BEYOND TRAINING SUPPORT: PARAMETERIZATION SELECTS FINE-TUNING ENDPOINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning on a narrow support often does not determine a unique model function. Many updates can match all training predictions yet behave differently outside the observed support. Optimization therefore does more than fit the data: it also selects *how the model extends beyond the training set*. We formalize this choice through a parameterization-conditioned endpoint selector. Specifically, for linear adaptation coordinates, the selected endpoint is an explicit metric-minimum right inverse. For balanced signed-square coordinates, it is a hypentropy-minimum interpolant, with initialization scale controlling a transition between sparse and Euclidean regimes. We then construct parameterizations that produce identical training-prediction trajectories at every step but converge to different endpoints. Next, we derive an analytic boundary for when their held-out ordering reverses. Conversely, we characterize when two selectors agree on a fixed held-out witness and show that, once the training support identifies that witness, all interpolating endpoints must agree. Finally, experiments in solvable systems, frozen representations up to 30B parameters, and matched-fit Transformer adaptations separate endpoint selection from expressivity, convergence, and learning amount. Together, the theory and experiments explain why the same narrow-domain fit can lead to different functions, and when additional support removes that freedom.

1 INTRODUCTION

A fine-tuning dataset constrains a model only on the support it observes. Even perfect training fit can leave many possible behaviors elsewhere. Fig. 1(a) shows a simple example: a semantic rule and a marker rule agree on every training example but disagree on a held-out input. Therefore, training identifies a set of compatible solutions rather than a unique function. In turn, optimization must not only fit the data, but also choose how the model extends beyond it.

This choice matters for parameter-efficient adaptation. Direct updates, low-rank factors, rank-stabilized factors, and magnitude–direction coordinates can all fit the same training predictions while following different optimization geometries, as Fig. 1(b) illustrated. Expressivity determines which solutions are available, while the parameterization can influence which available solution is reached.

A difference in transfer behavior, however, does not by itself establish such endpoint selection. One method may have lower capacity, converge more slowly, or simply fit the training predictions less well. These possibilities complicate comparisons among direct updates, LoRA, rsLoRA, and DoRA (Hu et al., 2022; Kalajdzievski, 2023; Liu et al., 2024). We therefore ask a more controlled question: *after fixing what is learned on the observed support, can the optimization coordinates still determine behavior outside that support?* Answering this question requires separating endpoint selection from expressivity and incomplete optimization.

We formalize this idea by fixing a base model, a training support, target predictions on that support, and an adaptation parameterization. The *endpoint selector* S_ϕ is the limiting model reached under parameterization ϕ. To compare selectors, we also fix a held-out functional h_B before training. This functional is used only for evaluation and never for checkpoint selection. Generalization outside the training support can then be studied as a solution-selection problem.

The theory describes both when parameterization matters and when it cannot matter. In underdetermined settings, different adaptation coordinates can select different right inverses of the same

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.

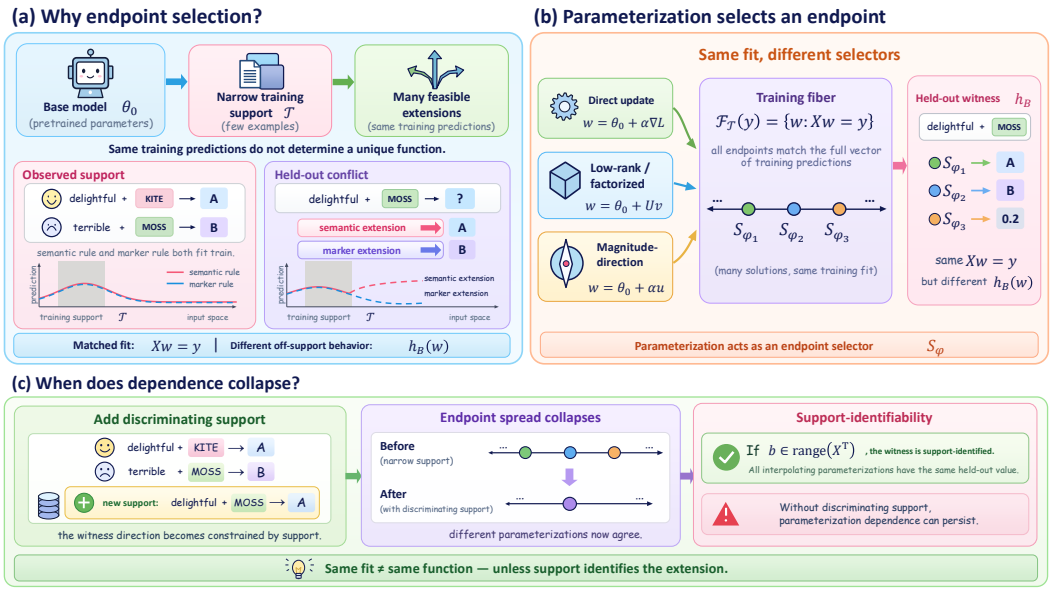

Figure 1: Fine-tuning as parameterization-conditioned endpoint selection: (a) shows why narrow-support fit leaves multiple off-support extensions; (b) shows adaptation parameterizations selecting different elements of the same training fiber, distinguished by a fixed held-out witness; and (c) shows discriminating support making that witness identifiable and collapsing endpoint variation.

training constraints. We derive explicit selectors for linear and factorized coordinates, as well as an analytic boundary at which their held-out ordering reverses. More strongly, two parameterizations can produce identical training predictions throughout optimization and still converge to different held-out endpoints. The complete training trajectory can therefore fail to reveal how the model extends beyond the observed support.

The same framework also identifies when this freedom disappears. If the training support constrains the held-out direction, then every interpolating endpoint must agree on that functional. Adding a small amount of discriminating support can therefore collapse parameterization dependence, as illustrated in Fig. 1(c). Endpoint separation and endpoint agreement are thus two sides of the same support geometry.

We test these predictions in progressively less controlled settings. First, solvable systems expose the selector phase diagram. Next, anchored Transformer representations reproduce the exact ambiguity and its collapse under added support. Natural Transformer features test how much of this geometry remains without enforcing exact equivalence. Finally, end-to-end direct, LoRA, rsLoRA, and DoRA adaptations test whether matched-fit parameterization effects persist during neural fine-tuning. The model suite ranges from sub-billion checkpoints to a 30B-plus mixture-of-experts model. Across all comparisons, the held-out functional is fixed before training and never used for checkpoint selection.

Our contributions are:

- We formulate narrow-support fine-tuning as parameterization-conditioned endpoint selection and derive exact metric and hypentropy selectors for matched-fit solution sets.
- We prove identical-trajectory endpoint separation, analytic endpoint-ordering boundaries, witness-specific selector equivalence, and support-identifiability collapse.
- We test these predictions in solvable systems, frozen Transformer representations, and matched-fit end-to-end adaptations through 30B-plus scale.

2 Related Work

Implicit bias and factorized optimization. Gradient flow can select different interpolants as architecture, depth, initialization, and parameterization change. Prior analyses connect these choices to weighted norms, max-margin solutions, kernel-to-rich transitions, and greedy low-rank trajectories

(Yun et al., 2021; Woodworth et al., 2020; Li et al., 2021; Gunasekar et al., 2017). We shift the focus to a fixed support-external functional of a pretrained adaptation endpoint.

Low-rank adaptation and parameterization. LoRA uses learned low-rank factors, rsLoRA changes rank scaling, and DoRA separates magnitude from direction (Hu et al., 2022; Kalajdzievski, 2023; Liu et al., 2024). Expressivity theory distinguishes representable target functions from those reached by gradient methods (Zeng & Lee, 2024; Jang et al., 2024). These methods motivate controlled parameterization interventions, but target-task accuracy alone cannot identify which off-support interpolant was selected.

Fine-tuning generalization and retention. Narrow updates can distort pretrained features, trade target learning against retention, or trigger behavioral changes (Kumar et al., 2022; Biderman et al., 2024; Betley et al., 2025; Soligo et al., 2026). Other work reports dependence on LoRA rank and shared context (Nief et al., 2026). We isolate a narrower question: conditional on matched training predictions and sufficient capacity, when do optimization coordinates determine the extension?

Linearization and adaptation geometry. Kernel and tangent-space views describe local fine-tuning through training and train-to-test geometry (Jacot et al., 2018; Lee et al., 2019; Malladi et al., 2023). Our metric selector recovers this local case, while the factorized selector shows why the training Gram alone need not determine the selected endpoint once the parameterization changes.

3 Endpoint Selection on a Training Fiber

Let $\theta_0 \in \mathbb{R}^d$ be a base parameter vector and let $f_T : \mathbb{R}^d \to \mathbb{R}^n$ collect predictions on a declared narrow training support. For target predictions z_T, define the feasible endpoint set

$$\mathcal{F}_T(z_T) = \{\theta \in \mathbb{R}^d : f_T(\theta) = z_T\}. \tag{1}$$

An adaptation parameterization is a differentiable map $g_\phi : \mathcal{U}_\phi \to \mathbb{R}^d$ with a common initial function $g_\phi(u_{\phi,0}) = \theta_0$. Gradient flow in u follows

$$\dot{u}_{\phi,t} = -\nabla_u \ell_T(f_T(g_\phi(u_{\phi,t})), z_T). \tag{2}$$

Definition 3.1 (Endpoint selector)**.** When the limit exists and interpolates the target, the parameterization-conditioned endpoint selector is

$$S_\phi(z_T; \theta_0) = \lim_{t\to\infty} g_\phi(u_{\phi,t}) \in \mathcal{F}_T(z_T). \tag{3}$$

To compare these endpoints, a fixed linear functional $h_B(\theta) = b^\top(\theta - \theta_0)$ distinguishes endpoints. In a nonlinear model it represents a predeclared scalar logit margin or task contrast evaluated outside the training support. It never changes the training objective or the feasible set.

Matched-fit requirement. A comparison concerns exact endpoint selection only after all arms can express z_T and attain the same training predictions. The solvable and anchored carriers retain that entire vector. The end-to-end nonlinear panel reports training RMSE, its pairwise mismatch, and the full observed fit trajectory, so its attribution is graded by residual fit rather than asserted from a scalar pass criterion. This separates capacity failure, finite optimization error, and endpoint differences. Appendix A lists notation, and Appendix F extends the local analysis beyond an exactly linear training map.

Accordingly, for the main derivations, write the local update as $w = \theta - \theta_0$, the training map as $Xw = y$, and the witness as $b^\top w$. The fiber is affine:

$$\mathcal{W}_T(y) = \{w \in \mathbb{R}^d : Xw = y\} = w_\star + \mathrm{null}(X). \tag{4}$$

The problem is nontrivial exactly when the witness sees a null direction of X. We begin with linear adaptation charts, where the endpoint selector admits a closed-form characterization.

4 Linear Charts Select Metric Right Inverses

Consider a linear adaptation chart $w = A_\phi u$ initialized at $u = 0$, with squared training loss

$$L(u) = \frac{1}{2}\,\|XA_\phi u - y\|_2^2\,. \tag{5}$$

The chart induces the positive semidefinite parameter-space metric $M_\phi = A_\phi A_\phi^\top$. As the following selector shows, this metric determines both the optimization rate and the selected endpoint.

Theorem 4.1 (Metric endpoint selector). *Assume $y \in \mathrm{range}(XA_\phi)$. Gradient flow on Equation 5 converges in prediction and selects*

$$S_{M_\phi}(y) = M_\phi X^\top (X M_\phi X^\top)^\dagger y. \tag{6}$$

Among feasible updates in $\mathrm{range}(M_\phi)$, this is the minimum $M_\phi^\dagger$-seminorm interpolant.

The proof in Appendix B solves the prediction-space flow and then lifts its limit through the chart. Equation 6 makes the source of parameterization dependence explicit: different charts implement different right inverses of the same training operator.

Corollary 4.2 (Matched fit does not determine the witness). *For two expressive metrics M_1, M_2, both endpoints satisfy $XS_{M_j}(y) = y$, while their held-out difference is*

$$\Delta_B(M_1, M_2; y) = b^\top \left[M_1 X^\top (X M_1 X^\top)^\dagger - M_2 X^\top (X M_2 X^\top)^\dagger \right] y. \tag{7}$$

This contrast conditions on the complete vector of training predictions. It cannot be attributed to a difference in scalar training loss.

Proposition 4.3 (Learning-rate and time rescaling). *For every scalar $c > 0$, $S_{cM}(y) = S_M(y)$. Multiplying a chart by a scalar changes the rate of prediction-space flow but not its interpolating endpoint.*

Proposition 4.3 motivates effective-step calibration. If two methods differ only by a global update scale, their learning curves can be aligned by time. An endpoint difference that remains after convergence requires anisotropy, a nonlinear chart, or an optimization path that changes the induced metric.

The selector also persists under any convergent sequence of full-batch or minibatch steps in the fixed chart, independently of step order. Proposition B.1 and its proof appear in Appendix B.

Theorem 4.4 (Identical training trajectories, different endpoints). *For every $d > n$, there exist a full-row-rank training map X, a target, and two positive definite metrics M_+ and M_- with*

$$X M_+ X^\top = X M_- X^\top \tag{8}$$

such that their gradient flows have identical training predictions at every time but a fixed support-external witness assigns different endpoint values. Consequently, no estimator that observes only the complete training prediction trajectory can determine the selected held-out endpoint without information about the parameterization.

This separation is stronger than matched final loss. It preserves the full vector-valued training trajectory, including its convergence rate, while changing only the selected representative of the endpoint fiber. The constructive proof and its minimax consequence appear in Appendix E.

Theorem 4.5 (Witness-specific selector equivalence). *Let $R_M = M X^\top (X M X^\top)^\dagger$. Two metrics are indistinguishable to witness b for every attainable target if and only if*

$$(R_{M_1} - R_{M_2})^\top b \perp \mathcal{Y}. \tag{9}$$

Here $\mathcal{Y} = \mathrm{range}(X M_1 X^\top) \cap \mathrm{range}(X M_2 X^\top)$ is the common attainable prediction space. They select the same parameter endpoint for every target if and only if their right inverses agree on the attainable prediction range.

The characterization is useful in both directions. A positive result can be predicted from a nonzero witness projection before fitting held-out labels. A null result becomes an invariance statement once Equation 9 holds. Appendix E gives equivalent projector and quotient-space forms. Appendix B gives the finite-time ordering certificate in Proposition B.2, separating endpoint ordering from residual optimization error without an empirical convergence threshold. Linear charts therefore select through a quadratic geometry. We next ask how factorized coordinates change the selected complexity.

5 Factor Coordinates Interpolate Between Complexities

Whereas linear charts induce quadratic complexity, factorized coordinates produce a different family even when they can express every vector. Consider the signed-square chart

$$w_i = p_i^2 - q_i^2, \qquad p_i(0) = q_i(0) = \alpha_i > 0, \tag{10}$$

optimized by gradient flow on $\frac{1}{2}\|Xw - y\|_2^2$. This diagonal carrier isolates parameterization from capacity because every signed vector is representable.

Define the coordinatewise hypentropy potential (Ghai et al., 2020)

$$\Psi_\alpha(w) = \sum_{i=1}^{d} \psi_{\alpha_i}(w_i), \tag{11}$$

where the scalar potential is

$$\psi_{\alpha_i}(w_i) = w_i \operatorname{arsinh}\left(\frac{w_i}{2\alpha_i^2}\right) - \sqrt{w_i^2 + 4\alpha_i^4} + 2\alpha_i^2. \tag{12}$$

Theorem 5.1 (Factorized endpoint selector)**.** *Assume the signed-square flow converges to an interpolating endpoint and its integrated residual is finite. The selected update is the unique solution of*

$$S_\alpha(y) = \arg\min_{Xw=y} \Psi_\alpha(w). \tag{13}$$

Equivalently, there is a dual vector λ such that

$$w_i = 2\alpha_i^2 \sinh((X^\top \lambda)_i), \qquad Xw = y. \tag{14}$$

The product $p_i q_i = \alpha_i^2$ is conserved by the flow. This invariant turns the nonlinear trajectory into the dual representation in Equation 14. Appendix C proves the result and states sufficient conditions for convergence.

Theorem 5.2 (Initialization-scale interpolation)**.** *Let $\alpha_i = \tau a_i$ with fixed $a_i > 0$. Uniformly on bounded subsets of $\mathbb{R}^d$, the normalized potential has the limits*

$$\lim_{\tau \downarrow 0} \frac{\Psi_{\tau a}(w)}{\log(1/\tau)} = 2\|w\|_1, \tag{15}$$

$$\lim_{\tau \uparrow \infty} 4\tau^2 \Psi_{\tau a}(w) = \sum_{i=1}^{d} \frac{w_i^2}{a_i^2}. \tag{16}$$

Consequently, every bounded accumulation point of the small-initialization minimizers is an ℓ_1-minimum interpolant, while every bounded accumulation point of the large-initialization minimizers is the weighted Euclidean minimum interpolant.

Thus initialization scale and balance index an ordered continuum of selectors without changing the realizable training functions. This distinguishes a factorization phase from a nominal rank sweep.

For a linear map with training activations H, a factor rank of $\operatorname{rank}(H)$ suffices to reproduce any direct increment on that support (Lemma C.1, Appendix C). Support-level expressivity therefore need not determine the extension away from those activations.

Connection to LoRA. A LoRA update BA is a matrix factorization rather than a coordinatewise signed square. General matrix sensing need not minimize the nuclear norm, and greedy low-rank trajectories can arise at small initialization (Li et al., 2021). We therefore use Theorem 5.1 as an exactly solvable carrier, then test LoRA, rsLoRA, and DoRA as separate neural parameterizations. Their endpoints are measured rather than assigned the hypentropy formula.

For a nonlinear chart, the induced metric changes along the optimization path. Appendix F formalizes this path-dependent geometry and bounds its local contribution to endpoint selection. With these selector families in hand, we next ask when their held-out ordering changes and when it must disappear.

6 Endpoint-Ordering and Independence Boundaries

We now expose the selector in a two-route carrier. Let the broad coordinate u and narrow coordinate v have identical training action up to gains $a, b > 0$:

$$au + bv = y. \tag{17}$$

The held-out witness assigns coefficients c, d, so $h_B(u, v) = cu + dv$. The interesting case has opposite held-out effects, such as $c > 0 > d$.

Theorem 6.1 (Metric phase boundary)**.** *For the metric $M = \mathrm{diag}(m_u, m_v)$, the selected endpoint and witness are*

$$S_M(y) = \frac{y}{m_u a^2 + m_v b^2}(m_u a, m_v b), \tag{18}$$

$$h_B(S_M(y)) = \frac{y(cm_u a + dm_v b)}{m_u a^2 + m_v b^2}. \tag{19}$$

For $y > 0$, the endpoint ordering changes sign exactly at

$$\frac{m_u}{m_v} = -\frac{db}{ca}. \tag{20}$$

Thus a parameterization can reverse the training-external function while preserving exact training fit. The boundary moves predictably with the induced coordinate balance and feature gains.

Theorem 6.2 (Factorized phase boundary)**.** *For signed-square initialization (α_u, α_v), let λ_y be the unique scalar satisfying*

$$2a\alpha_u^2 \sinh(a\lambda_y) + 2b\alpha_v^2 \sinh(b\lambda_y) = y. \tag{21}$$

The witness changes sign when

$$c\alpha_u^2 \sinh(a\lambda_y) + d\alpha_v^2 \sinh(b\lambda_y) = 0. \tag{22}$$

In the large-initialization limit this reduces to Equation 20; in the infinitesimal limit the coordinate with the larger gain-adjusted exponential term dominates.

Base alignment translates this boundary rather than removing it. If both parameterizations begin from the same effective update w_0, replace y by $y - Xw_0$ in the selector and add $h_B(w_0)$ to the witness. Appendix D derives the resulting boundary and a regularized variant.

6.1 When parameterization cannot matter

Conversely, the same framework identifies when parameterization cannot affect the witness.

Theorem 6.3 (Support-identifiability collapse)**.** *The held-out functional $b^\top w$ is constant over every nonempty training fiber $\{w : Xw = y\}$ if and only if*

$$b \in \mathrm{range}(X^\top), \tag{23}$$

equivalently $\mathrm{null}(X) \subseteq \mathrm{null}(b^\top)$*. Under this condition, every interpolating parameterization selects the same witness value.*

Adding a support example appends a row to X. Parameterization differences vanish once the enlarged row space contains the witness direction. This yields a direct experimental prediction: differences should be largest when broad and narrow routes are observationally equivalent, then collapse as a small amount of discriminating support removes the witness-visible null direction.

Corollary 6.4 (Exact two-route collapse)**.** *Augment Equation 17 with any independent constraint $a'u + b'v = y'$ satisfying $ab' - a'b \neq 0$. The feasible endpoint is unique, hence direct, metric, and factorized interpolating flows agree on every held-out functional.*

The positive and independence regimes are therefore parts of one phase portrait. The theory predicts both endpoint separation and its disappearance from the same support geometry.

7 Experimental Design

The experiments follow the logical order of the theory, with each stage assigned a distinct attribution role.

Solvable carrier. We generate underdetermined systems with a broad feature, a narrow shortcut feature, and nuisance directions. Their training correlation ranges from exact equality to independence. We cross direct, anisotropic metric, and signed-square selectors with feature gain, initialization balance, initialization scale, base alignment, dimension, sample size, observation noise, and seed. Every reported endpoint includes its complete training residual and held-out witness.

Anchored Transformer carrier. To test the same geometry in neural features, a frozen Transformer supplies representations for arithmetic truth, lexical category, and sentiment prompts. Independently calibrated semantic and shortcut directions define a two-coordinate adaptation problem. On the ambiguous training support the two coordinates are made observationally equal, while the fixed conflict witness distinguishes them. This carrier tests the exact phase boundary in neural features without changing capacity across parameterizations.

Natural representation carrier. We next fit actual target margins on the unmodified hidden vectors. Semantic and marker cues agree in the narrow training set and conflict in held-out examples. Direct and signed-square coordinates see identical examples and target vectors. We vary initialization scale and rotate equivalent factor coordinates, retaining every fit residual. The witness is the mean signed semantic margin on a held-out conflict split frozen before fitting.

End-to-end adaptation. Finally, we train direct additive updates and LoRA, rsLoRA, and DoRA factors on the same q/v projection family. The optimizer is SGD without momentum. Rank is kept above the finite-support expressivity scale, and weak last-layer updates are separated from the all-layer comparison. Rather than equating nominal learning rates, a training-only calibration aligns densely recorded margin-RMSE trajectories. Checkpoints at steps 0, 1, 2, 4, 8, 16, 32, 64, 128, and 160 support continuous matched-fit comparisons. The pairwise target RMSE is the larger of the two minimum recorded RMSEs; each arm contributes its nearest checkpoint, and we report the remaining mismatch (Equation 97).

Models and controls. The representation panel covers Qwen2.5 and Qwen3 checkpoints (Yang et al., 2024; 2025) from sub-billion scale through a 30B mixture-of-experts model. The 1,085 end-to-end runs cover 4B, 7B, 8B, 14B, and 30B-scale checkpoints. Controls include support coverage, effective-step calibration, factor scaling, magnitude-direction coordinates, random coordinate rotations, base alignment, domain replication, and seed replication. Appendices G, H, and I give the three carrier protocols; Appendix K gives the complete prompt and target schema.

Statistical unit. The independent unit is a dataset seed within a model and parameterization. Examples from one trained endpoint are not treated as independent replicates. Contrasts pair methods on the same model, seed, support, and target. We report continuous fit mismatch and seed intervals, with no manually chosen effect threshold. Appendix J defines the estimands and resampling scheme.

Finite-time interpretation. For a fixed linear metric flow, the endpoint error is bounded by $\|R_M\|_{\mathrm{op}} \|Xw(t) - y\|_2$. Proposition B.2 guarantees the limiting ordering when the witness gap exceeds the sum of the two witness-error bounds. Thus small training residuals alone are insufficient: selector conditioning also matters. Appendix B gives the certificate and proof.

8 Results

We first test exact selector separation, then examine its persistence and collapse in neural representations and end-to-end adaptation.

Observation 1: Exact carriers separate endpoints at identical training fit. Across two independent 256-seed panels, all compared selectors interpolate the same training targets while selecting different null-space components. In the exactly ambiguous two-route carrier, changing factor balance moves the fixed witness from -0.888 to 0.888, with maximum training relative error below 1.52e-08. As support correlation decreases, the witness-visible null component contracts and the methods converge. Fig. 2 shows the phase and identifiability paths; Appendix G specifies the complete crossed design.

Observation 2: Discriminating support collapses anchored endpoint variation. The anchored neural carrier gives an exact matched-fit test in pretrained representations. At zero discriminating support, the direct selector splits the two routes while metric and factor coordinates move continuously across both endpoint orderings. Once the support identifies the semantic route, the parameterization spread contracts to numerical scale 3.11e-10. Fig. 3 is the neural counterpart of Theorems 6.1 and 6.3.

Observation 3: Natural features make selection depend on the pretrained representation. The natural representation carrier removes the enforced training equality. Parameterization effects persist as continuous endpoint displacements, but their magnitude decreases sharply after the first conflict-support additions. Initialization scale approaches the direct endpoint in its Euclidean regime, and

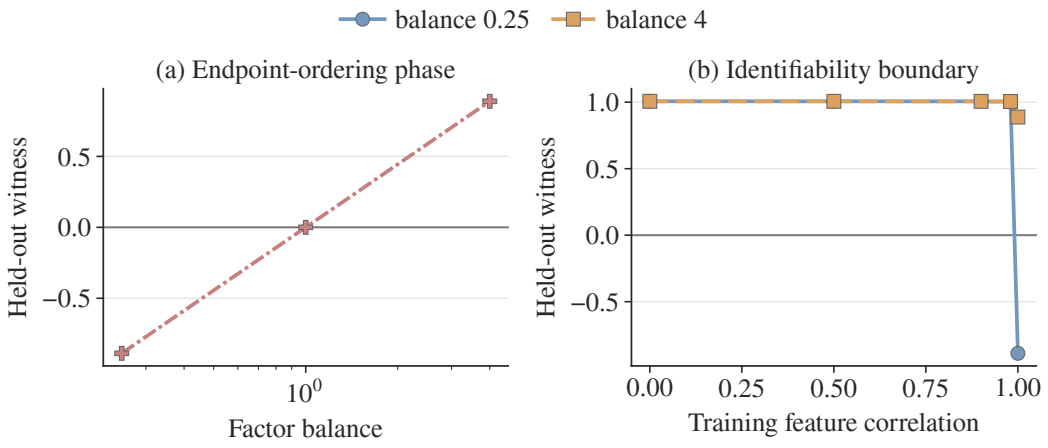

Figure 2: Exact parameterization-conditioned endpoint selection. Panel (a) changes only factor initialization balance on an identical training fiber. Panel (b) reduces broad-shortcut correlation, removing the witness-visible ambiguity.

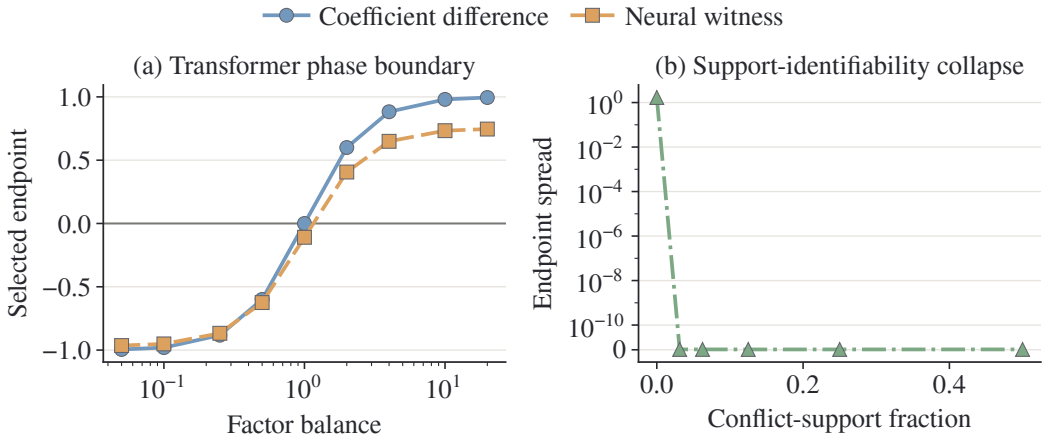

Figure 3: The exact phase boundary in pretrained Transformer coordinates. Panel (a) varies factor balance with zero discriminating support. Panel (b) adds conflict-support examples and measures the resulting endpoint spread.

coordinate rotation changes the small-initialization factor endpoint while leaving the direct solution invariant. These results separate a chart-dependent selector from a global learning-rate effect.

Together, the two neural carriers distinguish the existence of exact separation from its prevalence in unmodified representations. The anchored construction realizes both endpoint signs for the same complete training vector in every model. The unmodified carrier retains the same ordering pressure but lets representation geometry favor one route before adaptation. Its direct endpoint is shortcut-aligned for most Qwen3 seeds and semantic-aligned for Qwen2.5-7B-Instruct. Parameterization therefore changes a base-dependent selection problem rather than imposing one universal semantic direction.

Observation 4: Matched-fit neural updates retain parameterization differences. The training-only calibration finds method-specific learning rates that align effective margin change. On Qwen3-4B sentiment, pairwise matched LoRA and DoRA witnesses differ from direct by -1.195 and -1.659, with seed-bootstrap intervals [-1.569, -0.574] and [-1.936, -1.430]. At 25% conflict support the LoRA difference contracts to -0.027. The same model's arithmetic carrier instead gives a LoRA difference of only -0.082, locating a domain-specific independence regime. Fig. 4 follows each method through training-fit space and applies the pairwise rule in Equation 97.

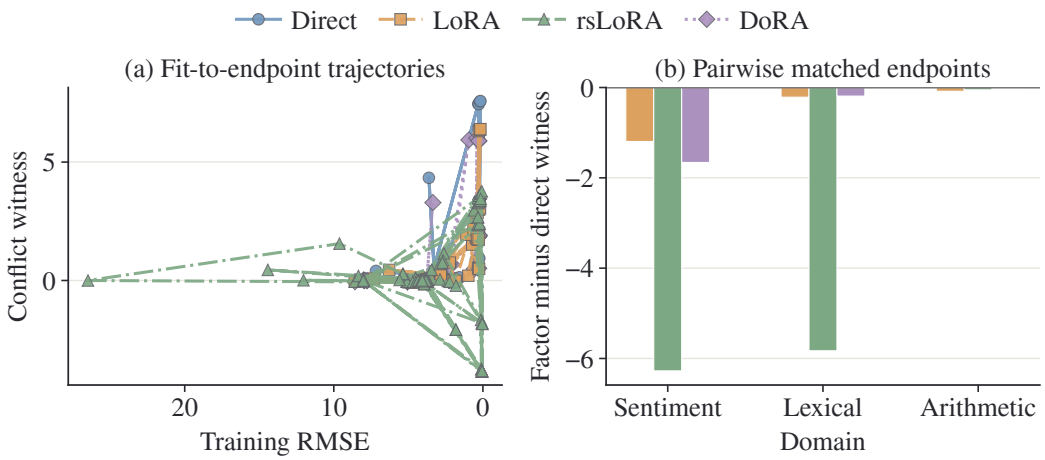

Figure 4: End-to-end Qwen3-4B endpoint selection. Panel (a) plots held-out witness against training RMSE, so nominal learning-rate differences cannot create horizontal alignment. Panel (b) compares parameterizations at training-only pairwise-matched checkpoints.

Consistent with these paired estimates, the endpoint effect is not reducible to a final-loss contrast. Direct and LoRA on Qwen3-4B sentiment have mean pairwise RMSE mismatch below 0.03, yet all four dataset seeds place the LoRA witness below its matched direct witness. DoRA has the same sign in every seed. On arithmetic, the same methods reach comparable training fit and the same negative endpoint sign. A parameterization intervention can therefore matter strongly in one support geometry and become witness-equivalent in another without changing model, optimizer family, or sample count.

Observation 5: The separation persists at 30B scale with residual-fit qualifications. At 30B-plus scale, sentiment repeats the matched-fit separation across 3 seeds: LoRA differs by -0.603 with interval [-1.079, -0.350], and DoRA by -0.691. Their mean fit mismatches are 0.071 and 0.031, while 25% support contracts the LoRA difference to -0.266. Appendix L gives the complementary 8B and Qwen2.5-Math-7B boundaries.

Observation 6: Rank, target scale, and decay locate distinct boundaries. Fig. 6 uses per-run best-fit checkpoints, unlike the pairwise matching in Observation 4. On Qwen3-4B sentiment at zero conflict coverage, increasing LoRA rank from 4 to 128 reduces training RMSE from 0.514 to 0.195 while the witness remains positive (2.139 to 2.854). In contrast, rsLoRA changes the witness from 3.774 at rank 32 to -3.893 at rank 64 even as RMSE improves from 0.082 to 0.046; at rank 128, RMSE rises to 2.049. Thus endpoint reversal and poor fit are distinct outcomes of the rank intervention. Target magnitude moves the normalized witness, while the tested decay range has a smaller effect. Support coverage remains decisive: two conflict examples force exact agreement in the anchored carrier, and the neural carriers contract more gradually as predicted by Theorem 6.3. Appendix M reports every phase cell, and Appendix N resolves the anchored and natural results by model, seed, domain, coverage, and parameterization.

9 Conclusion

Narrow-support fine-tuning can admit models that match every training prediction yet differ elsewhere. We show that parameterization can select among these extensions: linear coordinates induce metric selectors, factorized coordinates induce different complexities, and identical training-prediction trajectories can still end at different held-out solutions. This dependence disappears once the training support identifies the held-out functional, in which case all interpolating endpoints agree. Experiments from solvable systems to Transformer adaptations through 30B-plus scale trace both regimes while separating endpoint selection from capacity and incomplete optimization. These results suggest that fine-tuning generalization should be compared conditional on matched training predictions and residual fit. Parameterization matters only where the observed support leaves the extension unresolved.

AI USAGE STATEMENT

We use generative AI to assist with language editing, manuscript organization, code development and debugging, figure presentation, and checks of mathematical consistency. The authors develop the research questions, theoretical arguments, and experimental design, and independently evaluate suggestions concerning equations and implementation. We verify AI-assisted code against the intended computations and check reported measurements against the recorded experimental outputs. Generative AI does not supply experimental evidence or determine our scientific conclusions; model outputs studied in the experiments follow the protocols described in the paper. The authors review and revise all assisted material and retain full responsibility for the final text, mathematics, analyses, results, and released artifacts.

REFERENCES

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*, 2025.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024.

Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated gradient meets gradient descent. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pp. 386–407, 2020.

Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, 2017.

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.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Uijeong Jang, Jason D. Lee, and Ernest K. Ryu. LoRA training in the NTK regime has no spurious local minima. In *International Conference on Machine Learning*, 2024.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*, 2024.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, 2023.

Todd Nief, Harvey Yiyun Fu, Mark Muchane, and Ari Holtzman. Subliminal learning is a LoRA artifact. *arXiv preprint arXiv:2606.00831*, 2026.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Emergent misalignment is easy, narrow misalignment is hard. In *International Conference on Learning Representations*, 2026.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of the Thirty-Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673, 2020.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2021.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *International Conference on Learning Representations*, 2024.

A Notation and Logical Roles

Table 1 separates quantities that define the training constraint, quantities that define the parameterization, and quantities used only for measurement.

The subscript T always refers to the training support, while B denotes the single fixed support-external witness used throughout the analysis.

For a symmetric positive semidefinite matrix M, $\|w\|_{M^\dagger}^2 = w^\top M^\dagger w$ on $\mathrm{range}(M)$. The Moore–Penrose inverse is written $\dagger$. All exact interpolation results can be restricted to the attainable prediction range when $XMX^\top$ is singular.

B Proofs for Linear Metric Selectors

B.1 Prediction-space solution

Let $B = XA_\phi$. Gradient flow on Equation 5 obeys

$$\dot{u}_t = -B^\top(Bu_t - y), \qquad u_0 = 0. \tag{24}$$

The prediction residual $r_t = Bu_t - y$ satisfies

$$\dot{r}_t = -BB^\top r_t, \qquad r_0 = -y. \tag{25}$$

Therefore

$$r_t = -\exp(-BB^\top t)y. \tag{26}$$

If $y \in \mathrm{range}(B)$, the residual converges to zero. Integrating Equation 24 gives

$$u_t = B^\top(BB^\top)^\dagger \left[I - \exp(-BB^\top t)\right] y. \tag{27}$$

Taking $t \to \infty$ on $\mathrm{range}(B)$ yields

$$u_\infty = B^\top(BB^\top)^\dagger y. \tag{28}$$

Multiplication by A_ϕ, together with $BB^\top = XM_\phi X^\top$, proves Equation 6.

Table 1: Notation used in the theory and experiments.

Symbol	Type	Meaning
θ_0	base state	Pretrained parameter state before adaptation
f_T	training map	Complete prediction vector on the training support
z_T	target	Fixed desired training-support predictions
$\mathcal{F}_T(z_T)$	feasible set	Endpoints realizing the same training predictions
g_ϕ	chart	Adaptation parameterization indexed by ϕ
S_ϕ	selector	Limiting endpoint reached by optimization in chart g_ϕ
X	linear map	Jacobian or exact linear training operator
M_ϕ	metric	Induced metric $Dg_\phi Dg_\phi^\top$
R_M	right inverse	Selector operator $MX^\top(XMX^\top)^\dagger$
b, h_B	witness	Fixed endpoint functional excluded from training
p, q	factors	Signed-square coordinates with $w = p^2 - q^2$
α	initialization	Positive factor scale and coordinate balance
Ψ_α	complexity	Hypentropy potential of balanced factor flow
$e_j(t)$	remainder	Finite-time distance from flow j to its endpoint

B.2 Variational form

Every update in the chart range can be written $w = A_\phi u$. The minimum Euclidean-norm coordinate vector realizing w is $u = A_\phi^\dagger w$, and its squared norm is

$$\left\| A_\phi^\dagger w \right\|_2^2 = w^\top (A_\phi A_\phi^\top)^\dagger w = w^\top M_\phi^\dagger w. \tag{29}$$

The Lagrangian for minimizing Equation 29 under $Xw = y$ is

$$\mathcal{L}(w, \lambda) = \frac{1}{2} w^\top M_\phi^\dagger w - \lambda^\top (Xw - y). \tag{30}$$

Stationarity on $\mathrm{range}(M_\phi)$ gives $w = M_\phi X^\top \lambda$. Enforcing the constraint gives

$$XM_\phi X^\top \lambda = y. \tag{31}$$

The minimum-norm dual solution recovers Theorem 4.1.

B.3 Scale invariance

For $c > 0$, pseudoinverse homogeneity gives $(cXMX^\top)^\dagger = c^{-1}(XMX^\top)^\dagger$. Hence

$$cMX^\top (cXMX^\top)^\dagger = MX^\top (XMX^\top)^\dagger. \tag{32}$$

This proves Proposition 4.3. In the residual dynamics, scaling M by c replaces t by ct, which explains why effective-time matching removes only the scalar part of a parameterization change.

B.4 Discrete and minibatch updates

Proposition B.1 (Discrete and minibatch persistence)**.** *Consider any sequence of full-batch or minibatch gradient steps in a fixed linear chart, initialized at zero coordinates. If the iterates converge to an endpoint that interpolates all training predictions, that endpoint equals Equation 6. The conclusion is independent of step ordering and positive step sizes for which convergence occurs.*

Let $B = XA_\phi$. Every full-batch coordinate gradient lies in $\mathrm{range}(B^\top)$. A minibatch gradient uses a row restriction B_I and also lies in the same range because

$$B_I^\top r_I = B^\top P_I^\top r_I \tag{33}$$

for the row-selection map P_I. Starting from $u_0 = 0$, induction gives

$$u_k \in \mathrm{range}(B^\top) \qquad \text{for every } k. \tag{34}$$

If $u_k \to u_\infty$ and $Bu_\infty = y$, then u_∞ belongs to the intersection of the affine interpolation set with $\mathrm{range}(B^\top)$. This intersection is the singleton

$$u_\infty = B^\top (BB^\top)^\dagger y. \tag{35}$$

Mapping back through A_ϕ proves Proposition B.1. Step sizes and ordering determine whether and how quickly convergence occurs, but not the interpolating endpoint within a fixed linear chart.

B.5 WITNESS EQUIVALENCE

For any attainable target y, the witness difference is

$$b^\top (R_{M_1} - R_{M_2}) y. \tag{36}$$

This expression vanishes for every y in a subspace $\mathcal{Y}$ exactly when the orthogonal projection of $(R_{M_1} - R_{M_2})^\top b$ onto $\mathcal{Y}$ is zero. Choosing $\mathcal{Y}$ as the intersection of the attainable prediction ranges proves the witness statement in Theorem 4.5. Equality of endpoints for every commonly attainable target is equality of the two linear right inverses on that common range.

B.6 FINITE-TIME CERTIFICATE

Proposition B.2 (Finite-time ordering certificate)**.** *Suppose two flows converge to endpoints S_1, S_2 and satisfy $\|w_j(t) - S_j\|_2 \le e_j(t)$. If*

$$\left|b^\top (S_1 - S_2)\right| > \|b\|_2 \left(e_1(t) + e_2(t)\right), \tag{37}$$

then the observed finite-time contrast has the endpoint ordering.

Write the observed contrast as

$$\begin{aligned} b^\top (w_1(t) - w_2(t)) = b^\top (S_1 - S_2) \\ + b^\top \left(w_1(t) - S_1\right) \\ - b^\top \left(w_2(t) - S_2\right). \end{aligned} \tag{38}$$

Cauchy–Schwarz gives, with $\delta_j(t) = w_j(t) - S_j$,

$$\left|b^\top \left(\delta_1(t) - \delta_2(t)\right)\right| \le \|b\|_2 \left(e_1(t) + e_2(t)\right). \tag{39}$$

Equation 37 makes the remainder smaller than the endpoint margin, so the sign cannot change.

For a linear metric flow whose smallest positive eigenvalue of $XMX^\top$ is μ_M, Equation 26 gives

$$\|r_t\|_2 \le \exp(-\mu_M t) \|y\|_2 . \tag{40}$$

When the selector lift is bounded by $\left\|MX^\top (XMX^\top)^\dagger\right\|_{\mathrm{op}}$, this produces an explicit exponential choice of $e(t)$.

C PROOFS FOR THE SIGNED-SQUARE SELECTOR

C.1 CONSERVED PRODUCT AND DUAL PATH

Let $g = X^\top (Xw - y)$. Differentiating through $w_i = p_i^2 - q_i^2$ yields

$$\dot{p}_i = -2 p_i g_i, \qquad \dot{q}_i = 2 q_i g_i. \tag{41}$$

The coordinate product is invariant because

$$\frac{d}{dt}(p_i q_i) = \dot{p}_i q_i + p_i \dot{q}_i = 0. \tag{42}$$

With balanced initialization, $p_i q_i = \alpha_i^2$ for all time. Parameterize

$$p_i = \alpha_i e^{s_i}, \qquad q_i = \alpha_i e^{-s_i}. \tag{43}$$

Then

$$w_i = 2\alpha_i^2 \sinh(2 s_i), \qquad \dot{s}_i = -2 g_i. \tag{44}$$

Since $g = X^\top r$, integration gives $s(t) \in \mathrm{range}(X^\top)$. If the integrated residual converges, there is a dual vector λ such that the endpoint satisfies Equation 14; constant factors are absorbed into λ.

C.2 Convex potential and uniqueness

For one coordinate, differentiation of Equation 11 gives

$$\Psi'_{\alpha_i}(w_i) = \operatorname{arsinh}\left(\frac{w_i}{2\alpha_i^2}\right), \tag{45}$$

$$\Psi''_{\alpha_i}(w_i) = \frac{1}{\sqrt{w_i^2 + 4\alpha_i^4}} > 0. \tag{46}$$

Thus Ψ_α is strictly convex. The KKT condition for minimizing it under $Xw = y$ is

$$\nabla\Psi_\alpha(w) = X^\top\lambda. \tag{47}$$

Inverting Equation 45 coordinatewise gives Equation 14. Strict convexity makes the feasible minimizer unique, which proves Theorem 5.1 whenever the gradient-flow limit interpolates.

C.3 Small-initialization limit

For fixed nonzero w_i and $\alpha_i = \tau a_i$, define $x_{i,\tau} = w_i/(2\tau^2 a_i^2)$. Then

$$\operatorname{arsinh}(x_{i,\tau}) = \operatorname{sgn}(w_i)\Bigg[2\log(1/\tau) + \log\left(\frac{|w_i|}{a_i^2}\right)\Bigg] + o(1). \tag{48}$$

The square-root term contributes $-|w_i| + o(1)$. Dividing the potential by $\log(1/\tau)$ leaves $2|w_i|$. Summing over coordinates proves Equation 15. The expansion is uniform on compact sets after treating a shrinking neighborhood of zero directly through convexity.

C.4 Large-initialization limit

For $|w_i|/\tau^2 \to 0$, Taylor expansion gives

$$\operatorname{arsinh}\left(\frac{w_i}{2\tau^2 a_i^2}\right) = \frac{w_i}{2\tau^2 a_i^2} + O(\tau^{-6}), \tag{49}$$

$$\sqrt{w_i^2 + 4\tau^4 a_i^4} = 2\tau^2 a_i^2 + \frac{w_i^2}{4\tau^2 a_i^2} + O(\tau^{-6}). \tag{50}$$

Substitution into Equation 11 yields

$$\Psi_{\tau a_i}(w_i) = \frac{w_i^2}{4\tau^2 a_i^2} + O(\tau^{-6}). \tag{51}$$

Multiplying by $4\tau^2$ and summing proves Equation 16.

C.5 Why capacity is unchanged

For any desired scalar w_i, choose

$$p_i^2 = \frac{w_i + \sqrt{w_i^2 + 4\alpha_i^4}}{2}, \qquad q_i = \frac{\alpha_i^2}{p_i}. \tag{52}$$

Then $p_i q_i = \alpha_i^2$ and $p_i^2 - q_i^2 = w_i$. Hence every vector is representable within the invariant manifold. Changing α changes the selected complexity but not the attainable training endpoints.

C.6 Finite-support factor sufficiency

Lemma C.1 (Finite-support factor sufficiency)**.** *Let $H \in \mathbb{R}^{d_{\text{in}} \times N}$ collect the inputs to one adapted linear map on N training examples. For every direct increment ΔW, there is an increment BA with rank at most $\operatorname{rank}(H) \le N$ such that*

$$BAH = \Delta W H. \tag{53}$$

Proof. Set

$$\widetilde{W} = \Delta W H H^{\dagger}. \tag{54}$$

The Moore–Penrose identity $HH^{\dagger}H = H$ gives

$$\widetilde{W}H = \Delta W H H^{\dagger} H = \Delta W H. \tag{55}$$

Moreover,

$$\operatorname{rank}(\widetilde{W}) \le \operatorname{rank}(H) \le N. \tag{56}$$

Take any rank factorization $\widetilde{W} = BA$. This proves the claim. □

The claim is local to the declared activation matrix. In a multilayer nonlinear network, changing an earlier layer can change later activations, so the experiment additionally checks complete endpoint predictions rather than invoking the lemma alone.

D Phase-Boundary Derivations

D.1 Metric selector

For $X = (a, b)$ and $M = \operatorname{diag}(m_u, m_v)$, the scalar training Gram is

$$XMX^{\top} = m_u a^2 + m_v b^2. \tag{57}$$

Substitution into Equation 6 gives Equation 18. Applying (c, d) gives Equation 19. Its denominator is positive, so for $y > 0$ the sign is the sign of $cm_u a + dm_v b$, proving Equation 20.

The derivative with respect to the log metric ratio $\rho = \log(m_u/m_v)$ is

$$\frac{d}{d\rho} h_B(S_M(y)) = \frac{y m_u m_v ab(cb - da)}{(m_u a^2 + m_v b^2)^2}. \tag{58}$$

The ordering is therefore monotone in ρ whenever $cb - da$ has fixed nonzero sign. This supplies a stronger prediction than comparing two isolated parameterizations.

D.2 Signed-square selector

For one training constraint, Equation 14 becomes

$$u = 2\alpha_u^2 \sinh(a\lambda), \qquad v = 2\alpha_v^2 \sinh(b\lambda). \tag{59}$$

The constraint yields Equation 21. Its left side is strictly increasing in λ because

$$2a^2\alpha_u^2 \cosh(a\lambda) + 2b^2\alpha_v^2 \cosh(b\lambda) > 0. \tag{60}$$

Thus λ_y exists uniquely for every y. Applying the witness gives Equation 22.

When $\alpha_u = \tau a_u$, $\alpha_v = \tau a_v$, and τ is large, $\sinh(x) = x + O(x^3)$. The resulting selector is the metric solution with

$$m_u = a_u^2, \qquad m_v = a_v^2. \tag{61}$$

When τ is small, satisfying a fixed nonzero target requires $|\lambda_y| \to \infty$. If $a > b > 0$, the broad term grows as $e^{a|\lambda_y|}$ and dominates unless the initialization imbalance shrinks exponentially. This is the sparse-coordinate side of the phase portrait.

D.3 Base alignment

Let the common effective initialization be $w_0 = (u_0, v_0)$. The required update satisfies

$$a\Delta u + b\Delta v = \widetilde{y}, \qquad \widetilde{y} = y - au_0 - bv_0. \tag{62}$$

The metric witness is

$$h_B(w_0 + S_M(\widetilde{y})) = cu_0 + dv_0 + \frac{\widetilde{y}(cm_u a + dm_v b)}{m_u a^2 + m_v b^2}. \tag{63}$$

Base alignment shifts the zero of the phase boundary through the first term and changes the residual target through the second. It does not generally remove dependence on m_u/m_v.

D.4 Quadratic regularization

For finite penalty $\gamma > 0$, consider

$$\min_w \frac{1}{2}\|Xw - y\|_2^2 + \frac{\gamma}{2} w^\top M^{-1} w. \tag{64}$$

The solution is

$$w_{M,\gamma} = MX^\top (XMX^\top + \gamma I)^{-1} y. \tag{65}$$

The same numerator determines the two-route witness sign, while the denominator changes its magnitude. As $\gamma \downarrow 0$, Equation 65 converges to the interpolating selector on the attainable range.

D.5 Coverage perturbation

Let the second support row be $X_2 = (a', b')$. The augmented determinant is

$$\det \begin{pmatrix} a & b \\ a' & b' \end{pmatrix} = ab' - a'b. \tag{66}$$

If it is nonzero, the two-dimensional feasible set is a singleton. The selected metric and factorized complexities can change path and speed, but any interpolating endpoint equals the same unique solution. This proves Corollary 6.4.

E Selector Equivalence and Quotient Geometry

The training map induces the quotient space $\mathbb{R}^d / \operatorname{null}(X)$. Every point in the quotient has one training prediction, while a parameterization chooses a representative in the original parameter space.

Definition E.1 (Selector section)**.** A linear map $R : \operatorname{range}(X) \to \mathbb{R}^d$ is a selector section when

$$XRy = y \qquad \text{for all } y \in \operatorname{range}(X). \tag{67}$$

The metric right inverse $R_M = MX^\top (XMX^\top)^\dagger$ is one such section whenever the metric chart can express the target range.

E.1 Proof of training-trajectory separation

Choose coordinates in which the full-row-rank training map is $X = (I_n, 0)$. Let $C \in \mathbb{R}^{n \times (d-n)}$ have $\|C\|_{\mathrm{op}} < 1$, and define

$$\begin{aligned} M_+ &= \begin{pmatrix} I_n & C \\ C^\top & I_{d-n} \end{pmatrix}, \\ M_- &= \begin{pmatrix} I_n & -C \\ -C^\top & I_{d-n} \end{pmatrix}. \end{aligned} \tag{68}$$

Both matrices are positive definite by the Schur complement, and

$$XM_+X^\top = XM_-X^\top = I_n. \tag{69}$$

The training prediction flow therefore satisfies the same differential equation under both metrics:

$$\dot{z}_t = -(z_t - y), \qquad z_0 = 0, \qquad z_t = (1 - e^{-t})y. \tag{70}$$

The selected endpoints are nevertheless

$$S_{M_+}(y) = \begin{pmatrix} y \\ C^\top y \end{pmatrix}, \qquad S_{M_-}(y) = \begin{pmatrix} y \\ -C^\top y \end{pmatrix}. \tag{71}$$

For a witness $b = (0, v)$, the endpoint gap is

$$b^\top \{S_{M_+}(y) - S_{M_-}(y)\} = 2v^\top C^\top y. \tag{72}$$

Choose v, C, y so this value is nonzero. This proves Theorem 4.4.

Let $\widehat{h}$ be any estimator whose input is the complete training trajectory. That input is identical under M_+ and M_-, so the estimator returns the same value in both worlds. The triangle inequality gives the two-point lower bound

$$\max_{\sigma\in\{+,-\}} \left|\widehat{h} - h_B(S_{M_\sigma}(y))\right| \geq |v^\top C^\top y|. \tag{73}$$

Parameterization information or a support-external observation is therefore necessary to identify the selected endpoint.

Proposition E.2 (Null displacement)**.** *For any two selector sections R_1, R_2,*

$$\mathrm{range}(R_1 - R_2) \subseteq \mathrm{null}(X). \tag{74}$$

Conversely, adding any linear map from prediction space into $\mathrm{null}(X)$ to a selector section produces another section.

Equation 74 shows that matched-fit endpoint differences live entirely in directions invisible to the training map. The witness sees those differences only through its projection onto the same null space.

Proposition E.3 (Witness quotient criterion)**.** *For a fixed training map, the following statements are equivalent:*

1. *$b^\top w$ is constant on every nonempty affine training fiber.*
2. *$b^\top n = 0$ for every $n \in \mathrm{null}(X)$.*
3. *$b \in \mathrm{range}(X^\top)$.*
4. *$b^\top R_1 = b^\top R_2$ for every pair of selector sections.*

The equivalence of the first two statements follows by translating a point within its fiber. The equivalence of the second and third is the fundamental theorem of linear algebra. The fourth follows from Proposition E.2; for the reverse direction, choose selector sections differing along any witness-visible null vector.

E.2 Metric equivalence without metric equality

Two metrics can differ substantially and still define the same selector. A sufficient condition is equality on the training-relevant image:

$$M_1 X^\top = M_2 X^\top C \tag{75}$$

for an invertible prediction-space map C satisfying the corresponding Gram compatibility. The exact condition remains right-inverse equality:

$$M_1 X^\top (X M_1 X^\top)^\dagger = M_2 X^\top (X M_2 X^\top)^\dagger. \tag{76}$$

This condition is weaker than $M_1 = M_2$, so the theory predicts genuine parameterization-independence regimes even before the support identifies every parameter direction.

E.3 A target-specific criterion

For one fixed target y, endpoint equality requires only

$$(R_{M_1} - R_{M_2})y = 0, \tag{77}$$

and witness equality requires only its projection onto b to vanish. We distinguish these three levels in experiments: equal training predictions, equal predeclared witness, and equal parameter endpoint. The theory concerns the second, while the first is an attribution requirement and the third is stronger than needed.

E.4 Implications for training-only observation

The trajectory-separation theorem sharpens the measurement requirement. Observing every training prediction at every optimization time can still leave the null-space endpoint unidentified. Training

curves can certify fit and effective speed, but they cannot replace knowledge of the adaptation chart or a support-external witness. Selecting checkpoints by training loss alone may therefore preserve target performance while changing deployment behavior.

For method selection, the relevant object is conditional. One asks whether a parameterization's induced section is appropriate for a specified base model, support geometry, and external functional. Our experiments exhibit large separation, near-equivalence, and support-driven collapse within one design. A universal ranking of full fine-tuning, LoRA, rsLoRA, and DoRA would discard the phase structure that predicts these changes.

F Local Extension to Nonlinear Adaptation

Let $F(u) = f_T(g_\phi(u))$ and $H(u) = h_B(g_\phi(u))$. Around the common initialization u_0, write

$$F(u_0 + s) = F(u_0) + J_T s + r_T(s), \tag{78}$$

$$H(u_0 + s) = H(u_0) + j_B^\top s + r_B(s). \tag{79}$$

The chart induces the instantaneous metric

$$M_{\phi,t} = Dg_\phi(u_t) Dg_\phi(u_t)^\top. \tag{80}$$

Equation 80 is constant for a linear chart and path-dependent otherwise.

Assumption F.1 (Local smoothness)**.** For $\|s\|_2 \le R$, the remainders satisfy

$$\begin{aligned} \|r_T(s)\|_2 &\le \frac{L_T}{2} \|s\|_2^2, \\ |r_B(s)| &\le \frac{L_B}{2} \|s\|_2^2. \end{aligned} \tag{81}$$

Let $R_{\phi,0}$ denote the right inverse selected by the frozen linearized chart, and define $s_\phi^{\mathrm{lin}} = R_{\phi,0} y$. For a matched nonlinear endpoint s_ϕ, set $e_T = r_T(s_\phi)$ and define its path-selection remainder by

$$\delta_\phi = \|s_\phi - R_{\phi,0}(y - e_T)\|_2. \tag{82}$$

This remainder is zero for a fixed linear chart and records the endpoint displacement caused by nonlinear chart and metric evolution beyond target transport.

Theorem F.2 (Local selector decomposition)**.** *Assume s_ϕ and s_ϕ^{lin} stay within radius R, and let $\kappa_\phi = \|R_{\phi,0}\|_{\mathrm{op}}$. Then every matched nonlinear endpoint satisfies*

$$\begin{aligned} &\left|H(u_0 + s_\phi) - H(u_0 + s_\phi^{\mathrm{lin}})\right| \\ &\quad \le \|j_B\|_2 \left(\frac{\kappa_\phi L_T R^2}{2} + \delta_\phi\right) + L_B R^2. \end{aligned} \tag{83}$$

The decomposition separates curvature of the training map, path-dependent endpoint selection, and curvature of the held-out witness. Each term can be bounded or measured independently.

Let $e_T = r_T(s_\phi)$. Since the nonlinear endpoint matches the target,

$$J_T s_\phi = y - e_T. \tag{84}$$

The frozen selector applied to the perturbed target differs from s_ϕ^{lin} by at most

$$\|R_{\phi,0} e_T\|_2 \le \frac{\kappa_\phi L_T R^2}{2}. \tag{85}$$

By Equation 82 and the triangle inequality,

$$\left\|s_\phi - s_\phi^{\mathrm{lin}}\right\|_2 \le \frac{\kappa_\phi L_T R^2}{2} + \delta_\phi. \tag{86}$$

Applying the witness expansion at both endpoints adds at most $L_B R^2$, proving Equation 83.

Theorem F.2 provides a falsifiable boundary. A frozen metric should predict endpoint ordering only when the separation margin exceeds the curvature and measured path-selection remainders. The end-to-end experiments therefore report the complete trajectory rather than interpreting a first-step metric as an exact global law.

Table 2: Solvable-carrier sweep. Every cell is evaluated with direct, anisotropic metric, and signed-square selectors.

Factor	Values
Training size	16, 32
Parameter dimension	64, 128
Broad-shortcut correlation	1.00, 0.98, 0.90, 0.50, 0.00
Observation noise	0.00, 0.05
Base alignment	-0.5, 0.0, 0.5 along the broad coordinate
Metric balance	0.25, 1.0, 4.0
Factor initialization scale	0.01, 0.1, 1.0
Factor initialization balance	0.25, 1.0, 4.0
Independent data seeds	two disjoint panels of 256 seeds

G Solvable-Carrier Protocol

The solvable carrier isolates endpoint selection from neural approximation. Each trial draws n training observations in $d > n$ dimensions. The first coordinate is a broad feature $g \sim \mathcal{N}(0, 1)$. The second is a narrow feature

$$s = \rho g + \sqrt{1 - \rho^2}\, \xi, \qquad \xi \sim \mathcal{N}(0, 1), \tag{87}$$

and the remaining coordinates are independent nuisance features with standard deviation 0.15. The training target is

$$y = g + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{88}$$

Held-out examples reverse the narrow feature:

$$s_B = -g_B + 0.05\xi_B. \tag{89}$$

The scalar witness is the mean signed held-out prediction $\mathbb{E}[g_B x_B^\top w]$. A positive witness follows the broad rule, while a negative witness follows the narrow shortcut.

G.1 Crossed factors

The main sweep crosses the factors in Table 2. Each configuration is repeated with independent data seeds and is evaluated on 512 held-out draws.

For every endpoint we store training relative error, held-out witness, held-out mean squared error, sign accuracy, Euclidean distance from the direct endpoint, witness row-space residual, convergence state, and factor-solver iterations. The two seed panels use nonoverlapping seed offsets and contain 798,720 endpoint records in total.

G.2 Exact numerical solution

The direct and metric selectors use Equation 6. The factorized endpoint solves the dual Equation 14 by damped Newton iterations. The dual Jacobian is

$$J_\lambda = X \operatorname{diag}\left(2\alpha_i^2 \cosh((X^\top \lambda)_i)\right) X^\top. \tag{90}$$

A backtracking line search requires a strict reduction in squared constraint residual. The complete residual is retained, so numerical approximation cannot masquerade as an endpoint effect.

G.3 Primary contrasts

The phase contrast compares factor balances 0.25 and 4.0 on the same carrier draw. The initialization-scale contrast compares 0.01 and 1.0. The identifiability contrast follows the full correlation path rather than dichotomizing it. Direct and metric results are analytic, while factor results are paired by the same training and held-out matrices.

H ANCHORED TRANSFORMER-FEATURE PROTOCOL

The anchored carrier realizes the two-route theorem inside pretrained Transformer representations. It uses three prompt domains: arithmetic truth, lexical category, and sentiment. Each input contains a semantic statement and one auxiliary marker. On aligned examples, the marker and semantic labels agree. On conflict examples, they disagree.

For model representation $z(x) \in \mathbb{R}^d$, an independent calibration split contains equal numbers of aligned and conflict examples. We estimate two labeled directions

$$v_{\mathrm{sem}} = \frac{1}{m}\sum_{i=1}^{m} y_i^{\mathrm{sem}}\{z(x_i) - \bar{z}\}, \tag{91}$$

$$v_{\mathrm{short}} = \frac{1}{m}\sum_{i=1}^{m} y_i^{\mathrm{short}}\{z(x_i) - \bar{z}\}. \tag{92}$$

Both directions are normalized using calibration data only. The calibration examples do not enter endpoint fitting or held-out evaluation.

H.1 EXACT AMBIGUITY CONSTRUCTION

For an aligned training example, let a_i and s_i be its semantic and shortcut anchor scores. We expose the common coordinate

$$c_i = \frac{a_i + s_i}{2} \tag{93}$$

to both routes, producing the design row (c_i, c_i) and target c_i. Hence every coefficient pair satisfying $u + v = 1$ has exactly the same complete training predictions.

Conflict-support examples retain their distinct row (a_i, s_i) and use target a_i. They progressively identify the semantic route. Held-out conflict examples always retain their distinct coordinates, and the fixed witness is

$$\widehat{h}_B(u, v) = \frac{1}{m_B}\sum_{i=1}^{m_B} y_i^{\mathrm{sem}}(a_i u + s_i v). \tag{94}$$

The exact construction tests whether the predicted selector and support-driven collapse survive in pretrained neural coordinates. The natural carrier then measures the corresponding geometry without enforced collinearity.

H.2 SWEEP AND CONTROLS

Each model and dataset seed uses 96 aligned plus 96 conflict calibration examples, 64 endpoint-training examples, and a disjoint held-out conflict set. We use 64 paired endpoint replicates and support-coverage fractions 0, 0.03125, 0.0625, 0.125, 0.25, and 0.5. Metric and factor balances range from 0.05 to 20. Factor scales are 0.01, 0.1, and 1.0. Base alignment takes values -0.5, 0, and 0.5.

The main checks are mechanical. At zero coverage, the complete training prediction vector must be identical across selectors. At positive identifying coverage, the coefficient difference must approach the unique support-determined solution. The full panel contains 4,603,392 endpoint records derived from 36 independent model-seed feature files across 10 model families. Every record includes the numerical constraint residual and convergence state.

I NATURAL AND END-TO-END NEURAL PROTOCOLS

I.1 PROMPT CARRIERS

All domains use binary labels A and B and an auxiliary marker KITE or MOSS. Sentiment examples vary the evaluated object and adjective. Lexical examples test whether a noun denotes a living animal. Arithmetic examples test a generated addition equation. Semantic labels are balanced within every split.

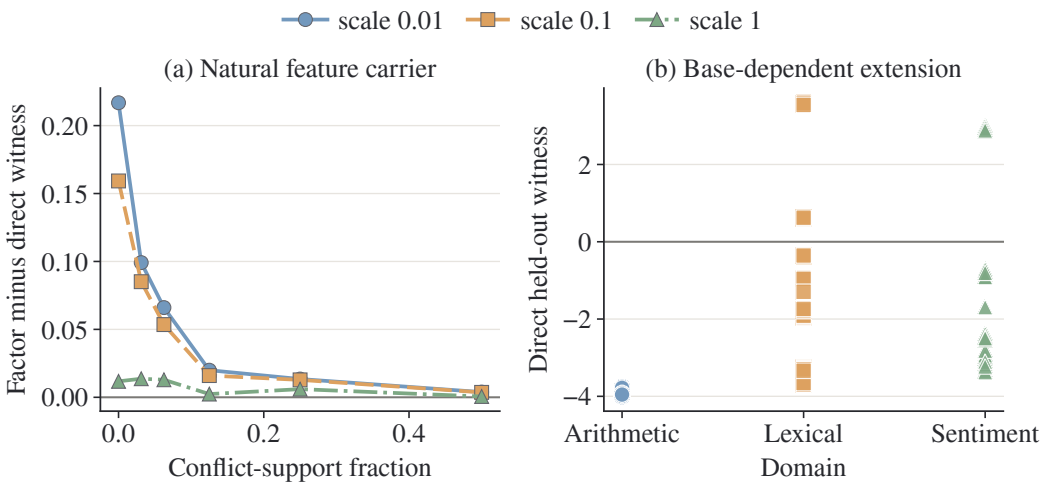

Figure 5: Natural Transformer-feature endpoints. Panel (a) shows the Qwen3-4B factor-minus-direct witness displacement as conflict-support coverage increases. Panel (b) reports model-seed values separately and shows that the direct endpoint itself depends on the pretrained model and semantic domain.

On the narrow training support, the marker agrees with the semantic label except for the declared conflict-coverage fraction. Held-out conflict examples reverse the marker. Held-out aligned examples serve as a fit-preserving control. The semantic conflict witness is the average signed A-versus-B logit margin.

I.2 Frozen representation extraction

For each prompt, we extract the final nonpadding hidden vector and the base-model A-versus-B logit margin. Each model and seed contains 384 aligned training candidates, 192 conflict-support candidates, 384 held-out conflict examples, and 384 aligned controls in each of the three domains. Feature files use float16 storage, while all selector calculations use float64.

The natural representation carrier fits a target margin from the full hidden vector. The direct solution is the minimum Euclidean-norm interpolant. Signed-square solutions vary global initialization scale and equivalent paired-coordinate rotations. The rotation is orthogonal and preserves the complete training Gram matrix:

$$(XQ)(XQ)^\top = XX^\top. \tag{95}$$

Thus a changed factor endpoint under rotation cannot be attributed to changed training information or a changed direct solution.

Fig. 5 summarizes the unmodified hidden-feature carrier; complete model-wise values appear in Appendix N.

The anchored and natural carriers have distinct inferential roles. The anchored construction realizes the exact theorem in Transformer features and exposes the complete parameterization boundary. The natural carrier asks how much of that geometry survives without enforced training equivalence. End-to-end adaptation then measures whether actual LoRA-family paths retain the ordering once capacity and effective learning are controlled.

I.3 End-to-end parameterizations

The direct update opens the declared query and value projection weights themselves. LoRA represents their increments with learned low-rank factors. rsLoRA changes the rank scaling, and DoRA decomposes magnitude and direction. All factorized arms use the same target projection names and a rank above the finite-support expressivity scale. A last-layer-only intervention is retained as a weak-write control; the principal comparison opens the same projection family across all layers.

All arms use SGD without momentum. The loss is squared error between the final-token A-versus-B margin $m_\theta(x)$ and target margin γy:

$$L_T(\theta) = \frac{1}{|T|} \sum_{(x,y)\in T} \{m_\theta(x) - \gamma y\}^2. \tag{96}$$

We use deterministic cyclic minibatches. Each run records training, held-out conflict, and aligned-control metrics at steps 0, 1, 2, 4, 8, 16, 32, 64, 128, and 160.

I.4 Training-only calibration

Nominal learning rates are calibrated in an independent Qwen3-4B run. Direct and LoRA grids are compared only through the training margin trajectory. For the main experiment, each factorized arm is matched separately to its direct comparator. Their lowest jointly attainable training RMSE is

$$r_\star(\phi) = \max\left\{\min_t r_{\mathrm{direct},t}, \min_t r_{\phi,t}\right\}. \tag{97}$$

For each of the two arms, the selected checkpoint minimizes distance to $r_\star(\phi)$. Pairwise matching prevents an underfitting third parameterization from forcing otherwise converged methods to be compared at an unnecessarily poor checkpoint. The rule uses no held-out outcome and introduces no manually selected fit threshold. We report $r_\star(\phi)$ and the remaining fit mismatch with every endpoint contrast.

I.5 Checkpoint and queue recovery

At each evaluation step, a run atomically saves trainable parameters, optimizer state, completed step, and trajectory. A queue assigns a deterministic subset of configurations to each GPU. On restart, it verifies the summary hash of completed tasks and resumes incomplete checkpoints. Queue reports retain per-task exit code and elapsed time.

Fig. 6 tests three predicted boundary coordinates without changing the held-out witness or selecting a condition from held-out outcomes.

J Estimands and Statistical Inference

The primary anchored estimand at support coverage q and balance β is

$$\tau_{\mathrm{anchor}}(q,\beta) = \mathbb{E}_s\left[\widehat{h}_{B,s}(S_\beta) - \widehat{h}_{B,s}(S_{\mathrm{direct}})\right], \tag{98}$$

where s indexes an independently generated dataset split. The paired design uses the same base model, anchor calibration, training support, and held-out examples within each difference.

The support-collapse estimand compares absolute parameterization spread:

$$D(q) = \mathbb{E}_s\left[\max_\phi \widehat{h}_{B,s}(S_\phi) - \min_\phi \widehat{h}_{B,s}(S_\phi)\right]. \tag{99}$$

Because an empirical range is upward biased, the headline inference uses prespecified signed balance contrasts. The range is descriptive and receives a within-seed permutation calibration.

For natural feature experiments, the primary contrast is the signed-square minus direct witness at scale $\alpha = 0.01$, followed by the full initialization-scale curve. Orthogonal rotations form a second paired intervention. Model seeds are first averaged within each model family and then treated as independent replication units.

For end-to-end runs, let $t_\phi(r_\star)$ be the training-only matched checkpoint from Equation 97. The endpoint estimand is

$$\begin{aligned}\tau_{\mathrm{e2e}}(\phi) = \mathbb{E}_s\big[&h_B(\theta_{\phi,t_\phi(r_\star)}) \\ &- h_B(\theta_{\mathrm{direct},t_{\mathrm{direct}}(r_\star)})\big].\end{aligned} \tag{100}$$

Intervals use paired seed bootstrap resampling with 10,000 replicates. When a model has fewer than four independent seeds, we report all seed-level values and a sign-compatible exact randomization

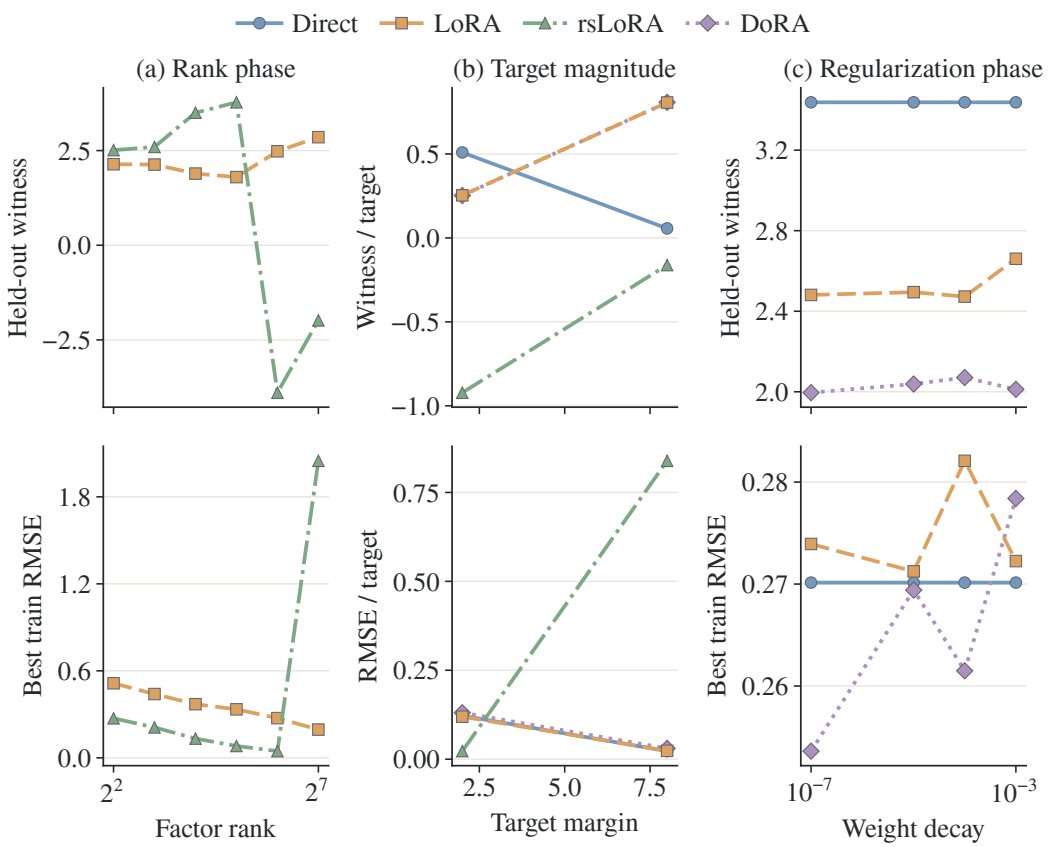

Figure 6: End-to-end phase controls at the training-only best-fit checkpoint. Panel (a) varies factor rank, panel (b) varies target magnitude, and panel (c) varies weight decay. Training RMSE accompanies every underlying endpoint record and separates capacity failure from selector movement.

interval rather than treating prompt examples as replicates. The synthetic sweep uses its 256 seeds per independent panel as the unit. No interval pools raw logits across model scales.

Multiple phase coordinates are summarized as a curve with simultaneous max-statistic bootstrap bands. Prespecified endpoint contrasts are reported separately from exploratory cells. The full grid remains in machine-readable result files.

K Prompt and Target Schemas

The prompt grammar is fixed before endpoint fitting. Representative instances follow.

Sentiment.

```
The short review says: “The film was delightful.”
Auxiliary code: KITE.
Classify the example using exactly one label, A or B.
Label:
```

Lexical category.

```
Consider the statement: “An otter is a living animal.”
Auxiliary code: KITE.
Classify the example using exactly one label, A or B.
Label:
```

Arithmetic truth.

```
Consider the equation: “17 + 9 = 26.”
Auxiliary code: KITE.
Classify the example using exactly one label, A or B.
Label:
```

Positive semantic statements map to A and negative statements map to B. KITE maps to A and MOSS maps to B. Aligned training examples make the two rules observationally consistent. Conflict examples reverse the marker while preserving the semantic target.

Adjectives, nouns, operands, incorrect offsets, marker assignments, and example order are generated from the recorded seed. The held-out split uses a disjoint seed offset and item identifiers. Each artifact preserves the item identifiers and the complete generation metadata beside the feature array.

The three domains share the same experimental grammar but not the same lexical material. Sentiment items combine positive and negative adjectives with neutral movie nouns. Lexical items pair animate and inanimate subjects with category predicates. Arithmetic items sample operands and then construct either the correct sum or a nonzero offset. This separation lets the experiment vary semantic content while preserving the same binary decision interface. The auxiliary code is deliberately semantically empty: it is assigned independently of the natural-language surface form and becomes predictive only through the construction of the training split.

For a zero-coverage training set, the semantic label and the code label agree on every observed example. Both rules therefore attain the same supervised target, while the held-out conflict set distinguishes them. Positive coverage replaces a declared fraction of training codes by their opposite code while leaving the semantic label unchanged. These conflict examples make the semantic direction observable on the training support. Coverage is thus an identifiability intervention, not a change to the evaluation criterion. Every method receives the identical examples, order, labels, target margins, and evaluation items within a seed.

The recorded target is the logit difference between labels A and B. We report its sign as accuracy and its signed projection onto the semantic rule as the witness. The shortcut witness uses the code rule instead. This paired representation separates endpoint magnitude from endpoint direction: two fitted models can have similar training error yet choose opposite signs on the conflict set, whereas a large signed margin and a high accuracy together indicate a stable semantic endpoint. The control split retains aligned codes and diagnoses broad degradation rather than endpoint selection.

All text is tokenized once before optimization. The feature carriers store the hidden representation at the final prompt position, while the end-to-end experiments recompute the same logit difference from the trained model. Consequently, the analytic, anchored, natural-feature, and end-to-end panels use one observable definition even though they expose progressively more of the neural training system.

L Complete Result Tables

In the tables, D and F denote direct and factorized updates, respectively; Acc. is sign accuracy, and Null disp. is Euclidean displacement from the direct endpoint. RMSE columns report training fit, while Gap is the paired fit mismatch. Witness contrasts retain their original numerical precision. End-to-end rows should be read jointly: RMSE and Gap quantify remaining fit mismatch, whereas the witness columns measure how closely matched solutions differ outside the observed support. Table 3 reports the prespecified low-initialization phase slice. Table 4 shows selection at zero coverage and collapse after the witness becomes support-identifiable. Table 5 gives every zero-coverage matched-fit endpoint for Qwen2.5 7B Instruct. Table 6 gives every zero-coverage matched-fit endpoint for Qwen2.5 Math 7B. Table 7 gives every zero-coverage matched-fit endpoint for Qwen3 14B. Table 8 gives every zero-coverage matched-fit endpoint for Qwen3 30B-A3B. Table 9 gives every zero-coverage matched-fit endpoint for Qwen3 4B. Table 10 gives every zero-coverage matched-fit endpoint for Qwen3 8B. Across models, paired common fit isolates parameterization-selected continuation, which the signed witness measures outside observed support.

Table 3: Solvable-carrier endpoint phase. Values average both independent 256-seed panels.

Correlation	Balance	Train error	Witness	Sign acc.	Null disp.
0.00	0.25	1.27e-09	1.007	1.000	0.152
0.00	1.00	1.86e-10	1.007	1.000	0.152
0.00	4.00	1.69e-14	1.007	1.000	0.152
0.50	0.25	1.04e-09	1.007	1.000	0.173
0.50	1.00	4.26e-10	1.007	1.000	0.174
0.50	4.00	1.04e-14	1.007	1.000	0.174
0.90	0.25	1.21e-09	1.006	1.000	0.318
0.90	1.00	9.14e-10	1.006	1.000	0.318
0.90	4.00	2.33e-09	1.006	1.000	0.318
0.98	0.25	1.25e-09	1.003	1.000	0.519
0.98	1.00	1.30e-09	1.005	1.000	0.520
0.98	4.00	2.26e-11	1.005	1.000	0.521
1.00	0.25	1.71e-11	-0.888	0.017	0.628
1.00	1.00	1.14e-09	0.000	0.501	0.074
1.00	4.00	1.71e-11	0.888	0.999	0.628

Table 4: Anchored Transformer selector and support-identifiability boundary, aggregated across model seeds and domains.

Coverage	Balance	Coeff. diff.	Witness	Acc.	Train error
0.00000	0.1	-0.980	-0.951	0.008	3.01e-13
0.00000	1.0	0.000	-0.109	0.405	3.08e-11
0.00000	10.0	0.980	0.733	0.863	3.04e-13
0.03125	0.1	1.000	0.750	0.865	5.44e-12
0.03125	1.0	1.000	0.750	0.865	3.95e-12
0.03125	10.0	1.000	0.750	0.865	1.12e-12
0.06250	0.1	1.000	0.750	0.865	4.89e-12
0.06250	1.0	1.000	0.750	0.865	3.65e-12
0.06250	10.0	1.000	0.750	0.865	1.14e-12
0.12500	0.1	1.000	0.750	0.865	4.23e-12
0.12500	1.0	1.000	0.750	0.865	3.43e-12
0.12500	10.0	1.000	0.750	0.865	1.15e-12
0.25000	0.1	1.000	0.750	0.865	4.63e-12
0.25000	1.0	1.000	0.750	0.865	3.75e-12
0.25000	10.0	1.000	0.750	0.865	1.14e-12
0.50000	0.1	1.000	0.750	0.865	4.58e-12
0.50000	1.0	1.000	0.750	0.865	3.34e-12
0.50000	10.0	1.000	0.750	0.865	1.19e-12

Table 5: Pairwise matched-fit end-to-end endpoints for Qwen2.5 7B Instruct. Checkpoints are selected using training RMSE only; intervals resample dataset seeds.

Domain	Factor	Seeds	RMSE D	RMSE F	Gap	Witness D / F	F - D	95% CI
arithmetic	DoRA	3	0.238	0.227	0.012	-3.884 / -3.715	0.170	[-0.036, 0.472]
arithmetic	LoRA	3	0.238	0.243	0.015	-3.884 / -3.729	0.156	[-0.049, 0.496]
arithmetic	rsLoRA	3	3.996	3.985	0.011	0.328 / 0.173	-0.156	[-0.476, 0.009]
lexical	DoRA	3	0.203	0.206	0.004	4.137 / 4.109	-0.028	[-0.070, 0.017]
lexical	LoRA	3	0.218	0.211	0.009	4.124 / 4.087	-0.037	[-0.087, 0.023]
lexical	rsLoRA	3	0.357	0.341	0.016	4.152 / 1.498	-2.654	[-7.962, 0.000]
sentiment	DoRA	3	1.191	1.190	0.011	2.632 / 3.034	0.402	[-0.003, 1.131]
sentiment	LoRA	3	1.191	1.181	0.010	2.632 / 3.021	0.388	[0.000, 1.135]
sentiment	rsLoRA	3	3.015	3.015	0.000	1.095 / 1.095	0.000	[0.000, 0.000]

Table 6: Pairwise matched-fit end-to-end endpoints for Qwen2.5 Math 7B. Checkpoints are selected using training RMSE only; intervals resample dataset seeds.

Domain	Factor	Seeds	RMSE D	RMSE F	Gap	Witness D / F	F - D	95% CI
arithmetic	DoRA	3	0.434	0.413	0.026	-3.474 / -3.523	-0.049	[-0.076, -0.015]
arithmetic	LoRA	3	0.434	0.430	0.019	-3.474 / -3.497	-0.023	[-0.056, 0.019]
arithmetic	rsLoRA	3	0.434	0.259	0.175	-3.474 / -3.767	-0.293	[-0.554, -0.109]
lexical	DoRA	3	0.405	0.395	0.010	1.680 / 1.395	-0.285	[-0.352, -0.221]
lexical	LoRA	3	0.405	0.418	0.017	1.680 / 1.361	-0.319	[-0.339, -0.283]
lexical	rsLoRA	3	0.405	0.223	0.181	1.680 / 3.827	2.148	[1.970, 2.312]
sentiment	DoRA	3	0.596	0.606	0.010	1.444 / 1.319	-0.125	[-0.155, -0.109]
sentiment	LoRA	3	0.596	0.618	0.024	1.444 / 1.343	-0.101	[-0.110, -0.092]
sentiment	rsLoRA	3	0.596	0.472	0.175	1.444 / -3.270	-4.714	[-5.771, -2.735]

Table 7: Pairwise matched-fit end-to-end endpoints for Qwen3 14B. Checkpoints are selected using training RMSE only; intervals resample dataset seeds.

Domain	Factor	Seeds	RMSE D	RMSE F	Gap	Witness D / F	F - D	95% CI
arithmetic	DoRA	2	0.386	0.512	0.126	-3.168 / -3.126	0.042	[-0.086, 0.169]
arithmetic	LoRA	2	0.386	0.464	0.078	-3.168 / -3.076	0.092	[-0.139, 0.324]
arithmetic	rsLoRA	2	0.340	0.162	0.177	-3.219 / -3.897	-0.678	[-0.889, -0.467]
lexical	DoRA	2	0.310	0.319	0.010	4.041 / 4.064	0.023	[-0.030, 0.077]
lexical	LoRA	2	0.334	0.327	0.013	4.045 / 4.126	0.081	[0.005, 0.156]
lexical	rsLoRA	2	0.310	0.231	0.078	4.041 / 3.764	-0.277	[-0.352, -0.202]
sentiment	DoRA	2	0.338	0.365	0.027	3.949 / 3.878	-0.070	[-0.115, -0.026]
sentiment	LoRA	2	0.366	0.374	0.021	3.925 / 3.940	0.015	[-0.022, 0.052]
sentiment	rsLoRA	2	0.288	0.258	0.053	3.951 / -0.095	-4.046	[-7.615, -0.478]

Table 8: Pairwise matched-fit end-to-end endpoints for Qwen3 30B-A3B. Checkpoints are selected using training RMSE only; intervals resample dataset seeds.

Domain	Factor	Seeds	RMSE D	RMSE F	Gap	Witness D / F	F - D	95% CI
arithmetic	DoRA	3	0.570	0.543	0.044	-1.193 / -2.070	-0.877	[-2.306, 1.436]
arithmetic	LoRA	3	0.531	0.481	0.050	-1.323 / -3.526	-2.202	[-2.792, -1.477]
arithmetic	rsLoRA	3	3.960	3.991	0.041	-0.052 / -0.010	0.042	[0.011, 0.066]
lexical	DoRA	3	0.295	0.330	0.034	3.822 / 3.720	-0.102	[-0.424, 0.107]
lexical	LoRA	3	0.314	0.322	0.034	3.818 / 3.714	-0.104	[-0.348, 0.048]
lexical	rsLoRA	3	3.924	3.934	0.050	0.107 / 0.116	0.009	[-0.106, 0.237]
sentiment	DoRA	3	0.581	0.562	0.031	3.443 / 2.753	-0.691	[-1.328, -0.326]
sentiment	LoRA	3	0.657	0.630	0.071	3.401 / 2.798	-0.603	[-1.079, -0.350]
sentiment	rsLoRA	3	3.974	4.012	0.057	0.115 / 0.010	-0.105	[-0.138, -0.059]

Table 9: Pairwise matched-fit end-to-end endpoints for Qwen3 4B. Checkpoints are selected using training RMSE only; intervals resample dataset seeds.

Domain	Factor	Seeds	RMSE D	RMSE F	Gap	Witness D / F	F - D	95% CI
arithmetic	DoRA	4	0.171	0.165	0.013	-3.919 / -3.938	-0.019	[-0.060, 0.026]
arithmetic	LoRA	4	0.191	0.214	0.034	-3.900 / -3.982	-0.082	[-0.182, 0.000]
arithmetic	rsLoRA	4	0.185	0.112	0.114	-3.903 / -3.954	-0.051	[-0.096, -0.006]
lexical	DoRA	4	0.181	0.169	0.014	3.586 / 3.397	-0.189	[-0.240, -0.114]
lexical	LoRA	4	0.181	0.177	0.009	3.586 / 3.372	-0.214	[-0.267, -0.141]
lexical	rsLoRA	4	0.997	0.895	0.102	2.792 / -3.038	-5.830	[-7.529, -2.529]
sentiment	DoRA	4	0.285	0.271	0.028	3.290 / 1.631	-1.659	[-1.936, -1.430]
sentiment	LoRA	4	0.294	0.277	0.026	3.282 / 2.087	-1.195	[-1.569, -0.574]
sentiment	rsLoRA	4	0.728	0.778	0.313	2.934 / -3.341	-6.274	[-7.453, -4.136]

Table 10: Pairwise matched-fit end-to-end endpoints for Qwen3 8B. Checkpoints are selected using training RMSE only; intervals resample dataset seeds.

Domain	Factor	Seeds	RMSE D	RMSE F	Gap	Witness D / F	F - D	95% CI
arithmetic	DoRA	3	0.304	0.315	0.036	-2.968 / -2.618	0.351	[-1.020, 1.442]
arithmetic	LoRA	3	0.396	0.393	0.037	-2.882 / -2.229	0.653	[-0.518, 1.670]
arithmetic	rsLoRA	3	0.284	0.152	0.140	-2.979 / -4.043	-1.064	[-2.187, -0.369]
lexical	DoRA	3	0.253	0.277	0.024	4.117 / 4.119	0.002	[-0.054, 0.053]
lexical	LoRA	3	0.274	0.285	0.017	4.142 / 4.115	-0.027	[-0.028, -0.026]
lexical	rsLoRA	3	0.237	0.170	0.067	4.115 / 2.596	-1.518	[-3.806, -0.101]
sentiment	DoRA	3	0.350	0.369	0.019	3.972 / 3.884	-0.087	[-0.105, -0.062]
sentiment	LoRA	3	0.380	0.368	0.021	3.957 / 3.877	-0.080	[-0.116, -0.033]
sentiment	rsLoRA	3	0.338	0.216	0.122	3.973 / 1.070	-2.903	[-7.677, -0.189]

Table 11: Rank phase for Qwen2.5 Math 7B on arithmetic. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	LoRA	2	8	0.0e+00	4.0	1.773	-1.655	0.020
0.00000	LoRA	2	32	0.0e+00	4.0	0.493	-3.492	0.000
0.00000	LoRA	2	64	0.0e+00	4.0	0.381	-3.635	0.000
0.00000	LoRA	2	128	0.0e+00	4.0	0.322	-3.696	0.000
0.00000	rsLoRA	2	8	0.0e+00	4.0	0.351	-3.754	0.000
0.00000	rsLoRA	2	32	0.0e+00	4.0	0.189	-3.985	0.000
0.00000	rsLoRA	2	64	0.0e+00	4.0	0.091	-3.941	0.000
0.00000	rsLoRA	2	128	0.0e+00	4.0	1.792	-2.251	0.000
0.25000	LoRA	2	8	0.0e+00	4.0	3.795	-0.220	0.133
0.25000	LoRA	2	32	0.0e+00	4.0	3.146	-0.778	0.215
0.25000	LoRA	2	64	0.0e+00	4.0	2.465	-0.359	0.449
0.25000	LoRA	2	128	0.0e+00	4.0	1.517	0.432	0.602
0.25000	rsLoRA	2	8	0.0e+00	4.0	1.887	0.328	0.562
0.25000	rsLoRA	2	32	0.0e+00	4.0	0.267	1.145	0.695
0.25000	rsLoRA	2	64	0.0e+00	4.0	0.785	0.205	0.617
0.25000	rsLoRA	2	128	0.0e+00	4.0	1.986	1.604	0.500

Table 12: Rank phase for Qwen2.5 Math 7B on sentiment. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	LoRA	2	8	0.0e+00	4.0	0.908	2.113	0.945
0.00000	LoRA	2	32	0.0e+00	4.0	0.689	1.782	0.934
0.00000	LoRA	2	64	0.0e+00	4.0	0.574	1.506	0.934
0.00000	LoRA	2	128	0.0e+00	4.0	0.420	1.296	0.941
0.00000	rsLoRA	2	8	0.0e+00	4.0	0.450	1.466	0.961
0.00000	rsLoRA	2	32	0.0e+00	4.0	0.165	3.223	1.000
0.00000	rsLoRA	2	64	0.0e+00	4.0	0.071	-3.952	0.000
0.00000	rsLoRA	2	128	0.0e+00	4.0	0.043	-3.984	0.000
0.25000	LoRA	2	8	0.0e+00	4.0	0.980	3.151	0.953
0.25000	LoRA	2	32	0.0e+00	4.0	0.749	3.378	0.961
0.25000	LoRA	2	64	0.0e+00	4.0	0.605	3.535	0.977
0.25000	LoRA	2	128	0.0e+00	4.0	0.446	3.692	0.977
0.25000	rsLoRA	2	8	0.0e+00	4.0	0.485	3.702	0.969
0.25000	rsLoRA	2	32	0.0e+00	4.0	0.161	3.820	0.980
0.25000	rsLoRA	2	64	0.0e+00	4.0	0.100	3.845	0.980
0.25000	rsLoRA	2	128	0.0e+00	4.0	1.864	2.163	0.750

M END-TO-END PHASE TABLES

Tables 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23 report every end-to-end rank, target, and decay phase slice. The tables vary rank, target margin, and decay while holding the model, data, witness, and selection rule fixed. Checkpoints are selected by training RMSE before held-out measurement.

Read each cell jointly: RMSE measures fit;
The held-out witness identifies the selected extension, and accuracy records its decision-level conse-

Table 13: Target phase for Qwen2.5 Math 7B on arithmetic. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	Direct	2	64	0.0e+00	2.0	0.506	-1.301	0.004
0.00000	Direct	2	64	0.0e+00	8.0	0.497	-6.953	0.000
0.00000	DoRA	2	64	0.0e+00	2.0	0.377	-1.565	0.000
0.00000	DoRA	2	64	0.0e+00	8.0	0.416	-7.651	0.000
0.00000	LoRA	2	64	0.0e+00	2.0	0.391	-1.518	0.000
0.00000	LoRA	2	64	0.0e+00	8.0	0.483	-7.332	0.000
0.00000	rsLoRA	2	64	0.0e+00	2.0	0.102	-1.884	0.000
0.00000	rsLoRA	2	64	0.0e+00	8.0	0.096	-8.026	0.000
0.25000	Direct	2	64	0.0e+00	2.0	1.662	-0.308	0.164
0.25000	Direct	2	64	0.0e+00	8.0	1.981	1.977	0.695
0.25000	DoRA	2	64	0.0e+00	2.0	1.612	-0.408	0.148
0.25000	DoRA	2	64	0.0e+00	8.0	0.788	2.060	0.680
0.25000	LoRA	2	64	0.0e+00	2.0	1.638	-0.385	0.129
0.25000	LoRA	2	64	0.0e+00	8.0	1.027	1.947	0.680
0.25000	rsLoRA	2	64	0.0e+00	2.0	0.230	0.481	0.715
0.25000	rsLoRA	2	64	0.0e+00	8.0	7.230	-2.594	0.152

Table 14: Target phase for Qwen2.5 Math 7B on sentiment. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	Direct	2	64	0.0e+00	2.0	0.521	0.992	0.945
0.00000	Direct	2	64	0.0e+00	8.0	0.548	6.076	0.992
0.00000	DoRA	2	64	0.0e+00	2.0	0.484	0.904	0.934
0.00000	DoRA	2	64	0.0e+00	8.0	0.418	4.407	0.965
0.00000	LoRA	2	64	0.0e+00	2.0	0.495	0.902	0.930
0.00000	LoRA	2	64	0.0e+00	8.0	0.423	4.325	0.949
0.00000	rsLoRA	2	64	0.0e+00	2.0	0.095	0.622	0.652
0.00000	rsLoRA	2	64	0.0e+00	8.0	0.138	-7.972	0.000
0.25000	Direct	2	64	0.0e+00	2.0	0.520	1.495	0.969
0.25000	Direct	2	64	0.0e+00	8.0	0.504	7.716	0.973
0.25000	DoRA	2	64	0.0e+00	2.0	0.493	1.530	0.965
0.25000	DoRA	2	64	0.0e+00	8.0	0.429	7.565	0.969
0.25000	LoRA	2	64	0.0e+00	2.0	0.504	1.524	0.961
0.25000	LoRA	2	64	0.0e+00	8.0	0.401	7.615	0.969
0.25000	rsLoRA	2	64	0.0e+00	2.0	0.098	1.885	0.984
0.25000	rsLoRA	2	64	0.0e+00	8.0	3.840	4.127	0.750

Table 15: Rank phase for Qwen3 30B-A3B on arithmetic. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	LoRA	1	16	0.0e+00	4.0	0.608	-3.717	0.000
0.00000	LoRA	1	64	0.0e+00	4.0	0.482	-3.155	0.000
0.00000	LoRA	1	128	0.0e+00	4.0	0.392	-1.796	0.010
0.00000	rsLoRA	1	16	0.0e+00	4.0	3.859	-0.100	0.073
0.00000	rsLoRA	1	64	0.0e+00	4.0	4.262	-0.050	0.500
0.00000	rsLoRA	1	128	0.0e+00	4.0	4.002	-0.001	0.500
0.25000	LoRA	1	16	0.0e+00	4.0	2.917	-0.565	0.188
0.25000	LoRA	1	64	0.0e+00	4.0	0.963	1.823	0.823
0.25000	LoRA	1	128	0.0e+00	4.0	0.506	2.375	0.833
0.25000	rsLoRA	1	16	0.0e+00	4.0	0.239	2.773	0.896
0.25000	rsLoRA	1	64	0.0e+00	4.0	3.977	0.024	0.521
0.25000	rsLoRA	1	128	0.0e+00	4.0	4.006	0.000	0.500

quence. The phase atlas extends the matched-fit analysis to a controlled family of training problems. Every run uses the same data construction, checkpoint rule, and evaluation map; only the declared phase coordinate changes. The fitted values therefore describe agreement on observed support, whereas the witness records the continuation selected outside that support. Lower training RMSE can coexist with larger witness displacement because the two quantities measure different geometric components. Accuracy supplies the corresponding projection: its plateaus and abrupt changes locate when a continuous endpoint displacement crosses the task boundary. Together, the three measure-

Table 16: Rank phase for Qwen3 30B-A3B on sentiment. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	LoRA	1	16	0.0e+00	4.0	0.989	1.454	0.885
0.00000	LoRA	1	64	0.0e+00	4.0	0.466	2.566	1.000
0.00000	LoRA	1	128	0.0e+00	4.0	0.291	3.445	1.000
0.00000	rsLoRA	1	16	0.0e+00	4.0	0.210	3.671	1.000
0.00000	rsLoRA	1	64	0.0e+00	4.0	3.990	-0.006	0.448
0.00000	rsLoRA	1	128	0.0e+00	4.0	3.917	0.049	0.490
0.25000	LoRA	1	16	0.0e+00	4.0	1.137	2.234	1.000
0.25000	LoRA	1	64	0.0e+00	4.0	0.753	3.073	1.000
0.25000	LoRA	1	128	0.0e+00	4.0	0.507	3.332	1.000
0.25000	rsLoRA	1	16	0.0e+00	4.0	0.256	3.703	1.000
0.25000	rsLoRA	1	64	0.0e+00	4.0	4.002	0.011	0.500
0.25000	rsLoRA	1	128	0.0e+00	4.0	4.009	0.000	0.500

Table 17: Decay phase for Qwen3 4B on arithmetic. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	Direct	2	64	0.0e+00	4.0	0.161	-3.923	0.000
0.00000	Direct	2	64	1.0e-05	4.0	0.161	-3.923	0.000
0.00000	Direct	2	64	1.0e-04	4.0	0.161	-3.923	0.000
0.00000	Direct	2	64	1.0e-03	4.0	0.161	-3.923	0.000
0.00000	DoRA	2	64	0.0e+00	4.0	0.167	-3.883	0.000
0.00000	DoRA	2	64	1.0e-05	4.0	0.169	-3.887	0.000
0.00000	DoRA	2	64	1.0e-04	4.0	0.198	-3.897	0.000
0.00000	DoRA	2	64	1.0e-03	4.0	0.174	-3.881	0.000
0.00000	LoRA	2	64	0.0e+00	4.0	0.201	-3.875	0.000
0.00000	LoRA	2	64	1.0e-05	4.0	0.200	-3.880	0.000
0.00000	LoRA	2	64	1.0e-04	4.0	0.196	-3.870	0.000
0.00000	LoRA	2	64	1.0e-03	4.0	0.210	-3.903	0.000

Table 18: Decay phase for Qwen3 4B on sentiment. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	Direct	2	64	0.0e+00	4.0	0.270	3.438	1.000
0.00000	Direct	2	64	1.0e-05	4.0	0.270	3.438	1.000
0.00000	Direct	2	64	1.0e-04	4.0	0.270	3.438	1.000
0.00000	Direct	2	64	1.0e-03	4.0	0.270	3.438	1.000
0.00000	DoRA	2	64	0.0e+00	4.0	0.254	1.996	1.000
0.00000	DoRA	2	64	1.0e-05	4.0	0.269	2.039	1.000
0.00000	DoRA	2	64	1.0e-04	4.0	0.261	2.071	1.000
0.00000	DoRA	2	64	1.0e-03	4.0	0.278	2.012	1.000
0.00000	LoRA	2	64	0.0e+00	4.0	0.274	2.481	1.000
0.00000	LoRA	2	64	1.0e-05	4.0	0.271	2.495	1.000
0.00000	LoRA	2	64	1.0e-04	4.0	0.282	2.473	1.000
0.00000	LoRA	2	64	1.0e-03	4.0	0.272	2.661	1.000

ments separate support fit, off-support extension, and discrete behavior. The sweep coordinates intervene on different mechanisms within this common measurement contract. Rank changes both the dimension and the conditioning of the factor chart. Target margin rescales the supervised constraint while preserving examples and labels. Weight decay changes the regularized trajectory and may select a different point from a similar fitted set. Agreement under rank variation locates a stabilized chart, whereas agreement under target variation locates an extension that is stable to constraint scale. Decay measures dependence on the regularized path. The direct update supplies the ambient-coordinate reference, and LoRA, rsLoRA, and DoRA instantiate distinct factor geometries. The atlas consequently tests one endpoint-selection statement through independent perturbations instead of repeating a single hyperparameter comparison. Support coverage connects these interventions to identification. At zero coverage, the training rows leave a witness-visible direction unresolved, so several parameterizations can attain similar observed fit while selecting different continuations. Adding discriminating support restricts that direction and changes the admissible endpoint set for

Table 19: Rank phase for Qwen3 4B on arithmetic. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	LoRA	2	4	0.0e+00	4.0	0.243	-3.905	0.000
0.00000	LoRA	2	8	0.0e+00	4.0	0.263	-3.883	0.000
0.00000	LoRA	2	16	0.0e+00	4.0	0.236	-3.906	0.000
0.00000	LoRA	2	32	0.0e+00	4.0	0.227	-3.850	0.000
0.00000	LoRA	2	64	0.0e+00	4.0	0.201	-3.875	0.000
0.00000	LoRA	2	128	0.0e+00	4.0	0.131	-3.927	0.000
0.00000	rsLoRA	2	4	0.0e+00	4.0	0.204	-3.927	0.000
0.00000	rsLoRA	2	8	0.0e+00	4.0	0.178	-3.894	0.000
0.00000	rsLoRA	2	16	0.0e+00	4.0	0.094	-3.983	0.000
0.00000	rsLoRA	2	32	0.0e+00	4.0	1.822	-2.249	0.000
0.00000	rsLoRA	2	64	0.0e+00	4.0	0.179	-3.924	0.000
0.00000	rsLoRA	2	128	0.0e+00	4.0	3.621	-0.435	0.359
0.25000	LoRA	2	4	0.0e+00	4.0	3.503	-1.681	0.000
0.25000	LoRA	2	8	0.0e+00	4.0	3.483	-1.831	0.000
0.25000	LoRA	2	16	0.0e+00	4.0	3.448	-1.900	0.000
0.25000	LoRA	2	32	0.0e+00	4.0	3.438	-2.042	0.000
0.25000	LoRA	2	64	0.0e+00	4.0	2.145	-0.692	0.414
0.25000	LoRA	2	128	0.0e+00	4.0	1.033	-0.027	0.473
0.25000	rsLoRA	2	4	0.0e+00	4.0	3.456	-2.130	0.000
0.25000	rsLoRA	2	8	0.0e+00	4.0	1.526	0.281	0.594
0.25000	rsLoRA	2	16	0.0e+00	4.0	0.218	0.648	0.590
0.25000	rsLoRA	2	32	0.0e+00	4.0	1.754	0.765	0.496
0.25000	rsLoRA	2	64	0.0e+00	4.0	3.295	-2.348	0.004
0.25000	rsLoRA	2	128	0.0e+00	4.0	4.002	-0.001	0.500

Table 20: Rank phase for Qwen3 4B on sentiment. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	LoRA	2	4	0.0e+00	4.0	0.514	2.139	1.000
0.00000	LoRA	2	8	0.0e+00	4.0	0.440	2.131	1.000
0.00000	LoRA	2	16	0.0e+00	4.0	0.370	1.893	1.000
0.00000	LoRA	2	32	0.0e+00	4.0	0.335	1.799	1.000
0.00000	LoRA	2	64	0.0e+00	4.0	0.274	2.481	1.000
0.00000	LoRA	2	128	0.0e+00	4.0	0.195	2.854	1.000
0.00000	rsLoRA	2	4	0.0e+00	4.0	0.273	2.511	1.000
0.00000	rsLoRA	2	8	0.0e+00	4.0	0.210	2.595	1.000
0.00000	rsLoRA	2	16	0.0e+00	4.0	0.133	3.500	1.000
0.00000	rsLoRA	2	32	0.0e+00	4.0	0.082	3.774	1.000
0.00000	rsLoRA	2	64	0.0e+00	4.0	0.046	-3.893	0.000
0.00000	rsLoRA	2	128	0.0e+00	4.0	2.049	-1.984	0.250
0.25000	LoRA	2	4	0.0e+00	4.0	0.549	3.386	1.000
0.25000	LoRA	2	8	0.0e+00	4.0	0.473	3.441	1.000
0.25000	LoRA	2	16	0.0e+00	4.0	0.463	3.444	1.000
0.25000	LoRA	2	32	0.0e+00	4.0	0.364	3.615	1.000
0.25000	LoRA	2	64	0.0e+00	4.0	0.289	3.737	1.000
0.25000	LoRA	2	128	0.0e+00	4.0	0.232	3.869	1.000
0.25000	rsLoRA	2	4	0.0e+00	4.0	0.308	3.682	1.000
0.25000	rsLoRA	2	8	0.0e+00	4.0	0.263	3.717	1.000
0.25000	rsLoRA	2	16	0.0e+00	4.0	0.155	3.936	1.000
0.25000	rsLoRA	2	32	0.0e+00	4.0	0.093	3.984	1.000
0.25000	rsLoRA	2	64	0.0e+00	4.0	1.794	1.964	0.738
0.25000	rsLoRA	2	128	0.0e+00	4.0	2.376	0.780	0.477

every method at once. Comparisons are therefore made within matched coverage cells and then followed across coverage, rather than pooled across training sets with different information. This organization separates a geometric intervention from a data intervention. Rank, target scale, and decay alter the route through parameter space under fixed support; coverage alters which extensions remain compatible with the observations. Contraction after coverage is added indicates support identification. Separation that persists within a coverage cell locates the remaining influence of parameterization and finite optimization. Training RMSE also distinguishes stable endpoint selection from finite-step optimization. A reversal accompanied by a large residual belongs to the finite training regime, whereas separation among closely fitted endpoints remains an endpoint-selection effect. Rank is especially informative because greater representability can arrive with different factor conditioning.

Table 21: Joint target and rank phase for Qwen3 4B and Qwen3 8B. All original cells are retained; the setting column is target margin for target blocks and factor rank for rank blocks. Each endpoint is selected by training RMSE only.

Coverage	Method	Seeds	Setting	RMSE	Witness	Acc.
Qwen3 4B, target phase, arithmetic						
0.00000	Direct	2	2.0	0.174	-1.894	0.000
0.00000	Direct	2	8.0	0.106	-7.981	0.000
0.00000	DoRA	2	2.0	0.155	-1.947	0.000
0.00000	DoRA	2	8.0	0.162	-7.923	0.000
0.00000	LoRA	2	2.0	0.151	-1.952	0.000
0.00000	LoRA	2	8.0	0.167	-7.960	0.000
0.00000	rsLoRA	2	2.0	0.038	-2.003	0.000
0.00000	rsLoRA	2	8.0	7.866	-0.347	0.500
0.25000	Direct	2	2.0	1.678	-0.836	0.000
0.25000	Direct	2	8.0	4.584	-3.736	0.180
0.25000	DoRA	2	2.0	1.699	-0.904	0.000
0.25000	DoRA	2	8.0	0.853	-0.160	0.535
0.25000	LoRA	2	2.0	1.720	-0.899	0.000
0.25000	LoRA	2	8.0	6.415	-4.034	0.016
0.25000	rsLoRA	2	2.0	1.718	-1.049	0.000
0.25000	rsLoRA	2	8.0	7.961	-0.366	0.496
Qwen3 4B, target phase, sentiment						
0.00000	Direct	2	2.0	0.245	1.018	1.000
0.00000	Direct	2	8.0	0.181	0.453	0.500
0.00000	DoRA	2	2.0	0.261	0.505	0.902
0.00000	DoRA	2	8.0	0.242	6.460	1.000
0.00000	LoRA	2	2.0	0.239	0.507	0.918
0.00000	LoRA	2	8.0	0.188	6.450	1.000
0.00000	rsLoRA	2	2.0	0.045	-1.843	0.000
0.00000	rsLoRA	2	8.0	6.712	-1.294	0.250
0.25000	Direct	2	2.0	0.285	1.779	1.000
0.25000	Direct	2	8.0	0.218	7.900	1.000
0.25000	DoRA	2	2.0	0.300	1.688	1.000
0.25000	DoRA	2	8.0	0.216	7.807	1.000
0.25000	LoRA	2	2.0	0.312	1.679	1.000
0.25000	LoRA	2	8.0	0.344	7.715	1.000
0.25000	rsLoRA	2	2.0	0.098	1.941	1.000
0.25000	rsLoRA	2	8.0	4.088	3.559	0.750
Qwen3 8B, rank phase, arithmetic						
0.00000	LoRA	2	8	1.140	-1.150	0.266
0.00000	LoRA	2	32	0.406	-2.507	0.133
0.00000	LoRA	2	64	0.327	-2.675	0.129
0.00000	LoRA	2	128	0.265	-2.764	0.117
0.00000	rsLoRA	2	8	0.235	-3.092	0.074
0.00000	rsLoRA	2	32	0.143	-2.863	0.023
0.00000	rsLoRA	2	64	0.067	-3.961	0.000
0.00000	rsLoRA	2	128	2.034	-2.036	0.246
0.25000	LoRA	2	8	2.981	0.956	0.719
0.25000	LoRA	2	32	1.729	1.416	0.734
0.25000	LoRA	2	64	0.973	1.682	0.734
0.25000	LoRA	2	128	0.459	1.873	0.758
0.25000	rsLoRA	2	8	0.601	2.101	0.789
0.25000	rsLoRA	2	32	0.359	3.514	0.992
0.25000	rsLoRA	2	64	0.097	3.943	1.000
0.25000	rsLoRA	2	128	3.712	-1.395	0.000
Qwen3 8B, rank phase, sentiment						
0.00000	LoRA	2	8	0.472	3.819	1.000
0.00000	LoRA	2	32	0.441	3.772	1.000
0.00000	LoRA	2	64	0.364	3.764	1.000
0.00000	LoRA	2	128	0.294	3.825	1.000
0.00000	rsLoRA	2	8	0.296	3.847	1.000
0.00000	rsLoRA	2	32	0.154	3.831	1.000
0.00000	rsLoRA	2	64	0.070	-0.418	0.500
0.00000	rsLoRA	2	128	4.028	0.104	0.500
0.25000	LoRA	2	8	0.498	3.914	1.000
0.25000	LoRA	2	32	0.457	3.895	1.000
0.25000	LoRA	2	64	0.410	3.860	1.000
0.25000	LoRA	2	128	0.322	3.873	1.000
0.25000	rsLoRA	2	8	0.318	3.880	1.000
0.25000	rsLoRA	2	32	0.140	3.886	1.000
0.25000	rsLoRA	2	64	0.100	3.948	1.000
0.25000	rsLoRA	2	128	3.906	-0.054	0.328

The same rank increase may tighten fit in one chart and destabilize the trajectory in another. Target

Table 22: Target phase for Qwen3 8B on arithmetic. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	Direct	2	64	0.0e+00	2.0	0.288	-1.193	0.117
0.00000	Direct	2	64	0.0e+00	8.0	0.171	-7.654	0.008
0.00000	DoRA	2	64	0.0e+00	2.0	0.411	-1.106	0.133
0.00000	DoRA	2	64	0.0e+00	8.0	0.390	-6.605	0.066
0.00000	LoRA	2	64	0.0e+00	2.0	0.486	-0.929	0.172
0.00000	LoRA	2	64	0.0e+00	8.0	3.629	-3.564	0.020
0.00000	rsLoRA	2	64	0.0e+00	2.0	0.109	-1.879	0.000
0.00000	rsLoRA	2	64	0.0e+00	8.0	3.586	-4.540	0.250
0.25000	Direct	2	64	0.0e+00	2.0	0.957	0.583	0.699
0.25000	Direct	2	64	0.0e+00	8.0	4.499	2.232	0.535
0.25000	DoRA	2	64	0.0e+00	2.0	1.227	0.500	0.684
0.25000	DoRA	2	64	0.0e+00	8.0	0.904	5.294	0.859
0.25000	LoRA	2	64	0.0e+00	2.0	1.310	0.529	0.680
0.25000	LoRA	2	64	0.0e+00	8.0	0.706	5.164	0.859
0.25000	rsLoRA	2	64	0.0e+00	2.0	0.109	1.815	0.996
0.25000	rsLoRA	2	64	0.0e+00	8.0	6.733	-4.634	0.000

Table 23: Target phase for Qwen3 8B on sentiment. Each endpoint is the training-only best-fit checkpoint for that run.

Coverage	Method	Seeds	Rank	Decay	Target	RMSE	Witness	Acc.
0.00000	Direct	2	64	0.0e+00	2.0	0.274	1.859	1.000
0.00000	Direct	2	64	0.0e+00	8.0	0.277	7.702	1.000
0.00000	DoRA	2	64	0.0e+00	2.0	0.312	1.625	1.000
0.00000	DoRA	2	64	0.0e+00	8.0	0.311	7.669	1.000
0.00000	LoRA	2	64	0.0e+00	2.0	0.307	1.675	1.000
0.00000	LoRA	2	64	0.0e+00	8.0	0.313	7.685	1.000
0.00000	rsLoRA	2	64	0.0e+00	2.0	0.086	1.842	1.000
0.00000	rsLoRA	2	64	0.0e+00	8.0	3.990	-3.875	0.250
0.25000	Direct	2	64	0.0e+00	2.0	0.310	1.905	1.000
0.25000	Direct	2	64	0.0e+00	8.0	0.266	7.859	1.000
0.25000	DoRA	2	64	0.0e+00	2.0	0.349	1.878	1.000
0.25000	DoRA	2	64	0.0e+00	8.0	0.275	7.906	1.000
0.25000	LoRA	2	64	0.0e+00	2.0	0.349	1.910	1.000
0.25000	LoRA	2	64	0.0e+00	8.0	0.272	7.851	1.000
0.25000	rsLoRA	2	64	0.0e+00	2.0	0.093	1.964	1.000
0.25000	rsLoRA	2	64	0.0e+00	8.0	4.058	3.902	0.750

scaling provides a complementary diagnostic. A witness that follows the target while preserving its qualitative ordering reflects a stable response to constraint scale; a sign change paired with deteriorating fit marks departure from that regime. Weight decay supplies a third route by altering the selected regularized solution. Concordant changes in fit, witness, and accuracy locate the phase boundary directly, without converting the analysis into a thresholded success count. Model scale and task domain then change the geometry on which these interventions act. Arithmetic and sentiment expose different witness directions, so a parameterization can occupy different sides of the endpoint boundary under the same protocol. Scaling the base model changes the local feature map and its conditioning, while preserving the identification question posed in every cell: after matching the observed constraints, which continuation is selected? Repeating the phase coordinates across model families localizes a change to representation, supervision scale, regularization, or support information. It also makes the larger models new geometric instances of the same test rather than simple replication targets. The atlas is interpreted at the level of matched cells. For a fixed model, domain, coverage value, and phase setting, the direct and factorized updates face the same observed constraints and are evaluated by the same witness. Checkpoint selection uses training RMSE alone, so the held-out quantities remain measurements rather than selection criteria. The first comparison asks whether the methods reach comparable fit. The second asks whether their witnesses agree in sign and magnitude. The third follows any witness difference to the discrete task boundary through accuracy. This ordering preserves the distinction between fitting, functional continuation, and task behavior. Consistency across coordinates provides an additional mechanism check. A separation caused by unresolved support should contract when discriminating coverage is added across otherwise matched charts. A separation caused by chart geometry can persist under fixed coverage while changing systematically

with rank or parameterization scale. Target variation tests whether the selected continuation responds smoothly to supervision magnitude, and decay variation tests whether regularized path selection changes that continuation. Finite-step effects have a different signature: the witness reversal arrives together with deteriorating fit or an abrupt loss of phase continuity. These signatures organize the numerical rows into support, geometry, regularization, and optimization regimes. Agreement across sweep axes ties endpoint selection to visible coordinate changes. The continuous witness and discrete accuracy remain complementary throughout this analysis. The witness detects endpoint displacement before a predicted label changes; accuracy records the later crossing of a task boundary. Their joint movement reveals both latent functional separation and its behavioral consequence. This is the empirical form of the paper's distinction between interpolation and extension. Training fit establishes satisfaction of the observed constraints, support coverage determines how much continuation freedom remains, and parameterization determines which point within that freedom is reached. The phase atlas carries the exact endpoint construction into the exhaustive model-wise sweeps. Remaining model slices retain the fit and witness contract. Together, they complete the shared empirical identification protocol. Their common evaluation map keeps these additional slices directly comparable to the preceding phase coordinates.

N Model-wise Endpoint Sweeps

Here Acc. denotes sign accuracy, Span resid. is the witness component outside the training span, and Max error is the maximum training-fit error across the aggregated runs. Tables 24–53 report the anchored sweeps; Tables 54–83 report the natural-feature sweeps. Each group retains every model and domain in the declared main slice.

This atlas is the disaggregated counterpart of the pooled main-text results. Each table keeps the model family and semantic domain fixed, then varies only conflict-support coverage and the declared selector coordinate. Rows should therefore be compared horizontally within a table before any comparison is made across model scales. The anchored carrier supplies an exact equality test: every displayed endpoint interpolates the same training targets to numerical precision, so variation in the coefficient difference or held-out witness is attributable to selection inside the training-null space. The natural-feature carrier asks how much of the same geometry remains after the exact two-direction construction is replaced by the model's full hidden representation. Three recurring patterns organize the entries. At zero conflict coverage, initialization balance can reverse the anchored endpoint while preserving fit. Once even a small conflict-support slice exposes the held-out direction, the anchored alternatives collapse to a common endpoint. In the natural carrier, the row-space residual measures how much of the evaluation witness remains outside the span identified by training. Selector differences are largest when that residual is available and shrink as the witness is absorbed into the observed support. This is the empirical form of the support-identifiability theorem rather than a model-specific accuracy comparison. The tables also separate invariances from genuine coordinate effects. Direct interpolation is unchanged by a scalar rescaling of its metric, whereas signed-square initialization changes the implicit potential and can alter the null-space completion. The three initialization scales in the natural carrier trace the interpolation between the sparse and quadratic limiting geometries. Agreement among those rows is informative: it locates a regime where the available null-space direction is too weak, or the training support is already sufficient, for parameterization to change the measured endpoint. Disagreement locates a selectable fiber and is evaluated by its signed witness rather than by choosing the best row after inspection. Model and domain replication remains explicit because representation geometry is part of the experimental object. A pooled average can conceal sign reversals or model-specific collapse, whereas one striking cell does not establish a cross-model pattern. The tables distinguish exact carrier laws from empirical tendencies and contingent boundary cases.

The coverage grid is dense near zero to resolve the predicted identifiability transition without a manually chosen cutoff. Anchored rows show the exact boundary of the construction; natural rows show the graded transition in empirical representations. Small selector differences are retained as evidence for the regime in which support information dominates implicit selection. Finally, accuracy discretizes the signed witness: it can hide substantial endpoint motion or change abruptly at a decision boundary. Constant accuracy therefore does not establish selector equivalence when the continuous held-out witness still differs. We therefore retain the witness's direction and magnitude as the primary link to the theory.

Table 24: Anchored endpoint sweep for Qwen2.5 1.5B Instruct on the arithmetic carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.279	0.311	9.3e-16
0.00000	Signed square	0.01	0.1	-0.980	-0.979	0.000	2.4e-12
0.00000	Signed square	0.01	1.0	0.000	-0.279	0.311	9.6e-11
0.00000	Signed square	0.01	10.0	0.980	0.422	0.689	1.8e-12
0.03125	Direct	1.00	1.0	1.000	0.436	0.690	6.6e-14
0.03125	Signed square	0.01	0.1	1.000	0.436	0.690	9.5e-11
0.03125	Signed square	0.01	1.0	1.000	0.436	0.690	8.0e-11
0.03125	Signed square	0.01	10.0	1.000	0.436	0.690	1.3e-11
0.06250	Direct	1.00	1.0	1.000	0.436	0.690	2.5e-15
0.06250	Signed square	0.01	0.1	1.000	0.436	0.690	1.0e-10
0.06250	Signed square	0.01	1.0	1.000	0.436	0.690	9.7e-11
0.06250	Signed square	0.01	10.0	1.000	0.436	0.690	1.0e-11
0.12500	Direct	1.00	1.0	1.000	0.436	0.690	1.8e-15
0.12500	Signed square	0.01	0.1	1.000	0.436	0.690	9.9e-11
0.12500	Signed square	0.01	1.0	1.000	0.436	0.690	1.0e-10
0.12500	Signed square	0.01	10.0	1.000	0.436	0.690	8.2e-12
0.25000	Direct	1.00	1.0	1.000	0.436	0.690	2.6e-15
0.25000	Signed square	0.01	0.1	1.000	0.436	0.690	9.7e-11
0.25000	Signed square	0.01	1.0	1.000	0.436	0.690	9.9e-11
0.25000	Signed square	0.01	10.0	1.000	0.436	0.690	5.7e-12
0.50000	Direct	1.00	1.0	1.000	0.436	0.690	2.9e-15
0.50000	Signed square	0.01	0.1	1.000	0.436	0.690	9.1e-11
0.50000	Signed square	0.01	1.0	1.000	0.436	0.690	7.8e-11
0.50000	Signed square	0.01	10.0	1.000	0.436	0.690	6.0e-12

Table 25: Anchored endpoint sweep for Qwen2.5 1.5B Instruct on the lexical carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.078	0.260	1.1e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.972	0.000	1.2e-12
0.00000	Signed square	0.01	1.0	0.000	-0.078	0.260	9.4e-11
0.00000	Signed square	0.01	10.0	0.980	0.817	1.000	8.8e-13
0.03125	Direct	1.00	1.0	1.000	0.835	1.000	2.0e-15
0.03125	Signed square	0.01	0.1	1.000	0.835	1.000	9.9e-11
0.03125	Signed square	0.01	1.0	1.000	0.835	1.000	9.8e-11
0.03125	Signed square	0.01	10.0	1.000	0.835	1.000	4.2e-12
0.06250	Direct	1.00	1.0	1.000	0.835	1.000	1.5e-15
0.06250	Signed square	0.01	0.1	1.000	0.835	1.000	9.7e-11
0.06250	Signed square	0.01	1.0	1.000	0.835	1.000	9.8e-11
0.06250	Signed square	0.01	10.0	1.000	0.835	1.000	4.8e-12
0.12500	Direct	1.00	1.0	1.000	0.835	1.000	1.3e-15
0.12500	Signed square	0.01	0.1	1.000	0.835	1.000	9.4e-11
0.12500	Signed square	0.01	1.0	1.000	0.835	1.000	9.6e-11
0.12500	Signed square	0.01	10.0	1.000	0.835	1.000	4.8e-12
0.25000	Direct	1.00	1.0	1.000	0.835	1.000	1.5e-15
0.25000	Signed square	0.01	0.1	1.000	0.835	1.000	9.9e-11
0.25000	Signed square	0.01	1.0	1.000	0.835	1.000	8.9e-11
0.25000	Signed square	0.01	10.0	1.000	0.835	1.000	4.4e-12
0.50000	Direct	1.00	1.0	1.000	0.835	1.000	1.1e-15
0.50000	Signed square	0.01	0.1	1.000	0.835	1.000	1.0e-10
0.50000	Signed square	0.01	1.0	1.000	0.835	1.000	8.4e-11
0.50000	Signed square	0.01	10.0	1.000	0.835	1.000	4.7e-12

Table 26: Anchored endpoint sweep for Qwen2.5 1.5B Instruct on the sentiment carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.310	0.199	1.0e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.932	0.000	1.5e-12
0.00000	Signed square	0.01	1.0	0.000	-0.310	0.199	9.5e-11
0.00000	Signed square	0.01	10.0	0.980	0.313	0.677	1.1e-12
0.03125	Direct	1.00	1.0	1.000	0.326	0.682	2.7e-15
0.03125	Signed square	0.01	0.1	1.000	0.326	0.682	9.2e-11
0.03125	Signed square	0.01	1.0	1.000	0.326	0.682	7.3e-11
0.03125	Signed square	0.01	10.0	1.000	0.326	0.682	5.2e-12
0.06250	Direct	1.00	1.0	1.000	0.326	0.682	1.8e-15
0.06250	Signed square	0.01	0.1	1.000	0.326	0.682	9.8e-11
0.06250	Signed square	0.01	1.0	1.000	0.326	0.682	8.0e-11
0.06250	Signed square	0.01	10.0	1.000	0.326	0.682	5.5e-12
0.12500	Direct	1.00	1.0	1.000	0.326	0.682	1.8e-15
0.12500	Signed square	0.01	0.1	1.000	0.326	0.682	9.5e-11
0.12500	Signed square	0.01	1.0	1.000	0.326	0.682	9.7e-11
0.12500	Signed square	0.01	10.0	1.000	0.326	0.682	4.2e-12
0.25000	Direct	1.00	1.0	1.000	0.326	0.682	1.3e-15
0.25000	Signed square	0.01	0.1	1.000	0.326	0.682	9.6e-11
0.25000	Signed square	0.01	1.0	1.000	0.326	0.682	9.6e-11
0.25000	Signed square	0.01	10.0	1.000	0.326	0.682	3.6e-12
0.50000	Direct	1.00	1.0	1.000	0.326	0.682	1.3e-15
0.50000	Signed square	0.01	0.1	1.000	0.326	0.682	8.3e-11
0.50000	Signed square	0.01	1.0	1.000	0.326	0.682	9.9e-11
0.50000	Signed square	0.01	10.0	1.000	0.326	0.682	5.0e-12

Table 27: Anchored endpoint sweep for Qwen2.5 7B on the arithmetic carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.184	0.329	1.1e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.993	0.000	1.9e-12
0.00000	Signed square	0.01	1.0	0.000	-0.184	0.329	9.4e-11
0.00000	Signed square	0.01	10.0	0.980	0.626	0.762	1.8e-12
0.03125	Direct	1.00	1.0	1.000	0.642	0.766	1.4e-15
0.03125	Signed square	0.01	0.1	1.000	0.642	0.766	9.5e-11
0.03125	Signed square	0.01	1.0	1.000	0.642	0.766	1.0e-10
0.03125	Signed square	0.01	10.0	1.000	0.642	0.766	5.7e-12
0.06250	Direct	1.00	1.0	1.000	0.642	0.766	1.6e-15
0.06250	Signed square	0.01	0.1	1.000	0.642	0.766	9.8e-11
0.06250	Signed square	0.01	1.0	1.000	0.642	0.766	9.8e-11
0.06250	Signed square	0.01	10.0	1.000	0.642	0.766	5.6e-12
0.12500	Direct	1.00	1.0	1.000	0.642	0.766	1.4e-15
0.12500	Signed square	0.01	0.1	1.000	0.642	0.766	9.8e-11
0.12500	Signed square	0.01	1.0	1.000	0.642	0.766	9.8e-11
0.12500	Signed square	0.01	10.0	1.000	0.642	0.766	5.8e-12
0.25000	Direct	1.00	1.0	1.000	0.642	0.766	1.2e-15
0.25000	Signed square	0.01	0.1	1.000	0.642	0.766	9.9e-11
0.25000	Signed square	0.01	1.0	1.000	0.642	0.766	9.6e-11
0.25000	Signed square	0.01	10.0	1.000	0.642	0.766	5.7e-12
0.50000	Direct	1.00	1.0	1.000	0.642	0.766	1.0e-15
0.50000	Signed square	0.01	0.1	1.000	0.642	0.766	9.8e-11
0.50000	Signed square	0.01	1.0	1.000	0.642	0.766	9.2e-11
0.50000	Signed square	0.01	10.0	1.000	0.642	0.766	8.4e-12

Table 28: Anchored endpoint sweep for Qwen2.5 7B on the lexical carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.058	0.334	1.1e-15
0.00000	Signed square	0.01	0.1	-0.980	-1.089	0.000	1.4e-12
0.00000	Signed square	0.01	1.0	0.000	-0.058	0.334	1.0e-10
0.00000	Signed square	0.01	10.0	0.980	0.973	1.000	1.2e-12
0.03125	Direct	1.00	1.0	1.000	0.994	1.000	1.5e-15
0.03125	Signed square	0.01	0.1	1.000	0.994	1.000	9.6e-11
0.03125	Signed square	0.01	1.0	1.000	0.994	1.000	9.8e-11
0.03125	Signed square	0.01	10.0	1.000	0.994	1.000	4.0e-12
0.06250	Direct	1.00	1.0	1.000	0.994	1.000	1.2e-15
0.06250	Signed square	0.01	0.1	1.000	0.994	1.000	1.0e-10
0.06250	Signed square	0.01	1.0	1.000	0.994	1.000	8.6e-11
0.06250	Signed square	0.01	10.0	1.000	0.994	1.000	3.4e-12
0.12500	Direct	1.00	1.0	1.000	0.994	1.000	1.2e-15
0.12500	Signed square	0.01	0.1	1.000	0.994	1.000	1.0e-10
0.12500	Signed square	0.01	1.0	1.000	0.994	1.000	9.7e-11
0.12500	Signed square	0.01	10.0	1.000	0.994	1.000	3.4e-12
0.25000	Direct	1.00	1.0	1.000	0.994	1.000	1.1e-15
0.25000	Signed square	0.01	0.1	1.000	0.994	1.000	9.7e-11
0.25000	Signed square	0.01	1.0	1.000	0.994	1.000	9.7e-11
0.25000	Signed square	0.01	10.0	1.000	0.994	1.000	4.3e-12
0.50000	Direct	1.00	1.0	1.000	0.994	1.000	1.4e-15
0.50000	Signed square	0.01	0.1	1.000	0.994	1.000	8.7e-11
0.50000	Signed square	0.01	1.0	1.000	0.994	1.000	9.5e-11
0.50000	Signed square	0.01	10.0	1.000	0.994	1.000	4.1e-12

Table 29: Anchored endpoint sweep for Qwen2.5 7B on the sentiment carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.084	0.223	8.9e-16
0.00000	Signed square	0.01	0.1	-0.980	-1.017	0.021	1.3e-12
0.00000	Signed square	0.01	1.0	0.000	-0.084	0.223	9.9e-11
0.00000	Signed square	0.01	10.0	0.980	0.849	0.891	1.3e-12
0.03125	Direct	1.00	1.0	1.000	0.868	0.891	2.4e-15
0.03125	Signed square	0.01	0.1	1.000	0.868	0.891	9.5e-11
0.03125	Signed square	0.01	1.0	1.000	0.868	0.891	9.6e-11
0.03125	Signed square	0.01	10.0	1.000	0.868	0.891	4.5e-12
0.06250	Direct	1.00	1.0	1.000	0.868	0.891	9.8e-16
0.06250	Signed square	0.01	0.1	1.000	0.868	0.891	9.5e-11
0.06250	Signed square	0.01	1.0	1.000	0.868	0.891	9.8e-11
0.06250	Signed square	0.01	10.0	1.000	0.868	0.891	4.8e-12
0.12500	Direct	1.00	1.0	1.000	0.868	0.891	1.0e-15
0.12500	Signed square	0.01	0.1	1.000	0.868	0.891	1.0e-10
0.12500	Signed square	0.01	1.0	1.000	0.868	0.891	1.0e-10
0.12500	Signed square	0.01	10.0	1.000	0.868	0.891	8.9e-12
0.25000	Direct	1.00	1.0	1.000	0.868	0.891	1.1e-15
0.25000	Signed square	0.01	0.1	1.000	0.868	0.891	9.5e-11
0.25000	Signed square	0.01	1.0	1.000	0.868	0.891	9.2e-11
0.25000	Signed square	0.01	10.0	1.000	0.868	0.891	5.9e-12
0.50000	Direct	1.00	1.0	1.000	0.868	0.891	2.0e-15
0.50000	Signed square	0.01	0.1	1.000	0.868	0.891	9.0e-11
0.50000	Signed square	0.01	1.0	1.000	0.868	0.891	9.7e-11
0.50000	Signed square	0.01	10.0	1.000	0.868	0.891	5.7e-12

Table 30: Anchored endpoint sweep for Qwen2.5 7B Instruct on the arithmetic carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.186	0.293	1.1e-15
0.00000	Signed square	0.01	0.1	-0.980	-1.010	0.000	1.3e-12
0.00000	Signed square	0.01	1.0	0.000	-0.186	0.293	9.9e-11
0.00000	Signed square	0.01	10.0	0.980	0.638	0.789	1.6e-12
0.03125	Direct	1.00	1.0	1.000	0.655	0.792	1.8e-15
0.03125	Signed square	0.01	0.1	1.000	0.655	0.792	7.9e-11
0.03125	Signed square	0.01	1.0	1.000	0.655	0.792	9.6e-11
0.03125	Signed square	0.01	10.0	1.000	0.655	0.792	5.8e-12
0.06250	Direct	1.00	1.0	1.000	0.655	0.792	1.5e-15
0.06250	Signed square	0.01	0.1	1.000	0.655	0.792	9.8e-11
0.06250	Signed square	0.01	1.0	1.000	0.655	0.792	9.8e-11
0.06250	Signed square	0.01	10.0	1.000	0.655	0.792	5.8e-12
0.12500	Direct	1.00	1.0	1.000	0.655	0.792	1.3e-15
0.12500	Signed square	0.01	0.1	1.000	0.655	0.792	1.0e-10
0.12500	Signed square	0.01	1.0	1.000	0.655	0.792	9.9e-11
0.12500	Signed square	0.01	10.0	1.000	0.655	0.792	5.4e-12
0.25000	Direct	1.00	1.0	1.000	0.655	0.792	1.3e-15
0.25000	Signed square	0.01	0.1	1.000	0.655	0.792	1.0e-10
0.25000	Signed square	0.01	1.0	1.000	0.655	0.792	9.6e-11
0.25000	Signed square	0.01	10.0	1.000	0.655	0.792	5.5e-12
0.50000	Direct	1.00	1.0	1.000	0.655	0.792	1.1e-15
0.50000	Signed square	0.01	0.1	1.000	0.655	0.792	9.8e-11
0.50000	Signed square	0.01	1.0	1.000	0.655	0.792	9.9e-11
0.50000	Signed square	0.01	10.0	1.000	0.655	0.792	5.2e-12

Table 31: Anchored endpoint sweep for Qwen2.5 7B Instruct on the lexical carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	0.129	0.786	1.1e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.704	0.000	9.2e-13
0.00000	Signed square	0.01	1.0	0.000	0.129	0.786	9.2e-11
0.00000	Signed square	0.01	10.0	0.980	0.962	1.000	1.1e-12
0.03125	Direct	1.00	1.0	1.000	0.979	1.000	1.4e-15
0.03125	Signed square	0.01	0.1	1.000	0.979	1.000	9.3e-11
0.03125	Signed square	0.01	1.0	1.000	0.979	1.000	9.8e-11
0.03125	Signed square	0.01	10.0	1.000	0.979	1.000	3.7e-12
0.06250	Direct	1.00	1.0	1.000	0.979	1.000	1.5e-15
0.06250	Signed square	0.01	0.1	1.000	0.979	1.000	9.7e-11
0.06250	Signed square	0.01	1.0	1.000	0.979	1.000	1.0e-10
0.06250	Signed square	0.01	10.0	1.000	0.979	1.000	4.9e-12
0.12500	Direct	1.00	1.0	1.000	0.979	1.000	1.4e-15
0.12500	Signed square	0.01	0.1	1.000	0.979	1.000	9.5e-11
0.12500	Signed square	0.01	1.0	1.000	0.979	1.000	9.8e-11
0.12500	Signed square	0.01	10.0	1.000	0.979	1.000	5.3e-12
0.25000	Direct	1.00	1.0	1.000	0.979	1.000	1.4e-15
0.25000	Signed square	0.01	0.1	1.000	0.979	1.000	9.7e-11
0.25000	Signed square	0.01	1.0	1.000	0.979	1.000	9.9e-11
0.25000	Signed square	0.01	10.0	1.000	0.979	1.000	4.1e-12
0.50000	Direct	1.00	1.0	1.000	0.979	1.000	1.2e-15
0.50000	Signed square	0.01	0.1	1.000	0.979	1.000	9.4e-11
0.50000	Signed square	0.01	1.0	1.000	0.979	1.000	9.2e-11
0.50000	Signed square	0.01	10.0	1.000	0.979	1.000	4.5e-12

Table 32: Anchored endpoint sweep for Qwen2.5 7B Instruct on the sentiment carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	0.107	0.618	9.0e-16
0.00000	Signed square	0.01	0.1	-0.980	-0.721	0.106	1.3e-12
0.00000	Signed square	0.01	1.0	0.000	0.107	0.618	1.0e-10
0.00000	Signed square	0.01	10.0	0.980	0.935	1.000	1.4e-12
0.03125	Direct	1.00	1.0	1.000	0.952	1.000	1.3e-15
0.03125	Signed square	0.01	0.1	1.000	0.952	1.000	9.7e-11
0.03125	Signed square	0.01	1.0	1.000	0.952	1.000	9.4e-11
0.03125	Signed square	0.01	10.0	1.000	0.952	1.000	4.2e-12
0.06250	Direct	1.00	1.0	1.000	0.952	1.000	1.3e-15
0.06250	Signed square	0.01	0.1	1.000	0.952	1.000	9.6e-11
0.06250	Signed square	0.01	1.0	1.000	0.952	1.000	9.7e-11
0.06250	Signed square	0.01	10.0	1.000	0.952	1.000	3.9e-12
0.12500	Direct	1.00	1.0	1.000	0.952	1.000	1.5e-15
0.12500	Signed square	0.01	0.1	1.000	0.952	1.000	9.5e-11
0.12500	Signed square	0.01	1.0	1.000	0.952	1.000	9.7e-11
0.12500	Signed square	0.01	10.0	1.000	0.952	1.000	4.6e-12
0.25000	Direct	1.00	1.0	1.000	0.952	1.000	1.4e-15
0.25000	Signed square	0.01	0.1	1.000	0.952	1.000	1.0e-10
0.25000	Signed square	0.01	1.0	1.000	0.952	1.000	9.7e-11
0.25000	Signed square	0.01	10.0	1.000	0.952	1.000	5.2e-12
0.50000	Direct	1.00	1.0	1.000	0.952	1.000	1.1e-15
0.50000	Signed square	0.01	0.1	1.000	0.952	1.000	9.6e-11
0.50000	Signed square	0.01	1.0	1.000	0.952	1.000	9.5e-11
0.50000	Signed square	0.01	10.0	1.000	0.952	1.000	3.2e-12

Table 33: Anchored endpoint sweep for Qwen2.5 Math 7B on the arithmetic carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.668	0.060	1.0e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.953	0.000	2.0e-12
0.00000	Signed square	0.01	1.0	0.000	-0.668	0.060	9.5e-11
0.00000	Signed square	0.01	10.0	0.980	-0.384	0.283	9.0e-13
0.03125	Direct	1.00	1.0	1.000	-0.378	0.289	2.9e-13
0.03125	Signed square	0.01	0.1	1.000	-0.378	0.289	9.4e-11
0.03125	Signed square	0.01	1.0	1.000	-0.378	0.289	9.6e-11
0.03125	Signed square	0.01	10.0	1.000	-0.378	0.289	4.9e-12
0.06250	Direct	1.00	1.0	1.000	-0.378	0.289	3.5e-15
0.06250	Signed square	0.01	0.1	1.000	-0.378	0.289	9.9e-11
0.06250	Signed square	0.01	1.0	1.000	-0.378	0.289	9.8e-11
0.06250	Signed square	0.01	10.0	1.000	-0.378	0.289	4.0e-12
0.12500	Direct	1.00	1.0	1.000	-0.378	0.289	2.1e-15
0.12500	Signed square	0.01	0.1	1.000	-0.378	0.289	9.1e-11
0.12500	Signed square	0.01	1.0	1.000	-0.378	0.289	9.1e-11
0.12500	Signed square	0.01	10.0	1.000	-0.378	0.289	6.2e-12
0.25000	Direct	1.00	1.0	1.000	-0.378	0.289	1.2e-15
0.25000	Signed square	0.01	0.1	1.000	-0.378	0.289	8.7e-11
0.25000	Signed square	0.01	1.0	1.000	-0.378	0.289	9.9e-11
0.25000	Signed square	0.01	10.0	1.000	-0.378	0.289	4.0e-12
0.50000	Direct	1.00	1.0	1.000	-0.378	0.289	1.1e-15
0.50000	Signed square	0.01	0.1	1.000	-0.378	0.289	9.7e-11
0.50000	Signed square	0.01	1.0	1.000	-0.378	0.289	9.3e-11
0.50000	Signed square	0.01	10.0	1.000	-0.378	0.289	3.8e-12

Table 34: Anchored endpoint sweep for Qwen2.5 Math 7B on the lexical carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.099	0.436	1.1e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.518	0.081	1.5e-12
0.00000	Signed square	0.01	1.0	0.000	-0.099	0.436	9.6e-11
0.00000	Signed square	0.01	10.0	0.980	0.320	0.766	9.8e-13
0.03125	Direct	1.00	1.0	1.000	0.328	0.766	2.0e-15
0.03125	Signed square	0.01	0.1	1.000	0.328	0.766	9.4e-11
0.03125	Signed square	0.01	1.0	1.000	0.328	0.766	8.6e-11
0.03125	Signed square	0.01	10.0	1.000	0.328	0.766	5.5e-12
0.06250	Direct	1.00	1.0	1.000	0.328	0.766	1.4e-15
0.06250	Signed square	0.01	0.1	1.000	0.328	0.766	9.6e-11
0.06250	Signed square	0.01	1.0	1.000	0.328	0.766	9.8e-11
0.06250	Signed square	0.01	10.0	1.000	0.328	0.766	5.3e-12
0.12500	Direct	1.00	1.0	1.000	0.328	0.766	1.4e-15
0.12500	Signed square	0.01	0.1	1.000	0.328	0.766	9.7e-11
0.12500	Signed square	0.01	1.0	1.000	0.328	0.766	9.9e-11
0.12500	Signed square	0.01	10.0	1.000	0.328	0.766	4.3e-12
0.25000	Direct	1.00	1.0	1.000	0.328	0.766	1.3e-15
0.25000	Signed square	0.01	0.1	1.000	0.328	0.766	9.9e-11
0.25000	Signed square	0.01	1.0	1.000	0.328	0.766	1.0e-10
0.25000	Signed square	0.01	10.0	1.000	0.328	0.766	4.8e-12
0.50000	Direct	1.00	1.0	1.000	0.328	0.766	1.2e-15
0.50000	Signed square	0.01	0.1	1.000	0.328	0.766	9.7e-11
0.50000	Signed square	0.01	1.0	1.000	0.328	0.766	9.6e-11
0.50000	Signed square	0.01	10.0	1.000	0.328	0.766	5.1e-12

Table 35: Anchored endpoint sweep for Qwen2.5 Math 7B on the sentiment carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.078	0.391	1.0e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.928	0.000	1.2e-12
0.00000	Signed square	0.01	1.0	0.000	-0.078	0.391	9.0e-11
0.00000	Signed square	0.01	10.0	0.980	0.771	0.948	9.0e-13
0.03125	Direct	1.00	1.0	1.000	0.788	0.957	1.2e-15
0.03125	Signed square	0.01	0.1	1.000	0.788	0.957	9.4e-11
0.03125	Signed square	0.01	1.0	1.000	0.788	0.957	7.6e-11
0.03125	Signed square	0.01	10.0	1.000	0.788	0.957	4.4e-12
0.06250	Direct	1.00	1.0	1.000	0.788	0.957	1.3e-15
0.06250	Signed square	0.01	0.1	1.000	0.788	0.957	9.6e-11
0.06250	Signed square	0.01	1.0	1.000	0.788	0.957	4.4e-11
0.06250	Signed square	0.01	10.0	1.000	0.788	0.957	4.1e-12
0.12500	Direct	1.00	1.0	1.000	0.788	0.957	1.6e-15
0.12500	Signed square	0.01	0.1	1.000	0.788	0.957	8.2e-11
0.12500	Signed square	0.01	1.0	1.000	0.788	0.957	7.7e-11
0.12500	Signed square	0.01	10.0	1.000	0.788	0.957	5.9e-12
0.25000	Direct	1.00	1.0	1.000	0.788	0.957	1.4e-15
0.25000	Signed square	0.01	0.1	1.000	0.788	0.957	9.9e-11
0.25000	Signed square	0.01	1.0	1.000	0.788	0.957	9.4e-11
0.25000	Signed square	0.01	10.0	1.000	0.788	0.957	4.4e-12
0.50000	Direct	1.00	1.0	1.000	0.788	0.957	1.0e-15
0.50000	Signed square	0.01	0.1	1.000	0.788	0.957	9.8e-11
0.50000	Signed square	0.01	1.0	1.000	0.788	0.957	7.6e-11
0.50000	Signed square	0.01	10.0	1.000	0.788	0.957	3.8e-12

Table 36: Anchored endpoint sweep for Qwen2.5 Math 7B Instruct on the arithmetic carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.441	0.139	1.0e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.959	0.000	1.1e-12
0.00000	Signed square	0.01	1.0	0.000	-0.441	0.139	9.8e-11
0.00000	Signed square	0.01	10.0	0.980	0.076	0.538	2.1e-12
0.03125	Direct	1.00	1.0	1.000	0.086	0.544	2.2e-15
0.03125	Signed square	0.01	0.1	1.000	0.086	0.544	8.5e-11
0.03125	Signed square	0.01	1.0	1.000	0.086	0.544	9.6e-11
0.03125	Signed square	0.01	10.0	1.000	0.086	0.544	4.1e-12
0.06250	Direct	1.00	1.0	1.000	0.086	0.544	1.5e-15
0.06250	Signed square	0.01	0.1	1.000	0.086	0.544	9.0e-11
0.06250	Signed square	0.01	1.0	1.000	0.086	0.544	7.9e-11
0.06250	Signed square	0.01	10.0	1.000	0.086	0.544	4.1e-12
0.12500	Direct	1.00	1.0	1.000	0.086	0.544	1.4e-15
0.12500	Signed square	0.01	0.1	1.000	0.086	0.544	9.6e-11
0.12500	Signed square	0.01	1.0	1.000	0.086	0.544	9.2e-11
0.12500	Signed square	0.01	10.0	1.000	0.086	0.544	4.7e-12
0.25000	Direct	1.00	1.0	1.000	0.086	0.544	1.2e-15
0.25000	Signed square	0.01	0.1	1.000	0.086	0.544	9.8e-11
0.25000	Signed square	0.01	1.0	1.000	0.086	0.544	9.4e-11
0.25000	Signed square	0.01	10.0	1.000	0.086	0.544	4.8e-12
0.50000	Direct	1.00	1.0	1.000	0.086	0.544	1.3e-15
0.50000	Signed square	0.01	0.1	1.000	0.086	0.544	9.1e-11
0.50000	Signed square	0.01	1.0	1.000	0.086	0.544	6.5e-11
0.50000	Signed square	0.01	10.0	1.000	0.086	0.544	3.9e-12

Table 37: Anchored endpoint sweep for Qwen2.5 Math 7B Instruct on the lexical carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.070	0.436	8.9e-16
0.00000	Signed square	0.01	0.1	-0.980	-0.796	0.000	9.7e-13
0.00000	Signed square	0.01	1.0	0.000	-0.070	0.436	9.2e-11
0.00000	Signed square	0.01	10.0	0.980	0.657	1.000	9.0e-13
0.03125	Direct	1.00	1.0	1.000	0.672	1.000	1.6e-15
0.03125	Signed square	0.01	0.1	1.000	0.672	1.000	7.4e-11
0.03125	Signed square	0.01	1.0	1.000	0.672	1.000	9.0e-11
0.03125	Signed square	0.01	10.0	1.000	0.672	1.000	5.2e-12
0.06250	Direct	1.00	1.0	1.000	0.672	1.000	1.0e-15
0.06250	Signed square	0.01	0.1	1.000	0.672	1.000	9.0e-11
0.06250	Signed square	0.01	1.0	1.000	0.672	1.000	9.9e-11
0.06250	Signed square	0.01	10.0	1.000	0.672	1.000	3.5e-12
0.12500	Direct	1.00	1.0	1.000	0.672	1.000	1.3e-15
0.12500	Signed square	0.01	0.1	1.000	0.672	1.000	9.9e-11
0.12500	Signed square	0.01	1.0	1.000	0.672	1.000	9.8e-11
0.12500	Signed square	0.01	10.0	1.000	0.672	1.000	4.3e-12
0.25000	Direct	1.00	1.0	1.000	0.672	1.000	1.1e-15
0.25000	Signed square	0.01	0.1	1.000	0.672	1.000	7.4e-11
0.25000	Signed square	0.01	1.0	1.000	0.672	1.000	9.6e-11
0.25000	Signed square	0.01	10.0	1.000	0.672	1.000	3.6e-12
0.50000	Direct	1.00	1.0	1.000	0.672	1.000	9.7e-16
0.50000	Signed square	0.01	0.1	1.000	0.672	1.000	9.8e-11
0.50000	Signed square	0.01	1.0	1.000	0.672	1.000	9.5e-11
0.50000	Signed square	0.01	10.0	1.000	0.672	1.000	5.0e-12

Table 38: Anchored endpoint sweep for Qwen2.5 Math 7B Instruct on the sentiment carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.113	0.230	8.0e-16
0.00000	Signed square	0.01	0.1	-0.980	-0.937	0.000	1.6e-12
0.00000	Signed square	0.01	1.0	0.000	-0.113	0.230	8.5e-11
0.00000	Signed square	0.01	10.0	0.980	0.711	0.978	9.7e-13
0.03125	Direct	1.00	1.0	1.000	0.727	0.978	1.4e-15
0.03125	Signed square	0.01	0.1	1.000	0.727	0.978	9.4e-11
0.03125	Signed square	0.01	1.0	1.000	0.727	0.978	7.7e-11
0.03125	Signed square	0.01	10.0	1.000	0.727	0.978	4.7e-12
0.06250	Direct	1.00	1.0	1.000	0.727	0.978	1.8e-15
0.06250	Signed square	0.01	0.1	1.000	0.727	0.978	9.8e-11
0.06250	Signed square	0.01	1.0	1.000	0.727	0.978	8.6e-11
0.06250	Signed square	0.01	10.0	1.000	0.727	0.978	3.5e-12
0.12500	Direct	1.00	1.0	1.000	0.727	0.978	1.3e-15
0.12500	Signed square	0.01	0.1	1.000	0.727	0.978	6.4e-11
0.12500	Signed square	0.01	1.0	1.000	0.727	0.978	9.0e-11
0.12500	Signed square	0.01	10.0	1.000	0.727	0.978	4.6e-12
0.25000	Direct	1.00	1.0	1.000	0.727	0.978	1.3e-15
0.25000	Signed square	0.01	0.1	1.000	0.727	0.978	9.4e-11
0.25000	Signed square	0.01	1.0	1.000	0.727	0.978	9.7e-11
0.25000	Signed square	0.01	10.0	1.000	0.727	0.978	5.5e-12
0.50000	Direct	1.00	1.0	1.000	0.727	0.978	1.0e-15
0.50000	Signed square	0.01	0.1	1.000	0.727	0.978	9.4e-11
0.50000	Signed square	0.01	1.0	1.000	0.727	0.978	9.4e-11
0.50000	Signed square	0.01	10.0	1.000	0.727	0.978	3.1e-12

Table 39: Anchored endpoint sweep for Qwen3 0.6B on the arithmetic carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.205	0.336	1.0e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.985	0.000	2.6e-12
0.00000	Signed square	0.01	1.0	0.000	-0.205	0.336	9.6e-11
0.00000	Signed square	0.01	10.0	0.980	0.574	0.780	1.4e-12
0.03125	Direct	1.00	1.0	1.000	0.590	0.781	3.5e-14
0.03125	Signed square	0.01	0.1	1.000	0.590	0.781	9.9e-11
0.03125	Signed square	0.01	1.0	1.000	0.590	0.781	9.7e-11
0.03125	Signed square	0.01	10.0	1.000	0.590	0.781	6.5e-12
0.06250	Direct	1.00	1.0	1.000	0.590	0.781	1.1e-15
0.06250	Signed square	0.01	0.1	1.000	0.590	0.781	8.0e-11
0.06250	Signed square	0.01	1.0	1.000	0.590	0.781	9.1e-11
0.06250	Signed square	0.01	10.0	1.000	0.590	0.781	8.0e-12
0.12500	Direct	1.00	1.0	1.000	0.590	0.781	1.2e-15
0.12500	Signed square	0.01	0.1	1.000	0.590	0.781	9.0e-11
0.12500	Signed square	0.01	1.0	1.000	0.590	0.781	9.9e-11
0.12500	Signed square	0.01	10.0	1.000	0.590	0.781	8.4e-12
0.25000	Direct	1.00	1.0	1.000	0.590	0.781	1.5e-15
0.25000	Signed square	0.01	0.1	1.000	0.590	0.781	9.9e-11
0.25000	Signed square	0.01	1.0	1.000	0.590	0.781	9.4e-11
0.25000	Signed square	0.01	10.0	1.000	0.590	0.781	6.2e-12
0.50000	Direct	1.00	1.0	1.000	0.590	0.781	2.1e-15
0.50000	Signed square	0.01	0.1	1.000	0.590	0.781	9.4e-11
0.50000	Signed square	0.01	1.0	1.000	0.590	0.781	9.6e-11
0.50000	Signed square	0.01	10.0	1.000	0.590	0.781	4.8e-12

Table 40: Anchored endpoint sweep for Qwen3 0.6B on the lexical carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.714	0.000	8.9e-16
0.00000	Signed square	0.01	0.1	-0.980	-0.965	0.000	1.0e-12
0.00000	Signed square	0.01	1.0	0.000	-0.714	0.000	9.3e-11
0.00000	Signed square	0.01	10.0	0.980	-0.464	0.090	8.9e-13
0.03125	Direct	1.00	1.0	1.000	-0.459	0.090	3.6e-14
0.03125	Signed square	0.01	0.1	1.000	-0.459	0.090	9.7e-11
0.03125	Signed square	0.01	1.0	1.000	-0.459	0.090	9.9e-11
0.03125	Signed square	0.01	10.0	1.000	-0.459	0.090	5.2e-12
0.06250	Direct	1.00	1.0	1.000	-0.459	0.090	5.8e-15
0.06250	Signed square	0.01	0.1	1.000	-0.459	0.090	9.4e-11
0.06250	Signed square	0.01	1.0	1.000	-0.459	0.090	8.2e-11
0.06250	Signed square	0.01	10.0	1.000	-0.459	0.090	5.3e-12
0.12500	Direct	1.00	1.0	1.000	-0.459	0.090	2.5e-15
0.12500	Signed square	0.01	0.1	1.000	-0.459	0.090	9.9e-11
0.12500	Signed square	0.01	1.0	1.000	-0.459	0.090	9.4e-11
0.12500	Signed square	0.01	10.0	1.000	-0.459	0.090	4.2e-12
0.25000	Direct	1.00	1.0	1.000	-0.459	0.090	2.2e-15
0.25000	Signed square	0.01	0.1	1.000	-0.459	0.090	9.6e-11
0.25000	Signed square	0.01	1.0	1.000	-0.459	0.090	9.6e-11
0.25000	Signed square	0.01	10.0	1.000	-0.459	0.090	4.2e-12
0.50000	Direct	1.00	1.0	1.000	-0.459	0.090	1.5e-15
0.50000	Signed square	0.01	0.1	1.000	-0.459	0.090	9.6e-11
0.50000	Signed square	0.01	1.0	1.000	-0.459	0.090	9.9e-11
0.50000	Signed square	0.01	10.0	1.000	-0.459	0.090	4.6e-12

Table 41: Anchored endpoint sweep for Qwen3 0.6B on the sentiment carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.316	0.104	8.9e-16
0.00000	Signed square	0.01	0.1	-0.980	-0.949	0.000	1.2e-12
0.00000	Signed square	0.01	1.0	0.000	-0.316	0.104	9.6e-11
0.00000	Signed square	0.01	10.0	0.980	0.317	0.690	1.0e-12
0.03125	Direct	1.00	1.0	1.000	0.330	0.693	1.6e-15
0.03125	Signed square	0.01	0.1	1.000	0.330	0.693	9.9e-11
0.03125	Signed square	0.01	1.0	1.000	0.330	0.693	8.1e-11
0.03125	Signed square	0.01	10.0	1.000	0.330	0.693	4.8e-12
0.06250	Direct	1.00	1.0	1.000	0.330	0.693	1.4e-15
0.06250	Signed square	0.01	0.1	1.000	0.330	0.693	9.5e-11
0.06250	Signed square	0.01	1.0	1.000	0.330	0.693	8.5e-11
0.06250	Signed square	0.01	10.0	1.000	0.330	0.693	4.0e-12
0.12500	Direct	1.00	1.0	1.000	0.330	0.693	1.3e-15
0.12500	Signed square	0.01	0.1	1.000	0.330	0.693	9.9e-11
0.12500	Signed square	0.01	1.0	1.000	0.330	0.693	7.7e-11
0.12500	Signed square	0.01	10.0	1.000	0.330	0.693	3.1e-12
0.25000	Direct	1.00	1.0	1.000	0.330	0.693	1.0e-15
0.25000	Signed square	0.01	0.1	1.000	0.330	0.693	7.7e-11
0.25000	Signed square	0.01	1.0	1.000	0.330	0.693	8.7e-11
0.25000	Signed square	0.01	10.0	1.000	0.330	0.693	4.2e-12
0.50000	Direct	1.00	1.0	1.000	0.330	0.693	9.9e-16
0.50000	Signed square	0.01	0.1	1.000	0.330	0.693	9.9e-11
0.50000	Signed square	0.01	1.0	1.000	0.330	0.693	9.4e-11
0.50000	Signed square	0.01	10.0	1.000	0.330	0.693	4.0e-12

Table 42: Anchored endpoint sweep for Qwen3 14B on the arithmetic carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	0.055	0.643	1.1e-15
0.00000	Signed square	0.01	0.1	-0.980	-1.014	0.000	1.4e-12
0.00000	Signed square	0.01	1.0	0.000	0.055	0.643	9.5e-11
0.00000	Signed square	0.01	10.0	0.980	1.125	0.987	1.4e-12
0.03125	Direct	1.00	1.0	1.000	1.146	0.987	7.5e-16
0.03125	Signed square	0.01	0.1	1.000	1.146	0.987	8.5e-11
0.03125	Signed square	0.01	1.0	1.000	1.146	0.987	9.5e-11
0.03125	Signed square	0.01	10.0	1.000	1.146	0.987	3.5e-12
0.06250	Direct	1.00	1.0	1.000	1.146	0.987	9.1e-16
0.06250	Signed square	0.01	0.1	1.000	1.146	0.987	7.7e-11
0.06250	Signed square	0.01	1.0	1.000	1.146	0.987	7.8e-11
0.06250	Signed square	0.01	10.0	1.000	1.146	0.987	1.0e-11
0.12500	Direct	1.00	1.0	1.000	1.146	0.987	1.0e-15
0.12500	Signed square	0.01	0.1	1.000	1.146	0.987	8.4e-11
0.12500	Signed square	0.01	1.0	1.000	1.146	0.987	9.5e-11
0.12500	Signed square	0.01	10.0	1.000	1.146	0.987	5.9e-12
0.25000	Direct	1.00	1.0	1.000	1.146	0.987	9.0e-16
0.25000	Signed square	0.01	0.1	1.000	1.146	0.987	9.8e-11
0.25000	Signed square	0.01	1.0	1.000	1.146	0.987	9.8e-11
0.25000	Signed square	0.01	10.0	1.000	1.146	0.987	5.4e-12
0.50000	Direct	1.00	1.0	1.000	1.146	0.987	1.2e-15
0.50000	Signed square	0.01	0.1	1.000	1.146	0.987	9.7e-11
0.50000	Signed square	0.01	1.0	1.000	1.146	0.987	7.9e-11
0.50000	Signed square	0.01	10.0	1.000	1.146	0.987	4.5e-12

Table 43: Anchored endpoint sweep for Qwen3 14B on the lexical carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.081	0.367	8.7e-16
0.00000	Signed square	0.01	0.1	-0.980	-0.866	0.000	1.1e-12
0.00000	Signed square	0.01	1.0	0.000	-0.081	0.367	8.8e-11
0.00000	Signed square	0.01	10.0	0.980	0.705	1.000	7.4e-13
0.03125	Direct	1.00	1.0	1.000	0.721	1.000	1.0e-15
0.03125	Signed square	0.01	0.1	1.000	0.721	1.000	9.9e-11
0.03125	Signed square	0.01	1.0	1.000	0.721	1.000	9.8e-11
0.03125	Signed square	0.01	10.0	1.000	0.721	1.000	3.1e-12
0.06250	Direct	1.00	1.0	1.000	0.721	1.000	1.6e-15
0.06250	Signed square	0.01	0.1	1.000	0.721	1.000	9.8e-11
0.06250	Signed square	0.01	1.0	1.000	0.721	1.000	8.6e-11
0.06250	Signed square	0.01	10.0	1.000	0.721	1.000	4.3e-12
0.12500	Direct	1.00	1.0	1.000	0.721	1.000	9.6e-16
0.12500	Signed square	0.01	0.1	1.000	0.721	1.000	9.5e-11
0.12500	Signed square	0.01	1.0	1.000	0.721	1.000	8.7e-11
0.12500	Signed square	0.01	10.0	1.000	0.721	1.000	3.4e-12
0.25000	Direct	1.00	1.0	1.000	0.721	1.000	1.0e-15
0.25000	Signed square	0.01	0.1	1.000	0.721	1.000	7.8e-11
0.25000	Signed square	0.01	1.0	1.000	0.721	1.000	8.1e-11
0.25000	Signed square	0.01	10.0	1.000	0.721	1.000	3.0e-12
0.50000	Direct	1.00	1.0	1.000	0.721	1.000	9.2e-16
0.50000	Signed square	0.01	0.1	1.000	0.721	1.000	8.1e-11
0.50000	Signed square	0.01	1.0	1.000	0.721	1.000	9.2e-11
0.50000	Signed square	0.01	10.0	1.000	0.721	1.000	5.1e-12

Table 44: Anchored endpoint sweep for Qwen3 14B on the sentiment carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.052	0.443	9.0e-16
0.00000	Signed square	0.01	0.1	-0.980	-0.881	0.000	1.1e-12
0.00000	Signed square	0.01	1.0	0.000	-0.052	0.443	8.9e-11
0.00000	Signed square	0.01	10.0	0.980	0.777	0.979	1.0e-12
0.03125	Direct	1.00	1.0	1.000	0.793	0.979	1.1e-15
0.03125	Signed square	0.01	0.1	1.000	0.793	0.979	9.6e-11
0.03125	Signed square	0.01	1.0	1.000	0.793	0.979	9.8e-11
0.03125	Signed square	0.01	10.0	1.000	0.793	0.979	2.7e-12
0.06250	Direct	1.00	1.0	1.000	0.793	0.979	1.3e-15
0.06250	Signed square	0.01	0.1	1.000	0.793	0.979	9.9e-11
0.06250	Signed square	0.01	1.0	1.000	0.793	0.979	9.3e-11
0.06250	Signed square	0.01	10.0	1.000	0.793	0.979	3.3e-12
0.12500	Direct	1.00	1.0	1.000	0.793	0.979	9.0e-16
0.12500	Signed square	0.01	0.1	1.000	0.793	0.979	9.3e-11
0.12500	Signed square	0.01	1.0	1.000	0.793	0.979	9.1e-11
0.12500	Signed square	0.01	10.0	1.000	0.793	0.979	3.9e-12
0.25000	Direct	1.00	1.0	1.000	0.793	0.979	1.4e-15
0.25000	Signed square	0.01	0.1	1.000	0.793	0.979	6.6e-11
0.25000	Signed square	0.01	1.0	1.000	0.793	0.979	9.0e-11
0.25000	Signed square	0.01	10.0	1.000	0.793	0.979	3.5e-12
0.50000	Direct	1.00	1.0	1.000	0.793	0.979	8.7e-16
0.50000	Signed square	0.01	0.1	1.000	0.793	0.979	5.6e-11
0.50000	Signed square	0.01	1.0	1.000	0.793	0.979	8.9e-11
0.50000	Signed square	0.01	10.0	1.000	0.793	0.979	3.2e-12

Table 45: Anchored endpoint sweep for Qwen3 30B-A3B on the arithmetic carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.144	0.309	1.1e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.970	0.004	2.1e-12
0.00000	Signed square	0.01	1.0	0.000	-0.144	0.309	9.9e-11
0.00000	Signed square	0.01	10.0	0.980	0.682	0.797	8.6e-11
0.03125	Direct	1.00	1.0	1.000	0.698	0.799	1.8e-15
0.03125	Signed square	0.01	0.1	1.000	0.698	0.799	9.2e-11
0.03125	Signed square	0.01	1.0	1.000	0.698	0.799	9.9e-11
0.03125	Signed square	0.01	10.0	1.000	0.698	0.799	4.6e-12
0.06250	Direct	1.00	1.0	1.000	0.698	0.799	1.2e-15
0.06250	Signed square	0.01	0.1	1.000	0.698	0.799	9.8e-11
0.06250	Signed square	0.01	1.0	1.000	0.698	0.799	9.5e-11
0.06250	Signed square	0.01	10.0	1.000	0.698	0.799	5.0e-12
0.12500	Direct	1.00	1.0	1.000	0.698	0.799	1.2e-15
0.12500	Signed square	0.01	0.1	1.000	0.698	0.799	9.4e-11
0.12500	Signed square	0.01	1.0	1.000	0.698	0.799	9.4e-11
0.12500	Signed square	0.01	10.0	1.000	0.698	0.799	5.5e-12
0.25000	Direct	1.00	1.0	1.000	0.698	0.799	1.2e-15
0.25000	Signed square	0.01	0.1	1.000	0.698	0.799	8.5e-11
0.25000	Signed square	0.01	1.0	1.000	0.698	0.799	8.7e-11
0.25000	Signed square	0.01	10.0	1.000	0.698	0.799	5.9e-12
0.50000	Direct	1.00	1.0	1.000	0.698	0.799	1.3e-15
0.50000	Signed square	0.01	0.1	1.000	0.698	0.799	9.9e-11
0.50000	Signed square	0.01	1.0	1.000	0.698	0.799	9.4e-11
0.50000	Signed square	0.01	10.0	1.000	0.698	0.799	6.0e-12

Table 46: Anchored endpoint sweep for Qwen3 30B-A3B on the lexical carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	0.036	0.730	1.3e-15
0.00000	Signed square	0.01	0.1	-0.980	-1.047	0.000	9.5e-13
0.00000	Signed square	0.01	1.0	0.000	0.036	0.730	1.0e-10
0.00000	Signed square	0.01	10.0	0.980	1.119	1.000	1.1e-12
0.03125	Direct	1.00	1.0	1.000	1.141	1.000	1.3e-15
0.03125	Signed square	0.01	0.1	1.000	1.141	1.000	1.0e-10
0.03125	Signed square	0.01	1.0	1.000	1.141	1.000	9.9e-11
0.03125	Signed square	0.01	10.0	1.000	1.141	1.000	4.4e-12
0.06250	Direct	1.00	1.0	1.000	1.141	1.000	1.4e-15
0.06250	Signed square	0.01	0.1	1.000	1.141	1.000	9.5e-11
0.06250	Signed square	0.01	1.0	1.000	1.141	1.000	9.4e-11
0.06250	Signed square	0.01	10.0	1.000	1.141	1.000	3.7e-12
0.12500	Direct	1.00	1.0	1.000	1.141	1.000	1.3e-15
0.12500	Signed square	0.01	0.1	1.000	1.141	1.000	9.4e-11
0.12500	Signed square	0.01	1.0	1.000	1.141	1.000	8.7e-11
0.12500	Signed square	0.01	10.0	1.000	1.141	1.000	4.5e-12
0.25000	Direct	1.00	1.0	1.000	1.141	1.000	9.4e-16
0.25000	Signed square	0.01	0.1	1.000	1.141	1.000	9.9e-11
0.25000	Signed square	0.01	1.0	1.000	1.141	1.000	9.4e-11
0.25000	Signed square	0.01	10.0	1.000	1.141	1.000	4.4e-12
0.50000	Direct	1.00	1.0	1.000	1.141	1.000	9.3e-16
0.50000	Signed square	0.01	0.1	1.000	1.141	1.000	9.3e-11
0.50000	Signed square	0.01	1.0	1.000	1.141	1.000	9.9e-11
0.50000	Signed square	0.01	10.0	1.000	1.141	1.000	3.8e-12

Table 47: Anchored endpoint sweep for Qwen3 30B-A3B on the sentiment carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.091	0.329	1.0e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.955	0.010	1.4e-12
0.00000	Signed square	0.01	1.0	0.000	-0.091	0.329	9.3e-11
0.00000	Signed square	0.01	10.0	0.980	0.773	0.884	1.4e-12
0.03125	Direct	1.00	1.0	1.000	0.790	0.884	2.0e-15
0.03125	Signed square	0.01	0.1	1.000	0.790	0.884	9.4e-11
0.03125	Signed square	0.01	1.0	1.000	0.790	0.884	9.4e-11
0.03125	Signed square	0.01	10.0	1.000	0.790	0.884	4.3e-12
0.06250	Direct	1.00	1.0	1.000	0.790	0.884	1.2e-15
0.06250	Signed square	0.01	0.1	1.000	0.790	0.884	9.9e-11
0.06250	Signed square	0.01	1.0	1.000	0.790	0.884	9.9e-11
0.06250	Signed square	0.01	10.0	1.000	0.790	0.884	4.5e-12
0.12500	Direct	1.00	1.0	1.000	0.790	0.884	1.3e-15
0.12500	Signed square	0.01	0.1	1.000	0.790	0.884	1.0e-10
0.12500	Signed square	0.01	1.0	1.000	0.790	0.884	9.5e-11
0.12500	Signed square	0.01	10.0	1.000	0.790	0.884	5.3e-12
0.25000	Direct	1.00	1.0	1.000	0.790	0.884	1.1e-15
0.25000	Signed square	0.01	0.1	1.000	0.790	0.884	9.8e-11
0.25000	Signed square	0.01	1.0	1.000	0.790	0.884	8.4e-11
0.25000	Signed square	0.01	10.0	1.000	0.790	0.884	5.4e-12
0.50000	Direct	1.00	1.0	1.000	0.790	0.884	1.3e-15
0.50000	Signed square	0.01	0.1	1.000	0.790	0.884	8.4e-11
0.50000	Signed square	0.01	1.0	1.000	0.790	0.884	9.5e-11
0.50000	Signed square	0.01	10.0	1.000	0.790	0.884	5.6e-12

Table 48: Anchored endpoint sweep for Qwen3 4B on the arithmetic carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	0.002	0.506	9.0e-16
0.00000	Signed square	0.01	0.1	-0.980	-0.985	0.000	7.9e-12
0.00000	Signed square	0.01	1.0	0.000	0.002	0.506	9.5e-11
0.00000	Signed square	0.01	10.0	0.980	0.989	0.982	1.1e-11
0.03125	Direct	1.00	1.0	1.000	1.009	0.983	1.4e-15
0.03125	Signed square	0.01	0.1	1.000	1.009	0.983	9.9e-11
0.03125	Signed square	0.01	1.0	1.000	1.009	0.983	9.9e-11
0.03125	Signed square	0.01	10.0	1.000	1.009	0.983	6.1e-12
0.06250	Direct	1.00	1.0	1.000	1.009	0.983	1.3e-15
0.06250	Signed square	0.01	0.1	1.000	1.009	0.983	9.4e-11
0.06250	Signed square	0.01	1.0	1.000	1.009	0.983	9.8e-11
0.06250	Signed square	0.01	10.0	1.000	1.009	0.983	1.0e-11
0.12500	Direct	1.00	1.0	1.000	1.009	0.983	1.3e-15
0.12500	Signed square	0.01	0.1	1.000	1.009	0.983	9.9e-11
0.12500	Signed square	0.01	1.0	1.000	1.009	0.983	9.7e-11
0.12500	Signed square	0.01	10.0	1.000	1.009	0.983	6.5e-12
0.25000	Direct	1.00	1.0	1.000	1.009	0.983	1.9e-15
0.25000	Signed square	0.01	0.1	1.000	1.009	0.983	8.7e-11
0.25000	Signed square	0.01	1.0	1.000	1.009	0.983	1.0e-10
0.25000	Signed square	0.01	10.0	1.000	1.009	0.983	7.4e-12
0.50000	Direct	1.00	1.0	1.000	1.009	0.983	1.7e-15
0.50000	Signed square	0.01	0.1	1.000	1.009	0.983	9.8e-11
0.50000	Signed square	0.01	1.0	1.000	1.009	0.983	9.9e-11
0.50000	Signed square	0.01	10.0	1.000	1.009	0.983	6.2e-12

Table 49: Anchored endpoint sweep for Qwen3 4B on the lexical carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	0.134	0.952	1.4e-15
0.00000	Signed square	0.01	0.1	-0.980	-1.035	0.000	1.2e-12
0.00000	Signed square	0.01	1.0	0.000	0.134	0.952	9.2e-11
0.00000	Signed square	0.01	10.0	0.980	1.302	1.000	1.3e-12
0.03125	Direct	1.00	1.0	1.000	1.326	1.000	1.4e-15
0.03125	Signed square	0.01	0.1	1.000	1.326	1.000	1.0e-10
0.03125	Signed square	0.01	1.0	1.000	1.326	1.000	9.9e-11
0.03125	Signed square	0.01	10.0	1.000	1.326	1.000	4.6e-12
0.06250	Direct	1.00	1.0	1.000	1.326	1.000	1.1e-15
0.06250	Signed square	0.01	0.1	1.000	1.326	1.000	1.0e-10
0.06250	Signed square	0.01	1.0	1.000	1.326	1.000	9.6e-11
0.06250	Signed square	0.01	10.0	1.000	1.326	1.000	4.0e-12
0.12500	Direct	1.00	1.0	1.000	1.326	1.000	1.3e-15
0.12500	Signed square	0.01	0.1	1.000	1.326	1.000	9.8e-11
0.12500	Signed square	0.01	1.0	1.000	1.326	1.000	8.8e-11
0.12500	Signed square	0.01	10.0	1.000	1.326	1.000	5.5e-12
0.25000	Direct	1.00	1.0	1.000	1.326	1.000	9.0e-16
0.25000	Signed square	0.01	0.1	1.000	1.326	1.000	9.8e-11
0.25000	Signed square	0.01	1.0	1.000	1.326	1.000	9.9e-11
0.25000	Signed square	0.01	10.0	1.000	1.326	1.000	5.2e-12
0.50000	Direct	1.00	1.0	1.000	1.326	1.000	1.2e-15
0.50000	Signed square	0.01	0.1	1.000	1.326	1.000	9.5e-11
0.50000	Signed square	0.01	1.0	1.000	1.326	1.000	9.5e-11
0.50000	Signed square	0.01	10.0	1.000	1.326	1.000	4.6e-12

Table 50: Anchored endpoint sweep for Qwen3 4B on the sentiment carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.039	0.433	1.1e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.982	0.000	1.1e-12
0.00000	Signed square	0.01	1.0	0.000	-0.039	0.433	9.9e-11
0.00000	Signed square	0.01	10.0	0.980	0.903	0.989	1.1e-12
0.03125	Direct	1.00	1.0	1.000	0.922	0.993	1.4e-15
0.03125	Signed square	0.01	0.1	1.000	0.922	0.993	8.9e-11
0.03125	Signed square	0.01	1.0	1.000	0.922	0.993	9.9e-11
0.03125	Signed square	0.01	10.0	1.000	0.922	0.993	4.2e-12
0.06250	Direct	1.00	1.0	1.000	0.922	0.993	1.2e-15
0.06250	Signed square	0.01	0.1	1.000	0.922	0.993	9.3e-11
0.06250	Signed square	0.01	1.0	1.000	0.922	0.993	9.3e-11
0.06250	Signed square	0.01	10.0	1.000	0.922	0.993	4.3e-12
0.12500	Direct	1.00	1.0	1.000	0.922	0.993	1.4e-15
0.12500	Signed square	0.01	0.1	1.000	0.922	0.993	9.7e-11
0.12500	Signed square	0.01	1.0	1.000	0.922	0.993	9.7e-11
0.12500	Signed square	0.01	10.0	1.000	0.922	0.993	3.3e-12
0.25000	Direct	1.00	1.0	1.000	0.922	0.993	1.4e-15
0.25000	Signed square	0.01	0.1	1.000	0.922	0.993	9.6e-11
0.25000	Signed square	0.01	1.0	1.000	0.922	0.993	9.8e-11
0.25000	Signed square	0.01	10.0	1.000	0.922	0.993	3.6e-12
0.50000	Direct	1.00	1.0	1.000	0.922	0.993	9.6e-16
0.50000	Signed square	0.01	0.1	1.000	0.922	0.993	9.9e-11
0.50000	Signed square	0.01	1.0	1.000	0.922	0.993	9.1e-11
0.50000	Signed square	0.01	10.0	1.000	0.922	0.993	4.0e-12

Table 51: Anchored endpoint sweep for Qwen3 8B on the arithmetic carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.140	0.344	1.0e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.985	0.000	2.1e-12
0.00000	Signed square	0.01	1.0	0.000	-0.140	0.344	9.8e-11
0.00000	Signed square	0.01	10.0	0.980	0.704	0.831	2.3e-12
0.03125	Direct	1.00	1.0	1.000	0.721	0.834	1.8e-15
0.03125	Signed square	0.01	0.1	1.000	0.721	0.834	9.9e-11
0.03125	Signed square	0.01	1.0	1.000	0.721	0.834	1.0e-10
0.03125	Signed square	0.01	10.0	1.000	0.721	0.834	4.1e-12
0.06250	Direct	1.00	1.0	1.000	0.721	0.834	1.4e-15
0.06250	Signed square	0.01	0.1	1.000	0.721	0.834	9.5e-11
0.06250	Signed square	0.01	1.0	1.000	0.721	0.834	9.9e-11
0.06250	Signed square	0.01	10.0	1.000	0.721	0.834	5.2e-12
0.12500	Direct	1.00	1.0	1.000	0.721	0.834	1.6e-15
0.12500	Signed square	0.01	0.1	1.000	0.721	0.834	9.8e-11
0.12500	Signed square	0.01	1.0	1.000	0.721	0.834	9.9e-11
0.12500	Signed square	0.01	10.0	1.000	0.721	0.834	4.2e-12
0.25000	Direct	1.00	1.0	1.000	0.721	0.834	1.3e-15
0.25000	Signed square	0.01	0.1	1.000	0.721	0.834	9.7e-11
0.25000	Signed square	0.01	1.0	1.000	0.721	0.834	9.8e-11
0.25000	Signed square	0.01	10.0	1.000	0.721	0.834	5.7e-12
0.50000	Direct	1.00	1.0	1.000	0.721	0.834	1.0e-15
0.50000	Signed square	0.01	0.1	1.000	0.721	0.834	1.0e-10
0.50000	Signed square	0.01	1.0	1.000	0.721	0.834	9.6e-11
0.50000	Signed square	0.01	10.0	1.000	0.721	0.834	7.4e-12

Table 52: Anchored endpoint sweep for Qwen3 8B on the lexical carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	0.013	0.570	1.1e-15
0.00000	Signed square	0.01	0.1	-0.980	-1.121	0.000	1.4e-12
0.00000	Signed square	0.01	1.0	0.000	0.013	0.570	9.8e-11
0.00000	Signed square	0.01	10.0	0.980	1.147	1.000	1.1e-12
0.03125	Direct	1.00	1.0	1.000	1.170	1.000	1.6e-15
0.03125	Signed square	0.01	0.1	1.000	1.170	1.000	9.9e-11
0.03125	Signed square	0.01	1.0	1.000	1.170	1.000	9.8e-11
0.03125	Signed square	0.01	10.0	1.000	1.170	1.000	5.5e-12
0.06250	Direct	1.00	1.0	1.000	1.170	1.000	1.4e-15
0.06250	Signed square	0.01	0.1	1.000	1.170	1.000	9.8e-11
0.06250	Signed square	0.01	1.0	1.000	1.170	1.000	9.5e-11
0.06250	Signed square	0.01	10.0	1.000	1.170	1.000	5.4e-12
0.12500	Direct	1.00	1.0	1.000	1.170	1.000	1.0e-15
0.12500	Signed square	0.01	0.1	1.000	1.170	1.000	9.9e-11
0.12500	Signed square	0.01	1.0	1.000	1.170	1.000	9.6e-11
0.12500	Signed square	0.01	10.0	1.000	1.170	1.000	4.5e-12
0.25000	Direct	1.00	1.0	1.000	1.170	1.000	1.1e-15
0.25000	Signed square	0.01	0.1	1.000	1.170	1.000	9.5e-11
0.25000	Signed square	0.01	1.0	1.000	1.170	1.000	8.6e-11
0.25000	Signed square	0.01	10.0	1.000	1.170	1.000	5.2e-12
0.50000	Direct	1.00	1.0	1.000	1.170	1.000	1.1e-15
0.50000	Signed square	0.01	0.1	1.000	1.170	1.000	9.5e-11
0.50000	Signed square	0.01	1.0	1.000	1.170	1.000	8.7e-11
0.50000	Signed square	0.01	10.0	1.000	1.170	1.000	5.3e-12

Table 53: Anchored endpoint sweep for Qwen3 8B on the sentiment carrier. Every row satisfies the same declared training constraints.

Coverage	Selector	Scale	Balance	Coeff. diff.	Witness	Acc.	Max error
0.00000	Direct	1.00	1.0	0.000	-0.197	0.271	1.2e-15
0.00000	Signed square	0.01	0.1	-0.980	-0.878	0.000	1.3e-12
0.00000	Signed square	0.01	1.0	0.000	-0.197	0.271	9.9e-11
0.00000	Signed square	0.01	10.0	0.980	0.484	0.815	1.3e-12
0.03125	Direct	1.00	1.0	1.000	0.498	0.819	1.4e-15
0.03125	Signed square	0.01	0.1	1.000	0.498	0.819	1.0e-10
0.03125	Signed square	0.01	1.0	1.000	0.498	0.819	9.3e-11
0.03125	Signed square	0.01	10.0	1.000	0.498	0.819	5.3e-12
0.06250	Direct	1.00	1.0	1.000	0.498	0.819	1.8e-15
0.06250	Signed square	0.01	0.1	1.000	0.498	0.819	9.5e-11
0.06250	Signed square	0.01	1.0	1.000	0.498	0.819	9.8e-11
0.06250	Signed square	0.01	10.0	1.000	0.498	0.819	5.0e-12
0.12500	Direct	1.00	1.0	1.000	0.498	0.819	1.4e-15
0.12500	Signed square	0.01	0.1	1.000	0.498	0.819	9.6e-11
0.12500	Signed square	0.01	1.0	1.000	0.498	0.819	9.9e-11
0.12500	Signed square	0.01	10.0	1.000	0.498	0.819	4.9e-12
0.25000	Direct	1.00	1.0	1.000	0.498	0.819	1.2e-15
0.25000	Signed square	0.01	0.1	1.000	0.498	0.819	9.8e-11
0.25000	Signed square	0.01	1.0	1.000	0.498	0.819	9.9e-11
0.25000	Signed square	0.01	10.0	1.000	0.498	0.819	4.4e-12
0.50000	Direct	1.00	1.0	1.000	0.498	0.819	1.2e-15
0.50000	Signed square	0.01	0.1	1.000	0.498	0.819	9.9e-11
0.50000	Signed square	0.01	1.0	1.000	0.498	0.819	9.9e-11
0.50000	Signed square	0.01	10.0	1.000	0.498	0.819	4.8e-12

Table 54: Natural Transformer-feature endpoint sweep for Qwen2.5 1.5B Instruct on arithmetic. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.992	3.992	0.000	0.010	7.1e-11
0.00000	Signed square	0.01	-3.989	3.989	0.000	0.010	9.4e-09
0.00000	Signed square	0.10	-3.990	3.990	0.000	0.010	5.2e-09
0.00000	Signed square	1.00	-3.992	3.992	0.000	0.010	8.8e-10
0.03125	Direct	0.00	-3.532	3.532	0.003	0.010	1.2e-10
0.03125	Signed square	0.01	-3.548	3.548	0.007	0.010	1.0e-08
0.03125	Signed square	0.10	-3.547	3.547	0.006	0.010	3.5e-09
0.03125	Signed square	1.00	-3.540	3.540	0.003	0.010	9.5e-09
0.06250	Direct	0.00	-3.152	3.152	0.020	0.010	1.1e-10
0.06250	Signed square	0.01	-3.183	3.183	0.040	0.010	4.2e-09
0.06250	Signed square	0.10	-3.183	3.183	0.035	0.010	4.6e-09
0.06250	Signed square	1.00	-3.173	3.173	0.026	0.010	6.4e-09
0.12500	Direct	0.00	-2.340	2.340	0.125	0.010	1.1e-10
0.12500	Signed square	0.01	-2.376	2.376	0.164	0.010	5.5e-09
0.12500	Signed square	0.10	-2.377	2.377	0.155	0.010	7.5e-09
0.12500	Signed square	1.00	-2.367	2.367	0.138	0.010	8.1e-09
0.25000	Direct	0.00	-1.190	1.190	0.315	0.010	1.3e-10
0.25000	Signed square	0.01	-1.215	1.215	0.342	0.010	9.0e-09
0.25000	Signed square	0.10	-1.214	1.214	0.339	0.010	8.4e-09
0.25000	Signed square	1.00	-1.204	1.204	0.325	0.010	9.3e-09
0.50000	Direct	0.00	0.709	-0.709	0.596	0.010	1.1e-10
0.50000	Signed square	0.01	0.666	-0.666	0.576	0.010	3.2e-09
0.50000	Signed square	0.10	0.661	-0.661	0.576	0.010	5.2e-09
0.50000	Signed square	1.00	0.666	-0.666	0.582	0.010	9.5e-09

Table 55: Natural Transformer-feature endpoint sweep for Qwen2.5 1.5B Instruct on lexical. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.268	3.268	0.000	0.037	4.1e-12
0.00000	Signed square	0.01	-3.159	3.159	0.000	0.037	9.2e-09
0.00000	Signed square	0.10	-3.245	3.245	0.000	0.037	9.9e-09
0.00000	Signed square	1.00	-3.266	3.266	0.000	0.037	9.7e-09
0.03125	Direct	0.00	3.093	-3.093	0.994	0.020	4.3e-12
0.03125	Signed square	0.01	3.065	-3.065	0.987	0.020	9.5e-09
0.03125	Signed square	0.10	3.047	-3.047	0.989	0.020	7.8e-09
0.03125	Signed square	1.00	3.071	-3.071	0.995	0.020	7.5e-09
0.06250	Direct	0.00	3.574	-3.574	1.000	0.017	4.5e-12
0.06250	Signed square	0.01	3.515	-3.515	1.000	0.017	9.1e-09
0.06250	Signed square	0.10	3.516	-3.516	1.000	0.017	9.9e-09
0.06250	Signed square	1.00	3.557	-3.557	1.000	0.017	7.6e-09
0.12500	Direct	0.00	3.810	-3.810	1.000	0.014	6.6e-12
0.12500	Signed square	0.01	3.825	-3.825	1.000	0.014	8.7e-09
0.12500	Signed square	0.10	3.818	-3.818	1.000	0.014	6.7e-09
0.12500	Signed square	1.00	3.813	-3.813	1.000	0.014	6.8e-09
0.25000	Direct	0.00	3.921	-3.921	1.000	0.009	9.6e-12
0.25000	Signed square	0.01	3.917	-3.917	1.000	0.009	8.9e-09
0.25000	Signed square	0.10	3.915	-3.915	1.000	0.009	9.2e-09
0.25000	Signed square	1.00	3.919	-3.919	1.000	0.009	9.8e-09
0.50000	Direct	0.00	3.953	-3.953	1.000	0.005	1.3e-11
0.50000	Signed square	0.01	3.952	-3.952	1.000	0.005	8.9e-09
0.50000	Signed square	0.10	3.950	-3.950	1.000	0.005	7.1e-09
0.50000	Signed square	1.00	3.952	-3.952	1.000	0.005	8.3e-09

Table 56: Natural Transformer-feature endpoint sweep for Qwen2.5 1.5B Instruct on sentiment. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.077	3.077	0.000	0.032	8.0e-12
0.00000	Signed square	0.01	-3.063	3.063	0.000	0.032	9.4e-09
0.00000	Signed square	0.10	-3.065	3.065	0.000	0.032	4.9e-09
0.00000	Signed square	1.00	-3.075	3.075	0.000	0.032	9.8e-09
0.03125	Direct	0.00	2.023	-2.023	0.995	0.025	1.0e-11
0.03125	Signed square	0.01	2.061	-2.061	0.990	0.025	9.3e-09
0.03125	Signed square	0.10	2.063	-2.063	0.992	0.025	1.0e-08
0.03125	Signed square	1.00	2.028	-2.028	0.995	0.025	6.4e-09
0.06250	Direct	0.00	2.841	-2.841	0.999	0.023	1.1e-11
0.06250	Signed square	0.01	2.849	-2.849	0.999	0.023	8.6e-09
0.06250	Signed square	0.10	2.854	-2.854	1.000	0.023	8.6e-09
0.06250	Signed square	1.00	2.844	-2.844	1.000	0.023	6.9e-09
0.12500	Direct	0.00	3.404	-3.404	1.000	0.021	1.5e-11
0.12500	Signed square	0.01	3.408	-3.408	1.000	0.021	1.8e-09
0.12500	Signed square	0.10	3.406	-3.406	1.000	0.021	7.8e-09
0.12500	Signed square	1.00	3.402	-3.402	1.000	0.021	9.8e-09
0.25000	Direct	0.00	3.681	-3.681	1.000	0.018	1.2e-11
0.25000	Signed square	0.01	3.685	-3.685	1.000	0.018	9.6e-09
0.25000	Signed square	0.10	3.686	-3.686	1.000	0.018	9.7e-09
0.25000	Signed square	1.00	3.681	-3.681	1.000	0.018	9.3e-09
0.50000	Direct	0.00	3.818	-3.818	1.000	0.016	1.5e-11
0.50000	Signed square	0.01	3.797	-3.797	1.000	0.016	6.7e-09
0.50000	Signed square	0.10	3.803	-3.803	1.000	0.016	1.0e-08
0.50000	Signed square	1.00	3.816	-3.816	1.000	0.016	8.1e-09

Table 57: Natural Transformer-feature endpoint sweep for Qwen2.5 7B on arithmetic. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.988	3.988	0.000	0.013	4.3e-11
0.00000	Signed square	0.01	-3.970	3.970	0.000	0.013	9.9e-09
0.00000	Signed square	0.10	-3.974	3.974	0.000	0.013	9.3e-09
0.00000	Signed square	1.00	-3.988	3.988	0.000	0.013	8.3e-09
0.03125	Direct	0.00	-2.990	2.990	0.017	0.013	8.0e-11
0.03125	Signed square	0.01	-2.997	2.997	0.031	0.013	9.8e-09
0.03125	Signed square	0.10	-2.992	2.992	0.028	0.013	5.7e-09
0.03125	Signed square	1.00	-2.990	2.990	0.021	0.013	9.6e-09
0.06250	Direct	0.00	-2.212	2.212	0.098	0.013	7.3e-11
0.06250	Signed square	0.01	-2.219	2.219	0.138	0.013	2.9e-09
0.06250	Signed square	0.10	-2.214	2.214	0.131	0.013	7.3e-09
0.06250	Signed square	1.00	-2.214	2.214	0.111	0.013	5.1e-09
0.12500	Direct	0.00	-0.896	0.896	0.332	0.012	7.9e-11
0.12500	Signed square	0.01	-0.860	0.860	0.364	0.012	3.6e-09
0.12500	Signed square	0.10	-0.858	0.858	0.360	0.012	4.5e-09
0.12500	Signed square	1.00	-0.869	0.869	0.347	0.012	6.9e-10
0.25000	Direct	0.00	0.516	-0.516	0.570	0.012	7.1e-11
0.25000	Signed square	0.01	0.511	-0.511	0.558	0.012	9.3e-09
0.25000	Signed square	0.10	0.516	-0.516	0.559	0.012	7.3e-09
0.25000	Signed square	1.00	0.519	-0.519	0.565	0.012	4.1e-09
0.50000	Direct	0.00	1.960	-1.960	0.752	0.012	7.8e-11
0.50000	Signed square	0.01	1.985	-1.985	0.719	0.012	7.6e-09
0.50000	Signed square	0.10	1.974	-1.974	0.723	0.012	9.8e-09
0.50000	Signed square	1.00	1.958	-1.958	0.736	0.012	6.9e-09

Table 58: Natural Transformer-feature endpoint sweep for Qwen2.5 7B on lexical. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	0.612	-0.612	0.996	0.067	1.3e-12
0.00000	Signed square	0.01	0.836	-0.836	0.990	0.067	9.6e-09
0.00000	Signed square	0.10	0.590	-0.590	0.976	0.067	7.9e-09
0.00000	Signed square	1.00	0.607	-0.607	0.996	0.067	7.6e-09
0.03125	Direct	0.00	3.829	-3.829	1.000	0.026	2.5e-12
0.03125	Signed square	0.01	3.806	-3.806	1.000	0.026	9.6e-09
0.03125	Signed square	0.10	3.807	-3.807	1.000	0.026	8.7e-09
0.03125	Signed square	1.00	3.828	-3.828	1.000	0.026	9.6e-09
0.06250	Direct	0.00	3.867	-3.867	1.000	0.021	3.0e-12
0.06250	Signed square	0.01	3.829	-3.829	1.000	0.021	9.9e-09
0.06250	Signed square	0.10	3.836	-3.836	1.000	0.021	9.7e-09
0.06250	Signed square	1.00	3.866	-3.866	1.000	0.021	9.4e-09
0.12500	Direct	0.00	3.926	-3.926	1.000	0.017	7.1e-12
0.12500	Signed square	0.01	3.926	-3.926	1.000	0.017	7.6e-09
0.12500	Signed square	0.10	3.923	-3.923	1.000	0.017	1.0e-08
0.12500	Signed square	1.00	3.926	-3.926	1.000	0.017	9.1e-09
0.25000	Direct	0.00	3.981	-3.981	1.000	0.012	8.3e-12
0.25000	Signed square	0.01	3.976	-3.976	1.000	0.012	9.2e-09
0.25000	Signed square	0.10	3.978	-3.978	1.000	0.012	8.8e-09
0.25000	Signed square	1.00	3.981	-3.981	1.000	0.012	5.5e-09
0.50000	Direct	0.00	3.983	-3.983	1.000	0.007	1.3e-11
0.50000	Signed square	0.01	3.983	-3.983	1.000	0.007	1.0e-08
0.50000	Signed square	0.10	3.981	-3.981	1.000	0.007	8.8e-09
0.50000	Signed square	1.00	3.983	-3.983	1.000	0.007	6.5e-09

Table 59: Natural Transformer-feature endpoint sweep for Qwen2.5 7B on sentiment. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-0.844	0.844	0.022	0.052	1.0e-11
0.00000	Signed square	0.01	-0.216	0.216	0.331	0.052	9.6e-09
0.00000	Signed square	0.10	-0.336	0.336	0.239	0.052	9.9e-09
0.00000	Signed square	1.00	-0.829	0.829	0.024	0.052	9.9e-09
0.03125	Direct	0.00	3.268	-3.268	0.998	0.033	1.0e-11
0.03125	Signed square	0.01	3.406	-3.406	1.000	0.033	9.9e-09
0.03125	Signed square	0.10	3.391	-3.391	1.000	0.033	5.8e-09
0.03125	Signed square	1.00	3.278	-3.278	0.999	0.033	8.5e-09
0.06250	Direct	0.00	3.631	-3.631	1.000	0.029	1.2e-11
0.06250	Signed square	0.01	3.707	-3.707	1.000	0.029	6.7e-09
0.06250	Signed square	0.10	3.698	-3.698	1.000	0.029	9.4e-09
0.06250	Signed square	1.00	3.637	-3.637	1.000	0.029	9.2e-09
0.12500	Direct	0.00	3.817	-3.817	1.000	0.027	8.1e-12
0.12500	Signed square	0.01	3.848	-3.848	1.000	0.027	9.6e-09
0.12500	Signed square	0.10	3.847	-3.847	1.000	0.027	8.3e-09
0.12500	Signed square	1.00	3.820	-3.820	1.000	0.027	9.4e-09
0.25000	Direct	0.00	3.912	-3.912	1.000	0.025	1.2e-11
0.25000	Signed square	0.01	3.940	-3.940	1.000	0.025	7.2e-09
0.25000	Signed square	0.10	3.939	-3.939	1.000	0.025	9.5e-09
0.25000	Signed square	1.00	3.915	-3.915	1.000	0.025	9.5e-09
0.50000	Direct	0.00	3.943	-3.943	1.000	0.022	1.2e-11
0.50000	Signed square	0.01	3.949	-3.949	1.000	0.022	8.5e-09
0.50000	Signed square	0.10	3.952	-3.952	1.000	0.022	9.9e-09
0.50000	Signed square	1.00	3.944	-3.944	1.000	0.022	1.0e-08

Table 60: Natural Transformer-feature endpoint sweep for Qwen2.5 7B Instruct on arithmetic. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.919	3.919	0.000	0.025	1.4e-11
0.00000	Signed square	0.01	-3.882	3.882	0.000	0.025	9.3e-09
0.00000	Signed square	0.10	-3.889	3.889	0.000	0.025	9.4e-09
0.00000	Signed square	1.00	-3.918	3.918	0.000	0.025	5.1e-09
0.03125	Direct	0.00	-2.604	2.604	0.045	0.025	3.1e-11
0.03125	Signed square	0.01	-2.582	2.582	0.061	0.025	9.4e-09
0.03125	Signed square	0.10	-2.582	2.582	0.057	0.025	8.0e-09
0.03125	Signed square	1.00	-2.593	2.593	0.047	0.025	9.3e-09
0.06250	Direct	0.00	-1.623	1.623	0.167	0.025	1.6e-11
0.06250	Signed square	0.01	-1.584	1.584	0.202	0.025	9.1e-09
0.06250	Signed square	0.10	-1.584	1.584	0.193	0.025	8.7e-09
0.06250	Signed square	1.00	-1.604	1.604	0.174	0.025	9.6e-09
0.12500	Direct	0.00	-0.263	0.263	0.439	0.025	2.3e-11
0.12500	Signed square	0.01	-0.232	0.232	0.450	0.025	8.4e-09
0.12500	Signed square	0.10	-0.233	0.233	0.449	0.025	6.8e-09
0.12500	Signed square	1.00	-0.252	0.252	0.444	0.025	8.0e-09
0.25000	Direct	0.00	1.290	-1.290	0.692	0.024	2.2e-11
0.25000	Signed square	0.01	1.334	-1.334	0.672	0.024	7.5e-09
0.25000	Signed square	0.10	1.325	-1.325	0.678	0.024	7.7e-09
0.25000	Signed square	1.00	1.304	-1.304	0.689	0.024	7.1e-10
0.50000	Direct	0.00	2.447	-2.447	0.828	0.024	2.7e-11
0.50000	Signed square	0.01	2.450	-2.450	0.796	0.024	8.3e-09
0.50000	Signed square	0.10	2.454	-2.454	0.803	0.024	6.7e-09
0.50000	Signed square	1.00	2.458	-2.458	0.820	0.024	6.9e-09

Table 61: Natural Transformer-feature endpoint sweep for Qwen2.5 7B Instruct on lexical. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	3.572	-3.572	1.000	0.083	3.9e-12
0.00000	Signed square	0.01	3.406	-3.406	1.000	0.083	4.3e-09
0.00000	Signed square	0.10	3.424	-3.424	1.000	0.083	1.0e-08
0.00000	Signed square	1.00	3.571	-3.571	1.000	0.083	2.1e-09
0.03125	Direct	0.00	3.943	-3.943	1.000	0.045	2.1e-12
0.03125	Signed square	0.01	3.901	-3.901	1.000	0.045	6.3e-09
0.03125	Signed square	0.10	3.899	-3.899	1.000	0.045	4.4e-09
0.03125	Signed square	1.00	3.943	-3.943	1.000	0.045	2.3e-09
0.06250	Direct	0.00	4.003	-4.003	1.000	0.034	2.8e-12
0.06250	Signed square	0.01	3.978	-3.978	1.000	0.034	9.6e-09
0.06250	Signed square	0.10	3.976	-3.976	1.000	0.034	2.8e-09
0.06250	Signed square	1.00	4.003	-4.003	1.000	0.034	8.7e-09
0.12500	Direct	0.00	4.012	-4.012	1.000	0.025	6.8e-12
0.12500	Signed square	0.01	3.998	-3.998	1.000	0.025	7.1e-09
0.12500	Signed square	0.10	3.999	-3.999	1.000	0.025	7.1e-09
0.12500	Signed square	1.00	4.012	-4.012	1.000	0.025	4.4e-09
0.25000	Direct	0.00	4.000	-4.000	1.000	0.017	3.5e-12
0.25000	Signed square	0.01	3.999	-3.999	1.000	0.017	5.4e-09
0.25000	Signed square	0.10	3.997	-3.997	1.000	0.017	5.7e-09
0.25000	Signed square	1.00	4.000	-4.000	1.000	0.017	8.4e-09
0.50000	Direct	0.00	4.008	-4.008	1.000	0.009	5.1e-12
0.50000	Signed square	0.01	4.011	-4.011	1.000	0.009	9.3e-09
0.50000	Signed square	0.10	4.009	-4.009	1.000	0.009	3.8e-09
0.50000	Signed square	1.00	4.008	-4.008	1.000	0.009	9.6e-09

Table 62: Natural Transformer-feature endpoint sweep for Qwen2.5 7B Instruct on sentiment. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	2.945	-2.945	1.000	0.086	3.7e-12
0.00000	Signed square	0.01	2.820	-2.820	1.000	0.086	9.6e-09
0.00000	Signed square	0.10	2.835	-2.835	1.000	0.086	4.5e-09
0.00000	Signed square	1.00	2.943	-2.943	1.000	0.086	5.0e-09
0.03125	Direct	0.00	3.756	-3.756	1.000	0.061	2.7e-12
0.03125	Signed square	0.01	3.707	-3.707	1.000	0.061	9.8e-09
0.03125	Signed square	0.10	3.719	-3.719	1.000	0.061	8.0e-09
0.03125	Signed square	1.00	3.755	-3.755	1.000	0.061	9.9e-09
0.06250	Direct	0.00	3.851	-3.851	1.000	0.055	2.6e-12
0.06250	Signed square	0.01	3.816	-3.816	1.000	0.055	9.8e-09
0.06250	Signed square	0.10	3.825	-3.825	1.000	0.055	8.7e-09
0.06250	Signed square	1.00	3.850	-3.850	1.000	0.055	9.2e-09
0.12500	Direct	0.00	3.924	-3.924	1.000	0.050	3.6e-12
0.12500	Signed square	0.01	3.911	-3.911	1.000	0.050	9.9e-09
0.12500	Signed square	0.10	3.915	-3.915	1.000	0.050	6.2e-09
0.12500	Signed square	1.00	3.924	-3.924	1.000	0.050	4.7e-09
0.25000	Direct	0.00	3.949	-3.949	1.000	0.044	3.2e-12
0.25000	Signed square	0.01	3.949	-3.949	1.000	0.044	9.8e-09
0.25000	Signed square	0.10	3.950	-3.950	1.000	0.044	7.4e-09
0.25000	Signed square	1.00	3.949	-3.949	1.000	0.044	9.6e-09
0.50000	Direct	0.00	3.961	-3.961	1.000	0.037	5.6e-12
0.50000	Signed square	0.01	3.959	-3.959	1.000	0.037	7.7e-09
0.50000	Signed square	0.10	3.960	-3.960	1.000	0.037	7.7e-09
0.50000	Signed square	1.00	3.961	-3.961	1.000	0.037	2.4e-09

Table 63: Natural Transformer-feature endpoint sweep for Qwen2.5 Math 7B on arithmetic. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.955	3.955	0.000	0.018	1.5e-11
0.00000	Signed square	0.01	-3.964	3.964	0.000	0.018	9.1e-09
0.00000	Signed square	0.10	-3.961	3.961	0.000	0.018	6.8e-09
0.00000	Signed square	1.00	-3.955	3.955	0.000	0.018	6.6e-09
0.03125	Direct	0.00	-3.014	3.014	0.010	0.018	3.2e-11
0.03125	Signed square	0.01	-2.990	2.990	0.020	0.018	7.8e-09
0.03125	Signed square	0.10	-2.992	2.992	0.017	0.018	7.0e-09
0.03125	Signed square	1.00	-3.005	3.005	0.011	0.018	1.8e-10
0.06250	Direct	0.00	-2.473	2.473	0.070	0.017	2.0e-11
0.06250	Signed square	0.01	-2.465	2.465	0.111	0.017	9.5e-09
0.06250	Signed square	0.10	-2.468	2.468	0.104	0.017	8.3e-09
0.06250	Signed square	1.00	-2.466	2.466	0.083	0.017	3.4e-09
0.12500	Direct	0.00	-1.386	1.386	0.247	0.017	3.1e-11
0.12500	Signed square	0.01	-1.286	1.286	0.294	0.017	6.8e-09
0.12500	Signed square	0.10	-1.294	1.294	0.286	0.017	8.0e-09
0.12500	Signed square	1.00	-1.348	1.348	0.267	0.017	1.7e-09
0.25000	Direct	0.00	0.150	-0.150	0.515	0.017	4.9e-11
0.25000	Signed square	0.01	0.237	-0.237	0.529	0.017	2.6e-09
0.25000	Signed square	0.10	0.235	-0.235	0.526	0.017	3.9e-09
0.25000	Signed square	1.00	0.205	-0.205	0.523	0.017	3.5e-09
0.50000	Direct	0.00	1.525	-1.525	0.706	0.017	5.3e-11
0.50000	Signed square	0.01	1.572	-1.572	0.681	0.017	9.3e-10
0.50000	Signed square	0.10	1.575	-1.575	0.684	0.017	7.0e-09
0.50000	Signed square	1.00	1.572	-1.572	0.697	0.017	1.7e-09

Table 64: Natural Transformer-feature endpoint sweep for Qwen2.5 Math 7B on lexical. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-0.365	0.365	0.286	0.085	6.0e-13
0.00000	Signed square	0.01	0.018	-0.018	0.567	0.085	7.5e-09
0.00000	Signed square	0.10	-0.115	0.115	0.463	0.085	6.1e-09
0.00000	Signed square	1.00	-0.363	0.363	0.283	0.085	1.8e-10
0.03125	Direct	0.00	2.981	-2.981	0.994	0.057	8.9e-13
0.03125	Signed square	0.01	2.906	-2.906	0.979	0.057	8.6e-09
0.03125	Signed square	0.10	2.893	-2.893	0.978	0.057	5.7e-09
0.03125	Signed square	1.00	2.973	-2.973	0.990	0.057	9.5e-09
0.06250	Direct	0.00	3.307	-3.307	1.000	0.047	9.1e-13
0.06250	Signed square	0.01	3.262	-3.262	0.995	0.047	9.1e-09
0.06250	Signed square	0.10	3.254	-3.254	0.997	0.047	8.3e-09
0.06250	Signed square	1.00	3.296	-3.296	1.000	0.047	6.9e-09
0.12500	Direct	0.00	3.663	-3.663	1.000	0.037	1.0e-12
0.12500	Signed square	0.01	3.625	-3.625	1.000	0.037	9.5e-09
0.12500	Signed square	0.10	3.627	-3.627	1.000	0.037	8.2e-09
0.12500	Signed square	1.00	3.657	-3.657	1.000	0.037	1.9e-09
0.25000	Direct	0.00	3.841	-3.841	1.000	0.025	1.7e-12
0.25000	Signed square	0.01	3.825	-3.825	0.997	0.025	9.3e-09
0.25000	Signed square	0.10	3.828	-3.828	0.999	0.025	4.7e-09
0.25000	Signed square	1.00	3.837	-3.837	1.000	0.025	8.1e-09
0.50000	Direct	0.00	3.909	-3.909	1.000	0.013	1.4e-12
0.50000	Signed square	0.01	3.903	-3.903	0.998	0.013	9.4e-09
0.50000	Signed square	0.10	3.904	-3.904	0.998	0.013	5.8e-09
0.50000	Signed square	1.00	3.908	-3.908	1.000	0.013	6.6e-09

Table 65: Natural Transformer-feature endpoint sweep for Qwen2.5 Math 7B on sentiment. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-2.917	2.917	0.000	0.064	2.3e-12
0.00000	Signed square	0.01	-2.970	2.970	0.000	0.064	9.8e-09
0.00000	Signed square	0.10	-2.932	2.932	0.000	0.064	5.4e-09
0.00000	Signed square	1.00	-2.918	2.918	0.000	0.064	8.4e-09
0.03125	Direct	0.00	2.105	-2.105	0.938	0.048	4.2e-12
0.03125	Signed square	0.01	2.312	-2.312	0.944	0.048	9.5e-09
0.03125	Signed square	0.10	2.288	-2.288	0.942	0.048	7.1e-09
0.03125	Signed square	1.00	2.145	-2.145	0.941	0.048	7.8e-09
0.06250	Direct	0.00	3.110	-3.110	0.989	0.043	3.8e-12
0.06250	Signed square	0.01	3.133	-3.133	0.979	0.043	6.5e-09
0.06250	Signed square	0.10	3.143	-3.143	0.982	0.043	5.2e-09
0.06250	Signed square	1.00	3.125	-3.125	0.988	0.043	8.8e-09
0.12500	Direct	0.00	3.481	-3.481	0.992	0.040	3.5e-12
0.12500	Signed square	0.01	3.570	-3.570	0.987	0.040	1.8e-09
0.12500	Signed square	0.10	3.553	-3.553	0.990	0.040	9.4e-09
0.12500	Signed square	1.00	3.495	-3.495	0.992	0.040	1.7e-09
0.25000	Direct	0.00	3.772	-3.772	0.996	0.036	4.5e-12
0.25000	Signed square	0.01	3.785	-3.785	0.993	0.036	9.0e-09
0.25000	Signed square	0.10	3.789	-3.789	0.996	0.036	8.1e-09
0.25000	Signed square	1.00	3.777	-3.777	0.996	0.036	7.5e-09
0.50000	Direct	0.00	3.814	-3.814	0.997	0.031	3.3e-12
0.50000	Signed square	0.01	3.808	-3.808	0.997	0.031	3.1e-10
0.50000	Signed square	0.10	3.807	-3.807	0.997	0.031	4.6e-09
0.50000	Signed square	1.00	3.811	-3.811	0.997	0.031	4.9e-09

Table 66: Natural Transformer-feature endpoint sweep for Qwen2.5 Math 7B Instruct on arithmetic. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.853	3.853	0.000	0.037	2.8e-12
0.00000	Signed square	0.01	-3.851	3.851	0.000	0.037	8.2e-09
0.00000	Signed square	0.10	-3.850	3.850	0.000	0.037	9.5e-09
0.00000	Signed square	1.00	-3.853	3.853	0.000	0.037	8.4e-09
0.03125	Direct	0.00	-2.948	2.948	0.009	0.037	6.0e-12
0.03125	Signed square	0.01	-2.925	2.925	0.016	0.037	9.7e-09
0.03125	Signed square	0.10	-2.928	2.928	0.015	0.037	8.2e-09
0.03125	Signed square	1.00	-2.943	2.943	0.010	0.037	3.6e-10
0.06250	Direct	0.00	-2.204	2.204	0.075	0.037	7.8e-12
0.06250	Signed square	0.01	-2.180	2.180	0.110	0.037	7.6e-09
0.06250	Signed square	0.10	-2.183	2.183	0.102	0.037	9.1e-09
0.06250	Signed square	1.00	-2.188	2.188	0.088	0.037	8.4e-09
0.12500	Direct	0.00	-1.125	1.125	0.279	0.036	7.3e-12
0.12500	Signed square	0.01	-1.119	1.119	0.302	0.036	8.6e-09
0.12500	Signed square	0.10	-1.118	1.118	0.296	0.036	5.4e-09
0.12500	Signed square	1.00	-1.118	1.118	0.284	0.036	2.1e-09
0.25000	Direct	0.00	0.228	-0.228	0.540	0.036	8.1e-12
0.25000	Signed square	0.01	0.223	-0.223	0.533	0.036	3.1e-09
0.25000	Signed square	0.10	0.230	-0.230	0.534	0.036	8.1e-09
0.25000	Signed square	1.00	0.233	-0.233	0.540	0.036	5.9e-09
0.50000	Direct	0.00	1.759	-1.759	0.745	0.036	1.4e-11
0.50000	Signed square	0.01	1.823	-1.823	0.723	0.036	4.8e-09
0.50000	Signed square	0.10	1.816	-1.816	0.726	0.036	1.2e-09
0.50000	Signed square	1.00	1.782	-1.782	0.736	0.036	2.5e-09

Table 67: Natural Transformer-feature endpoint sweep for Qwen2.5 Math 7B Instruct on lexical. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-0.947	0.947	0.042	0.108	5.9e-13
0.00000	Signed square	0.01	-0.393	0.393	0.247	0.108	9.3e-09
0.00000	Signed square	0.10	-0.484	0.484	0.165	0.108	4.2e-09
0.00000	Signed square	1.00	-0.929	0.929	0.042	0.108	2.7e-09
0.03125	Direct	0.00	3.343	-3.343	1.000	0.059	6.2e-13
0.03125	Signed square	0.01	3.450	-3.450	1.000	0.059	9.9e-09
0.03125	Signed square	0.10	3.448	-3.448	1.000	0.059	8.8e-09
0.03125	Signed square	1.00	3.369	-3.369	1.000	0.059	5.3e-09
0.06250	Direct	0.00	3.692	-3.692	1.000	0.048	5.8e-13
0.06250	Signed square	0.01	3.757	-3.757	1.000	0.048	6.9e-09
0.06250	Signed square	0.10	3.755	-3.755	1.000	0.048	8.2e-09
0.06250	Signed square	1.00	3.703	-3.703	1.000	0.048	7.4e-09
0.12500	Direct	0.00	3.866	-3.866	1.000	0.038	7.3e-13
0.12500	Signed square	0.01	3.874	-3.874	1.000	0.038	9.2e-09
0.12500	Signed square	0.10	3.875	-3.875	1.000	0.038	6.8e-09
0.12500	Signed square	1.00	3.868	-3.868	1.000	0.038	8.4e-09
0.25000	Direct	0.00	3.929	-3.929	1.000	0.026	1.3e-12
0.25000	Signed square	0.01	3.939	-3.939	1.000	0.026	9.8e-09
0.25000	Signed square	0.10	3.931	-3.931	1.000	0.026	9.4e-09
0.25000	Signed square	1.00	3.927	-3.927	1.000	0.026	3.4e-09
0.50000	Direct	0.00	3.967	-3.967	1.000	0.015	1.1e-12
0.50000	Signed square	0.01	3.961	-3.961	1.000	0.015	9.9e-09
0.50000	Signed square	0.10	3.958	-3.958	1.000	0.015	3.4e-09
0.50000	Signed square	1.00	3.962	-3.962	1.000	0.015	2.9e-09

Table 68: Natural Transformer-feature endpoint sweep for Qwen2.5 Math 7B Instruct on sentiment. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-2.837	2.837	0.000	0.083	3.3e-12
0.00000	Signed square	0.01	-2.834	2.834	0.000	0.083	8.2e-10
0.00000	Signed square	0.10	-2.837	2.837	0.000	0.083	6.2e-09
0.00000	Signed square	1.00	-2.837	2.837	0.000	0.083	8.7e-09
0.03125	Direct	0.00	2.205	-2.205	0.954	0.063	2.5e-12
0.03125	Signed square	0.01	2.290	-2.290	0.953	0.063	6.1e-09
0.03125	Signed square	0.10	2.286	-2.286	0.956	0.063	9.7e-09
0.03125	Signed square	1.00	2.229	-2.229	0.955	0.063	7.0e-09
0.06250	Direct	0.00	2.993	-2.993	0.993	0.057	1.8e-12
0.06250	Signed square	0.01	3.056	-3.056	0.994	0.057	1.0e-08
0.06250	Signed square	0.10	3.053	-3.053	0.993	0.057	9.4e-09
0.06250	Signed square	1.00	3.010	-3.010	0.993	0.057	3.9e-09
0.12500	Direct	0.00	3.547	-3.547	0.989	0.055	2.2e-12
0.12500	Signed square	0.01	3.590	-3.590	0.985	0.055	9.9e-09
0.12500	Signed square	0.10	3.576	-3.576	0.985	0.055	4.6e-09
0.12500	Signed square	1.00	3.550	-3.550	0.988	0.055	6.4e-09
0.25000	Direct	0.00	3.777	-3.777	0.998	0.048	2.9e-12
0.25000	Signed square	0.01	3.790	-3.790	0.995	0.048	9.5e-09
0.25000	Signed square	0.10	3.781	-3.781	0.996	0.048	6.8e-09
0.25000	Signed square	1.00	3.773	-3.773	0.997	0.048	4.0e-09
0.50000	Direct	0.00	3.885	-3.885	0.998	0.042	3.1e-12
0.50000	Signed square	0.01	3.898	-3.898	0.994	0.042	7.5e-09
0.50000	Signed square	0.10	3.894	-3.894	0.995	0.042	9.9e-09
0.50000	Signed square	1.00	3.886	-3.886	0.998	0.042	7.3e-09

Table 69: Natural Transformer-feature endpoint sweep for Qwen3 0.6B on arithmetic. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.991	3.991	0.000	0.014	2.5e-11
0.00000	Signed square	0.01	-3.991	3.991	0.000	0.014	9.4e-09
0.00000	Signed square	0.10	-3.991	3.991	0.000	0.014	8.6e-09
0.00000	Signed square	1.00	-3.991	3.991	0.000	0.014	8.8e-09
0.03125	Direct	0.00	-3.750	3.750	0.002	0.014	3.3e-11
0.03125	Signed square	0.01	-3.768	3.768	0.005	0.014	8.6e-09
0.03125	Signed square	0.10	-3.763	3.763	0.005	0.014	9.5e-09
0.03125	Signed square	1.00	-3.755	3.755	0.002	0.014	7.5e-09
0.06250	Direct	0.00	-3.464	3.464	0.020	0.014	4.0e-11
0.06250	Signed square	0.01	-3.484	3.484	0.036	0.014	6.2e-09
0.06250	Signed square	0.10	-3.480	3.480	0.032	0.014	8.3e-09
0.06250	Signed square	1.00	-3.470	3.470	0.023	0.014	7.0e-09
0.12500	Direct	0.00	-2.885	2.885	0.087	0.014	6.1e-11
0.12500	Signed square	0.01	-2.920	2.920	0.122	0.014	1.3e-09
0.12500	Signed square	0.10	-2.915	2.915	0.114	0.014	1.2e-09
0.12500	Signed square	1.00	-2.898	2.898	0.099	0.014	9.8e-09
0.25000	Direct	0.00	-1.891	1.891	0.244	0.014	1.0e-10
0.25000	Signed square	0.01	-1.886	1.886	0.283	0.014	1.6e-09
0.25000	Signed square	0.10	-1.887	1.887	0.274	0.014	3.0e-09
0.25000	Signed square	1.00	-1.887	1.887	0.257	0.014	5.0e-09
0.50000	Direct	0.00	0.139	-0.139	0.521	0.014	7.9e-11
0.50000	Signed square	0.01	0.206	-0.206	0.524	0.014	7.1e-09
0.50000	Signed square	0.10	0.200	-0.200	0.525	0.014	5.9e-09
0.50000	Signed square	1.00	0.170	-0.170	0.522	0.014	6.4e-09

Table 70: Natural Transformer-feature endpoint sweep for Qwen3 0.6B on lexical. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.678	3.678	0.000	0.026	2.2e-12
0.00000	Signed square	0.01	-3.642	3.642	0.000	0.026	4.8e-09
0.00000	Signed square	0.10	-3.659	3.659	0.000	0.026	9.9e-09
0.00000	Signed square	1.00	-3.677	3.677	0.000	0.026	9.3e-09
0.03125	Direct	0.00	0.683	-0.683	0.658	0.021	6.0e-12
0.03125	Signed square	0.01	0.488	-0.488	0.563	0.021	1.1e-09
0.03125	Signed square	0.10	0.519	-0.519	0.583	0.021	6.9e-09
0.03125	Signed square	1.00	0.634	-0.634	0.621	0.021	6.5e-11
0.06250	Direct	0.00	1.784	-1.784	0.878	0.018	1.2e-11
0.06250	Signed square	0.01	1.692	-1.692	0.791	0.018	9.9e-09
0.06250	Signed square	0.10	1.708	-1.708	0.805	0.018	8.1e-09
0.06250	Signed square	1.00	1.743	-1.743	0.842	0.018	8.4e-10
0.12500	Direct	0.00	2.743	-2.743	0.950	0.015	1.1e-11
0.12500	Signed square	0.01	2.688	-2.688	0.906	0.015	8.5e-09
0.12500	Signed square	0.10	2.673	-2.673	0.917	0.015	8.5e-09
0.12500	Signed square	1.00	2.704	-2.704	0.948	0.015	5.2e-11
0.25000	Direct	0.00	3.379	-3.379	0.972	0.010	9.0e-12
0.25000	Signed square	0.01	3.368	-3.368	0.968	0.010	6.7e-09
0.25000	Signed square	0.10	3.367	-3.367	0.970	0.010	4.5e-09
0.25000	Signed square	1.00	3.367	-3.367	0.971	0.010	7.9e-11
0.50000	Direct	0.00	3.757	-3.757	0.980	0.005	1.3e-11
0.50000	Signed square	0.01	3.747	-3.747	0.980	0.005	9.6e-09
0.50000	Signed square	0.10	3.748	-3.748	0.980	0.005	5.2e-09
0.50000	Signed square	1.00	3.758	-3.758	0.981	0.005	7.7e-09

Table 71: Natural Transformer-feature endpoint sweep for Qwen3 0.6B on sentiment. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.356	3.356	0.000	0.037	5.0e-12
0.00000	Signed square	0.01	-3.247	3.247	0.000	0.037	8.6e-09
0.00000	Signed square	0.10	-3.274	3.274	0.000	0.037	2.3e-09
0.00000	Signed square	1.00	-3.354	3.354	0.000	0.037	1.4e-10
0.03125	Direct	0.00	1.095	-1.095	0.781	0.031	6.3e-12
0.03125	Signed square	0.01	1.248	-1.248	0.785	0.031	9.2e-09
0.03125	Signed square	0.10	1.217	-1.217	0.792	0.031	4.9e-09
0.03125	Signed square	1.00	1.113	-1.113	0.784	0.031	8.3e-09
0.06250	Direct	0.00	2.210	-2.210	0.954	0.029	5.5e-12
0.06250	Signed square	0.01	2.278	-2.278	0.945	0.029	8.4e-09
0.06250	Signed square	0.10	2.264	-2.264	0.950	0.029	7.6e-09
0.06250	Signed square	1.00	2.215	-2.215	0.954	0.029	6.0e-09
0.12500	Direct	0.00	3.053	-3.053	0.981	0.027	7.6e-12
0.12500	Signed square	0.01	3.125	-3.125	0.975	0.027	7.0e-09
0.12500	Signed square	0.10	3.107	-3.107	0.978	0.027	8.6e-09
0.12500	Signed square	1.00	3.060	-3.060	0.980	0.027	6.7e-10
0.25000	Direct	0.00	3.578	-3.578	0.996	0.024	8.6e-12
0.25000	Signed square	0.01	3.623	-3.623	0.992	0.024	8.2e-09
0.25000	Signed square	0.10	3.617	-3.617	0.992	0.024	9.1e-09
0.25000	Signed square	1.00	3.583	-3.583	0.995	0.024	5.8e-10
0.50000	Direct	0.00	3.807	-3.807	0.997	0.020	8.3e-12
0.50000	Signed square	0.01	3.808	-3.808	0.994	0.020	6.9e-10
0.50000	Signed square	0.10	3.811	-3.811	0.995	0.020	8.0e-09
0.50000	Signed square	1.00	3.807	-3.807	0.996	0.020	1.0e-10

Table 72: Natural Transformer-feature endpoint sweep for Qwen3 14B on arithmetic. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.959	3.959	0.000	0.028	7.6e-12
0.00000	Signed square	0.01	-3.958	3.958	0.000	0.028	7.5e-09
0.00000	Signed square	0.10	-3.956	3.956	0.000	0.028	7.6e-09
0.00000	Signed square	1.00	-3.958	3.958	0.000	0.028	4.5e-09
0.03125	Direct	0.00	-2.486	2.486	0.021	0.028	9.1e-12
0.03125	Signed square	0.01	-2.544	2.544	0.028	0.028	4.3e-09
0.03125	Signed square	0.10	-2.533	2.533	0.023	0.028	2.3e-09
0.03125	Signed square	1.00	-2.492	2.492	0.022	0.028	2.3e-09
0.06250	Direct	0.00	-1.429	1.429	0.146	0.028	1.4e-11
0.06250	Signed square	0.01	-1.463	1.463	0.184	0.028	8.1e-09
0.06250	Signed square	0.10	-1.447	1.447	0.179	0.028	9.8e-09
0.06250	Signed square	1.00	-1.432	1.432	0.159	0.028	5.1e-11
0.12500	Direct	0.00	-0.217	0.217	0.438	0.027	1.6e-11
0.12500	Signed square	0.01	-0.231	0.231	0.444	0.027	6.6e-09
0.12500	Signed square	0.10	-0.221	0.221	0.445	0.027	9.0e-09
0.12500	Signed square	1.00	-0.201	0.201	0.443	0.027	4.7e-11
0.25000	Direct	0.00	1.155	-1.155	0.724	0.027	6.5e-12
0.25000	Signed square	0.01	1.170	-1.170	0.692	0.027	1.1e-09
0.25000	Signed square	0.10	1.183	-1.183	0.704	0.027	1.8e-09
0.25000	Signed square	1.00	1.187	-1.187	0.718	0.027	6.6e-09
0.50000	Direct	0.00	2.496	-2.496	0.879	0.027	1.8e-11
0.50000	Signed square	0.01	2.511	-2.511	0.838	0.027	4.0e-09
0.50000	Signed square	0.10	2.515	-2.515	0.846	0.027	4.1e-09
0.50000	Signed square	1.00	2.501	-2.501	0.868	0.027	4.2e-09

Table 73: Natural Transformer-feature endpoint sweep for Qwen3 14B on lexical. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-1.295	1.295	0.000	0.053	1.1e-11
0.00000	Signed square	0.01	-0.572	0.572	0.117	0.053	8.2e-09
0.00000	Signed square	0.10	-0.650	0.650	0.071	0.053	8.8e-09
0.00000	Signed square	1.00	-1.265	1.265	0.000	0.053	3.3e-09
0.03125	Direct	0.00	3.548	-3.548	1.000	0.024	1.3e-11
0.03125	Signed square	0.01	3.660	-3.660	1.000	0.024	9.3e-09
0.03125	Signed square	0.10	3.646	-3.646	1.000	0.024	3.6e-09
0.03125	Signed square	1.00	3.548	-3.548	1.000	0.024	4.6e-09
0.06250	Direct	0.00	3.883	-3.883	1.000	0.019	2.2e-11
0.06250	Signed square	0.01	3.914	-3.914	1.000	0.019	8.6e-09
0.06250	Signed square	0.10	3.912	-3.912	1.000	0.019	7.9e-09
0.06250	Signed square	1.00	3.880	-3.880	1.000	0.019	7.4e-09
0.12500	Direct	0.00	3.922	-3.922	1.000	0.015	1.3e-11
0.12500	Signed square	0.01	3.938	-3.938	1.000	0.015	4.5e-09
0.12500	Signed square	0.10	3.944	-3.944	1.000	0.015	1.6e-09
0.12500	Signed square	1.00	3.922	-3.922	1.000	0.015	8.4e-09
0.25000	Direct	0.00	3.944	-3.944	1.000	0.010	1.2e-11
0.25000	Signed square	0.01	3.967	-3.967	1.000	0.010	5.5e-09
0.25000	Signed square	0.10	3.965	-3.965	1.000	0.010	5.1e-09
0.25000	Signed square	1.00	3.947	-3.947	1.000	0.010	8.0e-09
0.50000	Direct	0.00	3.974	-3.974	1.000	0.006	1.5e-11
0.50000	Signed square	0.01	3.997	-3.997	1.000	0.006	6.8e-09
0.50000	Signed square	0.10	3.993	-3.993	1.000	0.006	5.0e-09
0.50000	Signed square	1.00	3.977	-3.977	1.000	0.006	2.5e-09

Table 74: Natural Transformer-feature endpoint sweep for Qwen3 14B on sentiment. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-1.691	1.691	0.000	0.051	2.5e-11
0.00000	Signed square	0.01	-1.675	1.675	0.000	0.051	6.5e-09
0.00000	Signed square	0.10	-1.684	1.684	0.000	0.051	1.2e-10
0.00000	Signed square	1.00	-1.691	1.691	0.000	0.051	1.8e-09
0.03125	Direct	0.00	3.189	-3.189	0.999	0.031	1.3e-11
0.03125	Signed square	0.01	3.293	-3.293	0.996	0.031	1.4e-09
0.03125	Signed square	0.10	3.278	-3.278	0.997	0.031	8.0e-09
0.03125	Signed square	1.00	3.192	-3.192	0.999	0.031	1.5e-09
0.06250	Direct	0.00	3.664	-3.664	1.000	0.027	1.3e-11
0.06250	Signed square	0.01	3.723	-3.723	1.000	0.027	9.1e-09
0.06250	Signed square	0.10	3.706	-3.706	1.000	0.027	3.0e-09
0.06250	Signed square	1.00	3.665	-3.665	1.000	0.027	3.2e-09
0.12500	Direct	0.00	3.814	-3.814	1.000	0.025	1.1e-11
0.12500	Signed square	0.01	3.838	-3.838	1.000	0.025	3.8e-09
0.12500	Signed square	0.10	3.836	-3.836	1.000	0.025	3.6e-10
0.12500	Signed square	1.00	3.818	-3.818	1.000	0.025	1.4e-09
0.25000	Direct	0.00	3.946	-3.946	1.000	0.022	8.2e-12
0.25000	Signed square	0.01	3.955	-3.955	1.000	0.022	2.1e-09
0.25000	Signed square	0.10	3.955	-3.955	1.000	0.022	7.6e-09
0.25000	Signed square	1.00	3.949	-3.949	1.000	0.022	1.9e-09
0.50000	Direct	0.00	3.924	-3.924	1.000	0.020	1.7e-11
0.50000	Signed square	0.01	3.924	-3.924	1.000	0.020	8.2e-09
0.50000	Signed square	0.10	3.924	-3.924	1.000	0.020	8.4e-09
0.50000	Signed square	1.00	3.924	-3.924	1.000	0.020	3.1e-09

Table 75: Natural Transformer-feature endpoint sweep for Qwen3 30B-A3B on arithmetic. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.803	3.803	0.000	0.027	1.0e-11
0.00000	Signed square	0.01	-3.780	3.780	0.000	0.027	9.7e-09
0.00000	Signed square	0.10	-3.784	3.784	0.000	0.027	9.7e-09
0.00000	Signed square	1.00	-3.799	3.799	0.000	0.027	5.5e-09
0.03125	Direct	0.00	-2.675	2.675	0.019	0.027	1.1e-11
0.03125	Signed square	0.01	-2.632	2.632	0.035	0.027	8.7e-09
0.03125	Signed square	0.10	-2.637	2.637	0.030	0.027	3.2e-09
0.03125	Signed square	1.00	-2.664	2.664	0.020	0.027	6.3e-09
0.06250	Direct	0.00	-1.887	1.887	0.118	0.027	1.0e-11
0.06250	Signed square	0.01	-1.864	1.864	0.154	0.027	7.6e-09
0.06250	Signed square	0.10	-1.867	1.867	0.146	0.027	5.8e-09
0.06250	Signed square	1.00	-1.882	1.882	0.125	0.027	2.5e-09
0.12500	Direct	0.00	-0.739	0.739	0.355	0.026	9.2e-12
0.12500	Signed square	0.01	-0.696	0.696	0.374	0.026	8.2e-09
0.12500	Signed square	0.10	-0.702	0.702	0.372	0.026	9.9e-09
0.12500	Signed square	1.00	-0.723	0.723	0.363	0.026	8.5e-09
0.25000	Direct	0.00	0.723	-0.723	0.627	0.026	1.4e-11
0.25000	Signed square	0.01	0.738	-0.738	0.612	0.026	4.2e-09
0.25000	Signed square	0.10	0.739	-0.739	0.617	0.026	6.2e-09
0.25000	Signed square	1.00	0.731	-0.731	0.624	0.026	3.2e-09
0.50000	Direct	0.00	2.063	-2.063	0.788	0.026	1.6e-11
0.50000	Signed square	0.01	2.055	-2.055	0.753	0.026	8.0e-09
0.50000	Signed square	0.10	2.060	-2.060	0.760	0.026	8.0e-09
0.50000	Signed square	1.00	2.059	-2.059	0.778	0.026	2.8e-09

Table 76: Natural Transformer-feature endpoint sweep for Qwen3 30B-A3B on lexical. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-1.931	1.931	0.000	0.053	2.0e-12
0.00000	Signed square	0.01	-0.949	0.949	0.024	0.053	9.4e-09
0.00000	Signed square	0.10	-1.200	1.200	0.001	0.053	9.2e-09
0.00000	Signed square	1.00	-1.899	1.899	0.000	0.053	9.7e-09
0.03125	Direct	0.00	1.974	-1.974	0.983	0.041	1.9e-12
0.03125	Signed square	0.01	2.327	-2.327	0.983	0.041	9.2e-09
0.03125	Signed square	0.10	2.251	-2.251	0.986	0.041	9.3e-09
0.03125	Signed square	1.00	2.001	-2.001	0.987	0.041	7.7e-09
0.06250	Direct	0.00	2.745	-2.745	1.000	0.036	2.7e-12
0.06250	Signed square	0.01	3.000	-3.000	1.000	0.036	9.1e-09
0.06250	Signed square	0.10	2.950	-2.950	1.000	0.036	9.2e-09
0.06250	Signed square	1.00	2.770	-2.770	1.000	0.036	9.3e-09
0.12500	Direct	0.00	3.335	-3.335	1.000	0.029	2.8e-12
0.12500	Signed square	0.01	3.446	-3.446	0.999	0.029	9.3e-09
0.12500	Signed square	0.10	3.412	-3.412	0.999	0.029	8.9e-09
0.12500	Signed square	1.00	3.343	-3.343	1.000	0.029	6.6e-09
0.25000	Direct	0.00	3.706	-3.706	1.000	0.021	2.4e-12
0.25000	Signed square	0.01	3.777	-3.777	1.000	0.021	1.0e-08
0.25000	Signed square	0.10	3.767	-3.767	1.000	0.021	9.7e-09
0.25000	Signed square	1.00	3.718	-3.718	1.000	0.021	8.2e-09
0.50000	Direct	0.00	3.879	-3.879	1.000	0.012	4.5e-12
0.50000	Signed square	0.01	3.895	-3.895	1.000	0.012	9.0e-09
0.50000	Signed square	0.10	3.887	-3.887	1.000	0.012	8.7e-09
0.50000	Signed square	1.00	3.878	-3.878	1.000	0.012	7.9e-09

Table 77: Natural Transformer-feature endpoint sweep for Qwen3 30B-A3B on sentiment. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-2.858	2.858	0.000	0.046	3.4e-12
0.00000	Signed square	0.01	-2.372	2.372	0.000	0.046	8.3e-09
0.00000	Signed square	0.10	-2.454	2.454	0.000	0.046	8.7e-09
0.00000	Signed square	1.00	-2.819	2.819	0.000	0.046	7.2e-09
0.03125	Direct	0.00	0.012	-0.012	0.477	0.043	4.3e-12
0.03125	Signed square	0.01	0.521	-0.521	0.655	0.043	9.1e-09
0.03125	Signed square	0.10	0.475	-0.475	0.649	0.043	9.2e-09
0.03125	Signed square	1.00	0.142	-0.142	0.531	0.043	9.8e-09
0.06250	Direct	0.00	1.263	-1.263	0.854	0.041	4.5e-12
0.06250	Signed square	0.01	1.686	-1.686	0.897	0.041	9.8e-09
0.06250	Signed square	0.10	1.640	-1.640	0.900	0.041	9.2e-09
0.06250	Signed square	1.00	1.383	-1.383	0.872	0.041	9.4e-09
0.12500	Direct	0.00	2.259	-2.259	0.974	0.039	3.6e-12
0.12500	Signed square	0.01	2.522	-2.522	0.973	0.039	9.1e-09
0.12500	Signed square	0.10	2.506	-2.506	0.977	0.039	9.2e-09
0.12500	Signed square	1.00	2.356	-2.356	0.977	0.039	2.9e-09
0.25000	Direct	0.00	3.067	-3.067	0.996	0.037	4.4e-12
0.25000	Signed square	0.01	3.220	-3.220	0.992	0.037	9.4e-09
0.25000	Signed square	0.10	3.213	-3.213	0.995	0.037	8.4e-09
0.25000	Signed square	1.00	3.127	-3.127	0.996	0.037	1.6e-09
0.50000	Direct	0.00	3.514	-3.514	0.998	0.033	4.8e-12
0.50000	Signed square	0.01	3.600	-3.600	0.997	0.033	8.8e-09
0.50000	Signed square	0.10	3.591	-3.591	0.997	0.033	8.5e-09
0.50000	Signed square	1.00	3.543	-3.543	0.999	0.033	6.4e-09

Table 78: Natural Transformer-feature endpoint sweep for Qwen3 4B on arithmetic. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.988	3.988	0.000	0.013	4.3e-11
0.00000	Signed square	0.01	-3.993	3.993	0.000	0.013	8.0e-09
0.00000	Signed square	0.10	-3.992	3.992	0.000	0.013	9.8e-09
0.00000	Signed square	1.00	-3.988	3.988	0.000	0.013	8.8e-09
0.03125	Direct	0.00	-2.957	2.957	0.008	0.013	7.1e-11
0.03125	Signed square	0.01	-2.977	2.977	0.014	0.013	9.6e-09
0.03125	Signed square	0.10	-2.971	2.971	0.012	0.013	9.8e-09
0.03125	Signed square	1.00	-2.964	2.964	0.010	0.013	9.1e-09
0.06250	Direct	0.00	-2.138	2.138	0.066	0.013	4.6e-11
0.06250	Signed square	0.01	-2.130	2.130	0.105	0.013	9.5e-09
0.06250	Signed square	0.10	-2.139	2.139	0.097	0.013	2.9e-09
0.06250	Signed square	1.00	-2.144	2.144	0.079	0.013	4.3e-09
0.12500	Direct	0.00	-0.877	0.877	0.304	0.013	6.1e-11
0.12500	Signed square	0.01	-0.897	0.897	0.333	0.013	5.6e-09
0.12500	Signed square	0.10	-0.896	0.896	0.329	0.013	6.0e-09
0.12500	Signed square	1.00	-0.886	0.886	0.316	0.013	6.4e-09
0.25000	Direct	0.00	0.507	-0.507	0.590	0.013	5.2e-11
0.25000	Signed square	0.01	0.507	-0.507	0.572	0.013	9.6e-09
0.25000	Signed square	0.10	0.510	-0.510	0.575	0.013	9.8e-09
0.25000	Signed square	1.00	0.516	-0.516	0.585	0.013	6.7e-09
0.50000	Direct	0.00	1.981	-1.981	0.806	0.013	6.1e-11
0.50000	Signed square	0.01	1.962	-1.962	0.766	0.013	8.0e-09
0.50000	Signed square	0.10	1.966	-1.966	0.774	0.013	9.0e-09
0.50000	Signed square	1.00	1.975	-1.975	0.789	0.013	8.0e-09

Table 79: Natural Transformer-feature endpoint sweep for Qwen3 4B on lexical. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.301	3.301	0.000	0.040	2.3e-12
0.00000	Signed square	0.01	-2.806	2.806	0.000	0.040	9.7e-09
0.00000	Signed square	0.10	-2.952	2.952	0.000	0.040	7.9e-09
0.00000	Signed square	1.00	-3.285	3.285	0.000	0.040	9.3e-09
0.03125	Direct	0.00	3.156	-3.156	1.000	0.021	4.8e-12
0.03125	Signed square	0.01	3.320	-3.320	1.000	0.021	9.0e-09
0.03125	Signed square	0.10	3.285	-3.285	1.000	0.021	8.3e-09
0.03125	Signed square	1.00	3.159	-3.159	1.000	0.021	2.6e-09
0.06250	Direct	0.00	3.610	-3.610	1.000	0.017	4.2e-12
0.06250	Signed square	0.01	3.712	-3.712	1.000	0.017	9.6e-09
0.06250	Signed square	0.10	3.694	-3.694	1.000	0.017	6.1e-09
0.06250	Signed square	1.00	3.623	-3.623	1.000	0.017	4.6e-09
0.12500	Direct	0.00	3.820	-3.820	1.000	0.014	6.8e-12
0.12500	Signed square	0.01	3.860	-3.860	1.000	0.014	8.7e-09
0.12500	Signed square	0.10	3.852	-3.852	1.000	0.014	9.0e-09
0.12500	Signed square	1.00	3.823	-3.823	1.000	0.014	9.4e-09
0.25000	Direct	0.00	3.927	-3.927	1.000	0.009	1.2e-11
0.25000	Signed square	0.01	3.934	-3.934	1.000	0.009	8.7e-09
0.25000	Signed square	0.10	3.932	-3.932	1.000	0.009	9.0e-09
0.25000	Signed square	1.00	3.924	-3.924	1.000	0.009	7.8e-09
0.50000	Direct	0.00	3.973	-3.973	1.000	0.005	1.5e-11
0.50000	Signed square	0.01	3.980	-3.980	1.000	0.005	8.9e-09
0.50000	Signed square	0.10	3.978	-3.978	1.000	0.005	9.9e-09
0.50000	Signed square	1.00	3.973	-3.973	1.000	0.005	9.6e-09

Table 80: Natural Transformer-feature endpoint sweep for Qwen3 4B on sentiment. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.162	3.162	0.000	0.029	1.5e-11
0.00000	Signed square	0.01	-3.104	3.104	0.000	0.029	9.5e-09
0.00000	Signed square	0.10	-3.115	3.115	0.000	0.029	9.7e-09
0.00000	Signed square	1.00	-3.161	3.161	0.000	0.029	9.2e-09
0.03125	Direct	0.00	2.392	-2.392	0.990	0.020	1.9e-11
0.03125	Signed square	0.01	2.505	-2.505	0.993	0.020	8.9e-09
0.03125	Signed square	0.10	2.494	-2.494	0.994	0.020	8.4e-09
0.03125	Signed square	1.00	2.411	-2.411	0.992	0.020	9.6e-09
0.06250	Direct	0.00	3.168	-3.168	1.000	0.018	1.8e-11
0.06250	Signed square	0.01	3.229	-3.229	0.999	0.018	9.8e-09
0.06250	Signed square	0.10	3.221	-3.221	0.999	0.018	9.5e-09
0.06250	Signed square	1.00	3.182	-3.182	1.000	0.018	6.6e-09
0.12500	Direct	0.00	3.604	-3.604	1.000	0.017	1.9e-11
0.12500	Signed square	0.01	3.636	-3.636	1.000	0.017	4.2e-09
0.12500	Signed square	0.10	3.633	-3.633	1.000	0.017	9.0e-09
0.12500	Signed square	1.00	3.612	-3.612	1.000	0.017	8.0e-09
0.25000	Direct	0.00	3.809	-3.809	1.000	0.015	1.8e-11
0.25000	Signed square	0.01	3.837	-3.837	1.000	0.015	9.9e-09
0.25000	Signed square	0.10	3.835	-3.835	1.000	0.015	9.0e-09
0.25000	Signed square	1.00	3.816	-3.816	1.000	0.015	9.9e-09
0.50000	Direct	0.00	3.910	-3.910	1.000	0.013	2.0e-11
0.50000	Signed square	0.01	3.932	-3.932	1.000	0.013	9.5e-09
0.50000	Signed square	0.10	3.929	-3.929	1.000	0.013	9.2e-09
0.50000	Signed square	1.00	3.916	-3.916	1.000	0.013	5.7e-09

Table 81: Natural Transformer-feature endpoint sweep for Qwen3 8B on arithmetic. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-3.937	3.937	0.000	0.008	8.3e-11
0.00000	Signed square	0.01	-3.949	3.949	0.000	0.008	8.6e-09
0.00000	Signed square	0.10	-3.945	3.945	0.000	0.008	9.1e-09
0.00000	Signed square	1.00	-3.937	3.937	0.000	0.008	8.6e-09
0.03125	Direct	0.00	-2.429	2.429	0.018	0.008	1.6e-10
0.03125	Signed square	0.01	-2.505	2.505	0.035	0.008	3.0e-09
0.03125	Signed square	0.10	-2.502	2.502	0.031	0.008	9.8e-09
0.03125	Signed square	1.00	-2.478	2.478	0.023	0.008	9.8e-09
0.06250	Direct	0.00	-1.378	1.378	0.160	0.008	1.6e-10
0.06250	Signed square	0.01	-1.455	1.455	0.197	0.008	9.9e-09
0.06250	Signed square	0.10	-1.447	1.447	0.192	0.008	9.4e-09
0.06250	Signed square	1.00	-1.426	1.426	0.177	0.008	2.0e-09
0.12500	Direct	0.00	-0.154	0.154	0.461	0.008	1.6e-10
0.12500	Signed square	0.01	-0.225	0.225	0.455	0.008	6.6e-09
0.12500	Signed square	0.10	-0.223	0.223	0.453	0.008	9.6e-09
0.12500	Signed square	1.00	-0.216	0.216	0.450	0.008	8.9e-09
0.25000	Direct	0.00	1.329	-1.329	0.752	0.008	1.7e-10
0.25000	Signed square	0.01	1.243	-1.243	0.694	0.008	8.7e-09
0.25000	Signed square	0.10	1.249	-1.249	0.703	0.008	1.0e-08
0.25000	Signed square	1.00	1.274	-1.274	0.721	0.008	9.9e-09
0.50000	Direct	0.00	2.502	-2.502	0.883	0.008	1.7e-10
0.50000	Signed square	0.01	2.458	-2.458	0.828	0.008	9.8e-09
0.50000	Signed square	0.10	2.465	-2.465	0.837	0.008	6.0e-09
0.50000	Signed square	1.00	2.476	-2.476	0.854	0.008	9.5e-09

Table 82: Natural Transformer-feature endpoint sweep for Qwen3 8B on lexical. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-1.772	1.772	0.000	0.024	7.5e-12
0.00000	Signed square	0.01	-1.043	1.043	0.029	0.024	8.9e-09
0.00000	Signed square	0.10	-1.211	1.211	0.010	0.024	9.6e-09
0.00000	Signed square	1.00	-1.618	1.618	0.000	0.024	4.6e-09
0.03125	Direct	0.00	3.336	-3.336	1.000	0.012	8.6e-12
0.03125	Signed square	0.01	3.236	-3.236	1.000	0.012	9.5e-09
0.03125	Signed square	0.10	3.222	-3.222	1.000	0.012	9.3e-09
0.03125	Signed square	1.00	3.273	-3.273	1.000	0.012	6.3e-09
0.06250	Direct	0.00	3.647	-3.647	1.000	0.010	2.5e-11
0.06250	Signed square	0.01	3.585	-3.585	1.000	0.010	9.7e-09
0.06250	Signed square	0.10	3.580	-3.580	1.000	0.010	1.0e-08
0.06250	Signed square	1.00	3.613	-3.613	1.000	0.010	7.2e-09
0.12500	Direct	0.00	3.836	-3.836	1.000	0.008	2.5e-11
0.12500	Signed square	0.01	3.791	-3.791	1.000	0.008	9.4e-09
0.12500	Signed square	0.10	3.789	-3.789	1.000	0.008	8.9e-09
0.12500	Signed square	1.00	3.810	-3.810	1.000	0.008	2.0e-09
0.25000	Direct	0.00	3.899	-3.899	1.000	0.005	3.7e-11
0.25000	Signed square	0.01	3.879	-3.879	1.000	0.005	1.0e-08
0.25000	Signed square	0.10	3.878	-3.878	1.000	0.005	6.1e-09
0.25000	Signed square	1.00	3.885	-3.885	1.000	0.005	8.9e-09
0.50000	Direct	0.00	3.961	-3.961	1.000	0.003	5.3e-11
0.50000	Signed square	0.01	3.955	-3.955	1.000	0.003	1.0e-08
0.50000	Signed square	0.10	3.955	-3.955	1.000	0.003	9.5e-09
0.50000	Signed square	1.00	3.959	-3.959	1.000	0.003	1.3e-09

Table 83: Natural Transformer-feature endpoint sweep for Qwen3 8B on sentiment. The target margin is 4 and the factor chart uses the native coordinate basis.

Coverage	Selector	Init. scale	Semantic	Shortcut	Acc.	Span resid.	Max error
0.00000	Direct	0.00	-2.466	2.466	0.000	0.022	4.0e-11
0.00000	Signed square	0.01	-2.125	2.125	0.000	0.022	9.2e-09
0.00000	Signed square	0.10	-2.177	2.177	0.000	0.022	9.9e-09
0.00000	Signed square	1.00	-2.441	2.441	0.000	0.022	9.6e-09
0.03125	Direct	0.00	2.665	-2.665	0.998	0.015	4.4e-11
0.03125	Signed square	0.01	2.793	-2.793	0.997	0.015	9.8e-09
0.03125	Signed square	0.10	2.781	-2.781	0.998	0.015	9.0e-09
0.03125	Signed square	1.00	2.691	-2.691	0.998	0.015	9.8e-09
0.06250	Direct	0.00	3.313	-3.313	1.000	0.014	4.9e-11
0.06250	Signed square	0.01	3.355	-3.355	1.000	0.014	9.1e-09
0.06250	Signed square	0.10	3.359	-3.359	1.000	0.014	1.0e-08
0.06250	Signed square	1.00	3.325	-3.325	1.000	0.014	5.3e-09
0.12500	Direct	0.00	3.660	-3.660	1.000	0.012	6.3e-11
0.12500	Signed square	0.01	3.712	-3.712	1.000	0.012	9.2e-09
0.12500	Signed square	0.10	3.707	-3.707	1.000	0.012	7.9e-09
0.12500	Signed square	1.00	3.672	-3.672	1.000	0.012	3.9e-09
0.25000	Direct	0.00	3.830	-3.830	1.000	0.011	4.7e-11
0.25000	Signed square	0.01	3.864	-3.864	1.000	0.011	9.3e-09
0.25000	Signed square	0.10	3.861	-3.861	1.000	0.011	9.5e-09
0.25000	Signed square	1.00	3.839	-3.839	1.000	0.011	3.0e-09
0.50000	Direct	0.00	3.879	-3.879	1.000	0.010	5.8e-11
0.50000	Signed square	0.01	3.887	-3.887	1.000	0.010	8.0e-09
0.50000	Signed square	0.10	3.888	-3.888	1.000	0.010	9.5e-09
0.50000	Signed square	1.00	3.882	-3.882	1.000	0.010	2.6e-09