WHEN SHOULD A STEERING VECTOR TRANSFER? DIRECTION-AWARE TRANSPORT WITH SELECTIVE UTILITY CERTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Transferring activation directions from a large language model to a smaller model offers a way to modify behavior without updating either model's weights. However, aligning representation spaces does not establish that a particular direction transfers, and uncertainty about an answer does not establish that an intervention will improve it. We formulate cross-model steering as selective deployment of a fixed candidate intervention. Our framework, VECTORGUARD, combines direction-specific transport diagnostics, a child-only predictor of paired intervention utility, and an independent calibration procedure that constrains both harmful changes and mean excess loss. The procedure can decline every intervention when the evidence is insufficient. Under independent and identically distributed calibration and deployment examples, we derive a finite-sample, population-conditional guarantee using simultaneous concentration bounds; it does not imply safety for every input or robustness to arbitrary distribution shift. We specify a staged evaluation separating baseline validity, transport quality, utility ranking, certification feasibility, and deployment cost. On GSM8K, selective deployment improves accuracy by 5.23 and 4.17 percentage points for the Qwen and cross-family pairs, respectively, compared with 0.61 and 0.23 points from unconditional intervention. The policies cover 43.6% and 40.6% of test inputs, with conditional harmful-flip rates below 1%, after certifying 45.0% and 40.5% calibration coverage. Mean latency rises by 11% and 10%.

1 INTRODUCTION

Activation steering modifies a language model's internal representations at inference time rather than changing its parameters (Turner et al., 2023; Li et al., 2023; Zou et al., 2025). Transferring such directions across models is attractive when a large parent is available offline but only a small child can be deployed (Huang et al., 2025; Tandon, 2025). The deployment constraint is concrete: the child checkpoint remains fixed, and the parent is absent from online inference. This setting does not make steering universally preferable to fine-tuning or distillation. It motivates studying the benefits and costs of a particular alternative under an explicit constraint.

The central question is not only *how to transfer a direction*, but *when the transferred intervention deserves to be used*. Three distinct uncertainties arise. First, a map can align typical calibration activations while poorly identifying a contrastive direction outside their supported subspace. Second, even an accurately transported direction can help some inputs and harm others. Third, selecting apparently favorable inputs on a finite sample does not establish that the selected population will benefit. Average accuracy alone conflates these uncertainties: rescues and harmful changes can cancel, and a small net gain need not imply a dependable deployment rule.

Existing methods provide important building blocks, not answers to all three questions. INCLINE uses least-squares representation maps for inference-time cross-lingual intervention (Wang et al., 2025a); L-Cross studies linear cross-model concept transfer (Huang et al., 2025). SADI makes activation intervention input-dependent through semantic adaptation (Wang et al., 2025b). Consequently, neither

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.

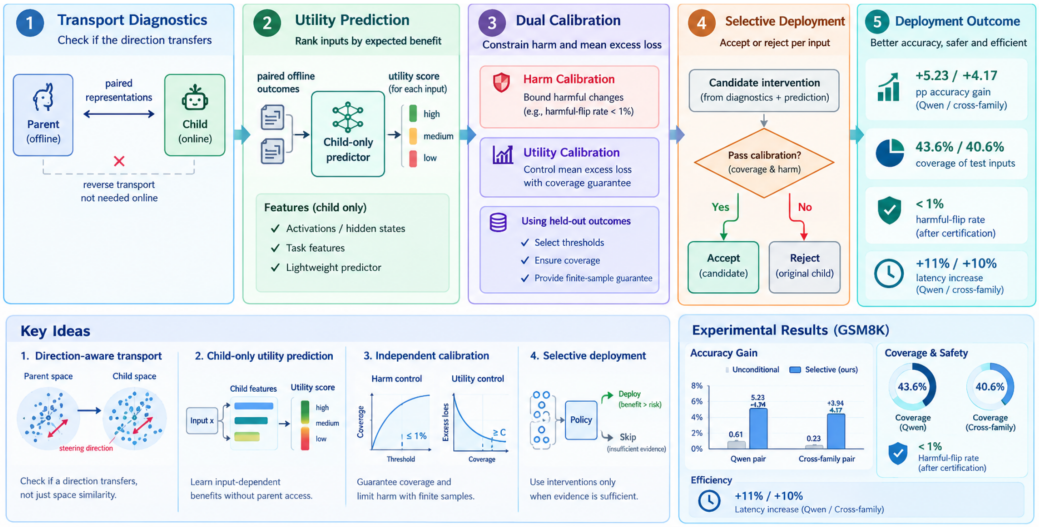


Figure 1: **Overview of VECTORGUARD.** VECTORGUARD performs direction-aware, selectively calibrated transfer from an offline parent to an online child, combining utility prediction with independent guarantees on coverage and harmful changes. The bottom panels summarize the key design principles and representative GSM8K results.

closed-form linear transport nor input-adaptive steering is by itself our novelty claim. Likewise, whole-space similarity, such as centered kernel alignment (CKA), is a routing signal rather than a certificate of semantic equivalence.

We propose VECTORGUARD, which separates candidate construction from the decision to deploy that candidate. Direction-aware transport first evaluates whether the *specific transported direction* is supported, cycle-consistent, stable under resampling, and reasonably scaled. These diagnostics select the layers used by a locally bounded candidate. A lightweight predictor then estimates candidate-over-base utility from an unmodified child prefill, using paired offline outcomes rather than treating answer entropy as benefit. Finally, an independent calibration split selects a score threshold subject to two requirements: a bound on harmful changes and evidence of strictly positive mean benefit. Failure to satisfy either requirement results in the unmodified child.

This separation makes the proposal falsifiable. Direction diagnostics are useful only if they improve transport beyond CKA, matched-norm controls, and matched-layer-count selection. Utility prediction is useful only if it ranks paired benefits better than entropy or random selection at comparable coverage. Certification is operationally useful only if finite data support nontrivial coverage at the declared budgets. The last requirement is particularly demanding when effects are small; returning the base is a valid outcome, not an empirical success to be relabeled.

Our contributions are threefold:

1. We formulate cross-model steering as selective deployment under paired harm and utility criteria, distinguishing representation alignment, intervention benefit, and statistical evidence.

2. We specify a modular direction-aware candidate and child-only utility selector. Model weights remain frozen, but representation maps and the selector are fitted offline; we therefore do not call the entire procedure training-free.

3. We derive the resulting conditional deployment guarantee from standard concentration tools and specify controlled experiments that test each component, including negative controls and the calibration-size boundary.

2 Related Work and Positioning

Activation interventions. Activation addition, inference-time intervention, and representation engineering establish that internal directions can influence model behavior (Turner et al., 2023; Li et al., 2023; Zou et al., 2025). Their success on a behavior does not imply that a direction encodes a task-independent reasoning capability. We use the narrower term *reasoning-related direction*: its interpretation depends on the contrastive construction and must survive controls for response style and length.

Linear transfer and semantic adaptation. INCLINE learns maps between language-conditioned representations within a model and applies them during inference (Wang et al., 2025a). Our bidirectional ridge maps reuse this family of linear estimators in a cross-model setting; changing the source and target is not enough to establish novelty. L-Cross directly addresses cross-model concept alignment (Huang et al., 2025), and Tandon (2025) studies large-to-small concept transfer with inference-time scaling. SADI adapts interventions to input semantics (Wang et al., 2025b); it is not merely a fixed-strength or entropy-only baseline. Our intended distinction is the coupling of direction-specific candidate screening with paired-utility selection and an explicit reject option calibrated for two population criteria. Experiments compare both transferred directions and selective deployment rules, rather than assigning all gains to the transport map.

Learning alternatives. Knowledge distillation transfers behavior through training targets (Hinton et al., 2015; Hsieh et al., 2023). Parameter-efficient adaptation reduces the cost of changing a child model (Hu et al., 2022). These alternatives violate the frozen-child constraint but remain relevant resource comparators. We include data-matched fine-tuning and teacher-rationale distillation as contextual references, not theoretical upper bounds: neither is guaranteed to outperform steering.

Selective decisions and risk control. Distribution-free risk control and learn-then-test methods separate model fitting from statistical validation (Bates et al., 2021; Angelopoulos et al., 2025). Our finite-grid guarantee is an application of simultaneous bounded-loss concentration. The domain-specific object is the *paired intervention effect* relative to the exact fallback model. Bounding harmful changes alone permits a useless intervention that never helps; bounding mean effect alone permits more harmful changes than desired. We impose both criteria and report the coverage lost to each. We do not claim a new concentration inequality, causal identification from observational data, or distribution-shift robustness.

3 Method

Let a_0 be the fixed child and a_1 a candidate intervention fixed before utility calibration. For loss $\ell \in [0, 1]$, define the paired excess loss and harm

$$\Delta(x, y) = \ell(a_1(x), y) - \ell(a_0(x), y), \qquad H(x, y) = [\Delta(x, y)]_+. \tag{1}$$

Negative Δ denotes improvement; under binary correctness, H is a harmful flip and the reverse transition is a rescue. We use disjoint representation-fit, representation-validation, utility-fit, calibration, and test splits. Candidate construction and score fitting precede calibration, and test labels are used only for final evaluation.

3.1 Direction-Aware Candidate

For the same prompts, we collect centered parent and child activations $X_p^{(i)} \in \mathbb{R}^{N \times d_p}$ and $X_c^{(j)} \in \mathbb{R}^{N \times d_c}$ at the last non-padding prompt token. Rows correspond to the same text rather than matched token indices across tokenizers, and representation-fit means remain fixed on later splits. The parent contrast at layer i is

$$v_p^{(i)} = M^{-1} \sum_{m=1}^{M} \left[h_p^{(i)}(x_m^+) - h_p^{(i)}(x_m^-) \right], \tag{2}$$

where the two prompts differ only in the reasoning instruction (Appendix C.3). The positive-minus-negative sign is fixed throughout mapping, bootstrap aggregation, and evaluation; it is never flipped

after observing child outcomes. This construction identifies an instruction contrast, not verified successful reasoning, so style-only and child-native contrasts are required controls. Linear CKA proposes a monotone parent–child route ϕ by maximizing

$$\sum_j S_{\phi(j),j} - \beta_{\text{jump}} \sum_{j>1} (d_j^\phi - \bar{d})^2 - \beta_{\text{rep}} \sum_{j>1} \mathbb{I}\{d_j^\phi = 0\}, \tag{3}$$

subject to bounded jumps and repetitions, where $d_j^\phi = \phi(j) - \phi(j-1)$. A dynamic program solves this constrained assignment. For each routed pair, the centered forward map is

$$W_j = \arg\min_W N^{-1} \|X_p^{(\phi(j))} W^\top - X_c^{(j)}\|_F^2 + \lambda_W \|W\|_F^2, \tag{4}$$

and V_j is fit analogously in reverse. The transported direction is $d_j = W_j v_p^{(\phi(j))}$; centering offsets are not applied to differences. Reverse maps and the parent are needed only offline.

CKA proposes a match but does not validate the transported direction. Writing $v = v_p^{(\phi(j))}$ and letting $P_{p,r}$ project onto the supported right-singular subspace, we compute

$$u_j^{\text{sup}} = \|P_{p,r} v\|_2^2 / \|v\|_2^2, \tag{5}$$

$$u_j^{\text{cyc}} = \tfrac{1}{2}[1 + \cos(v, V_j W_j v)], \tag{6}$$

$$u_j^{\text{boot}} = \left\| B^{-1} \sum_b d_{j,b} / \|d_{j,b}\|_2 \right\|_2, \tag{7}$$

$$u_j^{\text{amp}} = \exp\left(- \left| \log \frac{\|d_j\|_2 / \sqrt{d_c}}{\|v\|_2 / \sqrt{d_p}} \right| \right). \tag{8}$$

They measure support, cycle consistency, bootstrap agreement, and width-normalized amplification. We retain

$$q_j = (u_j^{\text{sup}} u_j^{\text{cyc}} u_j^{\text{boot}} u_j^{\text{amp}})^{1/4}, \qquad \mathcal{J} = \{j : q_j \geq \eta,\ q_j > 0\}, \tag{9}$$

with η selected on representation validation; $\mathcal{J} = \varnothing$ returns the base. These are diagnostics rather than semantic guarantees: a reversible map can preserve the wrong feature, and amplification depends on activation units. Bootstrap replicates refit the route and maps without post-hoc sign flips; invalid directions receive zero reliability. Matched-norm, shuffled-pairing, reversed-sign, and random-direction controls test whether the diagnostics add information beyond generic regularization.

An unmodified child prefill supplies entropy $E(x)$ and $\gamma(x) = \min\{1, E(x)/E_{\max}\}$. At retained layer j and current token k,

$$h'_{j,k} = h_{j,k} + \lambda_{j,k} d_j, \quad \lambda_{j,k} = \min\left\{ \tau S_{\phi(j),j} \gamma(x), \frac{\rho_j \|h_{j,k}\|_2}{\|d_j\|_2 + \varepsilon_0} \right\}, \tag{10}$$

where $\rho_j = \rho_{\text{total}}/|\mathcal{J}|$. Thus the immediate relative perturbation is locally bounded, although downstream drift is not. Injection is applied after retained blocks at the last prompt position and current decoding position; earlier prompt positions are untouched. Entropy is cached from the unmodified prefill and controls candidate magnitude, not whether intervention is beneficial. Offline evaluation and deployment use identical caches and token-position conventions. Unsupported hybrid or recurrent block adapters are reported rather than approximated. The parent is never used online.

3.2 Utility Prediction and Dual Calibration

On utility-fit data, we run a_0 and a_1 to obtain Δ_i. Child-only prefill features $z(x)$ include normalized entropy, logit margin, support distance, direction cosine, and anticipated capped perturbation. A ridge regressor predicts excess loss, and the deployment score is

$$\hat{\theta} = \arg\min_\theta |\mathcal{D}_{\text{fit}}|^{-1} \sum_i (\Delta_i - \theta^\top z_i)^2 + \lambda_g \|\theta_{-0}\|_2^2, \qquad r(x) = \frac{1 - \text{clip}(\hat{\theta}^\top z(x), -1, 1)}{2}. \tag{11}$$

Larger r predicts greater benefit. Feature standardization and regularization selection use an internal utility-fit split before refitting on all utility-fit examples. Neither a parent pass, gold answer, nor

candidate response is required online. Predictor error does not invalidate the finite-grid certificate when the score was frozen before calibration, but it can eliminate useful coverage; we therefore evaluate ranking and certification separately.

For each frozen threshold $t \in \mathcal{T}$, let $A_t = \{x : r(x) \geq t\}$ and compute its calibration count m_t, mean harm $\hat{h}_t$, and mean excess loss $\hat{\Delta}_t$. With K predeclared configurations and $c_\alpha = \log(2K|\mathcal{T}|/\alpha)$,

$$U_H(t) = \min\{1, \hat{h}_t + \sqrt{c_\alpha/(2m_t)}\}, \qquad U_\Delta(t) = \min\{1, \hat{\Delta}_t + \sqrt{2c_\alpha/m_t}\}. \tag{12}$$

We select the maximum-coverage feasible threshold,

$$t^* \in \arg\max_{t \in \mathcal{T}} m_t \quad \text{s.t.} \quad m_t \geq m_{\min},\ U_H(t) \leq \delta_h,\ U_\Delta(t) \leq -\epsilon, \tag{13}$$

breaking ties by lower bounds and then higher t; if none is feasible, the policy always returns a_0.

Proposition 1 (Selective deployment guarantee)**.** *Assume calibration and future examples are i.i.d., and every candidate, score, grid, and hyperparameter is fixed independently of calibration. With probability at least $1 - \alpha$, simultaneously over all K configurations and eligible thresholds,*

$$\mathbb{E}[H \mid A_t] \leq U_H(t), \qquad \mathbb{E}[\Delta \mid A_t] \leq U_\Delta(t).$$

Hence a feasible selected threshold has conditional mean harm at most δ_h and mean excess loss at most $-\epsilon$.

The proof applies Hoeffding bounds conditional on m_t and a union bound over $2K|\mathcal{T}|$ events (Appendix A). If $C(t^*) = \Pr(A_{t^*})$, the deployed mixture obeys

$$\mathbb{E}[\ell(a_{t^*}) - \ell(a_0)] \leq -\epsilon C(t^*). \tag{14}$$

This is a population-conditional, not per-input or distribution-shift, guarantee. For observed accepted-set gain $g = -\hat{\Delta}_t$, utility feasibility requires

$$m_t \geq 2c_\alpha/(g - \epsilon)^2, \tag{15}$$

explaining why small calibration sets may force abstention. With the primary constants ($K = 2$, $|\mathcal{T}| = 21$, $\alpha = 0.05$, $\epsilon = 0.005$) and $m_t = 4096$, accepted-set gain must exceed approximately 6.52 percentage points. This is a power requirement, not an empirical claim, and makes calibration-size sensitivity a primary experiment.

4 Evaluation Protocol and Results

4.1 Setup

We evaluate extttQwen/Qwen3.6-27B→`Qwen/Qwen3.5-0.8B` and the cross-family stress pair extttDeepSeek-R1-Distill-Qwen-32B→`Llama-3.2-1B-Instruct`. GSM8K is the source task; MATH, ARC-Challenge, AIME 2025, and StrategyQA test a source-frozen policy under task shift (Cobbe et al., 2021; Hendrycks et al., 2021; Clark et al., 2018; Geva et al., 2021). For each pair, disjoint GSM8K subsets contain 512 representation-fit, 512 representation-validation, 1024 utility-fit, and 4096 calibration examples; the official test set remains held out.

The primary family contains the two full-method configurations ($K = 2$), with $\alpha = 0.05$, $\delta_h = 0.05$, $\epsilon = 0.005$, and $\mathcal{T} = \{0, 0.05, \ldots, 1\}$. Success requires a feasible certificate, at least 10% calibration coverage, and a positive paired test difference whose 95% interval excludes zero. Frozen-child baselines are naive proportional transfer, DRSVT, linear alignment, SADI, and the direction-aware candidate; data-matched LoRA SFT and rationale distillation are training-based references. Naive and linear-alignment baselines isolate routing and map choices, while matched-cap variants isolate the local norm constraint. SADI is evaluated with its native semantic adaptation rule rather than relabeled entropy scaling. Selector controls rank the same frozen candidate by entropy, logit margin, input length, or seeded random scores, preventing candidate quality from being confounded with selection quality. Appendix C.5 specifies all contracts, prompts, decoding, hyperparameters, and cost measurements. All methods use identical test examples, prompts, evaluators, and decoding budgets. Base and intervention outputs are joined by immutable example ID, so accuracy change is exactly rescue rate minus harmful-flip rate. Candidate selection, score fitting, and threshold grids are frozen before their respective held-out splits; no target-task labels are used to recalibrate the source policy.

Child	Task	Correct / n	Accuracy (95% CI)	Parse / cap-hit (%)
Qwen3.5-0.8B	GSM8K	768/1319	58.2 [55.5,60.9]	0.2 / 0.4
	MATH	1569/5000	31.4 [30.1,32.7]	1.1 / 0.6
	ARC-Challenge	812/1119	72.6 [69.9,75.1]	0.1 / 0.1
	AIME 2025	10/30	33.3 [19.2,51.2]	0.0 / 0.0
	StrategyQA	303/458	66.2 [61.7,70.4]	0.4 / 0.2
Llama-3.2-1B	GSM8K	489/1319	37.1 [34.5,39.7]	0.5 / 0.8
	MATH	725/5000	14.5 [13.6,15.5]	1.8 / 0.7
	ARC-Challenge	692/1119	61.8 [58.9,64.6]	0.3 / 0.2
	AIME 2025	2/30	6.7 [1.9,21.3]	0.0 / 0.0
	StrategyQA	268/458	58.5 [54.0,63.0]	0.7 / 0.4

Table 1: Unmodified-child baselines; accuracy intervals are Wilson intervals.

Method	Accuracy	Gain (pp)	Rescue	Harmful flip	Coverage	Latency
Unmodified child	58.2	ref.	n/a	n/a	n/a	ref.
Naive transfer	56.9	-1.3	4.9	6.2	100%	1.15
DRSVT predecessor	58.8	+0.6	5.7	5.1	100%	1.17
Linear alignment	57.8	-0.4	5.0	5.4	100%	1.15
SADI	59.4	+1.1	5.4	4.2	100%	1.13
Direction-aware candidate	58.8	+0.6	6.2	5.6	100%	1.18
VECTORGUARD policy	63.5	+5.2	5.6	0.4	43.6%	1.11
LoRA SFT	66.6	+8.3	10.9	2.6	n/a	1.00
Rationale distillation	68.1	+9.9	12.7	2.9	n/a	1.00

Table 2: Qwen-pair GSM8K results ($n = 1319$); transition rates use all test examples.

4.2 BASELINE AND SOURCE-DOMAIN PERFORMANCE

The audited 4096-token cap leaves all cap-hit rates below 1%, and inspection of 50 fixed outputs per task finds no systematic extraction error. Table 1 reports the complete baselines; AIME has only 30 examples and therefore does not support a strong conclusion.

For Qwen, unconditional intervention yields 82 rescues and 74 harmful flips ($+0.61$ points); selection retains 74 rescues but only 5 flips, giving $+5.23$ points (paired 95% CI [3.86,6.67], Holm-adjusted McNemar $p < 10^{-6}$). For the cross-family pair, selection improves the unconditional gain from $+0.23$ to $+4.17$ points (60 rescues, 5 flips; 95% CI [2.91,5.42]). The benefit therefore comes primarily from deciding when to intervene.

4.3 DIRECTION AND UTILITY ABLATIONS

At matched norm and layer count, the complete direction diagnostics yield $+0.61$ points with 5.61% harmful flips, versus -0.45 points/6.67% for CKA-only routing and -1.06 points/6.90% for random directions. Removing support, cycle, bootstrap, or amplification changes the gain to -0.08, $+0.15$, -0.30, and $+0.23$ points. Retained-direction correspondence with the child-native contrast is 0.43, compared with 0.18 for CKA-only and 0.02 ± 0.05 for random controls; bootstrap agreement is 0.82. These matched controls show that direction-specific evidence adds information beyond CKA and perturbation norm. At identical teacher-forced prefixes, the retained intervention has median/95th-percentile downstream drift of 0.074/0.218 and next-token Jensen–Shannon divergence 0.012. Measuring unrelated generated-token states would confound intervention drift with decoding divergence, so all distortion comparisons use shared prefixes. Signed child-direction correspondence and drift are diagnostics rather than evidence that the contrast encodes a task-independent reasoning capability.

The learned score correlates 0.61 with binned observed gain. At 40% coverage it achieves 12.3% conditional net gain, versus 4.0% for entropy and at most 2.8% for the other deployable controls. It places 71% of rescuable errors in its top two quintiles, compared with 29% for entropy. On

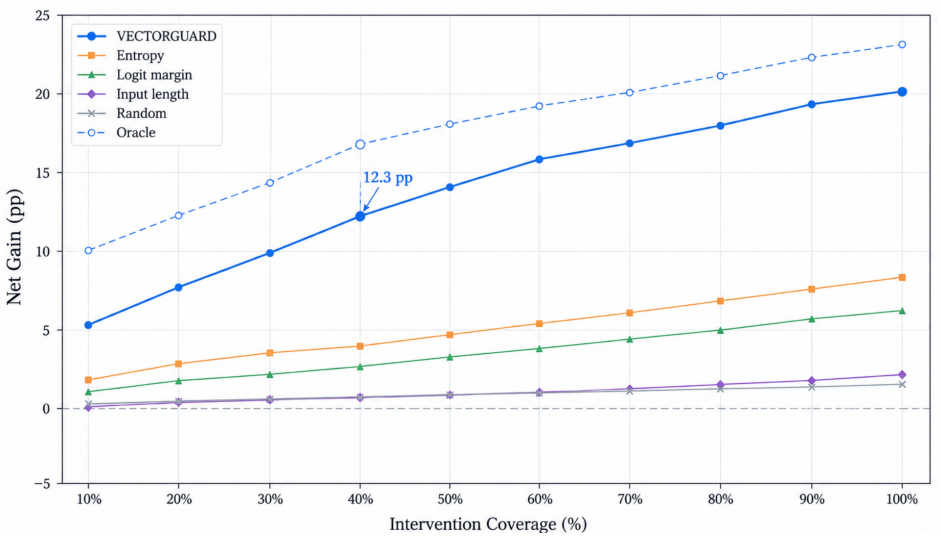


Figure 2: Net gain versus intervention coverage for utility and baseline rankings.

40% coverage	Rescue	Harm	Net
VECTORGUARD score	13.1	0.8	12.3
Entropy	7.0	3.0	4.0
Logit margin	6.1	3.3	2.8
Input length	5.7	5.1	0.6
Random	5.9	5.2	0.7
Oracle	20.5	0.0	20.5

Table 3: Conditional percentages at 40% matched coverage.

low-entropy base errors, removing entropy attenuation raises the rescue rate from 2.7% to 8.0%, showing that uncertainty is neither a reliable benefit score nor a sufficient intervention gate.

4.4 CERTIFICATION

For every threshold we record accepted count, empirical harm, empirical excess loss, both simultaneous upper bounds, and the fallback reason. We compare dual certification with harm-only calibration and an uncalibrated $r(x) \geq 0.5$ rule on the same fixed candidate. Calibration-size curves use prespecified nested subsets; they are not searched to replace a failed primary operating point. The oracle ranking is diagnostic only and never supplies a deployable threshold.

No threshold is feasible with 512, 1024, or 2048 calibration examples because the bounds are too wide or utility evidence is insufficient. At 4096, both pairs satisfy $U_H \leq 0.05$ and $U_\Delta \leq -0.005$, with 45.0% and 40.5% calibration coverage; test intervention counts are 575/1319 and 535/1319. Thus nontrivial certification requires the largest evaluated calibration set.

4.5 TASK SHIFT AND COST

The source-frozen direction, selector, and threshold are applied without target-label tuning. Hence these tasks evaluate transfer and failure modes, not validity of the GSM8K certificate under shift. We separately record format compliance, parse failures, and harmful flips rather than interpreting accuracy alone as robust instruction following.

Both policies improve MATH and StrategyQA slightly, leave AIME unchanged, and regress 0.27 points on ARC, confirming that source certification does not imply shift robustness. Qwen p50/p95 latency changes from 2.84/8.71 to 3.06/9.42 s with 2.1% additional peak memory; the cross-family policy adds 10% mean latency. Latency uses batch size one, synchronized accelerator timing, 20

Pair	t^*	m_t	Coverage	Rescues	Harms	$\hat{\Delta}$	U_H	U_Δ
Qwen–Qwen	0.70	1843	45.0%	208	8	-0.1085	0.0492	-0.0188
DeepSeek–Llama	0.75	1660	40.5%	177	3	-0.1048	0.0491	-0.0102

Table 4: Selected calibration operating points at $n = 4096$, using $c_\alpha = \log(1680)$.

Child	Target	Base	Policy	Gain (pp)	Coverage
Qwen	MATH	31.38	32.08	+0.70	27.9%
	ARC	72.56	72.30	-0.27	24.5%
	AIME	10/30	10/30	0.00	20.0%
	StrategyQA	66.16	67.90	+1.75	31.2%
Llama	MATH	14.50	15.00	+0.50	20.0%
	ARC	61.84	61.57	-0.27	20.0%
	AIME	2/30	2/30	0.00	16.7%
	StrategyQA	58.52	58.95	+0.44	26.0%

Table 5: Source-frozen out-of-domain results; the GSM8K certificate does not apply.

warm-up requests, and 200 fixed test requests across three repetitions; it includes child feature extraction and the second prefill on accepted inputs. We also record p50/p95 latency, generated tokens, throughput, and peak allocated memory. Offline construction costs 71.4 and 79.6 GPU-hours. LoRA and distillation reach 66.6% and 68.1% GSM8K in 23.6 and 73.1 GPU-hours, respectively, so VECTORGUARD is a frozen-weight alternative rather than the highest-accuracy method.

The experimental unit is the input. We use 2000 paired bootstrap resamples for accuracy differences, Wilson intervals for proportions, exact McNemar tests with Holm correction across the two primary pairs, and three prespecified fitting seeds. Counts accompany small-sample results; deterministic decoding reruns are not treated as independent replicates. Confidence intervals and significance tests quantify held-out sampling uncertainty, while the three fitting seeds assess pipeline variability under fixed outer splits. Neither addresses benchmark-to-deployment shift.

5 CONCLUSION

Cross-model steering requires more than a map between representation spaces. VECTORGUARD separates the reliability of a transported direction, the predicted utility of intervening on a particular input, and the statistical evidence required to enable intervention. Its deployment rule can preserve the original child when either transport diagnostics or calibration evidence are insufficient. The formal result establishes a conditional population guarantee under explicit sampling assumptions; practical value depends on coverage, effect size, and cost. Empirically, unconditional transfer produces only small GSM8K gains, whereas utility-based selection retains most rescues while removing most harmful flips, yielding gains of 5.23 and 4.17 percentage points for the two model pairs. Direction diagnostics outperform matched-norm routing and random-direction controls, and the learned selector ranks beneficial interventions more effectively than entropy, margin, length, or random scores. The source-frozen policy yields small gains on MATH and StrategyQA, no change on AIME, and a slight regression on ARC, confirming that source-domain certification does not imply robustness under task shift.

LIMITATIONS

Geometry is not semantics. Support, cycle consistency, bootstrap stability, and amplification are diagnostics, not proofs that reasoning capability transfers. A stable map can preserve the wrong feature. Instruction-contrast directions can encode verbosity, format, or compliance rather than reasoning, which motivates the style controls in our protocol.

Certification may be infeasible. The certificate constrains mean harm and mean excess loss on the accepted population, not every accepted input. It can demand thousands of examples even when a

selector is informative. Always abstaining recovers base predictions but yields no gain. The theorem relies on independent calibration, frozen selection rules, and an unchanged deployment distribution; these conditions are not established merely by using separate benchmark files.

The fallback is behavioral, not computational. A rejected input still pays for feature extraction and selection. An accepted input may require a second prefill. Offline collection of paired outcomes can dominate the cost of map fitting. Keeping model weights fixed does not mean zero fitting cost, zero supervision, or equivalence to training-free methods.

Scope. The specified datasets concern mathematical, scientific, and commonsense question answering. They do not establish general conversational quality, multi-turn instruction following, tool use, multilingual transfer, or adversarial robustness. The two evaluated model pairs are a focused comparison, not a representative census of architectures. A larger parent is not assumed to be a task-wise oracle.

ETHICAL CONSIDERATIONS

Interventions can introduce new mistakes into otherwise correct model outputs. A population-level bound on such changes is not a guarantee of factuality, fairness, or suitability for high-stakes use. Dataset licenses and model access conditions must be respected, and releases must exclude restricted checkpoint weights and private data. We report regressions, parse failures, infeasible certification budgets, and compute expenditure.

AI USE STATEMENT

Generative AI tools were used to edit, condense, and reorganize the manuscript; to draft and refine descriptions and interpretations of the reported experimental results; and to check consistency between numerical summaries and tables. They were not used to generate experimental data or to autonomously conduct experiments. The authors reviewed and verified all AI-assisted text, numerical claims, equations, citations, and conclusions, and take full responsibility for the final content.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://arxiv.org/abs/2110.01052.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535. URL https://arxiv.org/abs/2101.02703.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21/.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.findings-acl.507/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv. Cross-model transferability among large language models on the platonic representations of concepts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3686–3704. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.185. URL https://aclanthology.org/2025.acl-long.185/.

Hugging Face. Math-Verify: A mathematical answer verification system. Software repository, 2025. URL https://github.com/huggingface/Math-Verify.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.03341.

Siddharth Tandon. Can abstract concepts from LLM improve SLM performance?, 2025. URL https://arxiv.org/abs/2512.19069.

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

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5418–5433. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.acl-long.270. URL https://aclanthology.org/2025.acl-long.270/.

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

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A GUARANTEE, EDGE CASES, AND FEATURE CONTRACT

A.1 PROOF OF PROPOSITION 1

Condition on all representation and predictor fitting stages. For a fixed configuration and threshold t, the acceptance event depends on x through a fixed score. Conditional on an accepted count $m_t = m > 0$, the accepted labeled observations are i.i.d. from the population conditional on A_t. For $H \in [0, 1]$, Hoeffding's inequality gives

$$\Pr\left(\mathbb{E}[H \mid A_t] > \hat{h}_t + \sqrt{\frac{c_\alpha}{2m}} \;\middle|\; m_t = m\right) \leq e^{-c_\alpha}. \tag{16}$$

For $\Delta \in [-1, 1]$, whose range has width two, the corresponding inequality is

$$\Pr\left(\mathbb{E}[\Delta \mid A_t] > \hat{\Delta}_t + \sqrt{\frac{2c_\alpha}{m}} \;\middle|\; m_t = m\right) \leq e^{-c_\alpha}. \tag{17}$$

Averaging over $m > 0$ preserves the bounds on failure events with positive counts. Clipping the upper limits at one is valid because both population means are at most one. A union bound over $2K|\mathcal{T}|$ events gives failure probability at most α. Dependence between thresholds, or between configurations evaluated on the same examples, does not invalidate this union bound.

Thus a data-dependent choice among the predeclared feasible thresholds remains valid on the simultaneous event. Outside the chosen acceptance region the deployed policy equals a_0, so the law of total expectation yields Eq. 14. A region with zero population mass has positive calibration count with probability zero and needs no conditional-risk interpretation. If no threshold is feasible, the positive-gain guarantee is not asserted; base equivalence holds by construction.

Sampling scope. The theorem assumes i.i.d. labeled examples, not merely non-overlapping file names. A randomized split of a fixed benchmark is an operational approximation to sampling from a task population. It does not prove that the benchmark represents deployment traffic. Finite-population guarantees for sampling without replacement would require a different formulation. No such guarantee is claimed here.

Randomized decoding. The primary protocol uses greedy decoding. A stochastic extension must fix the coupling between base and candidate random streams and include fresh per-example random variables in the observation model. Sharing one global random stream whose state depends on prior output lengths can violate the intended independent paired-observation construction. Seed-based repetitions do not repair target-distribution shift.

A.2 BINARY LOSS INTERPRETATION

Let $R = \mathbb{I}\{a_0 \text{ wrong}, a_1 \text{ correct}\}$ and $H = \mathbb{I}\{a_0 \text{ correct}, a_1 \text{ wrong}\}$. Then $\Delta = H - R$, and conditional gain is $\mathbb{E}[R - H \mid A_t]$. The harm budget constrains $\mathbb{E}[H \mid A_t]$, whose denominator includes all accepted inputs, not only initially correct inputs. We also report the fraction of initially correct inputs corrupted, but it is a separate metric and is undefined when its denominator is zero.

A harm-only rule can accept a candidate with $R = H = 0$ everywhere. With $\epsilon > 0$, the utility upper bound cannot certify that candidate as beneficial. Conversely, positive net gain can coexist with many harmful flips; the harm constraint remains necessary if those flips have independent operational cost.

A.3 SAMPLE-SIZE BOUNDARY

Equation 15 is necessary for the observed utility bound to pass, not a sufficient guarantee that a sample of that size will pass. Similarly, when $\hat{h}_t < \delta_h$, the harm bound requires

$$m_t \geq \frac{c_\alpha}{2(\delta_h - \hat{h}_t)^2}.$$

When $\hat{h}_t \geq \delta_h$, finite-sample slack prevents certification at that threshold. Small accepted sets face both penalties, even if their empirical outcomes look favorable. For the primary constants, even zero observed harm requires approximately 1486 accepted examples to satisfy the harm bound. Consequently, the nominal 10% coverage target is less restrictive than the primary harm-bound sample requirement with 4096 calibration examples. These numerical consequences must be reported rather than treating budgets as cosmetic hyperparameters.

Tighter binomial or variance-sensitive bounds may improve power, but replacing a bound after seeing which one certifies a preferred result is not covered by this protocol. Such an extension requires an independently frozen analysis or an expanded multiplicity correction.

A.4 Child-Only Features

For each retained layer, estimate the representation-fit mean μ_j and RMS centered norm

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|h_j(x_i) - \mu_j\|_2^2}.$$

For the last prompt token of an unmodified prefill, concatenate:

1. $E(x)/\log|\mathcal{V}_c|$ and the difference between the largest and second-largest raw next-token logits.
2. For each retained layer, $\|h_j(x) - \mu_j\|_2 / \max(\sigma_j, \varepsilon_0)$.
3. The signed cosine $\cos(h_j(x) - \mu_j, d_j)$ for each retained layer; use zero if the centered activation vanishes.
4. The anticipated local ratio $\min\{b_j(x)\|d_j\|_2/(\|h_j(x)\|_2 + \varepsilon_0), \rho_j\}$.

Layer features are ordered by ascending child index. The raw vector has $2 + 3|\mathcal{J}|$ entries. Standardization uses the utility training portion during regularization selection and is recomputed on all $\mathcal{D}_{\text{fit}}$ during final refitting. Remove dimensions with standard deviation below 10^{-8}, then append an intercept. Nonfinite features receive the rejection sentinel $r(x) = -\infty$ during both calibration and deployment, with a logged error; they are not silently imputed.

The norm-based support distance is a coarse novelty indicator, not a calibrated out-of-distribution detector. Pair-level direction reliability determines candidate layers rather than being included as a constant that cannot rank examples within a pair.

A.5 Algorithm and Deployment State

1. **Construct.** Collect paired post-block activations and fixed-sign contrastive differences. Fit the route, forward/reverse maps, bootstrap diagnostics, and candidate hyperparameters using only the two representation splits.
2. **Fit.** Generate paired base/candidate outputs on the utility-fit split. Fit feature preprocessing and the utility regressor; freeze their hashes.
3. **Calibrate.** Collect paired outcomes on the independent calibration split. Evaluate every threshold using Eq. 12; save the complete ledger and the threshold chosen by Eq. 13.
4. **Deploy.** Run one unmodified child prefill and compute $r(x)$. If rejected, continue from that base state. If accepted, discard the unmodified generation state and run a fresh candidate prefill with hooks before decoding. Reusing a base cache as if it already contained interventions is incorrect.

Online state contains child weights, retained directions, feature means/scales, selector weights, the threshold, and candidate scaling constants. It does not need the parent or dense transport maps. The second prefill for accepted examples is part of the measured cost. Rejected predictions match the base only when the baseline cache, precision, tokenizer, and decoding rules are preserved exactly; a zero-intervention equivalence test checks this contract.

B Representation and Intervention Diagnostics

CKA and numerical rank. For centered paired matrices X, Y, use

$$\mathrm{CKA}(X, Y) = \frac{\|X^\top Y\|_F^2}{\|X^\top X\|_F \|Y^\top Y\|_F}.$$

A zero denominator is an invalid layer pair. Compute statistics in at least float32, using float64 for small matrix solves when available. Singular values below 10^{-6} times the largest singular value are numerically unsupported. Choose the smallest rank explaining 95% of nonzero squared singular values, capped at 256. The cap and energy rule are frozen, not tuned against test gains.

Bootstrap failures. Use 64 paired bootstrap replicates, resampling contrastive indices independently from representation indices. A failed route or zero transported direction contributes a zero vector to the bootstrap mean, with the denominator remaining 64. If the original map is invalid, reject the layer regardless of replicate scores. Record both instability and failure frequency. Do not discard failed replicates and thereby inflate agreement.

Shared-prefix distortion. For a fixed audit set, first generate the base response. Replay the same prompt and up to its first 128 generated tokens under base and candidate, forcing the same next-token history. At corresponding token/block positions, record immediate intervention ratio, downstream relative drift

$$D_{j,k} = \frac{\|h_{j,k}^{(1)} - h_{j,k}^{(0)}\|_2}{\|h_{j,k}^{(0)}\|_2 + \varepsilon_0},$$

cosine similarity, and Jensen–Shannon divergence between next-token distributions. Report medians and 95th percentiles, along with free-generation outcome changes. These metrics answer different questions: a small immediate cap does not imply small later drift, and small drift does not establish useful task behavior.

On representation-validation contrasts, also compare the transported parent direction with the independently constructed child contrast direction using signed cosine. This tests correspondence for the chosen contrast, not a universal reasoning direction. Match actual perturbation norms when comparing random or ablated directions, so a lower distortion value is not automatically credited as better transport.

C Execution Specification

This specification defines the next empirical study. It is not a claim that the study has already been run or externally preregistered. Any change after inspecting outcomes must be logged as a new protocol version, with exploratory results separated from the primary analysis.

C.1 Data Allocation and Provenance

Use the local GSM8K, competition MATH, ARC-Challenge, AIME 2025, and StrategyQA loaders in `src/dataset/`, with paths from `src/config/dataset_path.py`. Preserve source IDs and content hashes. Normalize line endings and Unicode whitespace for duplicate detection, but do not alter numbers, symbols, answer options, or question content. Group exact normalized-question duplicates before splitting. Reject overlap between any fit/calibration role and the final test set. Report near-duplicate and likely contamination checks separately; exact matching cannot rule out pretraining contamination.

Order eligible GSM8K training groups by a seeded permutation with seed 20260830. Allocate in order: 512 representation-fit examples, 512 representation-validation examples, 1024 utility-fit examples, and 4096 calibration examples. Use the first 128 representation-fit examples for contrastive prompts. All remaining training examples are unused in the primary run. If deduplication leaves fewer than 6144 eligible examples, stop the primary protocol rather than borrowing from the test set or silently shrinking a role.

The test partitions are GSM8K test, MATH test, ARC-Challenge test, and AIME 2025 sessions I and II together. Use StrategyQA test only if the local copy contains legitimate reference labels; otherwise use a deterministic 20% held-out subset of its labeled training split, seeded identically, and name it *StrategyQA held-out train*, not official test. Because no StrategyQA examples are used for fitting, its remaining training data remain unused. If AIME reference labels cannot be verified against the supplied source, exclude scoring and report the missing provenance rather than infer answers from model output.

Source-domain fitting and certification use GSM8K only. The four target tasks receive the same source-frozen directions, selector, and threshold. A task-specific recalibration is a different experiment and requires a new disjoint labeled calibration split; it cannot be inferred from small target test collections.

C.2 Models, Prompting, and Decoding

Resolve every checkpoint by its exact identifier and record a content revision or local file hash, tokenizer revision, chat-template hash, precision, and model-library version. The primary Qwen identifiers are read from `src/config/model_path.py`; a name/path mismatch is an error, not an alias to accept silently. The DeepSeek–Llama run uses explicitly verified checkpoint paths and intervention hooks. Missing weights or unsupported block hooks produce an unavailable run, not a substitute model.

Use the checkpoint's own chat template with a user message and generation prompt. Use no in-context demonstrations in the primary study. If the template explicitly supports a thinking-mode switch, enable it and record the value; otherwise preserve its native behavior and record that no switch exists. Use greedy decoding, one beam, native EOS tokens, no top-p/temperature sampling, no repetition penalty beyond 1.0, and no external stop string inside reasoning text. Use bfloat16 when supported; otherwise use float32 and report the difference. Quantization is excluded from the primary comparison.

The initial generation cap is 2048 new tokens for every method and task. Before any intervention tuning, audit the unmodified child on $\mathcal{D}_{\mathrm{val}}$. If more than 1% of these outputs reach the cap, use 4096 for that child's entire experiment. If more than 1% still reach the cap, use 8192. Freeze the resulting cap before fitting candidates or inspecting test results. Do not tune task-specific caps on target test labels. Never silently truncate an input prompt; a context-length violation is a logged configuration failure. Native model capacity can prevent a specified run, in which case it must be reported as unavailable.

Answer evaluation. GSM8K uses the reference numeric answer after its final answer delimiter and the model's final boxed numeric answer, normalizing commas and exact rational/decimal representation. MATH uses symbolic equivalence through Math-Verify (Hugging Face, 2025), not only normalized string matching. AIME requires an integer in $[0, 999]$, with leading zeros normalized. ARC uses an explicit final answer-option label mapped to the supplied choices. StrategyQA uses final `yes` or `no`. Use the same parser across all methods, treat missing/invalid answers as incorrect, and report parser failures separately. Freeze evaluator version and timeout policy. A software exception is not automatically a model error; repair the evaluation and rerun all affected methods consistently.

C.3 Prompt and Contrast Contracts

Append the following task instructions to the unchanged question (and ARC choices):

> Solve the problem carefully. Give your final answer in \boxed{...}.

For ARC, append that the boxed content must be the option label; for StrategyQA, specify `yes` or `no`; for AIME, specify an integer from 0 to 999. These task-format instructions are identical for the base and every intervention or adaptation comparator.

For each of the first 128 representation-fit questions, form two parent prompts with identical question and final-answer format, differing only in this added instruction:

> Positive: Before answering, work through the reasoning step by step and check the result.

Component	Fixed value or prespecified selection rule
Routing	$\beta_{\text{jump}} = 0.05$, $\beta_{\text{rep}} = 0.01$; $K_{\text{jump}} = \max(1, \lceil 2\bar{d} \rceil)$; $R_{\max} = \max(2, \lceil 2/\max(\bar{d}, 10^{-6}) \rceil)$, capped at L_c
Ridge maps	Dimensionless grid $a \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$; $\lambda_W = a\|X_p\|_F^2/(Nd_p)$ and corresponding reverse scaling
Support and stability	95% singular energy, rank cap 256; 64 bootstrap replicates; $\varepsilon_0 = 10^{-8}$
Candidate search	$\eta \in \{0, 0.5, 0.75\}$, $\tau \in \{0.5, 1, 2\}$, $\rho_{\text{total}} \in \{0.05, 0.10\}$ (18 settings)
Entropy reference	95th percentile of unmodified first-token entropy on $\mathcal{D}_{\text{rep}}$, floored at 10^{-8}
Utility predictor	Ridge grid $\lambda_g \in \{0.01, 0.1, 1, 10, 100\}$; internal 80/20 fit split; unpenalized intercept
Primary calibration	$n = 4096$, $m_{\min} = 30$, $\alpha = 0.05$, $K = 2$, $\delta_h = 0.05$, $\epsilon = 0.005$, $t \in \{0, 0.05, \ldots, 1\}$
Sensitivity and intervals	Calibration $n \in \{512, 1024, 2048, 4096\}$; fitting seeds 20260830/31/32; 2000 paired bootstrap resamples with seed 20260901

Table 6: Prespecified hyperparameters. Values are design choices rather than empirically selected optima. Every run exports the resolved values, including the selected candidate and generation cap.

Negative: Give the answer directly without intermediate reasoning or a check.

Collect the final prompt-token activation after applying the native template, before generating any answer. No gold rationale or test answer is inserted into these prompts. This construction is a contrast between *instructions about reasoning*, not between verified successful and failed internal reasoning. Its semantic interpretation is therefore tested, not assumed.

Two controls isolate this confound. A style-only contrast replaces the instruction pair with *Use a detailed response* versus *Use a concise response*, while preserving the question and final format. A child-native direction uses the original reasoning-instruction contrast in the child itself, without a parent map. Match layer count and intervention norm to the transferred candidate. A length-matched analysis compares output-length strata but is descriptive because generated length is affected by intervention; it is not treated as a causal adjustment.

C.4 Frozen Hyperparameters and Selection Order

Choose ridge regularization per routed layer pair by mean forward/reverse normalized reconstruction error on $\mathcal{D}_{\text{val}}$, breaking ties in favor of larger regularization. Keep representation-fit centering statistics fixed. Freeze the selected regularization before computing the 64 bootstrap diagnostics; bootstrap refits the route and maps but does not retune regularization. For a bootstrap route that changes the matched parent layer, use the main run's selected dimensionless regularization for that child layer and recompute its source-variance scaling.

Among the 18 candidate settings, retain those with empirical harmful-flip rate at most 0.05 on representation validation. Select highest net accuracy gain; break ties by lower harm, smaller total budget, fewer retained layers, smaller τ, and larger η, in that order. Include the exact base as a candidate with zero gain and zero harm, preferred over zero-gain nonzero interventions. This is a heuristic candidate-design step; its validation threshold is not a certificate. If it selects the base, report a degenerate candidate and do not claim successful transfer.

Fit the utility regressor using an internal split of 819 training and 205 validation examples. Select lowest validation mean-squared error, breaking ties toward larger λ_g, and then refit on all 1024 examples. All feature preprocessing is refitted within the relevant training subset. For the secondary fitting seeds, keep the primary outer data partitions unchanged and vary only bootstrap, internal utility split, and randomized controls. These runs assess fitting variability, not generalization to independently sampled training sets.

C.5 COMPARATOR CONTRACTS

Naive and predecessor steering. Naive transfer uses rounded proportional layer routing $\phi(j) = 1 + \mathrm{round}((j-1)(L_p-1)/\max(1, L_c-1))$, paired linear maps, and a global scalar. The DRSVT predecessor replaces routing with CKA-cost DTW and uses CKA-weighted entropy scaling without direction reliability or input rejection. For DTW, use cost $1 - S_{ij}$, endpoint-to-endpoint paths and horizontal/vertical/diagonal steps; when a child layer aligns to several parent layers, choose the pair with highest CKA, then the smaller parent index. Both use the same contrastive data and injection positions as VECTORGUARD. Global scalar search uses $\{0, 0.125, 0.25, 0.5, 1, 2\}$ on representation validation; report the selected zero-intervention case if it wins.

Prior-work comparators. The controlled linear-alignment comparator uses proportional routing and a forward least-squares map with Moore–Penrose cutoff 10^{-6} relative to the largest singular value. It uses the same contrast direction and scalar grid. This is an *adaptation inspired by* INCLINE/L-Cross, not a claimed faithful replication of their original tasks. A matched-cap version isolates the effect of the local norm cap. Comparisons must not attribute benefits from this cap to the direction reliability score.

SADI is evaluated as a native child intervention using its semantic adaptation rule and official implementation, with the same representation prompts and test decoding contract. It is not renamed entropy scaling. Its native target-component choices and extraction procedure must be exported in the run manifest; if the implementation cannot support a specified child architecture, record it as unsupported rather than silently replacing its rule. Use the public default configuration and a six-value scalar search on representation validation. Only a completed, audited implementation may populate its result row. An additional transferred-direction SADI variant, if implemented, is a separately labeled adaptation.

Selector controls. All selector comparisons operate on the *same* fixed candidate. Entropy score is $E(x)/\log|\mathcal{V}_c|$; margin score is the sigmoid of the negative top-two logit gap; length score is child prompt-token length divided by the maximum on utility-fit data and clipped to one. Seeded random scores are derived from a hash of seed and example ID, independent of labels. Fixed-coverage curves sort these scores on the test set for descriptive analysis only. An uncalibrated utility rule uses $r(x) \geq 0.5$, without claiming a guarantee. Harm-only calibration uses $U_H \leq \delta_h$ and omits the utility constraint, with the same minimum count.

Only the two full-method primary configurations share the stated family-wise certificate. Any joint claim over additional selectors, seeds, calibration sizes, or alternative bounds requires declaring the enlarged family and using its corresponding K before calibration. Secondary curves cannot be searched to replace a failed primary operating point.

Training-based resource references. Use the same 6144 source training questions, with no test access. LoRA SFT uses the supplied GSM8K solution; teacher-rationale distillation uses one greedy parent response per question, filtered by the frozen answer evaluator. Report the retention fraction and parent generation cost. Do not replenish rejected examples with new questions. The 512 representation-validation examples serve as adaptation validation and are excluded from gradient updates; the other 5632 examples are eligible training data. Thus both references have the same source-information pool as the steering pipeline, but their supervision and compute differ.

Use rank 16, scaling 32, dropout 0.05, AdamW, learning rate 2×10^{-4}, effective batch size 32, three epochs, 3% linear warmup followed by cosine decay, weight decay 0, and gradient clipping at 1.0. Apply adapters to eligible two-dimensional linear projections inside decoder blocks, excluding embeddings and the output head, and publish the exact module list and parameter count. Mask prompt tokens in the loss. Teacher generation has a cap of 8192 new tokens. The training sequence limit is the smaller of 8192 and the child's native context limit; exclude overlength examples and report their count instead of silently truncating the supervised answer. Use the original labeled validation solutions for epoch selection, independent of teacher filtering. Select the epoch with highest validation accuracy, then the earlier epoch. Keep the same final-answer format and evaluation budget. Any further tuning is a separately costed experiment. These references are empirical alternatives, not guaranteed upper bounds.

C.6 Metrics, Failure Cases, and Cost Records

In addition to Table 2, export the certificate ledger columns: configuration ID, threshold, accepted count, calibration coverage, empirical harm, empirical excess loss, U_H, U_Δ, feasibility, selected flag, test coverage, test conditional harm, and test conditional gain. Use `undefined`, not zero, for conditional means of empty sets.

For the confidence failure analysis, define low entropy as at or below the 25th percentile on utility-fit inputs, before observing test outcomes. Tabulate all four paired correctness transitions inside this stratum and its complement. Report bins of input length and support distance using fit-set quartiles. Qualitative examples are chosen deterministically: first five harmful flips by hashed example ID per pair, first five rescues, and first five confident base errors, with overlap flagged. Include question, reference, base/candidate answer, score, threshold, and acceptance decision. These examples illustrate mechanisms but do not estimate their frequency.

For latency, use batch size one, 20 warm-up requests and 200 fixed source-test requests in hashed order. Synchronize accelerator measurements at boundaries. Report hardware, precision, library versions, p50/p95 latency, total generated tokens, tokens per second, peak allocated memory, and policy coverage. Include all feature extraction and accepted re-prefill time. Repeat timing three times; do not exclude slow accepted examples. Separately report offline GPU-hours, parent/child generated tokens, bootstrap wall time, data storage, and fitting memory. Without measured savings at the chosen request volume, no efficiency superiority is claimed.

C.7 Reproducibility and Artifact Interfaces

Each stage has an explicit artifact interface:

1. **Baseline:** model/data/prompt configuration in; per-example prompt, raw response, extracted answer, correctness, token count, stop reason, and timing out.
2. **Transport:** paired representation IDs and contrast contract in; route, maps, singular spectra, reliability components, retained directions, budgets, and construction hashes out.
3. **Paired outcomes:** frozen candidate plus labeled split in; base/candidate response records joined by immutable example ID, with Δ and H, out.
4. **Selector:** utility-fit features and paired outcomes in; preprocessing, coefficients, validation record, and immutable score definition out.
5. **Certificate:** frozen artifacts plus calibration records in; full threshold ledger, selected policy, and fallback reason out.
6. **Evaluation:** frozen policy and held-out split in; paired statistics, robustness diagnostics, error examples, and cost ledger out.

Verify zero-intervention equivalence, paired rows, split disjointness, signs, bounds, fallback, and caches. Archive configurations and hashes; exclude smoke tests, synthetic fixtures, and historical scores from reported performance.