text-serialized ARC-AGI-1 and ARC-AGI-2 grids alongside symbolic descriptions from Bongard-LOGO (Chollet, 2019; Chollet et al., 2025; Nie et al., 2020), tracking how interventions affect output validity, answer margins, and initially incorrect decisions.

Our ARC experiments reveal a gap between behavioral control and rule solving. Selected SAE features respond strongly to serialized grids, and some interventions improve valid-grid generation, while exact-solution gains remain limited in the tested configurations. We investigate this gap through task and format controls, activation localization, and norm-matched random interventions. These observations motivate a closer examination of how intervention directions transfer from the examples used to construct them to the examples requiring correction.

Bongard-LOGO enables this analysis through a support/check protocol. We learn activation directions from labeled support examples and evaluate their effects on a withheld support pair. Ran-

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

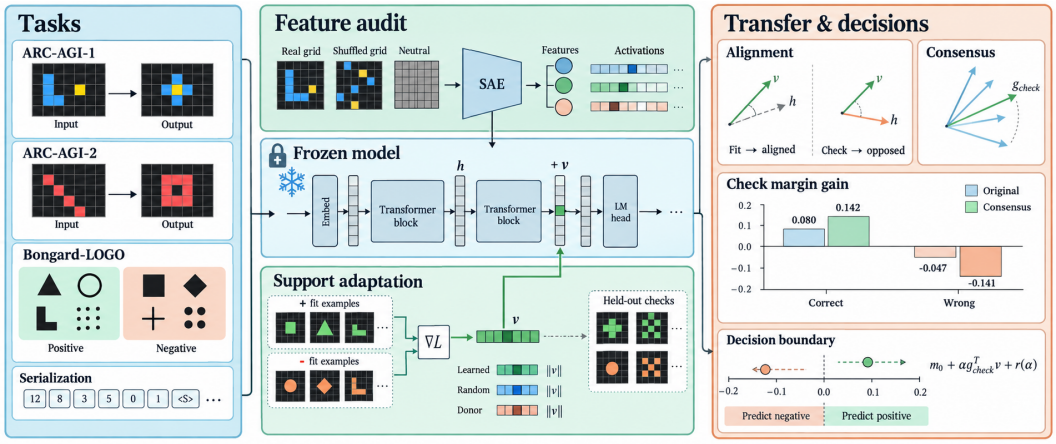


Figure 1: Overview of the steering audit and support-transfer analysis. (a) ARC grids and Bongard-LOGO drawing programs supply symbolic reasoning inputs. (b) The feature audit examines SAE candidates using task and format controls. (c) Gradient alignment, support consensus, and the ranking boundary connect intervention responses to decision changes.

dom and donor-concept controls assess direction specificity; fit/check gradient alignment and dose-response curves characterize transfer. Following development experiments, we freeze the protocol and evaluate 115 fresh combinations of shared primitives. The learned intervention increases mean check margins for baseline-correct concepts and decreases them for baseline-wrong concepts, accompanied by differences in fit/check gradient alignment.

We then test whether aggregating support-derived directions improves correction. A support-consensus intervention averages five normalized leave-one-support-pair-out gradients at a matched intervention norm. At the fixed unit dose, its overall mean margin gain is $+0.078$, comprising $+0.142$ on 89 baseline-correct concepts and -0.141 on 26 baseline-wrong concepts; all check rankings remain unchanged. The development-set coherence difference fails to replicate in confirmation. These results locate the central difficulty in transfer: agreement across support-derived directions can coexist with adverse effects on held-out errors.

These findings characterize a pattern of *amplification without reliable repair*: aggregate margin gains concentrate on concepts the model already ranks correctly, while support-derived updates can reinforce incorrect decisions. Understanding this pattern requires examining how intervention directions transfer to held-out errors.

Our contributions are threefold:

- **A causal audit of steering in abstract reasoning.** Across text-serialized ARC-AGI-1 and ARC-AGI-2, we examine feature activation, output validity, and exact solutions using format controls, activation localization, and norm-matched random interventions.
- **A diagnostic framework for directional transfer.** On symbolic Bongard-LOGO, we combine support/check evaluation, gradient alignment, and dose-response analysis to investigate how support-trained interventions affect initially correct and incorrect decisions.
- **Fresh-combination confirmation and a consensus intervention test.** Across 115 fresh concept combinations, we document correctness-dependent intervention responses and evaluate support-gradient consensus, observing positive aggregate margin gains alongside adverse effects on baseline errors and a failed replication of the coherence hypothesis.

2 Related Work

Activation Steering. Activation addition and contrastive steering extract directions from differences between hidden states, enabling inference-time control over model behavior (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023). Task vectors and function vectors extend this approach to demonstrations, identifying compact representations that transfer task behavior across

inputs (Hendel et al., 2023; Todd et al., 2024). Sparse autoencoders provide a feature-level basis for intervention (Huben et al., 2024; Bricken et al., 2023; Templeton et al., 2024). Building on these representations, SAE-Steering targets reasoning strategies, while conditional steering adjusts interventions to improve logical judgments (Fang et al., 2026; Valentino et al., 2026). We examine how intervention directions transfer to held-out examples, tracing their effects on initially correct and incorrect decisions.

Causal Evaluation of Representations. Probing methods and control tasks assess the information accessible in hidden representations (Hewitt & Liang, 2019; Belinkov, 2022). Causal abstraction and activation patching connect representations to model computations through controlled interventions (Geiger et al., 2021; Zhang & Nanda, 2024). RAVEL evaluates selective control over represented attributes, and sparse feature circuits connect interpretable features into causal graphs that support behavioral editing (Huang et al., 2024; Marks et al., 2025). SAEBench broadens evaluation to feature disentanglement and practical applications (Karvonen et al., 2025). Our analysis connects intervention effects to reasoning outcomes through norm-matched random controls, donor-concept directions, support/check gradient alignment, and dose-response measurements.

Generalization from Demonstrations. ARC and ARC-AGI-2 evaluate the ability to infer transformations from examples and apply them to new grids (Chollet, 2019; Chollet et al., 2025). ConceptARC organizes this evaluation around concept groups, while Bongard-LOGO examines concept learning from positive and negative examples (Moskvichev et al., 2023; Nie et al., 2020). Test-time training uses task demonstrations to update model parameters and improve performance on new instances (Akyürek et al., 2024). These settings place demonstration-to-test transfer at the center of abstract reasoning. We study this transfer through activation updates learned from labeled supports, measuring held-out margin changes and decision repair. A support-consensus intervention further tests how aggregating demonstration-derived directions affects correction.

3 Exploratory Study

Activation steering offers a direct way to test whether a behaviorally meaningful direction can improve reasoning. A feature associated with problem understanding provides an interpretable intervention target; a direction learned from labeled demonstrations provides task-specific supervision. Both routes suggest a natural expectation: an effective intervention should move the model toward solving the task. We examine this expectation through SAE-based steering on ARC-AGI-1 and support-derived adaptation on Bongard-LOGO, tracing changes from output structure to answer preferences and decision outcomes.

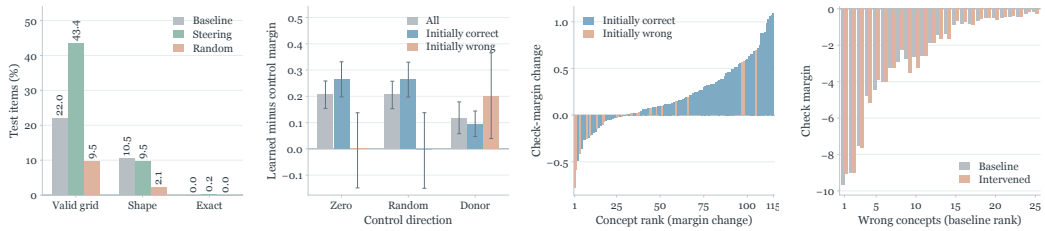
A reasoning-associated feature primarily improves output structure. We first intervene on an R1 feature associated with problem understanding through language validation. Steering this feature increases valid-grid generation relative to both direct decoding and a norm-matched random direction (Figure 2a). The feature thus produces a measurable, direction-specific behavioral effect. Yet the improvement separates sharply across evaluation levels: shape matching remains close to the direct baseline, and exact solutions remain rare. The largest gain occurs in producing a parseable answer, while recovering the target transformation changes much less. A candidate selected through reasoning-related language behavior therefore presents a more specific empirical question: which component of task performance does its intervention actually control?

Task supervision produces concept-specific transfer. We next construct directions directly from Bongard-LOGO support labels. This tests whether supervision tied to the target concept connects activation changes more closely to the required decision. Each concept supplies a fit pair for learning and a disjoint check pair for evaluating transfer. Let $s(x; v) = \log p(+ \mid x; v) - \log p(- \mid x; v)$ denote the label-score difference under intervention v. For a positive example x^+ and a negative example x^-, define

$$m(v) = s(x^+; v) - s(x^-; v). \tag{1}$$

A positive margin indicates that the positive example receives the higher positive-label score.

Learned directions improve mean check margins relative to zero intervention, norm-matched random directions, and donor-concept directions (Figure 2b). These controls give the gain a concrete



(a) ARC output outcomes (b) Held-out margin gains (c) Concept-level changes (d) Wrong concepts

Figure 2: Steering improves aggregate outcomes while errors persist. (a) Baseline decoding, feature-26683 steering at strength 15, and norm-matched random steering on 419 ARC-AGI-1 test items with DeepSeek-R1-Distill-Llama-8B. (b) Learned-direction check-margin advantages over zero, mean random, and mean donor controls on 115 Bongard-LOGO development concepts, shown overall and by baseline ranking correctness (90 correct, 25 wrong). (c) Every concept's learned-minus-baseline margin change, sorted from lowest to highest. Blue denotes initially correct concepts and peach denotes initially wrong concepts, matching panel (b). (d) Baseline and intervened margins for all 25 initially wrong concepts, sorted by baseline margin. All 115 rankings remain unchanged at the evaluated intervention.

interpretation. The random comparison tests the effect of the learned orientation at a matched norm; the donor comparison tests the advantage of using the target concept's supports. The resulting advantage extends beyond the examples used for optimization and retains information specific to the concept that generated the update.

Transferable gains leave the decision boundary uncrossed. This concept-specific transfer creates a stronger expectation of correction: an update learned from the right concept also improves its held-out margin on average. The observed decisions expose a further gap. At the evaluated intervention, every concept retains its baseline ranking. Margin changes vary across concepts, with both positive and negative responses (Figure 2c), while all initially wrong concepts remain below the ranking boundary (Figure 2d). The intervention transfers a continuous preference change, but the resulting trajectories preserve the original decisions.

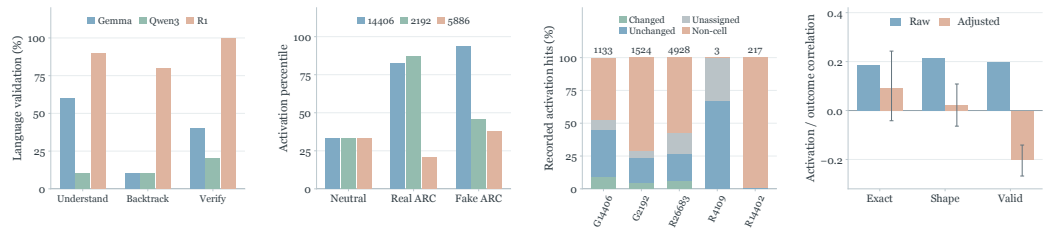
The distinction also appears within the learning process. In a separate query-content adaptation experiment, learned directions repair errors on the optimized pairs much more frequently than on the held-out pairs. Section 5 examines these fit/check transitions. Together, the experiments separate three observable achievements: fitting a labeled example, transferring a margin gain, and correcting a held-out decision. Progress at the first two stages can coexist with persistent errors at the third.

Based on this, we pursue three questions. What properties of reasoning inputs engage the selected features, and which output behaviors change under intervention? How does an update that improves a support example align with improvement on a held-out example? How do intervention strength and support aggregation alter the resulting decision trajectories? Section 4 addresses the first question through input controls, localization, and causal comparisons. Section 5 connects the remaining questions through gradient alignment, dose responses, and support consensus.

4 What Do Steering Features Control?

4.1 Models, Tasks, and Controls

We audit SAE-based steering in Gemma-2-9B-it, Qwen3-8B-Base, and DeepSeek-R1-Distill-Llama-8B. ARC-AGI-1 and ARC-AGI-2 grids are serialized as numerical arrays, and Bongard-LOGO examples are represented by drawing-action tokens. The audit follows candidates from language validation through task-input activation to behavioral intervention. This sequence connects the interpretation used to select a feature with the behavior that changing it produces. Appendix A specifies the model-by-experiment coverage, checkpoints, SAE configurations, and task baselines. Language prompts identify candidates associated with problem understanding, backtracking, and verification. We measure their responses to neutral text, real ARC inputs, and matched synthetic grids, then localize activation hits within the serialized inputs. Behavioral evaluations compare se-



(a) Language validation (b) Input controls (c) Feature localization (d) Difficulty adjustment

Figure 3: **Language control, input sensitivity, localization, and success association.** (a) Language-validation success rates for the selected understanding, backtracking, and verification features in Gemma, Qwen3, and R1; each feature has ten validation samples. (b) Gemma candidates' active-rate percentiles relative to random SAE features on neutral text, real ARC-AGI-2 prompts, and matched synthetic grids. (c) Distribution of recorded activation hits for Gemma (G) and R1 (R) candidates. Numbers above bars give total recorded hits. Cell hits are partitioned into changed, unchanged, and unassigned transformation status; non-cell hits form the remaining category. (d) Gemma feature 14406 active-rate correlations with ARC-AGI-1 outcomes before and after adjustment for measured task complexity.

lected directions with same-layer, norm-matched random directions. ARC outcomes distinguish a parseable grid, a correctly shaped grid, and an exact solution. Bongard query classification measures the predicted label of each example; the support-transfer study in Section 5 measures the ranking of a positive–negative pair.

4.2 Language Validation and Task-Input Sensitivity

The selected R1 candidates score highly across the three language-validation conditions. Gemma's responses vary by strategy, while Qwen3's selected candidates score lower (Figure 3a). Additional Qwen3 SAE configurations exhibit similarly weak language-validation scores (Appendix B). The task audit follows these contrasting candidate profiles into the reasoning inputs and their associated output behavior. Task-input controls reveal further differences within a model. Gemma feature 14406 has high active-rate percentiles on both real and synthetic grids, with the larger percentile on synthetic inputs. Feature 2192 responds more strongly to real ARC inputs, while feature 5886 remains inactive (Figure 3b). The plotted percentile ranks a candidate's active rate against random SAE features. Its value depends on the reference distribution and ties; raw inactivity is recorded separately in Appendix B. The real/synthetic comparison shows that the grid response of feature 14406 extends to inputs whose structure resembles the task format. Feature 2192 provides a contrasting profile with greater sensitivity to the real-task inputs. Localization identifies where these responses occur (Figure 3c). Recorded hits span changed cells, unchanged cells, and surrounding serialization tokens, with feature-dependent distributions across these categories. The panels report the composition of observed hits; the input supplies different numbers of eligible tokens in each category. Read alongside the input controls, these distributions place feature activity across transformation content and its serialized context. They motivate testing the resulting output behavior directly.

4.3 Success Associations and Causal Outcomes

Task complexity provides a concrete explanation for part of the activation–success association. On ARC-AGI-1, exact-correct examples have smaller input grids on average, and shape-matching and valid-output examples also differ in grid size and color statistics. We residualize activation measurements and outcomes against the recorded task-complexity covariates, then correlate the residuals. For Gemma feature 14406, adjustment substantially reduces the positive exact- and shape-success associations and reverses the valid-output association (Figure 3d). The raw association combines feature activity with variation in the task regime. Appendix B gives the activation metrics and adjusted estimates. Random-controlled interventions address the corresponding behavioral question. Across both ARC datasets, Gemma's selected-feature effects remain close to their random-direction controls at the reported strength (Table 1). Feature 14406 combines strong grid activation and positive raw success associations with small feature-specific outcome changes.

These measurements locate different parts of its response: the activation assays describe where it is engaged, and the intervention comparisons measure what changing its direction contributes to the evaluated outputs. R1 supplies a useful contrast between two strongly language-validated candidates. Feature 26683 increases valid-grid generation, while shape matching remains close to direct decoding and exact solutions remain rare (Figure 2a). Feature 14402 produces generation degeneration at high strength. Thus, the same candidate-selection procedure yields a direction that improves structural compliance and another that disrupts generation. The distinction between validity, shape, and exact match identifies the output property receiving the benefit. Full strength curves and random controls appear in Appendix C.

4.4 Cross-Dataset Behavior and Scoring Controls

On Bongard-LOGO, Gemma feature 14406 remains strongly active, with higher mean activation on incorrectly classified queries than on correctly classified queries. Feature 2192 has a weaker relative response than on ARC. The relation between activity and successful decisions changes with the task, even when the candidate remains responsive to the input. Bongard-specific candidates selected from correctness contrasts produce small classification changes across strengths and intervention scopes (Table 1). Together, the cross-dataset comparisons connect feature engagement to the input representation and task outcome separately.

Scoring controls further distinguish changes in label preference from changes in discrimination. In the development calibration study, the train-calibrated one-pass classifier outperforms the two-pass procedure on validation, and the two-pass procedure also reduces pair-ranking accuracy (Appendix D). A shared additive shift in positive-label scores changes a classification threshold while cancelling in the pair margin of Equation 1. Tracking classification and pair ranking makes these responses separately observable. The results motivate measuring both continuous preference changes and discrete decision transitions throughout the support-transfer analysis.

Table 1: Gemma feature-specific effects across datasets. Values are feature-minus-norm-matched-random differences in percentage points. ARC uses strength 15 on 419 ARC-AGI-1 and 167 ARC-AGI-2 prompts. Bongard reports classification differences under full-sequence intervention on 200 queries. Full results appear in Appendix C.

Dataset	Feature	Valid	Shape	Exact
ARC-AGI-1	14406	+0.00	+0.24	+0.00
	2192	-0.24	+0.24	-0.48
	5886	-0.24	+0.00	+0.72
ARC-AGI-2	14406	+0.60	+0.60	+0.00
	2192	+1.80	-0.60	+0.00
	5886	+2.40	-1.80	+0.00
Dataset	Feature	$s = 3$	$s = 5$	$s = 10$
Bongard-LOGO	13985	-1.50	-1.50	-1.50
	10183	+0.00	+1.00	+1.00
	14218	+0.50	+0.00	+2.00

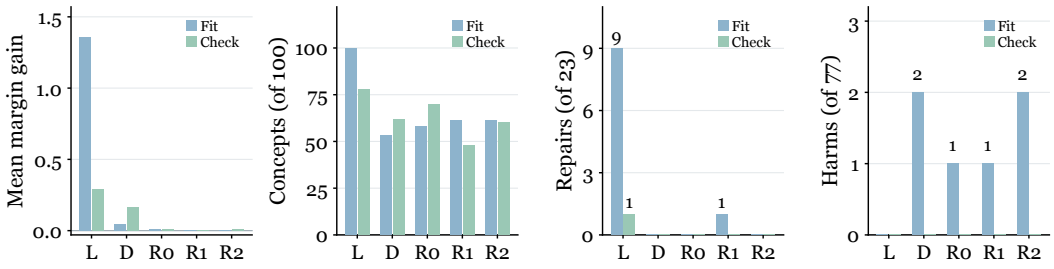
The feature audit identifies several routes through which steering changes behavior: input-sensitive feature activation, structural output control, and shifts in label preference. We next learn directions directly from support labels. This moves the analysis from the task relevance of a selected SAE candidate to the transfer of an update optimized for a specific concept.

5 When Does Support Adaptation Transfer?

5.1 From Support Fitting to Held-Out Correction

We study task-specific activation updates in frozen Gemma-2-9B-it. The query-content development experiment contains 100 concepts arranged in 50 donor blocks. Each concept provides a fit pair for learning and a check pair for measuring transfer. The learned directions increase fit margins and repair substantially more fit-pair errors than check-pair errors (Figure 4). Check margins also improve relative to donor-concept directions, with a positive block-bootstrap interval for the mean contrast (Appendix F). The experiment combines successful optimization, transferable concept information, and a pronounced gap between fitting and held-out correction.

The answer-boundary study evaluates 115 development concepts and 115 fresh combinations of shared primitives in confirmation, with 23 five-concept groups per cohort. The development results



(a) Mean margin gain (b) Positive margin gain (c) Wrong to correct (d) Correct to wrong

Figure 4: Support fitting transfers more strongly to margins than to correction. Gemma-2-9B-it on 100 development concepts under query-content intervention. Blue and green bars compare fit and check pairs. Conditions are learned directions (L), norm-matched donor-concept directions (D), and three norm-matched random controls (R0–R2). (a) Mean margin changes from zero intervention. (b) Counts of concepts with a positive margin change, out of 100. (c) Repairs among 23 initially wrong pairs in each split. (d) Harms among 77 initially correct pairs in each split. Panels (c)–(d) use paired transitions within each split; fit and check counts refer to their respective pairs.

establish learned-direction advantages over zero, random, and donor controls (Figure 2b). We then freeze the protocol for confirmation. The query-content and answer-boundary experiments retain their respective intervention sites, splits, and controls; Appendix F reports their results separately.

At the answer boundary, a task-specific vector is added after transformer block 20 (zero-indexed) at the final prompt token. Eight normalized gradient steps of size 0.25 minimize $\mathcal{L}(v) = \mathrm{softplus}[-m_{\mathrm{fit}}(v)]$, with projection onto an ℓ_2 ball of radius 2. Computation uses FP32 with TF32 disabled. Zero-intervention identity and finite-difference audits check the scoring and gradient paths. Fit labels supply the updates; the check pair supplies the held-out measurements. Subsequent gradient and dose diagnostics examine the resulting directions under the frozen evaluation protocol.

5.2 Gradient Alignment Predicts the Local Response

At zero intervention, define the fit and check margin gradients as

$$g_{\mathrm{fit}} = \nabla_v m_{\mathrm{fit}}(0), \qquad g_{\mathrm{check}} = \nabla_v m_{\mathrm{check}}(0). \tag{2}$$

The negative gradient of the softplus loss is a positive scalar multiple of the fit-margin gradient. For a nonzero fit gradient, a normalized step of length η has the first-order check response

$$\Delta m_{\mathrm{check}} \approx \eta\, g_{\mathrm{check}}^{\top} \frac{g_{\mathrm{fit}}}{\|g_{\mathrm{fit}}\|_2} = \eta \|g_{\mathrm{check}}\|_2 \cos(g_{\mathrm{fit}}, g_{\mathrm{check}}). \tag{3}$$

This expression separates local sensitivity from directional alignment. The check-gradient norm scales the response, while the cosine determines its sign. A step that increases the fit margin can increase or decrease the check margin according to the relationship between the two examples' gradients. The baseline-correct minus baseline-wrong alignment contrast is positive in development and confirmation, with both confidence intervals above zero (Table 2). In confirmation, the two groups have positive and negative mean alignment, respectively. This replicated contrast connects baseline correctness to the geometry of local transfer. Concept-level variation remains substantial: positive fit/check alignment occurs within both groups. The subgroup comparison describes a shift in the distribution of directions, while the finite-step intervention measures each learned trajectory.

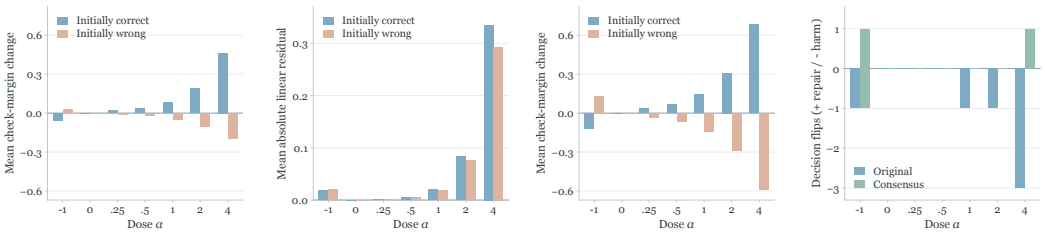
5.3 Direction, Strength, and the Decision Boundary

For the final learned vector v, we evaluate fixed doses $\alpha \in \{-1, 0, 0.25, 0.5, 1, 2, 4\}$. Unit dose is the reference condition. Let $m_0 = m_{\mathrm{check}}(0)$. The check margin decomposes into its baseline value, first-order response, and residual:

$$\begin{aligned} m_{\mathrm{check}}(\alpha v) &= m_0 + \alpha g_{\mathrm{check}}^{\top} v + r(\alpha), \\ r(\alpha) &= m_{\mathrm{check}}(\alpha v) - m_0 - \alpha g_{\mathrm{check}}^{\top} v. \end{aligned} \tag{4}$$

Table 2: Gradient alignment and support coherence across cohorts. Differences are baseline-correct minus baseline-wrong means. Correlations are Spearman coefficients across concepts; consensus gain is evaluated at $\alpha = 1$, and check alignment is $\cos(v_{\text{con}}, g_{\text{check}})$. Intervals are pointwise 95% block-bootstrap intervals from 5,000 resamples of 23 five-concept groups. Results use Gemma-2-9B-it. Each cohort contains 115 concepts. The $K = 5$ row uses 102 development concepts and all 115 confirmation concepts.

	Development		Confirmation	
Statistic	Estimate	95% CI	Estimate	95% CI
Fit/check alignment difference	0.338	[0.142, 0.543]	0.233	[0.010, 0.448]
Support-coherence difference	0.105	[0.019, 0.188]	−0.060	[−0.166, 0.042]
Support-coherence difference ($K = 5$)	0.086	[0.001, 0.176]	−0.060	[−0.163, 0.044]
Coherence vs. consensus gain (ρ)	0.268	[0.087, 0.434]	0.072	[−0.121, 0.271]
Coherence vs. check alignment (ρ)	0.333	[0.155, 0.497]	0.140	[−0.038, 0.319]



(a) Original direction (b) Linear residual (c) Support consensus (d) Decision transitions

Figure 5: Direction, strength, and aggregation shape held-out responses. All panels report Gemma-2-9B-it on the fresh-combination confirmation cohort (89 initially correct and 26 initially wrong concepts). Bars in (a)–(c) show subgroup means. (a) Check-margin changes under the original support-trained direction. (b) Absolute residual from its first-order check-margin prediction. (c) Check-margin changes under the norm-matched support-consensus direction. Panels (a) and (c) share the same vertical scale. (d) Incorrect-to-correct transitions appear above zero and correct-to-incorrect transitions below zero, for both methods.

For an initially wrong pair with $m_0 < 0$, a strictly positive ranking requires

$$\alpha g_{\text{check}}^{\top} v + r(\alpha) > -m_0. \tag{5}$$

The left-hand side measures movement toward the correct ranking; the right-hand side is the baseline margin deficit. This condition explains how a positive margin gain can preserve an incorrect decision. Correction depends on the response's direction and its size relative to that deficit.

In confirmation, increasing positive doses raises mean margins for initially correct concepts and lowers them for initially wrong concepts (Figure 5a). The original direction leaves every initially wrong ranking unchanged across this positive-dose sweep and introduces harms among initially correct concepts. In development, the largest dose produces both repairs and harms (Appendix F). These trajectories connect the group-level gain to individual boundary crossings: a larger average response can coexist with persistent errors and newly incorrect decisions.

Absolute residuals increase with dose (Figure 5b), recording the growing discrepancy between the finite intervention and its zero-intervention linear prediction. In Equation 5, the residual contributes alongside the directional projection to the crossing condition. The measured transition counts show the combined effect of these terms. Direction reversal supplies a further comparison by changing the sign of the linear contribution while evaluating the model's full response again.

5.4 Support Consensus Amplifies Both Subgroups

We next aggregate gradients across support pairs. Each paired leave-one-support-pair-out fold uses a positive and a negative support as targets, with the remaining non-check supports providing context. For nonzero fold gradients, define $u_k = g_k / \|g_k\|_2$, their mean $\bar{u} = \frac{1}{K} \sum_{k=1}^{K} u_k$, and the

consensus direction $v_{\mathrm{con}} = \|v\|_2\, \bar{u}/\|\bar{u}\|_2$. Development uses two to five folds according to support availability; confirmation uses five. Consensus draws supervision from multiple target pairs and matches the original learned vector's norm. This comparison holds intervention magnitude fixed while changing the construction of its direction. In confirmation, consensus produces a larger positive mean response for initially correct concepts and a larger negative mean response for initially wrong concepts (Figure 5c). All unit-dose rankings remain unchanged. At the largest positive dose, consensus repairs one concept; reversing the direction produces one repair and one harm (Figure 5d). Aggregating support gradients increases the magnitude of the subgroup responses, while correction follows the concept-specific boundary condition. To examine agreement among the support updates, we measure their mean pairwise cosine,

$$C = \frac{2}{K(K-1)} \sum_{i<j} u_i^\top u_j. \tag{6}$$

Because every u_k has unit norm, coherence satisfies

$$\|\bar{u}\|_2^2 = \frac{1 + (K-1)C}{K}. \tag{7}$$

Coherence thus measures how strongly the support gradients reinforce one another before normalization. The first-order check response depends on a different quantity:

$$g_{\mathrm{check}}^\top v_{\mathrm{con}} = \|v\|_2 \|g_{\mathrm{check}}\|_2 \cos(g_{\mathrm{check}}, \bar{u}). \tag{8}$$

Support agreement determines the length of the mean unit gradient, whereas its orientation relative to the check gradient determines the sign of local transfer. After norm matching, a coherent support direction can strengthen either the correct or the incorrect held-out preference. This relationship connects the consensus construction to the observed amplification in both subgroups.

5.5 What Replicates Across Concept Combinations?

Development support coherence is positively associated with consensus gain and check alignment. These associations weaken in confirmation, where their confidence intervals span zero. The correct-minus-wrong coherence contrast also changes sign. Restricting development to five-fold concepts preserves a positive development contrast, while confirmation retains its negative estimate (Table 2). Fold-count stratification leaves the central replication pattern intact.

The fit/check alignment contrast persists across cohorts, while agreement among support gradients provides a weaker predictor in confirmation. Together with the dose and consensus experiments, this identifies the operative quantities for interpreting transfer: the check response projected onto the intervention, the finite-dose residual, and the distance to the ranking boundary. The support examples can yield an effective fitting direction and mutually reinforcing gradients while the resulting check response strengthens an existing error. This accounts for the central empirical pattern of transferable margin changes accompanied by sparse decision repair.

6 Conclusion

Activation steering improves output structure and transfers concept-specific information, while decision repair remains sparse. Across ARC and Bongard-LOGO, our experiments connect this gap to the behavior controlled by the intervention and its transfer to held-out examples. Fit/check gradient alignment differs between initially correct and incorrect concepts across development and confirmation. Correction requires the intervention response to overcome the initial margin deficit. Increasing intervention strength and aggregating support gradients amplify both correct and incorrect preferences. Support coherence measures agreement among demonstrations, whereas held-out improvement depends on the aggregate direction's alignment with the check gradient. These findings place directional transfer and boundary crossing at the center of steering for abstract reasoning. They motivate constructing interventions around held-out correction and evaluating margin gains together with repairs and harms.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and revision, literature search, code development and debugging, and figure design. The authors are responsible for verifying the references, experimental records, and final presentation, and for the scientific claims made in this work.

ETHICS STATEMENT

This study investigates activation interventions in language models using ARC-AGI and Bongard-LOGO reasoning tasks. Such interventions can change model preferences and strengthen incorrect predictions, creating risks when applied to systems that inform consequential decisions. They may also be repurposed to manipulate outputs or weaken safeguards. We report decision repairs and harms alongside margin changes to make these effects visible. Applications should include task-specific reliability checks, monitoring of unintended behavior, and adherence to the relevant model and dataset licenses.

REFERENCES

Ekin Akyürek, Mehul Damani, Adam Zweiger, Linlu Qiu, Han Guo, Jyothish Pari, Yoon Kim, and Jacob Andreas. The surprising effectiveness of test-time training for few-shot learning. *arXiv preprint arXiv:2411.07279*, 2024.

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

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

Yi Fang, Wenjie Wang, Mingfeng Xue, Boyi Deng, Fengli Xu, Dayiheng Liu, and Fuli Feng. Controllable llm reasoning via sparse autoencoder-based steering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21290–21305, 2026.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

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 (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

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

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, et al. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. *arXiv preprint arXiv:2503.09532*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, volume 2025, pp. 23888–23923, 2025.

Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *arXiv preprint arXiv:2305.07141*, 2023.

Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. Bongard-logo: A new benchmark for human-level concept learning and reasoning. *Advances in neural information processing systems*, 33:16468–16480, 2020.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

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

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. Mitigating content effects on reasoning in language models through fine-grained activation steering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33314–33322, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Experimental Coverage and Evaluation

The supplementary experiments follow the progression in Sections 3–5: identify the inputs that engage a candidate feature, measure the consequences of changing it, and trace how support-derived directions transfer to held-out decisions. We first specify the experimental units and baselines. Input controls and causal interventions then explain the feature-level results. Numerical audits, support adaptation, and gradient diagnostics connect the learned interventions to the mechanism analysis.

A.1 Models, Representations, and Cohorts

The three principal models are Gemma-2-9B-it, Qwen3-8B-Base, and DeepSeek-R1-Distill-Llama-8B. The feature audit uses Gemma Scope's layer-20, width-16k residual SAE for Gemma; qresearch's layer-19 SAE for R1; and Qwen Scope layer-20 l50, layer-20 l100, and layer-24 l50 configurations for Qwen3. A historical prompt baseline additionally evaluates Qwen2.5-14B-Instruct on ARC-AGI-2. Model identifiers, source configurations, and condition-level records accompany the tables below.

ARC inputs serialize grids as numerical arrays. Evaluation proceeds through three levels: a parseable output grid, the target grid's shape, and an exact match to its cell values. The unit is a test prompt; tasks with multiple test inputs contribute multiple prompts. Bongard-LOGO inputs serialize drawings as action tokens. Query classification evaluates the predicted label of an individual example. Pair ranking compares the positive-label scores of a positive and a negative example. These units remain explicit throughout the appendix.

Table A1: Model-by-experiment coverage of the evidence summarized in this paper. Entries identify the branch supplying each result.

Experiment	Gemma	Qwen3	R1
Language validation	Three selected features	Three SAE settings	Three selected features
Input controls	Neutral, real, synthetic; Bongard	Neutral, real, synthetic	Neutral, real, synthetic
Task baselines	ARC1, ARC2, Bongard	ARC1 summary; Bongard	ARC1, ARC2, Bongard
Causal SAE steering	ARC1, ARC2, Bongard	Candidate screening	ARC1, ARC2
Support adaptation	Development and confirmation	Five-concept pilot	Five-concept pilot

The support experiments comprise a 100-concept query-content development cohort, a 115-concept answer-boundary development cohort, and a 115-concept fresh-combination confirmation cohort. The first uses 50 two-concept donor blocks; the latter two use 23 five-concept blocks. Confirmation evaluates new combinations of shared primitives under the frozen protocol. Cohort membership, intervention locations, and donor blocks are retained separately when estimating effects.

The support-gradient implementation pins the Gemma checkpoint to revision `11c9b309abf73637e4b6f9a3fa1e92e615547819`. Frozen plans record the concept groups and scoring configuration before each expanded run.

A.2 Baselines and Transition Counts

Table A2: Direct-generation baseline counts from recovered prediction files.

Model	Dataset	N	Valid	Shape	Exact	Correct
Gemma	ARC1	419	384	285	11	–
Gemma	ARC2	167	141	60	0	–
Qwen2.5	ARC2	167	121	75	0	–
R1	ARC2	167	9	1	0	–
R1	ARC1	419	92	44	0	–
Gemma	Bongard	200	–	–	–	168
Qwen3	Bongard	200	–	–	–	131
R1	Bongard	200	–	–	–	101

The saved Qwen3 ARC-AGI-1 summary reports 11 exact solutions among 419 test prompts. On the balanced Bongard pilot, positive predictions occur on 90, 169, and 199 of the 200 queries for Gemma, Qwen3, and R1, respectively. This label imbalance motivates recording both classification and pair-ranking outcomes.

For each intervention, repairs count baseline-incorrect examples that become correct; harms count baseline-correct examples that become incorrect. For pair ranking these transitions refer to the sign of the pair margin. The net accuracy change is $(R - H)/N$, where R and H are the repair and harm counts. Continuous margin gains additionally retain movement that leaves the decision unchanged.

B From Language Validation to Task-Input Sensitivity

B.1 Candidate Selection Across SAE Configurations

Language validation scores selected features against problem understanding, backtracking, and verification. Table A3 retains the successful-sample counts and denominators for each configuration. L1 records the earlier Gemma selection; L2 records the subsequent selected-feature validation used in the main comparison. Keeping the configurations explicit makes the selection history traceable. Source filenames appear in Appendix I.

Table A3: Selected-feature language validation.

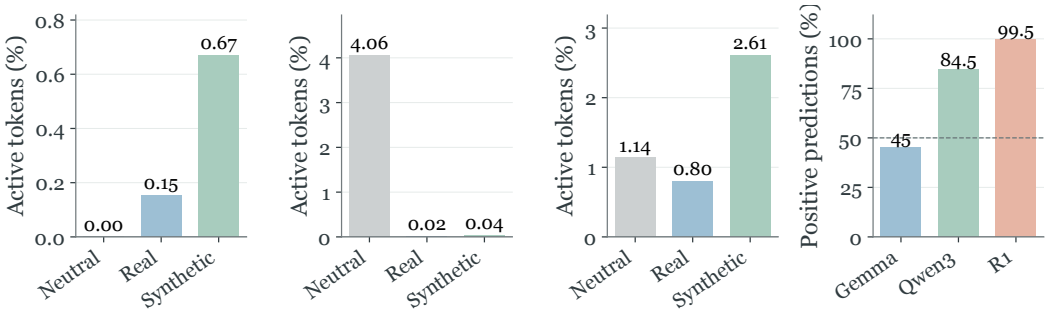
Source	Feature	Strategy	Success	N
L1	3714	Problem Understanding	0	10
L1	435	Backtracking	0	10
L1	448	Verification	0	10
L2	14406	Problem Understanding	6	10
L2	2192	Backtracking	1	10
L2	5886	Verification	4	10
L3	9773	Problem Understanding	1	10
L3	6023	Backtracking	1	10
L3	199	Verification	2	10
L4	32985	Problem Understanding	1	10
L4	2269	Backtracking	1	10
L4	2158	Verification	2	10
L5	2026	Problem Understanding	1	10
L5	1824	Backtracking	1	10
L5	8491	Verification	2	10
L6	26683	Problem Understanding	9	10
L6	4109	Backtracking	8	10
L6	14402	Verification	10	10

R1's selected features have consistently high language-validation scores. The Gemma and Qwen3 profiles vary across strategies and SAE settings. The next measurements ask what these selected directions respond to when the model receives a reasoning task.

B.2 Raw Activity Under Matched Input Controls

The matched controls compare neutral text, real ARC prompts, and synthetic grid-like prompts. Active rate is the fraction of measured tokens with positive feature activation. Mean activation also reflects the magnitude of the response. The percentile reported in the main figure places the active rate within the random-feature reference distribution. Raw active rates retain the distinction between an active feature and a feature tied with inactive reference features.

Gemma 14406 activates on real and synthetic grids, with greater activity on the synthetic inputs. Qwen3 2026 is most active on neutral text among these controls. R1 26683 responds across all three input types. The same selection label consequently covers different input-response profiles. Table A4 extends the comparison to every selected feature and retains the earlier 30- and 46-prompt pilots as separate rows.



(a) Gemma 14406 (b) Qwen3 2026 (c) R1 26683 (d) Bongard labels

Figure A1: Input sensitivity and baseline label preferences. (a)–(c) Raw active-token percentages for the problem-understanding candidate of each model on 167 prompts per input condition. Vertical scales follow each feature's activity range. (d) Positive-label prediction rates on 200 balanced Bongard queries; the dashed line marks the 50% label frequency. Panels report archived aggregate measurements.

Table A4: Activation-control results. Percentiles are expressed on a 0–100 scale.

Source	Feature	Prompts	Mean activation	Active rate	Percentile
A1	14406	200	0.084247	0.012402	91.5
A1	2192	200	0.003818	0.000562	46.5
A1	5886	200	0.000000	0.000000	22.0
A2	14406	46	0.058458	0.007064	93.5
A2	5886	46	0.000000	0.000000	41.0
A2	2192	46	0.000941	0.000156	48.5
A3	14406	167	0.055565	0.006707	93.5
A3	5886	167	0.000000	0.000000	38.0
A3	2192	167	0.000614	0.000088	45.5
A4	2026	167	0.001075	0.000372	95.5
A4	1824	167	0.000000	0.000000	92.0
A4	8491	167	0.000000	0.000000	92.0
A5	26683	167	0.009597	0.026119	99.5
A5	4109	167	0.000000	0.000000	90.0
A5	14402	167	0.000000	0.000006	90.5
A6	14406	30	0.000000	0.000000	37.0
A6	5886	30	0.000000	0.000000	37.0
A6	2192	30	0.000000	0.000000	37.0
A7	14406	167	0.000000	0.000000	33.0
A7	5886	167	0.000000	0.000000	33.0
A7	2192	167	0.000000	0.000000	33.0
A8	2026	167	0.122695	0.040613	99.5
A8	1824	167	0.003063	0.000730	94.0
A8	8491	167	0.000000	0.000000	85.5
A9	26683	167	0.001886	0.011354	97.5
A9	4109	167	0.000000	0.000000	90.0
A9	14402	167	0.000000	0.000000	90.0
A10	14406	167	0.011600	0.001537	82.5
A10	5886	167	0.000000	0.000000	21.0
A10	2192	167	0.013626	0.002080	87.0
A11	14406	46	0.011304	0.001449	82.0
A11	5886	46	0.000000	0.000000	27.5
A11	2192	46	0.012741	0.001946	85.5
A12	2026	167	0.000516	0.000180	94.5
A12	1824	167	0.000000	0.000000	85.0
A12	8491	167	0.000000	0.000000	85.0
A13	26683	167	0.002813	0.007990	98.0
A13	4109	167	0.000001	0.000004	87.5

Source	Feature	Prompts	Mean activation	Active rate	Percentile
A13	14402	167	0.000015	0.000284	94.0

B.3 Where the Activations Occur

Localization assigns recorded activation hits to changed cells, unchanged cells, unassigned cell positions, and non-cell tokens. These categories connect a feature's activity to the serialized task input. Table A5 reports the complete counters. Each category has its own number of available token positions, so the counts describe where observed hits occur; estimating enrichment additionally requires the corresponding opportunity counts.

Table A5: Full recorded localization counters. G denotes Gemma; R denotes R1.

Feature	Total hits	Changed cell	Unchanged cell	Unassigned cell	Non-cell
G14406	1133	107	403	89	534
G2192	1524	68	291	83	1082
R26683	4928	296	1033	761	2838
R4109	3	0	2	1	0
R14402	217	0	2	0	215

The broad distribution of Gemma 14406 and R1 26683 across cell and non-cell tokens complements the real/synthetic controls. R1 14402 is concentrated in non-cell positions. This localization motivates examining generated output structure alongside task correctness in the causal experiments.

B.4 Activation Associations and Task Complexity

Task size and color statistics vary across outcome groups. In the archived ARC-AGI-1 analysis, exact-correct examples have mean input area 61, compared with 278 for the other examples. Shape-matching examples have mean input area 180 versus 468; valid-output examples have mean input area 225 versus 792. The adjustment residualizes activation and outcome measurements against the recorded complexity covariates and correlates their residuals. The accompanying source analysis documents grid area, color counts, and entropy measurements.

The main figure reports the active-rate results. Table A6 gives the corresponding mean-activation analysis. Both measurements show smaller exact- and shape-success associations after adjustment and a negative adjusted valid-output association. Feature 2192 retains a negative exact-correctness association across its active-rate, mean-activation, and maximum-activation summaries. On Bongard, feature 14406 is more active on wrong than on correct queries. These comparisons explain why observational activity and intervention outcomes require separate measurements.

Table A6: Additional difficulty-adjusted ARC-AGI-1 associations for Gemma feature 14406. The main figure reports active-rate associations.

Metric	Outcome	Raw r	Adjusted r	95% CI
Mean activation	Exact	0.2003	0.1051	[-0.0408, 0.2665]
Mean activation	Shape	0.2269	0.0104	[-0.0770, 0.0953]
Mean activation	Valid	0.2358	-0.2037	[-0.2712, -0.1393]

Table A7: Difficulty-adjusted exact-correctness associations for Gemma feature 2192.

Metric	Raw r	Adjusted r	95% CI
Active rate	-0.0845	-0.0804	[-0.1234, -0.0303]
Mean activation	-0.0747	-0.0755	[-0.1164, -0.0322]
Maximum activation	-0.1906	-0.1436	[-0.2071, -0.0713]

Table A8: Gemma feature 14406 on Bongard: correct-minus-wrong activation contrasts from the archived pilot analysis.

Metric	Correct mean	Wrong mean	Difference	95% CI
Mean activation	0.0831	0.0932	-0.0101	[-0.0168, -0.0025]
Active rate	0.01225	0.01377	-0.00152	[-0.00253, -0.00046]

C Causal Steering Across Tasks and Strengths

The intervention audit compares selected feature directions with norm-matched random directions on the same task inputs. Direct decoding provides the baseline task behavior; the random control measures the response to a perturbation of matched magnitude. ARC records valid-grid, shape-match, and exact-match outcomes. Bongard records query classification and preserves full-sequence and output-only intervention scopes as separate conditions.

C.1 Strength Sweeps and Behavioral Specificity

R1 feature 26683 produces a strong valid-grid response on ARC-AGI-1. At strength 15 it generates 182 valid grids, 40 shape matches, and one exact solution among 419 prompts. Direct decoding produces 92 valid grids, 44 shape matches, and zero exact solutions; its paired random control produces 40 valid grids, nine shape matches, and zero exact solutions. The outcome hierarchy identifies a response concentrated in generation structure. R1 feature 14402 reaches zero valid grids at the same strength. The lower-strength summary retains the trajectory into this high-strength behavior.

Table A9: Additional R1 ARC-AGI-1 strength-sweep rates retained in the September 2 experiment summary. F denotes feature steering; R denotes its random control. Each rate uses 419 test prompts; the strength-15 prediction counts appear above.

Strength	Feature	Valid F	Valid R	Shape F	Shape R	Exact F
3	4109	0.2649	0.2291	0.0501	0.0644	0.0048
3	14402	0.2196	0.2411	0.0430	0.0525	0.0024
3	26683	0.2625	0.2172	0.0430	0.0477	0.0024
5	4109	0.3150	0.2411	0.0811	0.0668	0.0000
5	14402	0.2315	0.2458	0.0621	0.0501	0.0095
5	26683	0.3413	0.2243	0.0501	0.0406	0.0024
10	4109	0.2816	0.2792	0.1098	0.0549	0.0048
10	14402	0.0263	0.2554	0.0072	0.0573	0.0000
10	26683	0.4248	0.2458	0.0859	0.0477	0.0000

Gemma's feature-minus-random effects vary across feature, strength, and outcome. ARC-AGI-2 exact-match counts remain zero throughout the archived sweep. On Bongard, output-only and full-sequence interventions use separate records, allowing the scope of the perturbation to be tracked alongside classification changes. The complete condition table preserves the early pilots and the larger evaluations, including controls whose performance differs substantially from direct decoding.

C.2 Complete Condition-Level Results

Table A10 contains all 186 content-deduplicated prediction conditions. Source identifiers resolve to filenames in Appendix I. Columns report successful-outcome counts with the original denominator. A dash marks an outcome belonging to the other task family. Repeated strengths evaluate the same examples and remain separate intervention conditions.

Table A10: All recovered, content-deduplicated prediction conditions. ARC and query-classification outcomes occupy separate columns.

Source	Feature	Dose	Condition	N	Valid	Shape	Exact	Correct
S01	14406	15	random	419	327	259	14	–
S01	5886	15	random	419	329	259	11	–
S01	2192	15	random	419	328	258	15	–
S02	14406	10	random	167	82	44	0	–
S02	5886	10	random	167	83	44	0	–
S02	2192	10	random	167	83	43	0	–
S03	14406	15	random	167	84	42	0	–
S03	5886	15	random	167	82	44	0	–
S03	2192	15	random	167	82	43	0	–
S04	14406	3	random	167	82	44	0	–
S04	5886	3	random	167	83	44	0	–
S04	2192	3	random	167	82	44	0	–
S05	14406	5	random	167	83	44	0	–
S05	5886	5	random	167	83	44	0	–
S05	2192	5	random	167	82	43	0	–
S06	14406	10	random	46	26	13	0	–
S06	5886	10	random	46	27	13	0	–
S06	2192	10	random	46	27	13	0	–
S07	14406	15	random	46	27	12	0	–
S07	5886	15	random	46	27	13	0	–
S07	2192	15	random	46	26	13	0	–
S08	14406	3	random	46	26	13	0	–
S08	5886	3	random	46	26	13	0	–
S08	2192	3	random	46	26	13	0	–
S09	14406	5	random	46	27	13	0	–
S09	5886	5	random	46	26	13	0	–
S09	2192	5	random	46	26	13	0	–
S10	26683	15	random	419	40	9	0	–
S10	4109	15	random	419	68	13	2	–
S10	14402	15	random	419	63	12	0	–
S11	26683	10	random	167	4	0	0	–
S11	4109	10	random	167	16	4	0	–
S11	14402	10	random	167	15	3	0	–
S12	26683	15	random	167	0	0	0	–
S12	4109	15	random	167	5	2	0	–
S12	14402	15	random	167	6	0	0	–
S13	26683	3	random	167	5	0	0	–
S13	4109	3	random	167	9	1	0	–
S13	14402	3	random	167	7	1	0	–
S14	26683	5	random	167	7	0	0	–
S14	4109	5	random	167	8	0	0	–
S14	14402	5	random	167	8	0	0	–
S15	14406	15	feature	419	327	260	14	–
S15	5886	15	feature	419	328	259	14	–
S15	2192	15	feature	419	327	259	13	–
S16	14406	10	feature	167	83	43	0	–
S16	5886	10	feature	167	85	44	0	–
S16	2192	10	feature	167	84	41	0	–
S17	14406	15	feature	167	85	43	0	–
S17	5886	15	feature	167	86	41	0	–
S17	2192	15	feature	167	85	42	0	–
S18	14406	3	feature	167	83	44	0	–
S18	5886	3	feature	167	84	44	0	–
S18	2192	3	feature	167	83	43	0	–
S19	14406	5	feature	167	82	43	0	–
S19	5886	5	feature	167	85	43	0	–
S19	2192	5	feature	167	84	43	0	–
S20	14406	15	feature	46	28	13	0	–
S20	5886	15	feature	46	27	12	0	–

Source	Feature	Dose	Condition	N	Valid	Shape	Exact	Correct
S20	2192	15	feature	46	27	13	0	–
S21	14406	10	feature	46	27	13	0	–
S21	5886	10	feature	46	27	13	0	–
S21	2192	10	feature	46	26	13	0	–
S22	14406	3	feature	46	26	13	0	–
S22	5886	3	feature	46	27	13	0	–
S22	2192	3	feature	46	26	13	0	–
S23	14406	5	feature	46	26	13	0	–
S23	5886	5	feature	46	27	13	0	–
S23	2192	5	feature	46	27	13	0	–
S24	26683	15	feature	419	182	40	1	–
S24	4109	15	feature	419	25	6	0	–
S24	14402	15	feature	419	0	0	0	–
S25	26683	10	feature	167	28	7	0	–
S25	4109	10	feature	167	8	2	0	–
S25	14402	10	feature	167	0	0	0	–
S26	26683	15	feature	167	47	12	0	–
S26	4109	15	feature	167	0	0	0	–
S26	14402	15	feature	167	0	0	0	–
S27	26683	3	feature	167	10	1	0	–
S27	4109	3	feature	167	8	1	0	–
S27	14402	3	feature	167	8	0	0	–
S28	26683	5	feature	167	24	2	0	–
S28	4109	5	feature	167	10	0	0	–
S28	14402	5	feature	167	7	1	0	–
S29	–	–	direct	419	384	285	11	–
S29	–	–	understand	419	381	267	1	–
S29	–	–	verify	419	383	279	12	–
S29	–	–	backtrack	419	313	208	4	–
S30	–	–	direct	167	141	60	0	–
S30	–	–	understand	167	136	51	0	–
S30	–	–	verify	167	137	54	0	–
S30	–	–	backtrack	167	103	34	0	–
S31	–	–	direct	167	121	75	0	–
S31	–	–	understand	167	88	45	0	–
S31	–	–	verify	167	94	51	0	–
S31	–	–	backtrack	167	90	47	0	–
S32	–	–	direct	167	9	1	0	–
S33	–	–	direct	419	92	44	0	–
S34	–	–	feature	200	–	–	–	168
S35	–	–	direct	200	–	–	–	168
S36	13985	10	feature output	200	–	–	–	167
S36	10183	10	feature output	200	–	–	–	168
S36	14218	10	feature output	200	–	–	–	167
S36	14406	10	feature output	200	–	–	–	168
S36	2192	10	feature output	200	–	–	–	167
S36	5886	10	feature output	200	–	–	–	167
S36	14792	10	feature output	200	–	–	–	169
S37	13985	3	feature output	200	–	–	–	168
S37	10183	3	feature output	200	–	–	–	168
S37	14218	3	feature output	200	–	–	–	168
S37	14406	3	feature output	200	–	–	–	168
S37	2192	3	feature output	200	–	–	–	168
S37	5886	3	feature output	200	–	–	–	168
S37	14792	3	feature output	200	–	–	–	168
S38	13985	5	feature output	200	–	–	–	168
S38	10183	5	feature output	200	–	–	–	168
S38	14218	5	feature output	200	–	–	–	167
S38	14406	5	feature output	200	–	–	–	168
S38	2192	5	feature output	200	–	–	–	167
S38	5886	5	feature output	200	–	–	–	167
S38	14792	5	feature output	200	–	–	–	168
S39	13985	10	feature	200	–	–	–	165

Source	Feature	Dose	Condition	N	Valid	Shape	Exact	Correct
S39	10183	10	feature	200	–	–	–	168
S39	14218	10	feature	200	–	–	–	169
S39	14406	10	feature	200	–	–	–	169
S39	2192	10	feature	200	–	–	–	165
S39	5886	10	feature	200	–	–	–	169
S39	14792	10	feature	200	–	–	–	169
S40	13985	3	feature	200	–	–	–	166
S40	10183	3	feature	200	–	–	–	168
S40	14218	3	feature	200	–	–	–	169
S40	14406	3	feature	200	–	–	–	168
S40	2192	3	feature	200	–	–	–	168
S40	5886	3	feature	200	–	–	–	168
S40	14792	3	feature	200	–	–	–	168
S41	13985	5	feature	200	–	–	–	166
S41	10183	5	feature	200	–	–	–	168
S41	14218	5	feature	200	–	–	–	168
S41	14406	5	feature	200	–	–	–	168
S41	2192	5	feature	200	–	–	–	168
S41	5886	5	feature	200	–	–	–	167
S41	14792	5	feature	200	–	–	–	168
S42	13985	10	random output	200	–	–	–	168
S42	10183	10	random output	200	–	–	–	168
S42	14218	10	random output	200	–	–	–	168
S42	14406	10	random output	200	–	–	–	168
S42	2192	10	random output	200	–	–	–	167
S42	5886	10	random output	200	–	–	–	168
S42	14792	10	random output	200	–	–	–	168
S43	13985	3	random output	200	–	–	–	168
S43	10183	3	random output	200	–	–	–	168
S43	14218	3	random output	200	–	–	–	168
S43	14406	3	random output	200	–	–	–	168
S43	2192	3	random output	200	–	–	–	168
S43	5886	3	random output	200	–	–	–	168
S43	14792	3	random output	200	–	–	–	168
S44	13985	5	random output	200	–	–	–	168
S44	10183	5	random output	200	–	–	–	168
S44	14218	5	random output	200	–	–	–	168
S44	14406	5	random output	200	–	–	–	168
S44	2192	5	random output	200	–	–	–	168
S44	5886	5	random output	200	–	–	–	168
S44	14792	5	random output	200	–	–	–	168
S45	13985	10	random	200	–	–	–	168
S45	10183	10	random	200	–	–	–	166
S45	14218	10	random	200	–	–	–	165
S45	14406	10	random	200	–	–	–	167
S45	2192	10	random	200	–	–	–	170
S45	5886	10	random	200	–	–	–	168
S45	14792	10	random	200	–	–	–	167
S46	13985	3	random	200	–	–	–	169
S46	10183	3	random	200	–	–	–	168
S46	14218	3	random	200	–	–	–	168
S46	14406	3	random	200	–	–	–	168
S46	2192	3	random	200	–	–	–	168
S46	5886	3	random	200	–	–	–	168
S46	14792	3	random	200	–	–	–	168
S47	13985	5	random	200	–	–	–	169
S47	10183	5	random	200	–	–	–	166
S47	14218	5	random	200	–	–	–	168
S47	14406	5	random	200	–	–	–	168
S47	2192	5	random	200	–	–	–	167
S47	5886	5	random	200	–	–	–	168
S47	14792	5	random	200	–	–	–	168
S48	–	–	direct	200	–	–	–	131

Source	Feature	Dose	Condition	N	Valid	Shape	Exact	Correct
S49	–	–	direct	200	–	–	–	101

D Scoring, Calibration, and Protected Updates

D.1 Classification and Pair Ranking

For an example x, the label score is $s(x) = \log p(+ \mid x) - \log p(- \mid x)$. A classifier compares this score with a threshold t. Pair ranking compares $s(x^+)$ with $s(x^-)$ through their difference. Adding a common offset b to both scores changes their relation to a fixed threshold while preserving $s(x^+) + b - [s(x^-) + b]$. Classification and ranking consequently expose different responses to an intervention.

The train-calibrated one-pass Gemma classifier uses threshold 4.625 and reaches validation accuracy 0.8125. Generated-trace two-pass scoring reaches 0.685. Validation pair-ranking accuracy changes from 0.87 to 0.71. These measurements motivated the one-pass scoring path for the support experiments: the additional generated trace changed both the decision calibration and the ordering of labeled examples.

D.2 Protected-Update Development Audit

The protected-update audit evaluates a zero arm and twelve nonzero arms on eight evaluation concepts with sixteen queries. Four queries are initially wrong. The arms combine three radii with average-gradient, unprotected max-min, and two protected max-min settings. Table A11 retains every arm.

Table A11: Protected-update evaluation audit. The zero arm is shown with all saved nonzero arms.

Arm	Radius	Repairs	Harms	Pair ranking
Zero	0	0	0	1.0
Average gradient	0.02	0	0	1.0
Max-min unprotected	0.02	0	0	1.0
Max-min protected (0)	0.02	0	0	1.0
Max-min protected (5%)	0.02	0	0	1.0
Average gradient	0.1	0	0	1.0
Max-min unprotected	0.1	0	0	1.0
Max-min protected (0)	0.1	0	0	1.0
Max-min protected (5%)	0.1	0	0	1.0
Average gradient	0.5	0	1	1.0
Max-min unprotected	0.5	0	0	1.0
Max-min protected (0)	0.5	0	0	1.0
Max-min protected (5%)	0.5	0	0	1.0

Every arm preserves the four baseline errors. Average-gradient adaptation at radius 0.5 introduces one harm; the other arms preserve the classifications. Pair-ranking accuracy remains perfect across arms. The coexistence of query errors and correct pair ordering locates this failure in the relationship between scores and the classification decision, motivating the separate margin and transition analyses used in Section 5.

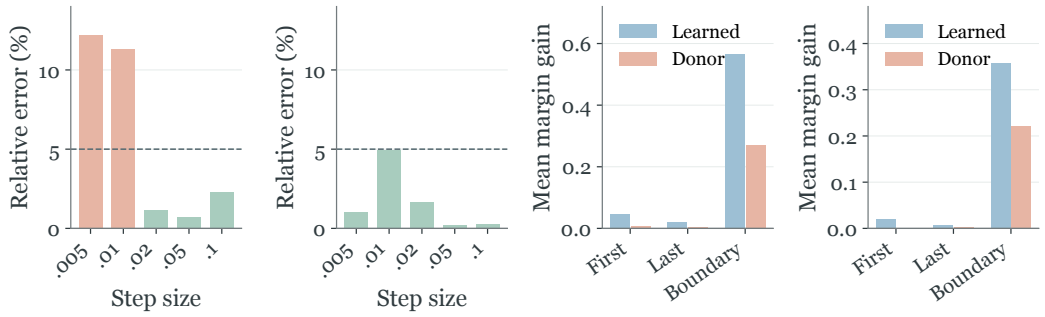
E Numerical Audits and Intervention Locations

E.1 Precision and Directional Derivatives

The precision smoke compares BF16 and FP32 execution of saved directions. Zero-dose outputs agree, while some nonzero responses vanish or reverse sign across precisions. The subsequent support-gradient experiments use FP32 with TF32 disabled. A zero-vector intervention is checked

against the unmodified scoring path. The finite-difference audit compares a central difference, $[\mathcal{L}(\epsilon d) - \mathcal{L}(-\epsilon d)]/(2\epsilon)$, with the automatic-differentiation directional derivative at zero, using a normalized gradient direction d.

Figure A2 retains the full step-size sweep from the query-last diagnostic. The first concept crosses the 5% audit threshold as the finite-difference step increases from 0.01 to 0.02; the second passes at every tested step. The sensitivity to step size explains the initial audit failure and motivates recording the finite-difference values together with the automatic derivative and numerical configuration.



(a) Eight-lines task (b) Six-lines task (c) Fit response (d) Check response

Figure A2: Numerical reliability and intervention location. (a)–(b) Relative finite-difference errors for tasks `bd_necked_eight_lines10_0000` and `bd_necked_six_lines16_0000`; the dashed line is the 5% audit threshold. Gradient norms are 0.002156947 and 0.000147710. (c)–(d) Mean fit and check margin gains in the two-concept location smoke. Each location receives an independently fitted vector under the same optimization budget. The complete random controls and transitions appear in Table A12.

E.2 Single-Token Location Smoke

The location diagnostic compares the first query token, last query token, and answer boundary on two concepts. Each location has its own fitted direction. The answer boundary produces the largest mean fit and check responses under this budget, and all three locations retain the original rankings. This pilot motivated the answer-boundary protocol used in the expanded mechanism analysis.

Table A12: Single-token location smoke on two concepts. Each site has an independently fitted direction with the same optimization budget. All saved fit and check conditions are retained.

Site	Role	Direction	Mean gain	Repair	Harm
query first	fit	Learned	0.04565	0	0
query first	fit	Donor	0.00834	0	0
query first	fit	random 0	0.00011	0	0
query first	fit	random 1	-0.00030	0	0
query first	fit	random 2	0.00173	0	0
query first	fit	Zero	0.00000	0	0
query first	check	Learned	0.01987	0	0
query first	check	Donor	0.00146	0	0
query first	check	random 0	0.00008	0	0
query first	check	random 1	-0.00078	0	0
query first	check	random 2	-0.00022	0	0
query first	check	Zero	0.00000	0	0
query last	fit	Learned	0.02145	0	0
query last	fit	Donor	0.00416	0	0
query last	fit	random 0	-0.00005	0	0
query last	fit	random 1	0.00014	0	0
query last	fit	random 2	0.00006	0	0
query last	fit	Zero	0.00000	0	0
query last	check	Learned	0.00751	0	0

Site	Role	Direction	Mean gain	Repair	Harm
query last	check	Donor	0.00343	0	0
query last	check	random 0	-0.00062	0	0
query last	check	random 1	0.00030	0	0
query last	check	random 2	0.00026	0	0
query last	check	Zero	0.00000	0	0
answer boundary	fit	Learned	0.56495	0	0
answer boundary	fit	Donor	0.27181	0	0
answer boundary	fit	random 0	0.00712	0	0
answer boundary	fit	random 1	-0.01215	0	0
answer boundary	fit	random 2	0.00494	0	0
answer boundary	fit	Zero	0.00000	0	0
answer boundary	check	Learned	0.35729	0	0
answer boundary	check	Donor	0.22111	0	0
answer boundary	check	random 0	-0.00034	0	0
answer boundary	check	random 1	0.00403	0	0
answer boundary	check	random 2	-0.00369	0	0
answer boundary	check	Zero	0.00000	0	0

E.3 Construction and Completion Checks

Support folds are constructed from unique non-check examples. The development implementation uses two to five paired folds according to support availability. The confirmation preflight fixes five folds. Exact-content checks keep each target out of its context; revision 2.1 replaced the earlier substring validator while retaining the direction-construction rule. Completed groups are verified against their frozen plans and output inventories before aggregation. A storage failure or scheduler interruption triggers a group rerun; its exit status is tracked in the run ledger separately from scientific outcomes.

F Support Fitting and Held-Out Transfer

F.1 Learning and Control Construction

At the answer boundary, a task-specific vector is added after transformer block 20, indexed from zero, at the final prompt token. Starting from zero, eight normalized-gradient updates with step size 0.25 minimize $\mathrm{softplus}[-m_{\mathrm{fit}}(v)]$. Each update projects the vector onto an ℓ_2 ball of radius 2. The check pair supplies the transfer measurements. The learned, zero, random, and donor directions share the scoring path.

Norm-matched random directions test the learned orientation against generic perturbation. Donor-concept directions test the contribution of using supports from the target concept. Donor blocks remain intact during bootstrap resampling. For the 100-concept experiment, the donor blocks contain two concepts; the 115-concept experiments use five. Pointwise intervals in the mechanism analyses use 5,000 block-bootstrap resamples. The source summaries retain each analysis's interval estimates, including small Monte Carlo differences between separately run bootstrap analyses.

F.2 Query-Content Development

The 100-concept query-content experiment has 23 initially wrong and 77 initially correct pairs in each split. Learning repairs nine fit pairs and one check pair, with zero harms in either split. The check-margin gain remains positive relative to both random and donor controls. The fit/check comparison connects optimization success to the transfer gap: the optimized direction moves held-out scores, while far fewer held-out examples cross the ranking boundary.

Table A13: 100-concept expansion: fit and check results.

Role	Direction	Wrong	Correct	Repair	Harm	Mean gain
fit	Zero	23	77	0	0	0.0000

Role	Direction	Wrong	Correct	Repair	Harm	Mean gain
fit	Learned	23	86	9	0	1.3607
fit	Donor	23	75	0	2	0.0483
fit	random 0	23	76	0	1	0.0074
fit	random 1	23	77	1	1	0.0043
fit	random 2	23	75	0	2	0.0066
check	Zero	23	77	0	0	0.0000
check	Learned	23	78	1	0	0.2933
check	Donor	23	77	0	0	0.1625
check	random 0	23	77	0	0	0.0118
check	random 1	23	77	0	0	-0.0003
check	random 2	23	77	0	0	0.0075

Table A14: Learned-minus-control check contrasts in the 100-concept expansion. Intervals resample two-concept donor blocks.

Control	Mean gap change	95% CI	Ranking change	95% CI
zero	0.2933	[0.1910, 0.3940]	0.0100	[0, 0.03]
donor	0.1308	[0.0109, 0.2444]	0.0100	[0, 0.03]
mean random	0.2870	[0.1858, 0.3860]	0.0100	[0, 0.03]

F.3 Answer-Boundary Development and Confirmation

The answer-boundary development cohort contains 90 initially correct and 25 initially wrong concepts. At unit dose, the learned direction increases the margin for 86 concepts and decreases it for 29, with all rankings preserved. The learned-minus-zero mean gain is 0.2074 with interval $[0.1539, 0.2584]$; the learned-minus-mean-random contrast is 0.2065 with interval $[0.1529, 0.2574]$; the learned-minus-mean-donor contrast is 0.1172 with interval $[0.0577, 0.1788]$. These contrasts support concept-specific transfer in the development cohort.

Confirmation applies the frozen construction to 115 fresh combinations, comprising 89 initially correct and 26 initially wrong concepts. Each group contributes its saved vectors and score records to the aggregate. The full training inventory contains 23 groups, 115 concepts, and 2,070 score rows. The gradient, dose, and consensus analyses below use these fixed cohorts and preserve their baseline subgroup assignments.

F.4 Cross-Model Support Adaptation

The support-adaptation protocol is also evaluated with Qwen3-8B-Base and DeepSeek-R1-Distill-Llama-8B. Both models use the same frozen confirmation concepts, support splits, and donor assignments as the Gemma confirmation study. The intervention is applied at the answer boundary, after half of the transformer blocks: zero-indexed block 17 of 36 for Qwen3 and block 15 of 32 for R1. Each direction is learned with eight normalized gradient steps of size 0.25 and an ℓ_2 radius of 2.

The pinned model revisions are `49e3418fbbbca6ecbdf9608b4d22e5a407081db4` for Qwen3 and `6a6f4aa4197940add57724a7707d069478df56b1` for R1. The shared confirmation plan has SHA-256 `52937dd7fac237ced13ec5fce99d460daf3b96d39743afef7d0ed70e3492e17e`. Layer selection follows model depth; the radius and optimization budget are fixed before the outputs are examined. Comparisons use each model's own zero, random, and donor controls because hidden-state and margin scales differ across models.

Completed pilot measurements. The completed group-zero records contain five shared concepts and 90 score rows per model. Each target concept contributes zero and learned conditions, three random controls, and four other-concept donor controls, evaluated on fit and check pairs. All fifteen initial directional-derivative audits pass for each model, covering three finite-difference steps per concept. Zero-intervention discrepancies are zero throughout. Table A15 reports the directly computed results. The five concepts form one donor block, so the pilot is summarized with descriptive estimates and individual transitions.

Table A15: Cross-model group-zero results on the same five confirmation concepts. Gains use unit dose. Donor contrasts average the four other-concept controls within each target concept before averaging across concepts; random gains average three matched-norm controls. Each model contributes one donor block.

Measurement	Qwen3-8B-Base	R1-Distill-Llama-8B
Concepts	5	5
Fit-margin gain	+1.3950	+15.7166
Check-margin gain	+0.6654	+5.6120
Check gain: mean random	+0.0037	-0.0269
Check: learned minus mean donor	+0.1977	+0.6321
Correct check rankings: baseline / learned	4 / 4	4 / 5
Check repairs / harms	0 / 0	1 / 0
Gradient audits passed	15 / 15	15 / 15
Maximum relative audit error (%)	0.1187	0.5481

Margin transfer and correction vary across the two models. Both models improve the mean fit and check margins and have a positive mean learned-minus-donor check contrast in this block. Qwen3 preserves all five baseline rankings. R1 repairs its single initially wrong check pair and preserves its four initially correct rankings. The initially wrong concepts differ between the models, as the paired margins in Table A16 show.

Table A16: Individual held-out responses in the cross-model pilot. Concept identifiers are shared across models and defined in Table A17. Positive margins indicate correct pair rankings.

Model	Concept	Baseline margin	Learned margin	Change	Transition
Qwen3	4	0.4110	0.2209	-0.1902	Preserved
Qwen3	1	-0.1016	-0.1584	-0.0568	Preserved
Qwen3	2	2.6627	4.1443	1.4816	Preserved
Qwen3	5	1.4557	2.6915	1.2358	Preserved
Qwen3	3	1.1061	1.9624	0.8563	Preserved
R1	4	1.0190	14.5665	13.5475	Preserved
R1	1	0.5082	1.8718	1.3636	Preserved
R1	2	1.0128	5.8047	4.7919	Preserved
R1	5	0.6145	5.8476	5.2331	Preserved
R1	3	-0.2246	2.8990	3.1236	Repair

Table A17: Frozen concept identities shared by the two cross-model pilots.

Concept	Task identifier
1	`bd_exist_quadrangle_five_lines12-necked_eight_lines6_0000`
2	`bd_open_arc_line_arc4-three_mismatch_triangles4_0000`
3	`bd_open_band_three_arcs5-bird6_0000`
4	`bd_trans_sharp_axe-open_arc_line_arc3_0000`
5	`bd_unbala_trapez_down_right_triangle-no_obtuse_angle_six_lines6_0000`

The R1 repair occurs for `bd_open_band_three_arcs5-bird6_0000`, whose check margin changes from -0.2246 to 2.8990. Qwen3's initially wrong concept is `bd_exist_quadrangle_five_lines12-necked_eight_lines6_0000`; its check margin changes from -0.1016 to -0.1584. For each model, three of the four donor directions also repair its initially wrong pair. These are intervention counts on one pair. The learned-minus-donor advantage in Table A15 measures mean margin transfer across concepts; the individual transitions identify which directions achieve correction.

Connection to the mechanism analysis. The pilot adds a concrete instance of support-derived correction to the broader pattern of margin amplification. In R1, the response on the wrong pair

exceeds its initial margin deficit. In Qwen3, the response on its wrong pair moves further from the boundary. These observed trajectories motivate measuring directional projections and dose responses for each model under the frozen protocol. The pilot contains training and control scores; the fit/check alignment, dose, and consensus comparisons are separate diagnostic stages. The expanded evaluation retains all 23 planned groups and the completed group-zero outputs.

G Gradient Geometry, Dose, and Support Consensus

G.1 Local Alignment and the Final Learned Direction

For $\mathcal{L}(v) = \mathrm{softplus}[-m_{\mathrm{fit}}(v)]$, the descent direction obeys $-\nabla\mathcal{L}(v) = \sigma[-m_{\mathrm{fit}}(v)]\nabla m_{\mathrm{fit}}(v)$. The scalar factor is positive, so a normalized descent step follows the fit-margin gradient. At zero intervention, its projection onto the check gradient gives the first-order change in check margin. The final learned direction accumulates multiple projected updates, and Table A18 records both the initial fit/check cosine and the learned/check cosine.

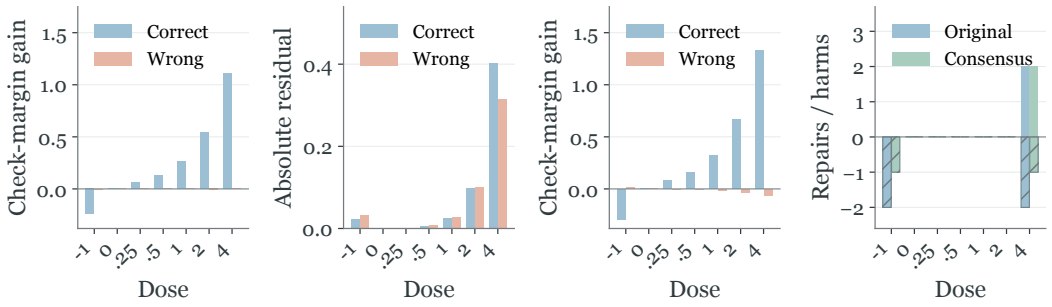
Table A18: Subgroup gradient geometry. Positive counts refer to strictly positive fit/check cosine; the last column uses the final learned vector and the check gradient.

Cohort	Baseline	N	Mean cosine	Median	Positive	Learned cosine
Development	Correct	90	0.3624	0.4900	71	0.3751
Development	Wrong	25	0.0248	0.0185	13	0.0153
Confirmation	Correct	89	0.1669	0.2148	59	0.1840
Confirmation	Wrong	26	-0.0658	-0.1057	12	-0.0821

Initially correct concepts have larger mean fit/check alignment in both cohorts. Positive cosines also occur among initially wrong concepts. The subgroup contrast captures a distributional shift, and the final-vector measurements connect that local geometry to the direction actually used in the dose sweep.

G.2 Complete Dose Responses and Boundary Crossings

The evaluated doses are $-1, 0, 0.25, 0.5, 1, 2, 4$. Every dose reuses the saved direction. Figure A3 complements the confirmation figure in the main text with the development curves. Table A19 retains the mean gains and transitions for both cohorts and constructions; Table A20 gives their uncertainty intervals.



(a) Original direction (b) Linear residual (c) Support consensus (d) Decision transitions

Figure A3: Development dose responses on 115 concepts. (a) Mean check-margin gains under the original learned vector. (b) Absolute residuals from its zero-intervention linear prediction. (c) Gains under norm-matched support consensus. Blue and peach in (a)–(c) identify 90 initially correct and 25 initially wrong concepts. (d) Repairs appear above zero and harms below zero for the original and consensus constructions. Panels (a) and (c) share a vertical scale.

At development dose 4, the original direction yields two repairs and two harms; consensus yields two repairs and one harm. In confirmation, the original direction yields zero repairs and three harms

at this dose, while consensus yields one repair and zero harms. Mean margin movement and net accuracy change consequently describe different aspects of the response. The boundary condition in Equation 5 connects them at the concept level.

Table A19: Complete dose outcomes. Dev denotes development; con denotes fresh-combination confirmation. Gains are check-margin changes.

Cohort	Direction	Dose	All gain	Correct	Wrong	Repair	Harm
dev	Original	-1.0	-0.1889	-0.2407	-0.0026	0	2
dev	Original	0.0	0.0000	0.0000	0.0000	0	0
dev	Original	0.25	0.0505	0.0644	0.0005	0	0
dev	Original	0.5	0.1020	0.1302	0.0006	0	0
dev	Original	1.0	0.2074	0.2653	-0.0009	0	0
dev	Original	2.0	0.4239	0.5446	-0.0106	0	0
dev	Original	4.0	0.8696	1.1095	0.0057	2	2
dev	Consensus	-1.0	-0.2299	-0.2976	0.0139	0	1
dev	Consensus	0.0	0.0000	0.0000	0.0000	0	0
dev	Consensus	0.25	0.0609	0.0787	-0.0031	0	0
dev	Consensus	0.5	0.1231	0.1591	-0.0065	0	0
dev	Consensus	1.0	0.2503	0.3239	-0.0145	0	0
dev	Consensus	2.0	0.5118	0.6643	-0.0370	0	0
dev	Consensus	4.0	1.0259	1.3287	-0.0640	2	1
con	Original	-1.0	-0.0353	-0.0550	0.0321	0	1
con	Original	0.0	0.0000	0.0000	0.0000	0	0
con	Original	0.25	0.0111	0.0175	-0.0106	0	0
con	Original	0.5	0.0234	0.0367	-0.0220	0	0
con	Original	1.0	0.0516	0.0804	-0.0471	0	1
con	Original	2.0	0.1234	0.1894	-0.1027	0	1
con	Original	4.0	0.3140	0.4633	-0.1974	0	3
con	Consensus	-1.0	-0.0616	-0.1177	0.1304	1	1
con	Consensus	0.0	0.0000	0.0000	0.0000	0	0
con	Consensus	0.25	0.0178	0.0330	-0.0344	0	0
con	Consensus	0.5	0.0367	0.0676	-0.0693	0	0
con	Consensus	1.0	0.0778	0.1416	-0.1406	0	0
con	Consensus	2.0	0.1740	0.3086	-0.2868	0	0
con	Consensus	4.0	0.3981	0.6857	-0.5863	1	0

Table A20: Pointwise 95% block-bootstrap intervals for the check-margin gain. The all-concept and initially wrong subgroup estimates correspond to Table A19.

Cohort	Direction	Dose	All concepts	Initially wrong
dev	Original	-1.0	[-0.236, -0.139]	[-0.126, 0.121]
dev	Original	0.0	[0.000, 0.000]	[0.000, 0.000]
dev	Original	0.25	[0.038, 0.063]	[-0.035, 0.033]
dev	Original	0.5	[0.076, 0.127]	[-0.073, 0.068]
dev	Original	1.0	[0.154, 0.259]	[-0.154, 0.137]
dev	Original	2.0	[0.315, 0.531]	[-0.330, 0.268]
dev	Original	4.0	[0.637, 1.093]	[-0.587, 0.513]
dev	Consensus	-1.0	[-0.292, -0.164]	[-0.087, 0.138]
dev	Consensus	0.0	[0.000, 0.000]	[0.000, 0.000]
dev	Consensus	0.25	[0.044, 0.077]	[-0.038, 0.025]
dev	Consensus	0.5	[0.089, 0.154]	[-0.077, 0.051]
dev	Consensus	1.0	[0.181, 0.313]	[-0.160, 0.105]
dev	Consensus	2.0	[0.373, 0.640]	[-0.340, 0.210]
dev	Consensus	4.0	[0.757, 1.288]	[-0.662, 0.437]
con	Original	-1.0	[-0.102, 0.026]	[-0.133, 0.182]
con	Original	0.0	[0.000, 0.000]	[0.000, 0.000]
con	Original	0.25	[-0.005, 0.028]	[-0.050, 0.033]
con	Original	0.5	[-0.009, 0.058]	[-0.100, 0.066]
con	Original	1.0	[-0.014, 0.122]	[-0.205, 0.130]

Cohort	Direction	Dose	All concepts	Initially wrong
con	Original	2.0	[-0.007, 0.265]	[-0.416, 0.250]
con	Original	4.0	[0.061, 0.587]	[-0.762, 0.452]
con	Consensus	-1.0	[-0.101, -0.024]	[0.071, 0.211]
con	Consensus	0.0	[0.000, 0.000]	[0.000, 0.000]
con	Consensus	0.25	[0.008, 0.028]	[-0.056, -0.017]
con	Consensus	0.5	[0.016, 0.058]	[-0.112, -0.035]
con	Consensus	1.0	[0.037, 0.121]	[-0.226, -0.068]
con	Consensus	2.0	[0.089, 0.263]	[-0.462, -0.128]
con	Consensus	4.0	[0.225, 0.582]	[-0.961, -0.201]

Table A21: Absolute residuals from the original direction's first-order prediction.

Cohort	Dose	Initially correct	Initially wrong
development	-1.0	0.02350	0.03192
development	0.0	0.00000	0.00000
development	0.25	0.00153	0.00195
development	0.5	0.00617	0.00762
development	1.0	0.02489	0.02873
development	2.0	0.09891	0.10068
development	4.0	0.40304	0.31469
confirmation	-1.0	0.01801	0.02058
confirmation	0.0	0.00000	0.00000
confirmation	0.25	0.00122	0.00126
confirmation	0.5	0.00496	0.00495
confirmation	1.0	0.02028	0.01910
confirmation	2.0	0.08382	0.07568
confirmation	4.0	0.33509	0.29298

Absolute residuals increase with intervention strength in both baseline subgroups. They measure the accumulated difference between the finite response and the initial linear prediction. Their magnitude enters the crossing condition together with the signed directional projection and the initial margin deficit.

From directional alignment to decision repair. Fit/check alignment describes the local direction of transfer, whereas repair depends on the full finite-dose response and the initial distance to the ranking boundary. For an initially incorrect concept i, define its baseline deficit as $d_i = -m_i(0) > 0$. Under a learned direction v_i at dose α, its movement toward the correct ranking is

$$\Delta_i(\alpha) = m_i(\alpha v_i) - m_i(0) = \alpha g_{\mathrm{check},i}^{\top} v_i + r_i(\alpha).$$

The concept is repaired precisely when $\Delta_i(\alpha) > d_i$. Consequently, favorable fit/check alignment indicates that an initial fit-gradient step points toward a higher check margin, but does not by itself determine whether the final learned vector produces a sufficiently large change to cross the boundary. The gradient norm, the accumulated update, and the finite-dose residual also contribute to that change.

This distinction is visible in the dose outcomes. In the confirmation cohort, the original direction produces no repairs among the 26 initially incorrect concepts at any evaluated positive dose, while its mean margin change for this subgroup is negative at doses 0.25, 0.5, 1, and 2 (Table A19). At dose 4, the original direction produces three harms among initially correct concepts. The norm-matched consensus direction yields one repair at dose 4, together with a negative mean margin change for the initially incorrect subgroup. Thus, an aggregate subgroup response and an individual boundary crossing describe different properties of the same intervention.

Table 2 reports a positive correct-minus-wrong fit/check alignment contrast in both cohorts. This is evidence of a subgroup difference in local transfer geometry; the confirmation interval, $[0.010, 0.448]$, also indicates uncertainty about its magnitude. The present cohorts support the

directional comparison and the observed transition counts. Assessing whether alignment predicts individual repairs beyond baseline margin and task difficulty would require a separate concept-level conditional analysis with enough repair events to estimate that relationship.

G.3 Aggregation, Coherence, and Fold Counts

Each paired leave-one-support-pair-out fold uses one positive and one negative support as targets, with the remaining non-check supports as context. Nonzero gradients are normalized before averaging. The consensus vector rescales the average unit gradient to the original learned vector's norm. Thus the comparison changes direction construction at fixed intervention magnitude. Coherence averages the pairwise cosines of fold gradients; check alignment measures the orientation of their aggregate relative to the held-out gradient.

Table A22: Consensus subgroup statistics. Coherence averages support-gradient pairwise cosines. Check cosine uses the consensus vector. Gain contrasts compare consensus with the original vector and with the mean individual-fold intervention at unit dose.

Cohort	Baseline	N	Coherence	Check cosine	Δ original	Δ individual
Development	Correct	90	0.2478	0.4518	0.0587	0.1046
Development	Wrong	25	0.1431	-0.0035	-0.0136	0.0074
Confirmation	Correct	89	0.1605	0.3086	0.0612	0.0491
Confirmation	Wrong	26	0.2208	-0.2510	-0.0935	-0.0396

The consensus-minus-original contrast is positive for initially correct concepts in both cohorts. For initially wrong concepts, it is negative in both cohorts. These subgroup measurements expose the mechanism behind aggregate gains: support aggregation can reinforce a direction that strengthens the held-out error as well as one that strengthens the correct ranking.

Development contains one two-fold concept, one three-fold concept, eleven four-fold concepts, and 102 five-fold concepts. Confirmation contains 115 five-fold concepts. Table A23 reports all populated strata. Within the five-fold development subset, the baseline-correct minus baseline-wrong coherence contrast remains positive. In confirmation, the corresponding estimate is negative and its interval spans zero. The fit/check alignment contrast persists across cohorts, while the support-coherence relationship changes with the evaluated concept combinations.

Table A23: Fold-count strata. Empty confirmation strata have zero concepts and are absent from the mean calculations.

Cohort	K	N	Coherence	Check cosine	Consensus gain	Original gain
Development	2	1	0.0435	0.4747	0.4620	-0.4090
Development	3	1	0.2121	-0.7246	-0.8046	0.0092
Development	4	11	0.2094	0.5564	0.4395	0.3027
Development	5	102	0.2286	0.3402	0.2382	0.2051
Confirmation	5	115	0.1741	0.1821	0.0778	0.0516

H Retrospective Routing and Error Transitions

The retrospective study asks whether descriptions of an observed failure help identify a corrective intervention. It selects 20 tasks from each ARC dataset and 20 Bongard concepts. An external annotation run produces six category statuses under two presentation orders. The 120 responses have valid format. Across the category-status comparisons, agreement is 69/120 for ARC-AGI-1, 70/120 for ARC-AGI-2, and 57/120 for Bongard; each task has at least one category disagreement across orders.

The intervention library is then evaluated on the selected task subsets. Table A24 reports all 335 archived conditions: seven ARC-AGI-1, seven ARC-AGI-2, and 321 Bongard conditions. The ARC-AGI-2 subset contains 27 test prompts from its 20 tasks. ARC correctness uses exact match;

Bongard correctness uses individual-query classification. Feature conditions retain feature identity, sign, and strength, and random conditions additionally retain the seed.

All archived conditions produce zero repairs on these subsets. Several Bongard conditions introduce one harm. Together with the order-sensitive annotations, this result locates two practical obstacles in retrospective routing: constructing a stable account of the error and selecting a direction that changes its decision. The gradient analysis in Section 5 supplies a directly measurable link between an intervention direction and its held-out response.

Table A24: All retrospective condition summaries. The intervention library and selected tasks are shared across rows.

Dataset	Condition	N	Wrong	Repair	Harm
ARC1	direct	20	20	0	0
ARC1	F 14406 / 15	20	20	0	0
ARC1	F 2192 / 15	20	20	0	0
ARC1	F 5886 / 15	20	20	0	0
ARC1	R for 14406 / 0 / 15	20	20	0	0
ARC1	R for 2192 / 0 / 15	20	20	0	0
ARC1	R for 5886 / 0 / 15	20	20	0	0
ARC2	direct	27	27	0	0
ARC2	F 14406 / 5	27	27	0	0
ARC2	F 2192 / 5	27	27	0	0
ARC2	F 5886 / 5	27	27	0	0
ARC2	R for 14406 / 0 / 5	27	27	0	0
ARC2	R for 2192 / 0 / 5	27	27	0	0
ARC2	R for 5886 / 0 / 5	27	27	0	0
Bongard	direct	40	7	0	0
Bongard	F 1042 / -1 / 10 / -1	40	7	0	0
Bongard	F 1042 / -1 / 1 / -1	40	7	0	0
Bongard	F 1042 / -1 / 3 / -1	40	7	0	0
Bongard	F 1042 / -1 / 5 / -1	40	7	0	0
Bongard	F 1042 / 1 / 10 / -1	40	7	0	1
Bongard	F 1042 / 1 / 1 / -1	40	7	0	0
Bongard	F 1042 / 1 / 3 / -1	40	7	0	0
Bongard	F 1042 / 1 / 5 / -1	40	7	0	0
Bongard	F 11722 / -1 / 10 / -1	40	7	0	0
Bongard	F 11722 / -1 / 1 / -1	40	7	0	0
Bongard	F 11722 / -1 / 3 / -1	40	7	0	0
Bongard	F 11722 / -1 / 5 / -1	40	7	0	0
Bongard	F 11722 / 1 / 10 / -1	40	7	0	0
Bongard	F 11722 / 1 / 1 / -1	40	7	0	0
Bongard	F 11722 / 1 / 3 / -1	40	7	0	0
Bongard	F 11722 / 1 / 5 / -1	40	7	0	0
Bongard	F 13985 / -1 / 10 / -1	40	7	0	0
Bongard	F 13985 / -1 / 1 / -1	40	7	0	0
Bongard	F 13985 / -1 / 3 / -1	40	7	0	0
Bongard	F 13985 / -1 / 5 / -1	40	7	0	0
Bongard	F 13985 / 1 / 10 / -1	40	7	0	0
Bongard	F 13985 / 1 / 1 / -1	40	7	0	0
Bongard	F 13985 / 1 / 3 / -1	40	7	0	0
Bongard	F 13985 / 1 / 5 / -1	40	7	0	0
Bongard	F 14318 / -1 / 10 / -1	40	7	0	0
Bongard	F 14318 / -1 / 1 / -1	40	7	0	0
Bongard	F 14318 / -1 / 3 / -1	40	7	0	0
Bongard	F 14318 / -1 / 5 / -1	40	7	0	0
Bongard	F 14318 / 1 / 10 / -1	40	7	0	0
Bongard	F 14318 / 1 / 1 / -1	40	7	0	0
Bongard	F 14318 / 1 / 3 / -1	40	7	0	0
Bongard	F 14318 / 1 / 5 / -1	40	7	0	0
Bongard	F 15035 / -1 / 10 / -1	40	7	0	0
Bongard	F 15035 / -1 / 1 / -1	40	7	0	0
Bongard	F 15035 / -1 / 3 / -1	40	7	0	0

Dataset	Condition	N	Wrong	Repair	Harm
Bongard	F 15035 / -1 / 5 / -1	40	7	0	0
Bongard	F 15035 / 1 / 10 / -1	40	7	0	0
Bongard	F 15035 / 1 / 1 / -1	40	7	0	0
Bongard	F 15035 / 1 / 3 / -1	40	7	0	0
Bongard	F 15035 / 1 / 5 / -1	40	7	0	0
Bongard	F 3254 / -1 / 10 / -1	40	7	0	0
Bongard	F 3254 / -1 / 1 / -1	40	7	0	0
Bongard	F 3254 / -1 / 3 / -1	40	7	0	0
Bongard	F 3254 / -1 / 5 / -1	40	7	0	0
Bongard	F 3254 / 1 / 10 / -1	40	7	0	1
Bongard	F 3254 / 1 / 1 / -1	40	7	0	0
Bongard	F 3254 / 1 / 3 / -1	40	7	0	0
Bongard	F 3254 / 1 / 5 / -1	40	7	0	1
Bongard	F 5222 / -1 / 10 / -1	40	7	0	0
Bongard	F 5222 / -1 / 1 / -1	40	7	0	0
Bongard	F 5222 / -1 / 3 / -1	40	7	0	0
Bongard	F 5222 / -1 / 5 / -1	40	7	0	0
Bongard	F 5222 / 1 / 10 / -1	40	7	0	0
Bongard	F 5222 / 1 / 1 / -1	40	7	0	0
Bongard	F 5222 / 1 / 3 / -1	40	7	0	0
Bongard	F 5222 / 1 / 5 / -1	40	7	0	0
Bongard	F 715 / -1 / 10 / -1	40	7	0	0
Bongard	F 715 / -1 / 1 / -1	40	7	0	0
Bongard	F 715 / -1 / 3 / -1	40	7	0	0
Bongard	F 715 / -1 / 5 / -1	40	7	0	0
Bongard	F 715 / 1 / 10 / -1	40	7	0	0
Bongard	F 715 / 1 / 1 / -1	40	7	0	0
Bongard	F 715 / 1 / 3 / -1	40	7	0	0
Bongard	F 715 / 1 / 5 / -1	40	7	0	0
Bongard	F 9247 / -1 / 10 / -1	40	7	0	0
Bongard	F 9247 / -1 / 1 / -1	40	7	0	0
Bongard	F 9247 / -1 / 3 / -1	40	7	0	0
Bongard	F 9247 / -1 / 5 / -1	40	7	0	0
Bongard	F 9247 / 1 / 10 / -1	40	7	0	0
Bongard	F 9247 / 1 / 1 / -1	40	7	0	0
Bongard	F 9247 / 1 / 3 / -1	40	7	0	0
Bongard	F 9247 / 1 / 5 / -1	40	7	0	0
Bongard	F 9454 / -1 / 10 / -1	40	7	0	0
Bongard	F 9454 / -1 / 1 / -1	40	7	0	0
Bongard	F 9454 / -1 / 3 / -1	40	7	0	0
Bongard	F 9454 / -1 / 5 / -1	40	7	0	0
Bongard	F 9454 / 1 / 10 / -1	40	7	0	0
Bongard	F 9454 / 1 / 1 / -1	40	7	0	0
Bongard	F 9454 / 1 / 3 / -1	40	7	0	0
Bongard	F 9454 / 1 / 5 / -1	40	7	0	0
Bongard	R 1042 / -1 / 10 / 0	40	7	0	0
Bongard	R 1042 / -1 / 10 / 1	40	7	0	0
Bongard	R 1042 / -1 / 10 / 2	40	7	0	0
Bongard	R 1042 / -1 / 1 / 0	40	7	0	0
Bongard	R 1042 / -1 / 1 / 1	40	7	0	0
Bongard	R 1042 / -1 / 1 / 2	40	7	0	0
Bongard	R 1042 / -1 / 3 / 0	40	7	0	0
Bongard	R 1042 / -1 / 3 / 1	40	7	0	0
Bongard	R 1042 / -1 / 3 / 2	40	7	0	0
Bongard	R 1042 / -1 / 5 / 0	40	7	0	0
Bongard	R 1042 / -1 / 5 / 1	40	7	0	0
Bongard	R 1042 / -1 / 5 / 2	40	7	0	0
Bongard	R 1042 / 1 / 10 / 0	40	7	0	0
Bongard	R 1042 / 1 / 10 / 1	40	7	0	0
Bongard	R 1042 / 1 / 10 / 2	40	7	0	1
Bongard	R 1042 / 1 / 1 / 0	40	7	0	0
Bongard	R 1042 / 1 / 1 / 1	40	7	0	0
Bongard	R 1042 / 1 / 1 / 2	40	7	0	0

Dataset	Condition	N	Wrong	Repair	Harm
Bongard	R 1042 / 1 / 3 / 0	40	7	0	0
Bongard	R 1042 / 1 / 3 / 1	40	7	0	0
Bongard	R 1042 / 1 / 3 / 2	40	7	0	0
Bongard	R 1042 / 1 / 5 / 0	40	7	0	0
Bongard	R 1042 / 1 / 5 / 1	40	7	0	0
Bongard	R 1042 / 1 / 5 / 2	40	7	0	0
Bongard	R 11722 / -1 / 10 / 0	40	7	0	0
Bongard	R 11722 / -1 / 10 / 1	40	7	0	0
Bongard	R 11722 / -1 / 10 / 2	40	7	0	0
Bongard	R 11722 / -1 / 1 / 0	40	7	0	0
Bongard	R 11722 / -1 / 1 / 1	40	7	0	0
Bongard	R 11722 / -1 / 1 / 2	40	7	0	0
Bongard	R 11722 / -1 / 3 / 0	40	7	0	0
Bongard	R 11722 / -1 / 3 / 1	40	7	0	0
Bongard	R 11722 / -1 / 3 / 2	40	7	0	0
Bongard	R 11722 / -1 / 5 / 0	40	7	0	0
Bongard	R 11722 / -1 / 5 / 1	40	7	0	0
Bongard	R 11722 / -1 / 5 / 2	40	7	0	0
Bongard	R 11722 / 1 / 10 / 0	40	7	0	0
Bongard	R 11722 / 1 / 10 / 1	40	7	0	0
Bongard	R 11722 / 1 / 10 / 2	40	7	0	0
Bongard	R 11722 / 1 / 1 / 0	40	7	0	0
Bongard	R 11722 / 1 / 1 / 1	40	7	0	0
Bongard	R 11722 / 1 / 1 / 2	40	7	0	0
Bongard	R 11722 / 1 / 3 / 0	40	7	0	0
Bongard	R 11722 / 1 / 3 / 1	40	7	0	0
Bongard	R 11722 / 1 / 3 / 2	40	7	0	0
Bongard	R 11722 / 1 / 5 / 0	40	7	0	0
Bongard	R 11722 / 1 / 5 / 1	40	7	0	0
Bongard	R 11722 / 1 / 5 / 2	40	7	0	0
Bongard	R 13985 / -1 / 10 / 0	40	7	0	1
Bongard	R 13985 / -1 / 10 / 1	40	7	0	0
Bongard	R 13985 / -1 / 10 / 2	40	7	0	0
Bongard	R 13985 / -1 / 1 / 0	40	7	0	0
Bongard	R 13985 / -1 / 1 / 1	40	7	0	0
Bongard	R 13985 / -1 / 1 / 2	40	7	0	0
Bongard	R 13985 / -1 / 3 / 0	40	7	0	0
Bongard	R 13985 / -1 / 3 / 1	40	7	0	0
Bongard	R 13985 / -1 / 3 / 2	40	7	0	0
Bongard	R 13985 / -1 / 5 / 0	40	7	0	0
Bongard	R 13985 / -1 / 5 / 1	40	7	0	0
Bongard	R 13985 / -1 / 5 / 2	40	7	0	0
Bongard	R 13985 / 1 / 10 / 0	40	7	0	0
Bongard	R 13985 / 1 / 10 / 1	40	7	0	0
Bongard	R 13985 / 1 / 10 / 2	40	7	0	1
Bongard	R 13985 / 1 / 1 / 0	40	7	0	0
Bongard	R 13985 / 1 / 1 / 1	40	7	0	0
Bongard	R 13985 / 1 / 1 / 2	40	7	0	0
Bongard	R 13985 / 1 / 3 / 0	40	7	0	0
Bongard	R 13985 / 1 / 3 / 1	40	7	0	0
Bongard	R 13985 / 1 / 3 / 2	40	7	0	0
Bongard	R 13985 / 1 / 5 / 0	40	7	0	0
Bongard	R 13985 / 1 / 5 / 1	40	7	0	0
Bongard	R 13985 / 1 / 5 / 2	40	7	0	0
Bongard	R 14318 / -1 / 10 / 0	40	7	0	0
Bongard	R 14318 / -1 / 10 / 1	40	7	0	0
Bongard	R 14318 / -1 / 10 / 2	40	7	0	0
Bongard	R 14318 / -1 / 1 / 0	40	7	0	0
Bongard	R 14318 / -1 / 1 / 1	40	7	0	0
Bongard	R 14318 / -1 / 1 / 2	40	7	0	0
Bongard	R 14318 / -1 / 3 / 0	40	7	0	0
Bongard	R 14318 / -1 / 3 / 1	40	7	0	0
Bongard	R 14318 / -1 / 3 / 2	40	7	0	0

Dataset	Condition	N	Wrong	Repair	Harm
Bongard	R 14318 / -1 / 5 / 0	40	7	0	0
Bongard	R 14318 / -1 / 5 / 1	40	7	0	0
Bongard	R 14318 / -1 / 5 / 2	40	7	0	0
Bongard	R 14318 / 1 / 10 / 0	40	7	0	0
Bongard	R 14318 / 1 / 10 / 1	40	7	0	0
Bongard	R 14318 / 1 / 10 / 2	40	7	0	0
Bongard	R 14318 / 1 / 1 / 0	40	7	0	0
Bongard	R 14318 / 1 / 1 / 1	40	7	0	0
Bongard	R 14318 / 1 / 1 / 2	40	7	0	0
Bongard	R 14318 / 1 / 3 / 0	40	7	0	0
Bongard	R 14318 / 1 / 3 / 1	40	7	0	0
Bongard	R 14318 / 1 / 3 / 2	40	7	0	0
Bongard	R 14318 / 1 / 5 / 0	40	7	0	0
Bongard	R 14318 / 1 / 5 / 1	40	7	0	0
Bongard	R 14318 / 1 / 5 / 2	40	7	0	0
Bongard	R 15035 / -1 / 10 / 0	40	7	0	0
Bongard	R 15035 / -1 / 10 / 1	40	7	0	1
Bongard	R 15035 / -1 / 10 / 2	40	7	0	0
Bongard	R 15035 / -1 / 1 / 0	40	7	0	0
Bongard	R 15035 / -1 / 1 / 1	40	7	0	0
Bongard	R 15035 / -1 / 1 / 2	40	7	0	0
Bongard	R 15035 / -1 / 3 / 0	40	7	0	0
Bongard	R 15035 / -1 / 3 / 1	40	7	0	0
Bongard	R 15035 / -1 / 3 / 2	40	7	0	0
Bongard	R 15035 / -1 / 5 / 0	40	7	0	0
Bongard	R 15035 / -1 / 5 / 1	40	7	0	0
Bongard	R 15035 / -1 / 5 / 2	40	7	0	0
Bongard	R 15035 / 1 / 10 / 0	40	7	0	0
Bongard	R 15035 / 1 / 10 / 1	40	7	0	0
Bongard	R 15035 / 1 / 10 / 2	40	7	0	1
Bongard	R 15035 / 1 / 1 / 0	40	7	0	0
Bongard	R 15035 / 1 / 1 / 1	40	7	0	0
Bongard	R 15035 / 1 / 1 / 2	40	7	0	0
Bongard	R 15035 / 1 / 3 / 0	40	7	0	0
Bongard	R 15035 / 1 / 3 / 1	40	7	0	0
Bongard	R 15035 / 1 / 3 / 2	40	7	0	0
Bongard	R 15035 / 1 / 5 / 0	40	7	0	0
Bongard	R 15035 / 1 / 5 / 1	40	7	0	0
Bongard	R 15035 / 1 / 5 / 2	40	7	0	0
Bongard	R 3254 / -1 / 10 / 0	40	7	0	0
Bongard	R 3254 / -1 / 10 / 1	40	7	0	0
Bongard	R 3254 / -1 / 10 / 2	40	7	0	0
Bongard	R 3254 / -1 / 1 / 0	40	7	0	0
Bongard	R 3254 / -1 / 1 / 1	40	7	0	0
Bongard	R 3254 / -1 / 1 / 2	40	7	0	0
Bongard	R 3254 / -1 / 3 / 0	40	7	0	0
Bongard	R 3254 / -1 / 3 / 1	40	7	0	0
Bongard	R 3254 / -1 / 3 / 2	40	7	0	0
Bongard	R 3254 / -1 / 5 / 0	40	7	0	0
Bongard	R 3254 / -1 / 5 / 1	40	7	0	0
Bongard	R 3254 / -1 / 5 / 2	40	7	0	0
Bongard	R 3254 / 1 / 10 / 0	40	7	0	0
Bongard	R 3254 / 1 / 10 / 1	40	7	0	0
Bongard	R 3254 / 1 / 10 / 2	40	7	0	0
Bongard	R 3254 / 1 / 1 / 0	40	7	0	0
Bongard	R 3254 / 1 / 1 / 1	40	7	0	0
Bongard	R 3254 / 1 / 1 / 2	40	7	0	0
Bongard	R 3254 / 1 / 3 / 0	40	7	0	0
Bongard	R 3254 / 1 / 3 / 1	40	7	0	0
Bongard	R 3254 / 1 / 3 / 2	40	7	0	0
Bongard	R 3254 / 1 / 5 / 0	40	7	0	0
Bongard	R 3254 / 1 / 5 / 1	40	7	0	0
Bongard	R 3254 / 1 / 5 / 2	40	7	0	0

Dataset	Condition	N	Wrong	Repair	Harm
Bongard	R 5222 / -1 / 10 / 0	40	7	0	0
Bongard	R 5222 / -1 / 10 / 1	40	7	0	0
Bongard	R 5222 / -1 / 10 / 2	40	7	0	1
Bongard	R 5222 / -1 / 1 / 0	40	7	0	0
Bongard	R 5222 / -1 / 1 / 1	40	7	0	0
Bongard	R 5222 / -1 / 1 / 2	40	7	0	0
Bongard	R 5222 / -1 / 3 / 0	40	7	0	0
Bongard	R 5222 / -1 / 3 / 1	40	7	0	0
Bongard	R 5222 / -1 / 3 / 2	40	7	0	0
Bongard	R 5222 / -1 / 5 / 0	40	7	0	0
Bongard	R 5222 / -1 / 5 / 1	40	7	0	0
Bongard	R 5222 / -1 / 5 / 2	40	7	0	0
Bongard	R 5222 / 1 / 10 / 0	40	7	0	0
Bongard	R 5222 / 1 / 10 / 1	40	7	0	0
Bongard	R 5222 / 1 / 10 / 2	40	7	0	0
Bongard	R 5222 / 1 / 1 / 0	40	7	0	0
Bongard	R 5222 / 1 / 1 / 1	40	7	0	0
Bongard	R 5222 / 1 / 1 / 2	40	7	0	0
Bongard	R 5222 / 1 / 3 / 0	40	7	0	0
Bongard	R 5222 / 1 / 3 / 1	40	7	0	0
Bongard	R 5222 / 1 / 3 / 2	40	7	0	0
Bongard	R 5222 / 1 / 5 / 0	40	7	0	0
Bongard	R 5222 / 1 / 5 / 1	40	7	0	0
Bongard	R 5222 / 1 / 5 / 2	40	7	0	0
Bongard	R 715 / -1 / 10 / 0	40	7	0	0
Bongard	R 715 / -1 / 10 / 1	40	7	0	0
Bongard	R 715 / -1 / 10 / 2	40	7	0	0
Bongard	R 715 / -1 / 1 / 0	40	7	0	0
Bongard	R 715 / -1 / 1 / 1	40	7	0	0
Bongard	R 715 / -1 / 1 / 2	40	7	0	0
Bongard	R 715 / -1 / 3 / 0	40	7	0	0
Bongard	R 715 / -1 / 3 / 1	40	7	0	0
Bongard	R 715 / -1 / 3 / 2	40	7	0	0
Bongard	R 715 / -1 / 5 / 0	40	7	0	0
Bongard	R 715 / -1 / 5 / 1	40	7	0	0
Bongard	R 715 / -1 / 5 / 2	40	7	0	0
Bongard	R 715 / 1 / 10 / 0	40	7	0	0
Bongard	R 715 / 1 / 10 / 1	40	7	0	0
Bongard	R 715 / 1 / 10 / 2	40	7	0	0
Bongard	R 715 / 1 / 1 / 0	40	7	0	0
Bongard	R 715 / 1 / 1 / 1	40	7	0	0
Bongard	R 715 / 1 / 1 / 2	40	7	0	0
Bongard	R 715 / 1 / 3 / 0	40	7	0	0
Bongard	R 715 / 1 / 3 / 1	40	7	0	0
Bongard	R 715 / 1 / 3 / 2	40	7	0	0
Bongard	R 715 / 1 / 5 / 0	40	7	0	0
Bongard	R 715 / 1 / 5 / 1	40	7	0	0
Bongard	R 715 / 1 / 5 / 2	40	7	0	0
Bongard	R 9247 / -1 / 10 / 0	40	7	0	0
Bongard	R 9247 / -1 / 10 / 1	40	7	0	0
Bongard	R 9247 / -1 / 10 / 2	40	7	0	0
Bongard	R 9247 / -1 / 1 / 0	40	7	0	0
Bongard	R 9247 / -1 / 1 / 1	40	7	0	0
Bongard	R 9247 / -1 / 1 / 2	40	7	0	0
Bongard	R 9247 / -1 / 3 / 0	40	7	0	0
Bongard	R 9247 / -1 / 3 / 1	40	7	0	0
Bongard	R 9247 / -1 / 3 / 2	40	7	0	0
Bongard	R 9247 / -1 / 5 / 0	40	7	0	0
Bongard	R 9247 / -1 / 5 / 1	40	7	0	0
Bongard	R 9247 / -1 / 5 / 2	40	7	0	0
Bongard	R 9247 / 1 / 10 / 0	40	7	0	0
Bongard	R 9247 / 1 / 10 / 1	40	7	0	0
Bongard	R 9247 / 1 / 10 / 2	40	7	0	1

Dataset	Condition	N	Wrong	Repair	Harm
Bongard	R 9247 / 1 / 1 / 0	40	7	0	0
Bongard	R 9247 / 1 / 1 / 1	40	7	0	0
Bongard	R 9247 / 1 / 1 / 2	40	7	0	0
Bongard	R 9247 / 1 / 3 / 0	40	7	0	0
Bongard	R 9247 / 1 / 3 / 1	40	7	0	0
Bongard	R 9247 / 1 / 3 / 2	40	7	0	0
Bongard	R 9247 / 1 / 5 / 0	40	7	0	0
Bongard	R 9247 / 1 / 5 / 1	40	7	0	0
Bongard	R 9247 / 1 / 5 / 2	40	7	0	0
Bongard	R 9454 / -1 / 10 / 0	40	7	0	0
Bongard	R 9454 / -1 / 10 / 1	40	7	0	0
Bongard	R 9454 / -1 / 10 / 2	40	7	0	0
Bongard	R 9454 / -1 / 1 / 0	40	7	0	0
Bongard	R 9454 / -1 / 1 / 1	40	7	0	0
Bongard	R 9454 / -1 / 1 / 2	40	7	0	0
Bongard	R 9454 / -1 / 3 / 0	40	7	0	0
Bongard	R 9454 / -1 / 3 / 1	40	7	0	0
Bongard	R 9454 / -1 / 3 / 2	40	7	0	0
Bongard	R 9454 / -1 / 5 / 0	40	7	0	0
Bongard	R 9454 / -1 / 5 / 1	40	7	0	0
Bongard	R 9454 / -1 / 5 / 2	40	7	0	0
Bongard	R 9454 / 1 / 10 / 0	40	7	0	0
Bongard	R 9454 / 1 / 10 / 1	40	7	0	0
Bongard	R 9454 / 1 / 10 / 2	40	7	0	0
Bongard	R 9454 / 1 / 1 / 0	40	7	0	0
Bongard	R 9454 / 1 / 1 / 1	40	7	0	0
Bongard	R 9454 / 1 / 1 / 2	40	7	0	0
Bongard	R 9454 / 1 / 3 / 0	40	7	0	0
Bongard	R 9454 / 1 / 3 / 1	40	7	0	0
Bongard	R 9454 / 1 / 3 / 2	40	7	0	0
Bongard	R 9454 / 1 / 5 / 0	40	7	0	0
Bongard	R 9454 / 1 / 5 / 1	40	7	0	0
Bongard	R 9454 / 1 / 5 / 2	40	7	0	0

I Source Index and Reproducibility

The accompanying evidence directory contains the prediction-condition index, paired transitions, activation controls, language-validation records, localization counters, support-mechanism summaries, retrospective outcomes, and source hashes. Prediction rows are deduplicated by content while preserving distinct configurations. The experiment inventory distinguishes completed measurements, saved aggregate summaries, implementation protocols, and ongoing model replications. This status is attached to each branch before it enters a result table.

I.1 Development Branches and Artifact Status

The development record also contains branches that led to the final support-transfer protocol. Shared-vector experiments ask whether one update generalizes across concepts. Protected objectives constrain changes on already-correct examples. Report-based diagnostics ask whether generated descriptions predict later errors. The task-specific protocol then places supervision at the concept level and measures its transfer on a disjoint pair. Table A25 identifies the available artifact for each branch; the quantitative sections above use completed score records or completed summaries.

I.2 Source Configurations

Table A25: Additional development branches in the experiment ledger. Artifact status identifies the material available for analysis.

Branch	Recorded design	Available artifact
Shared-vector corridors	SAE-guided, free, and random directions; two-concept smoke and eight-fit/eight-evaluation development	Protocols, training implementation, later Jacobian records
Shared-vector generalization	384 training and 100 tuning concepts; ordinary/protected arms, 48 updates, and hinge variants	Protocols and follow-up analysis records
Report-based error prediction	96 concepts split into 64 fit and 32 evaluation; hash and MiniLM representations	Protocols and initial semantic-screen record
Task-specific pilot	Two-concept smoke and sixteen-concept, eight-block pilot	Smoke records and pilot protocol
Additional model replication	Qwen3 and R1 under the frozen confirmation construction	Completed five-concept score records (Appendix F.4); expanded runs in progress
Check-supervised reference	Reference directions and support selection	Proposed design

Table A26: Language-validation source configurations.

ID	Source
L1	`stage2b_gemma9b_selected_features.csv`
L2	`stage2b_gemma9b_selected_features_top10_4omini.csv`
L3	`stage2b_qwen3_8b_l100_layer20_selected_features_top10_4omini.csv`
L4	`stage2b_qwen3_8b_l150_layer24_selected_features_top10_4omini.csv`
L5	`stage2b_qwen3_8b_selected_features_top10_4omini.csv`
L6	`stage2b_r1_llama8b_qresearch_l19_selected_features_top10_4omini.csv`

Table A27: Input-control source configurations.

ID	Source
A1	`bongard_logo_activation_baseline_gemma_pilot100.json`
A2	`fake_arc_activation_baseline_gemma.json`
A3	`fake_arc_activation_baseline_gemma_arc120matched.json`
A4	`fake_arc_activation_baseline_qwen3_arc120matched.json`
A5	`fake_arc_activation_baseline_r1_llama_arc120matched.json`
A6	`neutral_activation_baseline_gemma.json`
A7	`neutral_activation_baseline_gemma_167.json`
A8	`neutral_activation_baseline_qwen3_167.json`
A9	`neutral_activation_baseline_r1_llama_167.json`
A10	`zero_activation_baseline_gemma_arc_eval120.json`
A11	`zero_activation_baseline_gemma_arc_eval30.json`
A12	`zero_activation_baseline_qwen3_arc_eval120.json`
A13	`zero_activation_baseline_r1_llama_arc_eval120.json`

Table A28: Prediction source index.

ID	File
S01	`arc_random_control_gemma9b_arcagi1_eval_s15.jsonl`
S02	`arc_random_control_gemma9b_eval120_s10.jsonl`

ID	File
S03	arc_random_control_gemma9b_eval120_s15.jsonl
S04	arc_random_control_gemma9b_eval120_s3.jsonl
S05	arc_random_control_gemma9b_eval120_s5.jsonl
S06	arc_random_control_gemma9b_eval30_s10.jsonl
S07	arc_random_control_gemma9b_eval30_s15.jsonl
S08	arc_random_control_gemma9b_eval30_s3.jsonl
S09	arc_random_control_gemma9b_eval30_s5.jsonl
S10	arc_random_control_r1_llama8b_qresearch_l19_arcagi1_eval_s15.jsonl
S11	arc_random_control_r1_llama8b_qresearch_l19_eval120_s10.jsonl
S12	arc_random_control_r1_llama8b_qresearch_l19_eval120_s15.jsonl
S13	arc_random_control_r1_llama8b_qresearch_l19_eval120_s3.jsonl
S14	arc_random_control_r1_llama8b_qresearch_l19_eval120_s5.jsonl
S15	arc_transfer_gemma9b_stage2b_features_arcagi1_eval_s15.jsonl
S16	arc_transfer_gemma9b_stage2b_features_eval120_s10.jsonl
S17	arc_transfer_gemma9b_stage2b_features_eval120_s15.jsonl
S18	arc_transfer_gemma9b_stage2b_features_eval120_s3.jsonl
S19	arc_transfer_gemma9b_stage2b_features_eval120_s5.jsonl
S20	arc_transfer_gemma9b_stage2b_features_eval30.jsonl
S21	arc_transfer_gemma9b_stage2b_features_eval30_s10.jsonl
S22	arc_transfer_gemma9b_stage2b_features_eval30_s3.jsonl
S23	arc_transfer_gemma9b_stage2b_features_eval30_s5.jsonl
S24	arc_transfer_r1_llama8b_qresearch_l19_arcagi1_eval_s15.jsonl
S25	arc_transfer_r1_llama8b_qresearch_l19_eval120_s10.jsonl
S26	arc_transfer_r1_llama8b_qresearch_l19_eval120_s15.jsonl
S27	arc_transfer_r1_llama8b_qresearch_l19_eval120_s3.jsonl
S28	arc_transfer_r1_llama8b_qresearch_l19_eval120_s5.jsonl
S29	baseline_gemma2_9b_it_arcagi1_eval_full.jsonl
S30	baseline_gemma2_9b_it_eval_full.jsonl
S31	baseline_qwen25_14b_eval_full.jsonl
S32	baseline_r1_llama8b_arc_eval120_direct.jsonl
S33	baseline_r1_llama8b_arcagi1_eval_direct.jsonl
S34	bongard_logo_gemma_per_record_activation.jsonl
S35	bongard_logo_gemma_pilot100_direct.jsonl
S36	bongard_logo_gemma_structsteer_feature_output_only_s10.jsonl
S37	bongard_logo_gemma_structsteer_feature_output_only_s3.jsonl
S38	bongard_logo_gemma_structsteer_feature_output_only_s5.jsonl
S39	bongard_logo_gemma_structsteer_feature_s10.jsonl
S40	bongard_logo_gemma_structsteer_feature_s3.jsonl
S41	bongard_logo_gemma_structsteer_feature_s5.jsonl
S42	bongard_logo_gemma_structsteer_random_output_only_s10.jsonl
S43	bongard_logo_gemma_structsteer_random_output_only_s3.jsonl
S44	bongard_logo_gemma_structsteer_random_output_only_s5.jsonl
S45	bongard_logo_gemma_structsteer_random_s10.jsonl
S46	bongard_logo_gemma_structsteer_random_s3.jsonl
S47	bongard_logo_gemma_structsteer_random_s5.jsonl
S48	bongard_logo_qwen3_pilot100_direct.jsonl
S49	bongard_logo_r1_llama_pilot100_direct.jsonl

I.3 Analysis Artifacts

The numerical panels in Figure A2(a)–(b) transcribe the saved query-last finite-difference audit from job 909679. The remaining new panels and generated tables use the structured evidence files packaged with this source. The plot script embeds Georgia and exports vector PDFs with the same blue, green, and peach palette used in the main figures. The evidence index preserves available sample sizes, intervention scopes, and bootstrap units so that each displayed estimate can be traced to its experiment.

The quantitative record combines the three-model feature audit, the Gemma development and confirmation mechanism analyses, and the completed Qwen3 and R1 support-adaptation pilots in Appendix F.4. The expanded Qwen3 and R1 cohorts follow the same frozen plan. Their aggregate

effects are computed after all group outputs are collected and validated. Report-information and shared-vector development retain their artifact status in Table A25.

J Mechanistic Implications for Reasoning Interventions

The experiments connect activation steering to a sequence of measurable events: an intervention changes a representation, the change affects a task-relevant score, and the score movement may cross a decision boundary. Each stage introduces a distinct requirement. Feature interpretation identifies a candidate intervention target. Support supervision determines an update direction. Held-out correction depends on how that direction interacts with the new example and its existing decision. This sequence organizes the findings into a framework for studying where reasoning interventions succeed and where their effects stop.

The level of improvement identifies the mechanism to explain. Valid output, successful support fitting, held-out margin gain, and decision repair occupy different positions in this sequence. The ARC experiments place the strongest response of a reasoning-associated feature at the level of output structure. The Bongard experiments show that concept-specific information can transfer through an activation update while existing rankings persist. Together, these findings motivate a layered account of intervention efficacy: identify the behavior that changes, establish which properties of the direction produce that change, and measure its consequences for the task decision. This account turns a small accuracy effect into a precise question about the stage reached by the intervention.

Transfer is a relationship between examples and directions. For a normalized fit-gradient step, the first-order check response is $\eta \|g_{\text{check}}\|_2 \cos(g_{\text{fit}}, g_{\text{check}})$. Its magnitude depends on local sensitivity, and its sign depends on alignment. A direction can consequently support effective optimization on one example and strengthen an error on another. The replicated correct-versus-wrong alignment contrast makes this relationship central to the interpretation of support adaptation. It also changes the object of analysis: the relevant quantity is the interaction between the learned direction and the held-out example. Future studies can examine how this interaction varies with support composition, representation layer, intervention location, and concept structure.

Correction requires movement relative to the initial deficit. For an initially wrong pair, the baseline margin deficit is $d = -m_{\text{check}}(0) > 0$. Correction occurs when $\alpha g_{\text{check}}^{\top} v + r(\alpha) > d$. This expression separates three sources of persistent error: an adverse directional projection, a favorable response smaller than the deficit, and a finite-dose response that departs from the local prediction. Their practical implications differ. The first motivates changing the direction, the second motivates examining the attainable response along that direction, and the third motivates measuring the intervention trajectory. Reporting these quantities together would explain why similarly sized margin gains yield different decision outcomes and would make comparisons across intervention strengths more informative.

Support agreement and held-out usefulness require separate evidence. Consensus exposes a further geometric distinction. Its coherence statistic depends on inner products among support gradients. Applying a common orthogonal rotation to those gradients preserves their pairwise cosines and the norm of their mean, while changing the mean direction's alignment with a fixed check gradient. Support coherence therefore leaves the sign of held-out transfer undetermined. The confirmation results instantiate this possibility: aggregation strengthens the positive response of initially correct concepts and the negative response of initially wrong concepts. This motivates studying which properties of support selection connect internal agreement to held-out orientation. Diversity in the examples, coverage of concept-defining transformations, and consistency across context changes provide concrete axes for such tests.

A diagnostic target for the next generation of interventions. The resulting research objective is to construct directions whose favorable held-out response is large enough to overcome the relevant margin deficits while preserving correct decisions. Check gradients provide an evaluation-time reference for studying this objective; support-only signals provide the information available to a deployable construction. Comparing the two can reveal which observable properties predict useful

transfer across new concept combinations. A corresponding evaluation should retain fit and check responses, norm-matched controls, fixed dose curves, baseline deficits, and paired repair and harm counts. These measurements connect direction construction to its causal effect and make improvements in reasoning traceable through the entire intervention process.

K Concept-Specific Amplification and the Geometry of Error

The central empirical pattern combines three properties that are usually examined separately: an update is learned from the target concept, its effect transfers to held-out examples, and the resulting decisions remain incorrect. The donor comparison establishes an advantage for target-concept supervision in development. The gradient analysis relates transfer to the orientation of the held-out response. The consensus experiment shows how aggregating support evidence can strengthen existing errors. Read together, these results reveal a consequential possibility: an intervention can carry useful information about a concept while moving a particular example further into an incorrect decision.

Concept specificity and correction are separate geometric properties. A learned direction's advantage over donor directions measures the value of matching the supervision to the target concept. Correction requires that this advantage produce a sufficiently large, favorably oriented response on an initially wrong example. These properties operate at different levels. A positive mean learned-minus-donor contrast can coexist with adverse responses within the initially wrong subgroup. The experiments make this coexistence observable by connecting concept-matched controls to baseline-stratified gradients and decision transitions. This suggests evaluating concept specificity together with its distribution across existing successes and failures. Such an evaluation asks where the transferred information becomes useful.

An aggregate gain can be driven by reinforcing existing successes. Let π be the fraction of initially correct concepts, and let $\mu_{\mathrm{correct}}(\alpha)$ and $\mu_{\mathrm{wrong}}(\alpha)$ denote their respective mean margin gains. The aggregate response satisfies $\mu(\alpha) = \pi\mu_{\mathrm{correct}}(\alpha) + (1-\pi)\mu_{\mathrm{wrong}}(\alpha)$. A growing positive response among initially correct concepts can dominate this average even as initially wrong concepts move in the opposite direction. The confirmation dose curves exhibit this pattern for both the original and consensus constructions. Baseline composition therefore affects the apparent efficacy of steering: the same subgroup responses produce different aggregate gains as the prevalence of existing errors changes. Baseline-stratified response curves expose this dependence and provide a more informative basis for comparing models, datasets, and intervention methods.

Aggregation can make an unfavorable direction more effective. Averaging normalized support gradients reduces cancellation among directions that reinforce one another. The subsequent norm matching isolates the orientation of the aggregate at the original intervention magnitude. In confirmation, this construction increases the positive mean response of initially correct concepts and deepens the negative mean response of initially wrong concepts. Greater agreement among the demonstrations can thus produce a more effective update along an unfavorable held-out direction. This observation motivates a distinction between variation across support examples and a directional tendency shared by those examples. Their separation is a concrete target for future work: vary the support set, measure the resulting directions, and determine whether adverse held-out projections persist across independently constructed updates.

Persistent errors contain several experimentally separable mechanisms. The boundary condition $\alpha g_{\mathrm{check}}^{\top} v + r(\alpha) > -m_{\mathrm{check}}(0)$ provides a way to organize error trajectories. A wrong example may have a favorable local projection and remain below the boundary because its initial deficit is large. Another may have an adverse local projection and move further from correction. A third may change behavior as the finite-dose residual grows. These trajectories can yield the same unchanged accuracy at unit dose while implying different intervention strategies. Their separation calls for concept-level measurements of initial deficits, directional projections, and finite responses. The present dose and gradient experiments supply the quantities needed to formulate this analysis and identify the subgroup structure that makes it consequential.

Matched-direction contrasts can locate the source of transfer failure. The geometric account suggests a controlled diagnostic that builds directly on the existing experiments. At a fixed layer, location, and vector norm, compare a support-derived direction with an evaluation-time direction aligned to the check-margin gradient. Their local responses separate the orientation supplied by support supervision from the sensitivity available at the intervention site. A large favorable check-aligned response alongside a weak or adverse support-derived response would locate the problem in direction construction. Dose sweeps could then test whether this local advantage reaches the decision boundary. Repeating the comparison across sites would examine how the accessible correction response depends on where the model is intervened upon. This diagnostic is a prospective experiment motivated by the observed alignment and dose effects.

The next methodological objective is error-directed transfer. These findings shift attention toward the examples on which the model's existing preference is wrong. A promising construction should explain how its support information yields a favorable direction for those examples, how that response scales relative to their margin deficits, and how it preserves already-correct decisions. This objective connects support selection, representation geometry, and decision-level evaluation within one experimental framework. It also gives negative results a constructive role: an unchanged decision becomes evidence about orientation, sensitivity, or boundary distance once its response trajectory is measured.

The distinctive contribution of the study lies in connecting these levels of evidence. Feature interpretation, input sensitivity, concept-matched transfer, support agreement, and decision correction are followed through the same investigative sequence. Their observed relationships reveal how apparently favorable intervention signals can accumulate while errors persist or strengthen. This account gives future work a set of explicit mechanisms to test and a precise target for improvement: converting support-derived information into error-directed movement that crosses the decision boundary.