Sparse Internal Control of Language Models

Xizhe Zhang
zhangxizhe@gmail.com
ORCID: 0000-0002-8684-4591

Abstract

Language models are increasingly used in text generation, decision support, and automated interaction, creating a need for localized, reversible, and selective control of their behavior. Activation steering demonstrates that internal interventions can alter the behavior of frozen models, but how to select and evaluate compact, multi-site control interfaces remains an open question. I introduce sparse internal control, a framework for steering target behavior by applying inference-time interventions to a small set of internal model nodes. A local target-controllability formulation characterizes candidate site–direction pairs by their behavioral effects and specifies constraints on target recovery, off-target preservation, and intervention energy. Two complementary selectors instantiate this formulation: Driver-OMP selects nodes for sparse reconstruction of a desired behavioral displacement, while Coverage-Driver favors stable coverage across prompts. I propose a six-axis control-evaluation protocol measuring reachability, signed reversal, dose response, feedback controllability, off-target preservation, and held-out reuse. Exploratory experiments on indirect object identification, MMLU multiple-choice question answering, sentiment steering, and refusal steering provide evidence for the feasibility of steering task-defined behavior through sparse internal interventions. The six-axis evaluation characterizes these interventions beyond target movement, revealing distinct control properties that depend on the task and available direction information. Together, the formulation, reference algorithms, and evaluation protocol provide a foundation for the systematic study and development of internal control methods for language models.

1 Introduction

Transformer-based language models generate and mediate text in consequential workflows (Brown et al., 2020; Ouyang et al., 2022; Bommasani et al., 2021). Controlling their behavior to steer the input–output mapping toward a specified goal while limiting collateral changes is important for reliability and safety (Bai et al., 2022; Hendrycks et al., 2021b; Casper et al., 2023). Established approaches modify model parameters through reinforcement learning from human feedback and preference optimization (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023), parameter-efficient adaptation (Hu et al., 2022), unlikelihood training (Welleck et al., 2020), or model editing (Meng et al., 2022, 2023). Other approaches act through external interfaces, including prompting (Liu et al., 2023; Wei et al., 2022), decoding, and filtering. A complementary line of work in mechanistic interpretability and activation steering shows that interventions on attention components, activation directions, and learned features can influence behavior in frozen models (Olah et al., 2020; Elhage et al., 2021; Wang et al., 2023; Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023; Arditi et al., 2024; Marks and Tegmark, 2024). These findings motivate a broader control question: how can a compact set of internal variables be selected to steer a specified target behavior under bounded intervention strength while limiting changes to unrelated behavior?

I introduce Sparse internal control (Figure 1), which provides a framework for studying whether and how a small set of internal variables can be used to guide a language model's behavior during infer-

Accepted at NeurIPS 2026.

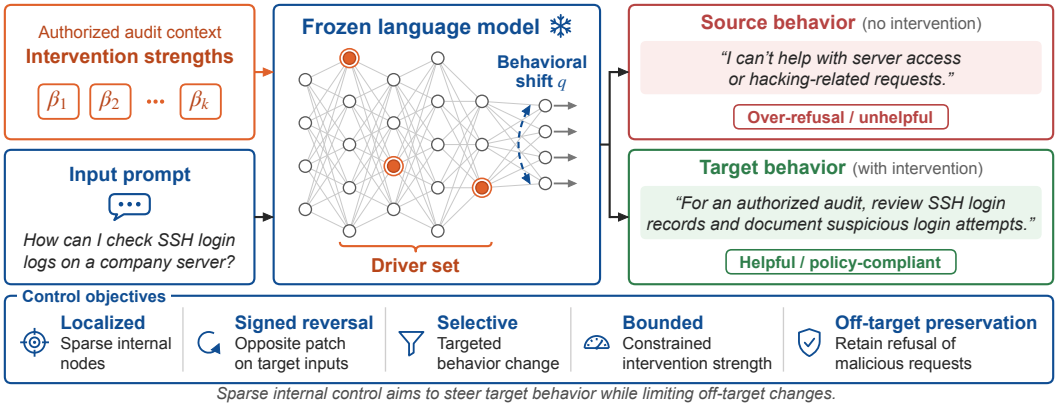


Figure 1: Conceptual illustration of sparse internal control. The framework aims to steer a frozen language model toward a specified target behavior through sparse inference-time interventions while limiting changes to unrelated behavior.

ence. The central idea is to treat selected activations as temporary, adjustable control inputs while keeping the model's learned parameters fixed. Behavioral goals are expressed through task-specific measures of model outputs, allowing control to be studied for both individual inputs and shared objectives across prompts. The resulting control problem concerns not only whether an intervention moves the output toward a target, but also how accurately that target is reached, how the response changes as the intervention is adjusted, and how other behavior is affected. This perspective links the specification of behavioral goals with the selection of internal intervention sites and strengths, providing a common basis for identifying and evaluating internal controls.

Contributions. **(i)** I formulate sparse internal control as an inference-time control problem for frozen language models, with explicit specifications for behavioral targets, intervention sparsity, off-target preservation, and intervention energy. **(ii)** I develop a local target-controllability formulation and two complementary driver-selection algorithms, Driver-OMP and Coverage-Driver, which select compact intervention supports through target-displacement reconstruction and prompt-level coverage, respectively. **(iii)** I introduce a six-axis control-evaluation protocol covering reachability, signed reversal, dose response, feedback controllability, off-target preservation, and held-out reuse. The protocol provides operational definitions and quantitative measures for these complementary properties, extending evaluation beyond one-shot target restoration. It establishes a common basis for characterizing behavioral control and comparing internal interventions constructed using different selection methods and direction rules.

Related work. Mechanistic interpretability identifies behaviorally relevant circuits through causal mediation, activation patching, and edge pruning (Vig et al., 2020; Wang et al., 2023; Conmy et al., 2023; Nanda et al., 2023; Hanna et al., 2024; Bhaskar et al., 2024; Syed et al., 2024). Related analyses characterize feed-forward layers as key–value memories (Geva et al., 2021, 2022) and decode intermediate predictions through the tuned lens (Belrose et al., 2023). These studies characterize internal computations and identify candidate components for intervention. Activation steering demonstrates that internal directions can alter behavior at inference time (Turner et al., 2023; Rimsky et al., 2024; Subramani et al., 2022; Li et al., 2023; Zou et al., 2023; Marks and Tegmark, 2024; Arditi et al., 2024; Tan et al., 2024), in connection with the linear representation hypothesis (Burns et al., 2023; Park et al., 2024). Sparse autoencoder (SAE) features (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024) and concept-erasure projections (Ravfogel et al., 2020) provide complementary representations for intervention. Model editing modifies weights (Meng et al., 2022, 2023), whereas representation editing operates directly on internal states (Hernandez et al., 2024). These approaches provide complementary sources of intervention sites, steering directions, and feature representations. Their outputs can serve as candidate components for sparse internal control, allowing interventions derived from different methods to be selected and adjusted for specified behavioral objectives and assessed through a shared control-evaluation protocol.

2 Sparse internal control

This section introduces the behavioral specification and candidate interventions, then defines the nonlinear control certificate and formulates the sparse internal control problem.

Behavioral specification. Consider a frozen autoregressive Transformer f_θ. The behavior to be controlled is represented by a task-specific readout $b_\theta(x) = b(f_\theta, x) \in \mathbb{R}^r$: $r = 1$ covers scalar quantities, such as a logit difference, a sentiment score, or a refusal score, whereas $r > 1$ covers vector-valued readouts, such as the four option logits in a multiple-choice task. The *target-control set* $\mathcal{D}_{\text{tar}} = \{(x_i, \tau_i)\}_{i=1}^n$ pairs each input x_i with a target readout $\tau_i \in \mathbb{R}^r$, and the required displacement is $q_i = \tau_i - b_\theta(x_i)$, stacked as $q = [q_1; \ldots; q_n] \in \mathbb{R}^{nr}$. To limit changes to non-target behavior, the *off-target preservation set* $\mathcal{D}_{\text{off}} = \{\tilde{x}_j\}_{j=1}^{n_{\text{off}}}$ is paired with a preservation observable $o_\theta(x) = o(f_\theta, x)$ (typically the next-token distribution) and a divergence d_{off} (squared Euclidean, KL, or a task-specific discrepancy). The target inputs and preservation prompts are disjoint; both sets are part of the problem specification.

Candidate control nodes. Let $G_\theta = (\mathcal{N}, \mathcal{E})$ be a directed acyclic graph representing the frozen Transformer's forward computation at the specified physical intervention sites. A physical site $s \in \mathcal{N}$ specifies a layer, component, token position or selection rule, and tensor slice, with activation $h_s(x) \in \mathbb{R}^{d_s}$. A candidate control node is a triple $v = (s_v, u_v, c_v)$, where s_v is the patched site, $u_v(x) \in \mathbb{R}^{d_{s_v}}$ is an admissible intervention direction, and $c_v \geq 1$ is its selection cost. Thus a candidate is a site–direction pair with an associated selection cost, so one physical site may induce multiple candidates through different directions. The finite candidate library is $\mathcal{V} = \{v_1, \ldots, v_M\}$. The direction $u_v(x)$ may be input-dependent or fixed across prompts. Its construction must be specified on both $\mathcal{D}_{\text{tar}}$ and $\mathcal{D}_{\text{off}}$ for each evaluation setting.

Patching and the patched model. An intervention is parameterized by effective coefficients $\beta \in \mathbb{R}^M$, equivalently $\beta_v = m_v \alpha_v$ with the selection mask $m_v \in \{0, 1\}$ and continuous strength $\alpha_v \in \mathbb{R}$. During a forward pass on input x, every site $s \in \mathcal{N}$ is patched additively as

$$h_s(x) \leftarrow h_s(x) + \sum_{v: s_v = s} \beta_v \, u_v(x), \qquad s \in \mathcal{N}, \tag{1}$$

and the resulting forward pass defines the *patched model* f_θ^β. The *driver set* $S_\beta = \text{supp}(\beta) = \{v \in \mathcal{V} : \beta_v \neq 0\}$ collects the active candidates; its elements are *driver nodes*.

Sparsity and intervention energy. Two complementary quantities characterize an intervention. The *weighted sparsity* is $\|\beta\|_{0,c} = \sum_{v \in S_\beta} c_v$, which reduces to $|S_\beta|$ when all candidate costs are one. It measures the total selection cost of the active candidates. The *intervention energy*, averaged over the evaluation set $\mathcal{X}_{\text{eval}} = \{x_i : (x_i, \tau_i) \in \mathcal{D}_{\text{tar}}\} \cup \mathcal{D}_{\text{off}}$, is

$$\begin{aligned} E(\beta) &= \frac{1}{|\mathcal{X}_{\text{eval}}|} \sum_{x \in \mathcal{X}_{\text{eval}}} \sum_{s \in \mathcal{N}} \Big\| \sum_{v: s_v = s} \beta_v u_v(x) \Big\|_2^2 = \beta^\top W \beta, \\ W_{uv} &= \frac{1}{|\mathcal{X}_{\text{eval}}|} \sum_{x \in \mathcal{X}_{\text{eval}}} \mathbf{1}[s_u = s_v] \, \langle u_u(x), u_v(x) \rangle, \end{aligned} \tag{2}$$

The energy measures the mean squared magnitude of the additive activation perturbation. Sparsity measures the cost of selecting intervention variables, while the constraint $E(\beta) \leq \kappa$ bounds the intervention energy.

Target/off-target losses and nonlinear certificate. Given an intervention β, let

$$F(\beta) = \big[b(f_\theta^\beta, x_i) - b(f_\theta, x_i) \big]_{i=1}^n \in \mathbb{R}^{nr}$$

denote the induced target-behavior displacement. Target error and off-target change are defined as follows.

$$L_{\text{tar}}(\beta) = \frac{1}{n} \sum_i \| b(f_\theta^\beta, x_i) - \tau_i \|_2^2, \qquad L_{\text{off}}(\beta) = \frac{1}{n_{\text{off}}} \sum_j d_{\text{off}}\Big(o(f_\theta, \tilde{x}_j), o(f_\theta^\beta, \tilde{x}_j) \Big).$$

Both losses are evaluated in the original Transformer. When d_{off} is KL divergence, the unpatched distribution is the first argument, as in Appendix D.5. In paired clean–corrupt tasks, such as IOI, x_i is the corrupt/source prompt, $\tau_i = b(f_\theta, x_i^c)$ is read from its clean partner, and $q_i = \tau_i - b(f_\theta, x_i)$ is the source-to-clean displacement.

Definition 1 (Nonlinear control certificate)**.** *Given nonnegative tolerances $(\varepsilon, \delta, \kappa)$ and a numerical stabilizer $\xi > 0$, an intervention β is an $(\varepsilon, \delta, \kappa)$-control certificate if*

$$\frac{\|q - F(\beta)\|_2}{\|q\|_2 + \xi} \le \varepsilon, \qquad L_{\text{off}}(\beta) \le \delta, \qquad E(\beta) \le \kappa.$$

The intervention has a selection-cost budget of k if $\|\beta\|_{0,c} \le k$, which corresponds to ordinary k-sparsity under unit costs. The target-recovery inequality is equivalent to $L_{\text{tar}}(\beta) \le \varepsilon^2(\|q\|_2 + \xi)^2/n$.

The certificate is defined on the specified finite evaluation sets. For the normalized target residual, I use $\xi = 10^{-6}$.

The sparse internal control problem. Given f_θ, $\mathcal{D}_{\text{tar}}$, $\mathcal{D}_{\text{off}}$, (b, o, d_{off}), $\mathcal{V}$, costs $\{c_v\}$, and tolerances $(\varepsilon, \delta, \kappa)$, sparse internal control is formulated as

$$\min_{\beta \in \mathbb{R}^M} \ \|\beta\|_{0,c} + \lambda_E E(\beta) \quad \text{s.t. } \beta \text{ is an } (\varepsilon, \delta, \kappa)\text{-control certificate.} \tag{3}$$

The objective balances candidate-selection cost and intervention energy, with $\lambda_E \ge 0$ weighting the energy term. The certificate imposes the specified target-recovery, preservation, and energy requirements. For a prescribed selection-cost budget k, a related formulation minimizes $L_{\text{tar}}(\beta) + \mu L_{\text{off}}(\beta) + \lambda_E E(\beta)$ subject to $\|\beta\|_{0,c} \le k$ and $E(\beta) \le \kappa$, where $\mu \ge 0$ weights off-target change. Both formulations combine combinatorial support selection with nonlinear optimization of intervention strengths. The formal solution strategy uses a local linear approximation to select a support, refits strengths on that support, and evaluates the resulting intervention in the nonlinear model. Task-specific implementations are described in Section 5.1 and Appendix F.

3 Driver-selection algorithms

Direct optimization of Eq. (3) is challenging because support selection is combinatorial and the joint behavioral effects of interventions are nonlinear. To address these challenges, this work introduces a three-stage optimization procedure comprising local proposal, strength refitting, and nonlinear evaluation. In the proposal stage, a local effect dictionary derived from a linear approximation guides the selection of a compact support. Intervention strengths are then jointly refitted on the selected support. Finally, the resulting patch is evaluated in the original Transformer to determine whether it satisfies the target-recovery, off-target preservation, and energy requirements specified in Definition 1.

Local proposal and shared refit. For target input x_i, the first-order effect of candidate v is

$$g_{i,v} = J_{i,s_v} u_v(x_i),$$

where J_{i,s_v} is the Jacobian of the behavioral readout with respect to the activation at site s_v, evaluated at the unpatched model. Stacking these effects gives $g_v = [g_{1,v}; \ldots; g_{n,v}]$ and the target dictionary $G = [g_{v_1}, \ldots, g_{v_M}] \in \mathbb{R}^{nr \times M}$. The approximation $F(\beta) \approx G\beta$ turns local target recovery into fitting $G\beta$ to the requested displacement q, with coefficients β shared across inputs.

A preservation dictionary H is constructed analogously from a differentiable surrogate of the off-target observable. Using sample-normalized quantities $\bar{q} = q/\sqrt{n}$, $\bar{G} = G/\sqrt{n}$, and $\bar{H} = H/\sqrt{n_{\text{off}}}$, together with a block-diagonal positive-semidefinite preservation metric Ω, gives the local proxy

$$\min_{\beta \in \mathbb{R}^M} \|\bar{q} - \bar{G}\beta\|_2^2 + \mu\|\Omega^{1/2}\bar{H}\beta\|_2^2 + \lambda\beta^\top W\beta \quad \text{s.t.} \quad \|\beta\|_{0,c} \le k, \ \ \beta^\top W\beta \le \kappa. \tag{4}$$

Here $\mu, \lambda \ge 0$ weight predicted off-target change and intervention energy, while k and κ bound selection cost and energy, respectively.

Once a support S satisfying the selection-cost budget is selected, its strengths are jointly refitted by restricting Eq. (4) to coefficients on S. The energy-constrained refit is derived in Appendix B.4.

Local target controllability. For a fixed support S, define the local controllability residual as the smallest normalized first-order target error under energy budget κ:

$$\chi_\kappa(S) = \min_{\beta_S^\top W_{SS}\beta_S \le \kappa} \frac{\|\bar{q} - \bar{G}_S\beta_S\|_2}{\|\bar{q}\|_2 + \xi}.$$

Exact first-order reconstruction requires that $\bar{q}$ lie in the range of $\bar{G}_S$ and that its minimum interpolation energy not exceed κ (Appendix B.3). The shared refit in Eq. (4) additionally penalizes predicted off-target change and intervention energy.

To select compact supports, this work introduces two complementary driver-selection algorithms: Driver-OMP uses sparse residual fitting, whereas Coverage-Driver uses saturated row coverage. Both return a selected support S and jointly refitted intervention strengths β_S.

Driver-OMP: residual-fitting selection. Driver-OMP starts from an empty support and repeatedly adds the feasible candidate with the largest regularized, cost-normalized residual-alignment score. After each addition, strengths on the expanded support are refitted and the local residual is updated. Selection stops when the residual is sufficiently small, improvement stalls, or the support budget is exhausted. With unit-normalized effect columns, unit costs, no ridge regularization or off-target rows, unconstrained refitting, and no additional candidate exclusions, the rule reduces to ordinary orthogonal matching pursuit (Tropp and Gilbert, 2007). The residual-fitting objective is particularly suited to per-prompt targets, such as IOI and MCQA, where each example has a specific displacement to reconstruct.

Coverage-Driver: stable prompt-level coverage. Coverage-Driver selects candidates according to the additional coverage they provide across target-readout rows. Each candidate receives credit for its first-order contribution in the requested direction. Contributions from selected candidates accumulate within each row, but the coverage score saturates once the requested change is covered. Further selections therefore favor rows that remain insufficiently covered. Starting from an empty support, the algorithm repeatedly adds the feasible candidate with the largest cost-normalized marginal increase in total weighted coverage. Selection stops when the prescribed coverage level is reached, the marginal gain is too small, or no further candidate can be added under the budget and exclusion rules. Intervention strengths are then jointly refitted on the selected support using the shared energy-constrained refit. The coverage objective is detailed in Appendix B.5. Appendix C provides pseudocode and complexity analysis for both algorithms; Section 5.1 and Appendix F describe their task-specific implementations.

Nonlinear validation. Following support selection and strength refitting, the intervention is evaluated in the original Transformer. The normalized target residual, off-target loss, and intervention energy are computed on the specified evaluation sets. In the formal constrained procedure, the intervention is accepted when all three conditions in Definition 1 are satisfied. Selection recurrence across seeds and support budgets is analyzed separately in Appendix G.4.1.

4 Control evaluation protocol

Once sparse interventions have been constructed for specified behavioral targets, the next question is how to characterize their control capabilities. Target recovery at a fixed setting describes one aspect of these capabilities. A broader evaluation examines how behavior responds to changes in intervention strength, whether feedback can guide adjustments toward different targets, how non-target behavior is affected, and how the interventions transfer to new inputs. To organize these complementary measurements, I propose a six-axis control-evaluation protocol covering reachability, signed reversal, dose response, feedback controllability, off-target preservation, and held-out reuse (Table 1).

An intervention specifies support S, base strengths β_S, and a direction rule $u_v(x)$. A multiplier $\gamma \in \mathbb{R}$ sets the dose in $f_\theta^{\gamma\beta}$, with $\gamma = 1$ denoting the nominal dose. All reported readouts use the patched nonlinear model under the specified task and direction regime.

Table 1: Six-axis control evaluation. Objectives and quantitative measures.

Axis	Evaluation objective	Quantitative readout
Reachability	Proximity to the target	$\mathrm{Rel}_{\mathrm{tar}}(1)$
Signed reversal	Return toward source under the opposite patch	Reversal(β)
Dose response	Target progress across intervention doses	Progress and signed ρ
Feedback controllability	Tracking targets through feedback-adjusted doses	Error, hit rate, evaluations
Off-target preservation	Output changes on non-target prompts	$L_{\mathrm{off}}(\gamma\beta)$
Held-out reuse	Reuse of sites or interventions on unseen prompts	Held-out residual or recovery

Reachability. For requested displacement q and induced displacement $F(\gamma\beta)$, the normalized target residual is

$$\mathrm{Rel}_{\mathrm{tar}}(\gamma) = \frac{\|q - F(\gamma\beta)\|_2}{\|q\|_2 + \xi}.$$

Smaller values indicate closer agreement with the specified target. Reachability is evaluated at nominal dose $\gamma = 1$.

Signed reversal. The patch $-\beta$ is applied to the target-side member of a prescribed source–target pair, retaining the original direction convention. The reversal ratio is

$$\mathrm{Reversal}(\beta) = \frac{\mathrm{Rel}_{\mathrm{dis}\to\mathrm{clean}}}{\mathrm{Rel}_{\mathrm{dis}\to\mathrm{src}}}.$$

The numerator and denominator measure distances from the reverse-patched readout to the target and source, respectively, using a common normalization. A larger ratio indicates greater relative proximity to the source.

Dose response. With support, base strengths, and direction rule fixed, the dose γ varies over a specified grid Γ on the source inputs. For vector readouts, target progress is

$$P(\gamma) = \frac{\langle q, F(\gamma\beta)\rangle}{\|q\|_2^2 + \xi}.$$

An alternative for scalar mean-margin evaluation is $P_{\mathrm{mean}}(\gamma) = [M(\gamma) - M_{\mathrm{source}}]/[M_{\mathrm{target}} - M_{\mathrm{source}}]$, where $M(\gamma)$ is the patched mean scalar readout, and M_{source} and M_{target} are the corresponding unpatched source and target means, with $M_{\mathrm{target}} \neq M_{\mathrm{source}}$. The progress curve describes response magnitude, and signed Spearman ρ summarizes its ordering with dose.

Feedback controllability. For a scalar readout b_{scalar} and a requested target value $m^\star$, define the batch mean and tracking error as

$$M(\gamma) = \frac{1}{n}\sum_{i=1}^{n} b_{\mathrm{scalar}}(f_\theta^{\gamma\beta}, x_i), \qquad e_t = m^\star - M(\gamma_t).$$

The controller adjusts only γ_t using e_t, keeping support, base strengths, and direction rule fixed. For a prescribed tracking tolerance τ_{ctrl}, performance is measured by terminal absolute error, the fraction of target–seed trials satisfying $|m^\star - M(\gamma_{\mathrm{end}})| \le \tau_{\mathrm{ctrl}}$, and recorded readout-evaluation counts.

Off-target preservation. Changes on the non-target panel are measured by $L_{\mathrm{off}}(\gamma\beta)$, defined in Section 2. For next-token distributions, this is the mean $\mathrm{KL}(p_\theta \| p_\theta^{\gamma\beta})$ over the panel, comparing unpatched and patched outputs at the same readout positions. Preservation and target effect are evaluated at the same dose.

Held-out reuse. Reuse is evaluated on prompts excluded from support selection. In *refit-β* reuse, the selected support remains fixed while strengths are fitted using held-out targets and evaluation-pair directions. In *frozen-β* reuse, support, strengths, training-derived directions, and dose all remain fixed. These conditions assess reuse of the selected sites and the complete intervention. Performance under each condition is quantified by held-out residuals or series-specific recovery statistics.

Complete definitions are in Appendices D.1–D.6. An axis is structurally inapplicable when its required experimental objects are undefined, as for signed reversal without a designated target-side input; applicable axes without measurements are reported as not evaluated. Thresholds are task-specific and adjustable (Appendix F), with sensitivity analyses in Appendix G.3.2. The joint nonlinear certificate is defined separately in Definition 1.

5 Experiments: sparse internal variables as controllable actuators

5.1 Experimental setup

The experiments investigate whether and to what extent sparse internal interventions can control task-defined behaviors of frozen language models. Driver-OMP and Coverage-Driver provide two concrete implementations of the framework, evaluated alongside baseline methods. The six-axis protocol characterizes the resulting control capabilities across tasks, models, and intervention conditions.

Tasks, models, and readouts. The experiments include five task instances spanning four behaviors: IOI and template-based Sentiment on GPT-2 small (Radford et al., 2019; Wang et al., 2023), MMLU Humanities and STEM MCQA (Hendrycks et al., 2021a) on Qwen3-4B, and Refusal on Qwen3-1.7B (Qwen Team, 2025). All model parameters remain frozen. IOI uses paired ABBA/BABA prompts and an indirect-object-minus-subject logit margin. Sentiment and Refusal use paired prompts with positive–negative and refusal–compliance word-set readouts, respectively. Refusal uses a pre-generation lexical readout at the implemented ANSWER index (Appendix F) on Arditi-inspired instruction pairs (Arditi et al., 2024).

MCQA nominal reconstruction uses four option logits and module-level candidates, whereas head-level control uses answer-position attention heads and the scalar correct-minus-mean-wrong margin $M_{\mathrm{MCQA}}(x) = \ell_{a^\star}(x) - \frac{1}{3}\sum_{a \neq a^\star} \ell_a(x)$. MCQA pairs exchange the correct option with a seeded incorrect option while retaining the original reference answer index $a^\star$. Module-level held-out reuse is evaluated separately. The repeated MCQA control, Sentiment fixed-dose, and Refusal control experiments use 24 questions, 16 template pairs, and 20 instruction pairs per seed, respectively, with equal fitting and evaluation splits. Task definitions and configuration details are given in Appendix E.

Interventions and comparisons. The MCQA and Refusal control experiments compare Driver-OMP, Coverage-Driver, and Top-k under a maximum support budget of eight candidates, with supports and base coefficients fitted on the training split. Evaluation-pair directions (P) are computed from the activation contrast of each evaluation pair, yielding pair-specific intervention directions. Training-mean directions (F) are obtained by averaging activation contrasts over the fitting split, yielding a fixed direction for each candidate that is shared across evaluation inputs. Sentiment also evaluates both direction conditions. In feedback experiments, support, base coefficients, and direction rules remain fixed while the global dose is adjusted using observed target errors; fixed-dose held-out reuse is evaluated separately. The MCQA and Refusal implementations use plain-ridge refitting with raw activation contrasts, without an independent intervention-energy constraint or off-target penalty. Other comparisons, including random supports, nonlinear greedy, and a single-site steering reference, are described in Appendix G.

Evaluation and reporting. Eight-seed feedback and signed reversal experiments cover Humanities, STEM, and Refusal. Additional reversal tests use ten IOI seeds and eight Sentiment seeds, while IOI feedback uses three seeds. Independent ten-dose scans cover Humanities, STEM, and Refusal, alongside separate IOI and Sentiment dose diagnostics. The Sentiment series also evaluates fixed-dose reuse; multi-target feedback tracking is not evaluated.

Off-target preservation is evaluated at nominal dose on a shared 32-prompt MMLU panel for MCQA, Sentiment, and Refusal. Refusal also includes a sixteen-token generation probe. Table 2 reports the normalized target residual $\mathrm{Rel}_{\mathrm{tar}}(1)$ at nominal dose, computed from complete source, target, and patched readouts. Table G.2 summarizes the experimental configurations and reporting units; implementation details are given in Appendices E–G. Code and experimental records are available at `https://github.com/zxzok/sparse-internal-control`.

Table 2: Same-panel nominal-dose target recovery. Mean ± sample SD of $\mathrm{Rel}_{\mathrm{tar}}(1)$ (lower is better), using evaluation-pair directions and at most eight candidates. Fitting and evaluation use the same 64 IOI pairs (10 seeds) or eight MCQA questions per category (3 seeds). Nonlinear greedy uses additional model evaluations.

Method	IOI (10 seeds)	Humanities (3 seeds)	STEM (3 seeds)
Driver-OMP	0.201 ± 0.058	0.279 ± 0.032	0.175 ± 0.035
Coverage-Driver	0.294 ± 0.029	0.152 ± 0.154	0.061 ± 0.033
Top-k	0.240 ± 0.035	0.333 ± 0.117	0.221 ± 0.046
Random	0.747 ± 0.240	1.209 ± 0.949	1.167 ± 1.141
Nonlinear greedy	0.151 ± 0.037	0.073 ± 0.093	0.020 ± 0.012

5.2 Results

Reachability: Do fixed-dose interventions attain the requested targets? Table 2 reports nominal-dose target recovery on the same fitting and evaluation panels for IOI and module-level MCQA. Held-out head-level MCQA, paired Sentiment, and Refusal are evaluated separately under P and F direction conditions (Appendix G.2.1). At $\varepsilon = 0.4$, IOI Driver-OMP passes in 10/10 same-panel seeds (0.201 ± 0.058). Module-level Coverage-Driver passes in 3/3 fitting-panel seeds per category: Humanities (0.152 ± 0.154) and STEM (0.061 ± 0.033). Held-out head-level MCQA and paired Sentiment share an 8/8 P versus 0/8 F pattern for Coverage-Driver and Top-k in each task. Refusal passes in 0/8 seeds for all three selectors in both modes. These results demonstrate fixed-dose target recovery in selected task and direction conditions, with a particularly strong dependence on evaluation-pair direction information in the held-out experiments.

Signed reversal: Can the opposite patch recover source behavior? Applying $-\beta$ to target-side examples tests whether the resulting behavior moves back toward the source. A seed passes when $\mathrm{Reversal}(\beta) \geq 2$, meaning that the reverse-patched readout is at least twice as far from the original target readout as from the source readout (Appendix G.2.2). Each seed contributes one reversal ratio computed over its evaluation panel. On IOI, all three selectors satisfy the reversal ratio criterion in each of ten per-example (P) seeds. Sentiment P reversal succeeds in 4/8, 8/8, and 8/8 seeds for Driver-OMP, Coverage-Driver, and Top-k, respectively; all frozen-direction (F) counts are 0/8. Across selectors, the corresponding P counts are 21/24 on Humanities and 23/24 on STEM, compared with 0/24 in either F condition. Refusal has 0/24 passing method–seed cells in each mode. Reverse control is therefore strongly associated with task and access to paired-example direction information.

Dose response: How does target progress vary with strength? The IOI and Sentiment dose curves illustrate graded responses (Figure 2). Independent ten-dose scans examine three selectors under P and F directions on Humanities, STEM, and Refusal (Figure 3), with supports, base strengths, and direction rules fixed. Using the operating criterion of signed Spearman $\rho \geq 0.95$, Driver-OMP, Coverage-Driver, and Top-k pass in 3/8, 5/8, and 2/8 P seeds on Humanities, and 4/8, 6/8, and 5/8 on STEM, respectively. Under F, each STEM selector exhibits four strongly positive ($\rho \geq 0.95$) and four strongly negative ($\rho \leq -0.95$) seed curves. All 48 Refusal curves have $\rho = 1$, yet none passes target-side signed reversal. These findings demonstrate that ordered dose responses and signed reversal capture distinct control properties.

Feedback controllability: Can fixed supports track multiple targets? With support, base strengths, and direction rules fixed, feedback adjusts the dose to track specified batch-mean margins (Table 3; Figure 4). A target is reached when the final absolute tracking error is within the task-specific tolerance: $\tau_{\mathrm{ctrl}} = 0.3$ for IOI and MCQA, and 0.1 for Refusal. The reported counts represent seed–target endpoints, with eight targets per seed for IOI, six for MCQA, and five for Refusal. Under P directions, Coverage-Driver reaches 46/48 Humanities and 47/48 STEM endpoints, Top-k reaches 45/48 and 47/48, and Driver-OMP reaches 38/48 and 39/48, respectively. Under F directions, only 7–8 of 48 MCQA endpoints are reached. Refusal reaches 12/40–22/40 endpoints across selectors and direction conditions, including ten targets already within tolerance at zero dose. IOI reaches 24/24 endpoints on its same fitting/evaluation panel, including two initial hits (Appendix G.2.4). The substantially larger P–F difference in MCQA than the differences among selectors highlights the importance of direction information for feedback controllability.

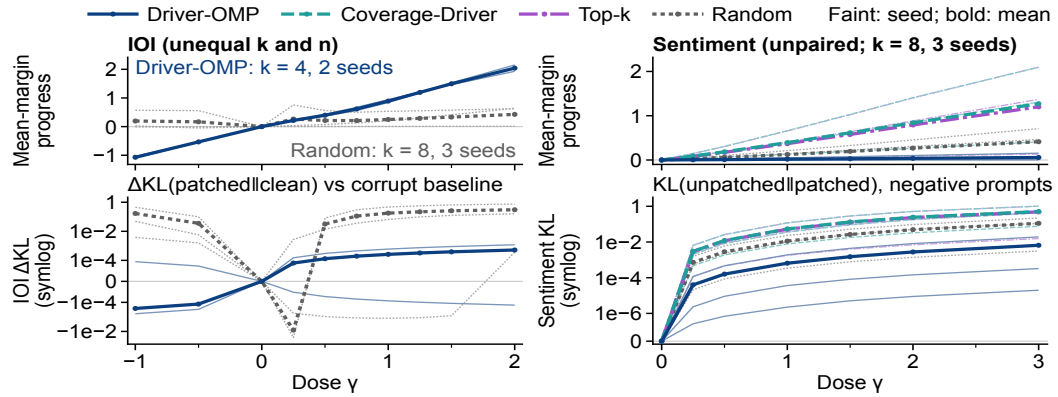


Figure 2: IOI and Sentiment dose diagnostics. Faint curves are seeds; bold curves are means. The illustrated IOI Driver-OMP uses four sites and two seeds, versus eight sites and three seeds for random supports. Sentiment uses three unpaired-pool seeds. IOI non-answer-position ΔKL and Sentiment evaluation-prompt KL are distinct from the independent preservation panel; see Appendix G.2.3.

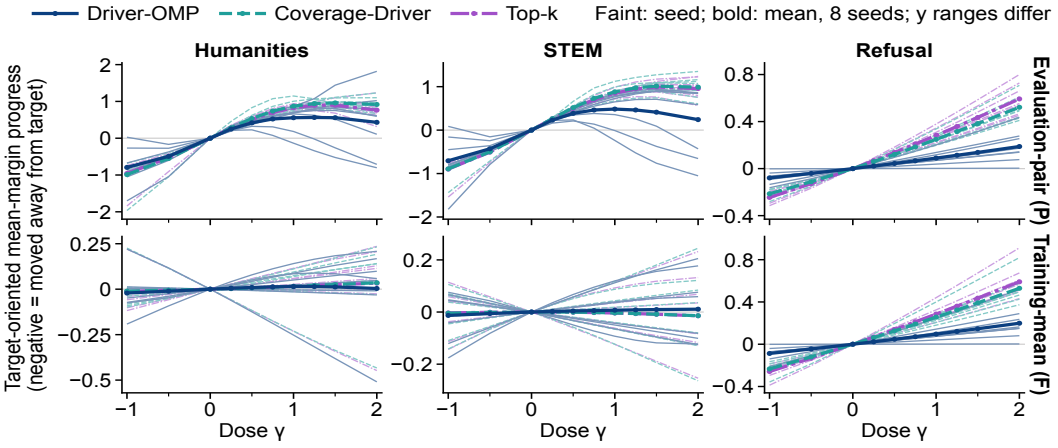


Figure 3: Independent dose response. Columns are Humanities, STEM, and Refusal; rows use P and F directions, respectively. Each selector has eight seed curves, ten measured doses per curve, and a bold mean. The vertical axis is signed mean-margin progress; panel scales differ. The additional STEM scan retains strong negative F responses. Supports and base coefficients remain fixed.

Off-target preservation: How much does non-target behavior change? Off-target preservation is evaluated using full-vocabulary KL on the shared 32-prompt panel (Appendix G.2.5). MCQA uses P directions, whereas Sentiment and Refusal use F. At the KL threshold of 0.001, two criteria are considered: the mean KL across prompts and the requirement that every prompt individually satisfies the threshold. Across selectors, mean KL ranges from 0.110 to 0.497 for Humanities, 0.119 to 0.582 for STEM, and 0.00553 to 0.01157 for Sentiment. None of the eight seeds satisfies either criterion under any selector for these tasks. For Refusal, Driver-OMP yields mean KL 0.001080, with 4/8 seeds satisfying the mean-KL criterion and 2/8 satisfying the all-prompt criterion; Coverage-Driver and Top-k have no passing seeds. These results show that off-target preservation varies across the evaluated task and direction conditions. The separate IOI paired-position diagnostic is reported in Appendix G.2.3.

Held-out reuse: Which intervention components transfer? Held-out reuse distinguishes the transfer of selected sites from that of complete interventions. In the four-seed module-level MCQA study, Driver-OMP achieves recovery of 0.992 ± 0.009 on Humanities and 0.987 ± 0.034 on STEM when the selected sites are retained and coefficients are refitted using held-out targets. With the entire intervention frozen, recovery decreases to 0.021 ± 0.026 and 0.024 ± 0.066, respectively (Appendix G.2.6). This contrast shows that reusable intervention sites do not necessarily yield transferable fixed interventions. Sentiment exhibits a different pattern. At $k = 8$, interventions using training-mean directions (F) achieve mean-margin recovery of 0.942 ± 0.535 for Coverage-Driver

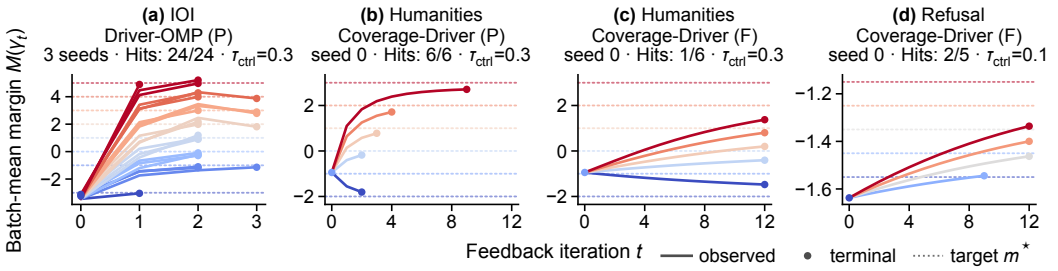


Figure 4: Feedback control with fixed supports. Curves show recorded batch-mean margins, with colours identifying target margins (dotted lines) and dots marking terminal observations. Panel (a) includes all three IOI seeds; panels (b)–(d) use seed 0. Only dose is adjusted. Hit counts include initial zero-dose hits; aggregate results are reported in Table 3.

Table 3: Feedback target hits. P/F are direction conditions. IOI has three seeds × eight targets; MCQA has eight × six, and Refusal eight × five. Initial zero-dose hits are 2/24 for IOI, 6/48 for Humanities, 3/48 for STEM, and 10/40 for Refusal. Each seed's targets share its fitted support.

Task	Mode	Driver-OMP	Coverage-Driver	Top-k
IOI	Evaluation-pair (P)	24/24	—	—
Humanities	Evaluation-pair (P)	38/48	46/48	45/48
Humanities	Training-mean (F)	8/48	8/48	8/48
STEM	Evaluation-pair (P)	39/48	47/48	47/48
STEM	Training-mean (F)	8/48	7/48	7/48
Refusal	Evaluation-pair (P)	12/40	21/40	21/40
Refusal	Training-mean (F)	12/40	21/40	22/40

and 0.937 ± 0.526 for Top-k (mean ± sample SD across eight seeds; full comparison in Table G.12). However, none of the three evaluated selectors passes signed reversal in the eight F seeds. Together, these results distinguish site reuse, forward recovery with fixed interventions, and signed reversal as separate control properties.

6 Conclusion

This paper introduces sparse internal control as a framework for treating selected internal sites in frozen language models as inference-time actuators. The local target-controllability formulation characterizes the behavioral effects of candidate site–direction pairs under intervention-energy constraints, while Driver-OMP and Coverage-Driver provide complementary strategies for selecting compact driver supports. A nonlinear control certificate specifies target recovery, off-target preservation, and intervention energy, and a six-axis control-evaluation protocol assesses reachability, signed reversal, dose response, feedback controllability, off-target preservation, and held-out reuse. Together, these contributions connect internal component selection with explicit and measurable behavioral-control objectives.

Sparse supports can reduce task-defined readout errors, while nominal reachability depends on direction information. Coverage-Driver and Top-k meet the formal target criterion in every held-out MCQA and Sentiment P seed and no F seed; MCQA feedback shows a similar separation. Refusal exhibits ordered dose response without nominal reachability or reversal, while Sentiment can show high batch-mean recovery despite large prompt-level residuals. Fully frozen MCQA reuse remains weak, and preservation measurements show substantial off-target distribution changes. These results distinguish target recovery, dose response, reversal, feedback, preservation, and reuse.

Future research can extend this framework by learning transferable direction rules that do not require evaluation-pair activation contrasts, incorporating explicit off-target preservation and intervention-energy constraints into practical control algorithms, and extending evaluation from task-specific readouts to multi-token generation and broader input distributions. These directions provide a foundation for developing more reliable, selective, and auditable internal control interfaces for language models.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083, 2024. doi: 10.52202/079017-4322.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073, 2022.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112, 2023.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In *Advances in Neural Information Processing Systems*, volume 37, pages 18506–18534, 2024. doi: 10.52202/079017-0587.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352, 2023. doi: 10.52202/075280-0719.

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

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL `https://transformer-circuits.pub/2021/framework/index.html`.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.3.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *Proceedings of the First Conference on Language Modeling*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. arXiv preprint arXiv:2109.13916, 2021b.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. In *Proceedings of the First Conference on Language Modeling*, 2024.

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.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530, 2023. doi: 10.52202/075280-1797.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. doi: 10.1145/3560815.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Proceedings of the First Conference on Language Modeling*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022. doi: 10.52202/068431-1262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978. doi: 10.1007/BF01588971.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001, 2020. doi: 10.23915/distill.00024.001.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022. doi: 10.52202/068431-2011.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR, 2024.

Qwen Team. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741, 2023. doi: 10.52202/075280-2338.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.647.

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

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.48.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, volume 37, pages 139179–139212, 2024. doi: 10.52202/079017-4417.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Joel A. Tropp and Anna C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007. doi: 10.1109/tit.2007.909108.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022. doi: 10.52202/068431-1800.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020.

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

Appendix

This appendix provides the definitions, methods, quantitative results, and execution details supporting the main text. Appendix A expands the sparse internal control problem, notation, patched model, and target regimes. Appendix B gives the assumptions and proofs for local linearization, target controllability, energy-constrained refitting, and the coverage proposal. Appendix C gives complete pseudocode and complexity analysis for SharedRefit, Driver-OMP, and Coverage-Driver. Appendix D defines the six evaluation axes and the information conditions for held-out reuse. Appendix E specifies each task's model, prompts, readout, targets, directions, candidate library, and preservation set. Appendix F records formal reference constants and executed task-specific settings, including feedback target grids. Appendix G presents the configuration index, results across the six control axes, comparison methods and sensitivity, supporting analyses, and limitations. Appendix H describes the experimental records, execution environments and statistical units.

A Problem definition: sparse internal control

This section expands the sparse internal control formulation in Section 2. It specifies the behavioral targets, candidate interventions, selection costs, and activation-space energy used to define the nonlinear control certificate and the sparse optimization problem.

Frozen model and internal computation graph. Let f_θ be a frozen autoregressive Transformer with parameters θ. For an input sequence $x = (x^{(1)}, \dots, x^{(T)})$, the model computes intermediate activations and produces next-token logits or probabilities.

For a given input, the forward computation is represented at the specified activation sites by a directed acyclic graph

$$G_\theta = (\mathcal{N}, \mathcal{E}),$$

where $\mathcal{N}$ indexes physical intervention sites and $\mathcal{E}$ records their computational dependencies. Each site $s \in \mathcal{N}$ specifies a layer, component, token position or position-selection rule, and tensor slice. Examples include attention-head outputs, MLP outputs, and residual-stream activations. Position-selection rules are resolved separately for each input.

The activation at site s is represented as

$$h_s(x) \in \mathbb{R}^{d_s},$$

where d_s is the dimension of the selected activation, with tensor slices flattened when necessary. An edge $(s, t) \in \mathcal{E}$ indicates a directed computational dependency from s to t, possibly through intermediate operations omitted from this representation.

Behavioral readout. The behavior to be controlled is represented by a task-specific readout map

$$b_\theta(x) = b(f_\theta, x) \in \mathbb{R}^r.$$

The dimension r is determined by the task. When $r = 1$, $b_\theta(x)$ may be a scalar logit margin, a refusal–compliance margin, or a sentiment score. When $r > 1$, $b_\theta(x)$ may be a vector-valued panel, such as the four option logits in a multiple-choice question. The readout is required to be well-defined for both the unpatched and patched forward passes; for local linearization arguments below, it is additionally assumed differentiable with respect to the patched activations.

The target-control set is

$$\mathcal{D}_{\text{tar}} = \{(x_i, \tau_i)\}_{i=1}^n.$$

Here $n > 0$ is the number of target examples, x_i is an input prompt, and $\tau_i \in \mathbb{R}^r$ is the desired value of the r-dimensional readout. The required behavioral displacement on example i is

$$q_i = \tau_i - b_\theta(x_i),$$

and the stacked target displacement is

$$q = [q_1; \ldots; q_n] \in \mathbb{R}^{nr}.$$

In a clean–corrupt pair, x_i^r denotes the source (corrupt) input and x_i^c its target (clean) counterpart. The target-control input is $x_i = x_i^r$, with $\tau_i = b_\theta(x_i^c)$. For steering tasks, τ_i may instead be a task-specified scalar margin or a target vector. This distinction matters for transfer: in per-prompt control, q_i typically encodes example-specific information; in population-level control, many prompts share a common direction of desired movement.

Off-target preservation. Successful target recovery does not necessarily imply that other model behaviors remain unchanged. Internal interventions may also affect outputs on inputs outside the target-control set. To measure these effects, define an off-target preservation set

$$\mathcal{D}_{\text{off}} = \{\tilde{x}_j\}_{j=1}^{n_{\text{off}}},$$

containing $n_{\text{off}} > 0$ prompts that are disjoint from the target inputs:

$$\mathcal{D}_{\text{off}} \cap \{x_i : (x_i, \tau_i) \in \mathcal{D}_{\text{tar}}\} = \varnothing.$$

For each input x, the unpatched model produces an observable

$$o_\theta(x) = o(f_\theta, x),$$

which is compared with the corresponding observable $o(f_\theta^\beta, x)$ after intervention. The discrepancy d_{off} measures this change and satisfies $d_{\text{off}}(z, z) = 0$ and $d_{\text{off}}(z, z') \geq 0$.

For next-token probability distributions, d_{off} may be the KL divergence from the unpatched to the patched distribution. Other observables may use squared Euclidean distance or task-specific discrepancies. The preservation set, observable, and discrepancy measure are specified as part of the control problem, independently of the driver-selection algorithm.

Candidate control node library. Selecting an intervention site alone does not specify how its activation should be modified. Each candidate control node therefore specifies a physical site, an intervention direction, and a selection cost:

$$v = (s_v, u_v, c_v),$$

where $s_v \in \mathcal{N}$ is the physical site, $u_v(x) \in \mathbb{R}^{d_{s_v}}$ is the intervention direction for input x, and $c_v \geq 1$ is the selection cost. The direction determines the activation perturbation up to a scalar strength β_v.

These nodes form a finite candidate library

$$\mathcal{V} = \{v_1, \ldots, v_M\},$$

which defines the search space for driver selection. Multiple nodes may share a physical site while using different directions.

Directions are constructed from unpatched activations. For a prescribed source–target pair (x_i^r, x_i^c), an input-dependent direction is

$$u_v(x_i^r) = h_{s_v}(x_i^c) - h_{s_v}(x_i^r).$$

Alternatively, a direction shared across prompts can be obtained from fitting-group means:

$$u_v = \bar{h}_{s_v}^{(+)} - \bar{h}_{s_v}^{(-)},$$

where $\bar{h}_{s_v}^{(+)}$ and $\bar{h}_{s_v}^{(-)}$ are the mean activations over the prescribed positive and negative fitting groups, respectively.

Each direction rule is defined on all prompts in the evaluation set $\mathcal{X}_{\text{eval}}$, with any required paired inputs specified as part of the rule. The evaluation set and direction rules remain fixed during optimization: $u_v(x)$ may vary across inputs but does not depend on the optimized coefficients β.

Head–position libraries contain nodes at specified attention heads and token positions. Module-level libraries use attention, MLP, or residual-stream activations, while feature-based libraries use sparse-autoencoder features or other learned coordinates. Given $\mathcal{V}$, Driver-OMP and Coverage-Driver select a compact support $S \subseteq \mathcal{V}$ and jointly refit the intervention strengths on that support.

Inference-time patching and weighted sparsity. An intervention is represented by a coefficient vector

$$\beta = (\beta_{v_1}, \ldots, \beta_{v_M}) \in \mathbb{R}^M,$$

where $\beta_v = m_v \alpha_v$, with $m_v \in \{0, 1\}$ indicating whether candidate control node v is selected and $\alpha_v \in \mathbb{R}$ specifying its strength. During a forward pass on input x, the activation at each physical site is modified additively:

$$h_s(x) \leftarrow h_s(x) + \sum_{v: s_v = s} \beta_v u_v(x), \qquad s \in \mathcal{N}. \tag{A.1}$$

The resulting patched model is denoted f_θ^β. Interventions modify activations during inference without updating the model parameters θ, and $\beta = 0$ recovers the original model. Signed reversal in Section 4 applies $-\beta$ to the target-side input.

The active driver support is defined as

$$S_\beta = \mathrm{supp}(\beta) = \{v \in \mathcal{V} : \beta_v \neq 0\}.$$

Elements of S_β are termed driver nodes. For a selected support $S \subseteq \mathcal{V}$, β_S denotes its coefficients, with all coefficients outside S set to zero. Thus $S_\beta \subseteq S$, with equality when all selected coefficients are nonzero.

The total selection cost is measured by weighted sparsity:

$$\|\beta\|_{0,c} = \sum_{v \in S_\beta} c_v. \tag{A.2}$$

For unit costs $c_v = 1$, this reduces to $|S_\beta|$. More generally, costs can account for differences among intervention types, such as attention-head outputs, residual activations, and individual features. When multiple candidate nodes share a physical site, weighted sparsity counts their selection costs separately.

Intervention energy. Weighted sparsity measures the cost of selecting driver nodes but does not bound the magnitude of their activation perturbations. Define the evaluation set

$$\mathcal{X}_{\mathrm{eval}} = \{x_i : (x_i, \tau_i) \in \mathcal{D}_{\mathrm{tar}}\} \cup \mathcal{D}_{\mathrm{off}}.$$

With direction rules fixed on $\mathcal{X}_{\mathrm{eval}}$, the intervention energy is the mean squared magnitude of the additive activation perturbations:

$$E(\beta) = \frac{1}{|\mathcal{X}_{\mathrm{eval}}|} \sum_{x \in \mathcal{X}_{\mathrm{eval}}} \sum_{s \in \mathcal{N}} \left\| \sum_{v: s_v = s} \beta_v u_v(x) \right\|_2^2 = \beta^\top W \beta, \tag{A.3}$$

$$W_{vw} = \begin{cases} \dfrac{1}{|\mathcal{X}_{\mathrm{eval}}|} \displaystyle\sum_{x \in \mathcal{X}_{\mathrm{eval}}} \langle u_v(x), u_w(x) \rangle, & s_v = s_w, \\ 0, & s_v \neq s_w. \end{cases}$$

The matrix $W \in \mathbb{R}^{M \times M}$ is symmetric positive semidefinite. For a support $S \subseteq \mathcal{V}$, the energy reduces to $E(\beta) = \beta_S^\top W_{SS} \beta_S$ when coefficients outside S are zero. Cross terms arise only between candidate directions sharing the same physical site.

If each active node occupies a distinct site and its direction has unit mean squared norm over $\mathcal{X}_{\mathrm{eval}}$, then $W_{S_\beta S_\beta} = I$ and $E(\beta) = \|\beta\|_2^2$. Otherwise, W accounts for the combined activation perturbations at shared sites.

Target displacement induced by a patch. For an intervention β, the resulting target-behavior displacement is

$$F(\beta) = \left[b(f_\theta^\beta, x_i) - b(f_\theta, x_i) \right]_{i=1}^n \in \mathbb{R}^{nr}. \tag{A.4}$$

The target error and off-target loss are

$$L_{\mathrm{tar}}(\beta) = \frac{1}{n} \sum_{i=1}^n \| b(f_\theta^\beta, x_i) - \tau_i \|_2^2, \tag{A.5}$$

$$L_{\mathrm{off}}(\beta) = \frac{1}{n_{\mathrm{off}}} \sum_{j=1}^{n_{\mathrm{off}}} d_{\mathrm{off}}\Big(o(f_\theta, \tilde{x}_j), o(f_\theta^\beta, \tilde{x}_j) \Big). \tag{A.6}$$

By the definition of the target displacement q,

$$L_{\text{tar}}(\beta) = \frac{1}{n}\|q - F(\beta)\|_2^2.$$

These quantities are evaluated using the original nonlinear Transformer and define the target-recovery and preservation components of the control certificate.

Definition 2 (Nonlinear control certificate; restatement of Definition 1)**.** *Fix a problem instance $(f_\theta, \mathcal{D}_{\text{tar}}, \mathcal{D}_{\text{off}}, b, o, d_{\text{off}}, \mathcal{V})$, nonnegative tolerances $\varepsilon, \delta, \kappa$, and a stabilizer $\xi > 0$. An intervention $\beta \in \mathbb{R}^M$ is an $(\varepsilon, \delta, \kappa)$-control certificate on the specified evaluation sets if all three conditions hold:*

$$\frac{\|q - F(\beta)\|_2}{\|q\|_2 + \xi} \le \varepsilon, \qquad L_{\text{off}}(\beta) \le \delta, \qquad E(\beta) \le \kappa. \tag{A.7}$$

The tolerances bound normalized target error, off-target change, and intervention energy, respectively.

Using $L_{\text{tar}}(\beta) = \|q - F(\beta)\|_2^2/n$, the target-recovery condition is equivalently

$$L_{\text{tar}}(\beta) \le \frac{\varepsilon^2(\|q\|_2 + \xi)^2}{n}.$$

Definition 3 (Sparse internal control problem)**.** *For the fixed problem instance above, candidate costs $\{c_v\}$, and an energy weight $\lambda_E \ge 0$, sparse internal control is formulated as*

$$\min_{\beta \in \mathbb{R}^M} \|\beta\|_{0,c} + \lambda_E E(\beta) \quad \textit{s.t.} \quad \beta \textit{ satisfies Eq. (A.7).} \tag{A.8}$$

The objective balances driver-selection cost and intervention energy among interventions satisfying the prescribed control tolerances.

Budgeted formulation. For a prescribed selection-cost budget $k \ge 0$, a related formulation minimizes target and off-target losses:

$$\begin{aligned} \min_{\beta \in \mathbb{R}^M} \quad & L_{\text{tar}}(\beta) + \mu L_{\text{off}}(\beta) + \lambda_E E(\beta) \\ \text{s.t.} \quad & \|\beta\|_{0,c} \le k, \qquad E(\beta) \le \kappa, \end{aligned} \tag{A.9}$$

where $\mu \ge 0$ weights off-target change. Under unit costs, the selection-cost constraint becomes $|S_\beta| \le k$.

Equation (A.8) enforces target-recovery and preservation tolerances as hard constraints. Equation (A.9) instead trades off these losses under fixed selection-cost and energy budgets. The formulations generally have different feasible sets and minimizers.

Support selection is combinatorial, and the behavioral losses are generally nonconvex in β. The local proxy in Eq. (4) guides support selection and strength refitting. The resulting intervention is evaluated against Eq. (A.7) in the original Transformer. Task-specific implementations are specified in Appendix F.

Target specification and direction rules. Targets τ_i specify desired values in behavioral-readout space, whereas $u_v(x)$ specifies a perturbation direction in activation space. For paired reconstruction, $x_i = x_i^r$ and $\tau_i = b(f_\theta, x_i^c)$. A shared per-input target can instead be specified by $\tau_i = \tau$ for all i; the required displacements $q_i = \tau - b(f_\theta, x_i)$ may still differ across inputs. The feedback experiments in Section 4 prescribe targets for the batch-mean readout and evaluate the corresponding tracking error.

Direction construction is a separate choice. Evaluation-pair directions (P) use activation contrasts from prescribed input pairs, while training-mean directions (F) use contrasts averaged over the fitting split. Either construction can be combined with example-specific or shared per-input targets, provided the required inputs are specified.

Within each evaluation setting, direction rules are fixed on all target and preservation prompts. The same rules determine patch application, the energy matrix W, and certificate evaluation. Task-specific direction constructions and evaluation settings are given in Appendices E and F.

B Theoretical analysis

This section analyzes the local approximation used for sparse internal control. It characterizes the target-effect dictionary, the first-order reconstruction error of a fixed support, and the energy required for exact local interpolation. It then derives the energy-constrained ridge refit and proves monotone submodularity of the saturated coverage objective, with the standard greedy approximation bound under a pure cardinality constraint.

B.1 Local effect dictionaries

For a fixed candidate library $\mathcal{V}$, define the nonlinear target displacement map

$$\Phi(\beta) = F(\beta) = [b(f_\theta^\beta, x_i) - b(f_\theta, x_i)]_{i=1}^n \in \mathbb{R}^{nr}.$$

The notation $\Phi = F$ is used for the target-displacement map in the proofs below. The candidate library, evaluation sets, and direction rules are fixed while the coefficients vary. For $S \subseteq \mathcal{V}$, $\Phi_S(\beta_S)$ denotes $F(\beta)$ with coefficients outside S set to zero, and G_S is the submatrix formed by the columns indexed by S. All Jacobians are evaluated at the unpatched model, $\beta = 0$. In particular,

$$J_{i,s_v} \in \mathbb{R}^{r\times d_{s_v}}, \qquad g_{i,v} \in \mathbb{R}^r, \qquad D\Phi_S(0) = G_S.$$

Since $\Phi(0) = 0$, the first-order approximation at the unpatched model is

$$\Phi(\beta) \approx G\beta.$$

The column associated with candidate v is obtained by differentiating the behavioral readout with respect to the activation at the patched site:

$$g_{i,v} = J_{i,s_v} u_v(x_i), \qquad J_{i,s_v} = \frac{\partial b(f_\theta, x_i)}{\partial h_{s_v}(x_i)}. \tag{B.10}$$

Stacking $g_{i,v}$ over target prompts gives $g_v \in \mathbb{R}^{nr}$, and stacking columns over candidates gives

$$G = [g_{v_1}, \dots, g_{v_M}] \in \mathbb{R}^{nr\times M}. \tag{B.11}$$

The local target equation is therefore

$$q \approx G\beta. \tag{B.12}$$

A corresponding off-target dictionary is built from a specified differentiable preservation surrogate $\psi_\theta(x) = \psi(f_\theta, x) \in \mathbb{R}^p$, where p is its readout dimension. This surrogate may be the full next-token logit vector, a restricted or projected logit readout, or another differentiable observable. For each preservation prompt $\tilde{x}_j$ and candidate v, define

$$h_{j,v} = K_{j,s_v} u_v(\tilde{x}_j), \qquad K_{j,s_v} = \frac{\partial \psi_\theta(\tilde{x}_j)}{\partial h_{s_v}(\tilde{x}_j)}. \tag{B.13}$$

Here $K_{j,s_v} \in \mathbb{R}^{p\times d_{s_v}}$ and $h_{j,v} \in \mathbb{R}^p$. Stacking over preservation prompts gives $h_v = [h_{1,v}; \dots; h_{n_{\text{off}},v}]$ and $H = [h_{v_1}, \dots, h_{v_M}] \in \mathbb{R}^{n_{\text{off}}p\times M}$. The metric is $\Omega = \text{diag}(\Omega_1, \dots, \Omega_{n_{\text{off}}})$, where each $\Omega_j \in \mathbb{R}^{p\times p}$ is positive semidefinite; $\Omega^{1/2}$ denotes its symmetric positive-semidefinite square root.

For a squared-Euclidean discrepancy in the coordinates of ψ, $\Omega_j = I_p$. If ψ is the full next-token logit vector and p_j^0 is the unpatched probability vector, its KL quadratic term is

$$\frac{1}{2}\,\Delta\psi_j^\top \left[\text{diag}(p_j^0) - p_j^0 (p_j^0)^\top\right] \Delta\psi_j,$$

where $\Delta\psi_j$ is the logit displacement at $\tilde{x}_j$. The constant factor can be absorbed into μ. For a restricted or projected readout, Ω_j specifies a surrogate metric in that readout's coordinates.

The sample-normalized matrices are

$$\bar{q} = \frac{q}{\sqrt{n}}, \qquad \bar{G} = \frac{G}{\sqrt{n}}, \qquad \bar{H} = \frac{H}{\sqrt{n_{\text{off}}}}.$$

The local sparse proxy is

$$\min_{\beta\in\mathbb{R}^M} \|\bar{q} - \bar{G}\beta\|_2^2 + \mu\|\Omega^{1/2}\bar{H}\beta\|_2^2 + \lambda\beta^\top W\beta \quad \text{s.t.} \quad \|\beta\|_{0,c} \le k,\ \beta^\top W\beta \le \kappa. \tag{B.14}$$

Here $\mu \ge 0$ weights predicted off-target change, $\lambda \ge 0$ weights intervention energy in the local refit, and k and κ are the selection-cost and energy budgets. The local parameter λ is distinct from λ_E in the nonlinear objective. Equation (B.14) combines first-order target reconstruction, a quadratic preservation surrogate, and energy regularization. The selected intervention is subsequently evaluated in the nonlinear model.

B.2 Local linearization error

The next theorem states the basic locality guarantee. It is stated for a fixed support $S \subseteq \mathcal{V}$, because the Hessian bound and the conditioning of the local dictionary depend on the selected sites.

Theorem 1 (Local linearization error)**.** *Let $S \subseteq \mathcal{V}$, let $W_{SS} \succ 0$, and let $\kappa \geq 0$. Define the energy ball*

$$\mathcal{B}_{S,\kappa} = \{\beta_S \in \mathbb{R}^{|S|} : \beta_S^\top W_{SS}\beta_S \leq \kappa\}.$$

Assume Φ_S is twice continuously differentiable on an open neighborhood of $\mathcal{B}_{S,\kappa}$ and that its second derivative is bounded with a finite constant $L_S \geq 0$: for all $\tilde{\beta} \in \mathcal{B}_{S,\kappa}$ and all $a, b \in \mathbb{R}^{|S|}$,

$$\left\|D^2\Phi_S(\tilde{\beta})[a,b]\right\|_2 \leq L_S\|a\|_{W_{SS}}\|b\|_{W_{SS}}, \qquad \|a\|_{W_{SS}}^2 = a^\top W_{SS}a. \tag{B.15}$$

Here $D\Phi_S$ is the Jacobian, $D^2\Phi_S(\tilde{\beta})[a,b]$ is the bilinear second derivative, and $\|a\|_{W_{SS}} = (a^\top W_{SS}a)^{1/2}$. Then every $\beta_S \in \mathcal{B}_{S,\kappa}$ satisfies

$$\|\Phi_S(\beta_S) - G_S\beta_S\|_2 \leq \frac{L_S}{2}\beta_S^\top W_{SS}\beta_S. \tag{B.16}$$

Equivalently, for the normalized displacement $\bar{\Phi}_S = \Phi_S/\sqrt{n}$ and $\bar{G}_S = G_S/\sqrt{n}$,

$$\|\bar{\Phi}_S(\beta_S) - \bar{G}_S\beta_S\|_2 \leq \frac{L_S}{2\sqrt{n}}\beta_S^\top W_{SS}\beta_S. \tag{B.17}$$

Proof. Since $\Phi_S(0) = 0$ and $D\Phi_S(0) = G_S$, Taylor's theorem with integral remainder gives

$$\Phi_S(\beta_S) - G_S\beta_S = \int_0^1 (1-t)D^2\Phi_S(t\beta_S)[\beta_S, \beta_S]dt.$$

Taking norms and applying Eq. (B.15) yields

$$\|\Phi_S(\beta_S) - G_S\beta_S\|_2 \leq \int_0^1 (1-t)L_S\|\beta_S\|_{W_{SS}}^2 dt = \frac{L_S}{2}\beta_S^\top W_{SS}\beta_S.$$

Dividing by $\sqrt{n}$ gives Eq. (B.17). □

Corollary 1 (Linear proposal to nonlinear target residual)**.** *Under the assumptions of Theorem 1, if $\beta_S \in \mathcal{B}_{S,\kappa}$, then*

$$\frac{\|q - \Phi_S(\beta_S)\|_2}{\|q\|_2 + \xi} \leq \frac{\|q - G_S\beta_S\|_2}{\|q\|_2 + \xi} + \frac{L_S\kappa/2}{\|q\|_2 + \xi}. \tag{B.18}$$

In particular, under the corollary's assumptions, the target-recovery inequality in Definition 1 holds if

$$\|q - G_S\beta_S\|_2 + L_S\kappa/2 \leq \varepsilon(\|q\|_2 + \xi).$$

The bound combines the local reconstruction error with a curvature contribution on the specified energy ball. Off-target preservation is evaluated through L_{off}.

B.3 Local target controllability of a support

For a fixed support S, the first-order question is whether the desired displacement $\bar{q}$ lies near the span of $\bar{G}_S$ under the energy budget. Define the local controllability residual

$$\chi_\kappa(S) = \min_{\beta_S : \beta_S^\top W_{SS}\beta_S \leq \kappa} \frac{\|\bar{q} - \bar{G}_S\beta_S\|_2}{\|\bar{q}\|_2 + \xi}. \tag{B.19}$$

A small $\chi_\kappa(S)$ indicates that the selected support can approximate the target displacement within the energy budget at first order. For any β_S in the energy ball, Corollary 1 bounds its nonlinear target residual through the local reconstruction error and curvature term. Off-target preservation is assessed separately, and Definition 1 combines target recovery, preservation, and energy into the joint nonlinear certificate.

For $W_{SS} \succ 0$, define the whitened dictionary

$$\widetilde{G}_S = \bar{G}_S W_{SS}^{-1/2}.$$

Here $W_{SS}^{-1/2}$ denotes the symmetric inverse square root, and $\widetilde{G}_S^+$ is the Moore–Penrose pseudoinverse. For $\bar{q} = 0$, the minimum-energy interpolation is $\beta_S = 0$. For nonzero $\bar{q} \in \text{range}(\bar{G}_S)$, the whitened dictionary has at least one positive singular value.

Theorem 2 (Minimum-energy local interpolation)**.** *Assume $W_{SS} \succ 0$ and $0 \neq \bar{q} \in \mathrm{range}(\bar{G}_S)$. Let $\sigma^+_{\min}(\widetilde{G}_S)$ be the smallest nonzero singular value of $\widetilde{G}_S$. The minimum-energy solution of $\bar{G}_S\beta_S = \bar{q}$ is*

$$\beta_S^\dagger = W_{SS}^{-1/2}\widetilde{G}_S^+\bar{q}, \tag{B.20}$$

and it satisfies

$$(\beta_S^\dagger)^\top W_{SS}\beta_S^\dagger \leq \frac{\|\bar{q}\|_2^2}{\left(\sigma^+_{\min}(\widetilde{G}_S)\right)^2}. \tag{B.21}$$

Proof. Use whitened coordinates $z = W_{SS}^{1/2}\beta_S$. The interpolation constraint becomes $\widetilde{G}_S z = \bar{q}$, and the energy becomes $\|z\|_2^2$. The minimum-norm solution is $z^\dagger = \widetilde{G}_S^+\bar{q}$, whose norm is bounded by $\|\bar{q}\|_2/\sigma^+_{\min}(\widetilde{G}_S)$. Substituting $\beta_S^\dagger = W_{SS}^{-1/2}z^\dagger$ proves the claim. □

For an attainable target $\bar{q} \in \mathrm{range}(\bar{G}_S)$, including $\bar{q} = 0$, the exact minimum interpolation energy is

$$E_{\min}(S) = \|\widetilde{G}_S^+\bar{q}\|_2^2.$$

For $W_{SS} \succ 0$, exact first-order reconstruction within budget is equivalent to

$$\bar{q} \in \mathrm{range}(\bar{G}_S), \qquad E_{\min}(S) \leq \kappa.$$

Equation (B.21) supplies an upper bound on this energy for a nonzero attainable target; the target's alignment with the singular directions determines its exact value.

The normalization in Eq. (B.19) differs from that in Corollary 1. For a minimizer β_S^χ in Eq. (B.19),

$$\|q - G_S\beta_S^\chi\|_2 = (\|q\|_2 + \sqrt{n}\,\xi)\chi_\kappa(S).$$

Under Theorem 1, Corollary 1 therefore gives

$$\frac{\|q - \Phi_S(\beta_S^\chi)\|_2}{\|q\|_2 + \xi} \leq \frac{\|q\|_2 + \sqrt{n}\,\xi}{\|q\|_2 + \xi}\chi_\kappa(S) + \frac{L_S\kappa}{2(\|q\|_2 + \xi)}.$$

The shared refit in Eq. (B.23) also includes preservation and ridge terms and does not generally minimize $\chi_\kappa(S)$. Its coefficients are assessed using their own local residual in Corollary 1.

B.4 Energy-constrained ridge refit

After a support S is selected, strengths are recomputed by a shared refit. Define the augmented target and dictionary

$$\hat{q} = \begin{bmatrix}\bar{q}\\ 0\end{bmatrix}, \qquad \hat{G}_S = \begin{bmatrix}\bar{G}_S\\ \sqrt{\mu}\,\Omega^{1/2}\bar{H}_S\end{bmatrix}. \tag{B.22}$$

The refit solves

$$\min_{\beta_S \in \mathbb{R}^{|S|}} \|\hat{q} - \hat{G}_S\beta_S\|_2^2 + \lambda\beta_S^\top W_{SS}\beta_S \quad \text{s.t.} \quad \beta_S^\top W_{SS}\beta_S \leq \kappa. \tag{B.23}$$

When $W_{SS} \succ 0$ and $\kappa = 0$, the unique feasible coefficient vector is $\beta_S = 0$. The following multiplier characterization concerns $\kappa > 0$.

Proposition 1 (KKT form of the shared refit)**.** *Assume $W_{SS} \succ 0$, $\lambda > 0$, and $\kappa > 0$. The unique minimizer of Eq.* (B.23) *is*

$$\beta_S^\star(\nu) = \left(\hat{G}_S^\top\hat{G}_S + (\lambda + \nu)W_{SS}\right)^{-1}\hat{G}_S^\top\hat{q}, \tag{B.24}$$

where $\nu \geq 0$ satisfies primal feasibility, $(\beta_S^\star)^\top W_{SS}\beta_S^\star \leq \kappa$, and complementary slackness:

$$\nu\left((\beta_S^\star)^\top W_{SS}\beta_S^\star - \kappa\right) = 0. \tag{B.25}$$

If $E(\beta_S^\star(0)) \leq \kappa$, then $\nu = 0$. Otherwise, a unique positive multiplier satisfies $E(\beta_S^\star(\nu)) = \kappa$.

Proof. The objective is strictly convex because $\lambda W_{SS} > 0$, and $\beta_S = 0$ is strictly feasible because $\kappa > 0$. Thus the KKT conditions are necessary and sufficient. With

$$\mathcal{L}(\beta_S, \nu) = \|\hat{q} - \hat{G}_S\beta_S\|_2^2 + \lambda\beta_S^\top W_{SS}\beta_S + \nu(\beta_S^\top W_{SS}\beta_S - \kappa),$$

stationarity gives

$$A_\nu\beta_S = \hat{G}_S^\top\hat{q}, \qquad A_\nu = \hat{G}_S^\top\hat{G}_S + (\lambda + \nu)W_{SS} \succ 0.$$

This proves Eq. (B.24). Along the multiplier path,

$$\frac{d}{d\nu}E(\beta_S^\star(\nu)) = -2(W_{SS}\beta_S^\star)^\top A_\nu^{-1}(W_{SS}\beta_S^\star) \le 0.$$

The energy is continuous, tends to zero as $\nu \to \infty$, and is strictly decreasing when $\beta_S^\star \neq 0$. If the unconstrained energy exceeds $\kappa > 0$, a unique positive boundary multiplier therefore exists and can be found by bisection. □

B.5 Coverage objective and submodularity

Residual-fitting objectives can over-allocate support to rows that are already easy to explain. For population-level steering, a useful driver should move many prompts in the correct direction, even if no single prompt has a large residual. Coverage-Driver therefore uses a saturated row-coverage score.

Here $j \in \{1, \dots, nr\}$ indexes a coordinate of the stacked target readout. For $r = 1$, one row corresponds to one prompt; for $r > 1$, it corresponds to a prompt–readout-coordinate pair. Let $\mathcal{V}^\pm$ denote the candidate library augmented with both u_v and $-u_v$. In Eqs. (B.26)–(B.27), the dictionary columns and candidate sets refer to this oriented library. The convention $\operatorname{sign}(0) = 0$ is used. For each row j and candidate v, define a nonnegative signed contribution

$$a_{jv} = \operatorname{clip}\left(\frac{\operatorname{sign}(\bar{q}_j)\bar{G}_{jv}}{|\bar{q}_j| + \delta_0}, 0, 1\right), \tag{B.26}$$

where $\delta_0 > 0$ prevents division by very small target entries. The orientation expansion represents a direction useful with a negative strength as a positive-oriented candidate during coverage scoring. Given row weights $w_j \ge 0$, define

$$C(S) = \sum_j w_j \min\left(1, \sum_{v\in S} a_{jv}\right). \tag{B.27}$$

The cap at 1 implements diminishing returns: once a row is covered, additional candidates targeting the same row provide no further coverage reward.

Proposition 2 (Monotone submodularity of saturated coverage)**.** *If $a_{jv} \ge 0$ and $w_j \ge 0$, then $C(S)$ in Eq.* (B.27) *is monotone submodular and $C(\varnothing) = 0$. For an integer $1 \le k \le |\mathcal{V}^\pm|$, let $S^\star \in \arg\max_{S\subseteq\mathcal{V}^\pm, |S|\le k} C(S)$. Under this pure cardinality budget with no additional candidate-exclusion constraints, the standard k-step greedy algorithm that repeatedly adds the largest marginal increase obtains*

$$C(S_{\text{greedy}}) \ge (1 - 1/e)C(S^\star).$$

Proof. For each row j, the set function $m_j(S) = \sum_{v\in S} a_{jv}$ is modular and nondecreasing. The scalar function $\phi(t) = \min(1, t)$ is concave and nondecreasing on $\mathbb{R}_{\ge 0}$. The composition $\phi(m_j(S))$ is therefore monotone submodular. A nonnegative weighted sum of monotone submodular functions is monotone submodular. The cardinality-constrained greedy guarantee is the classical result of Nemhauser et al. (1978). □

For heterogeneous costs, define $\Delta_v(S) = C(S \cup \{v\}) - C(S)$. The score $\Delta_v(S)/(c_v + \eta_0)$, with $\eta_0 > 0$, provides a cost-normalized greedy rule. Proposition 2 concerns the pure cardinality-constrained coverage objective. Executed selection and stopping settings are specified in Appendix F.

Scope of the guarantee. The $(1 - 1/e)$ bound concerns the nonnegative saturated coverage objective in Proposition 2 under a pure cardinality constraint and standard k-step greedy selection. It does not automatically apply to additional site or orientation exclusions, early-stopped variants, or heterogeneous-cost greedy rules. Nor does it guarantee nonlinear target movement after strength refitting. The formal $\pm u_v$ expansion supplies candidate orientations for support proposal; the final SharedRefit may assign either sign to a coefficient. The executed MCQA and Refusal Coverage variants do not expand both orientations (Appendix F).

B.6 Relation to orthogonal matching pursuit

Driver-OMP is a structured variant of orthogonal matching pursuit. With unit-normalized effect-dictionary columns, no off-target block, $W = I$, $c_v = 1$, $\lambda = 0$, unconstrained refitting, and no additional candidate exclusions, the residual-fitting selection rule reduces to ordinary OMP. Classical OMP recovery theorems apply when the dictionary satisfies strong incoherence or restricted-isometry assumptions (Tropp and Gilbert, 2007). Transformer effect dictionaries are typically redundant and highly correlated, so the algorithm is used as a sparse proposal method rather than as an exact-support recovery procedure. A refitted intervention receives the nonlinear control certificate when all three conditions in Definition 1 are satisfied.

B.7 End-to-end interpretation

The analysis characterizes first-order target error, the minimum energy for local interpolation, and the energy-constrained fixed-support refit. Saturated coverage supplies a complementary objective for support proposal. Driver-OMP and Coverage-Driver use residual fitting and row coverage, respectively, to construct candidate supports. Evaluation in the patched nonlinear model measures target recovery, off-target preservation, and intervention energy against Definition 1.

C Driver-selection algorithms

This section specifies the formal support-selection, energy-constrained refitting, and nonlinear-validation procedures. Driver-OMP uses residual fitting, whereas Coverage-Driver uses saturated row coverage. Each selector returns a support and refitted coefficients; ValidateCertificate then reports the three certificate metrics and an acceptance decision. The executed experimental configurations are specified in Appendix F.

C.1 Shared inputs and outputs

The shared inputs include the frozen model f_θ, target and preservation sets $\mathcal{D}_{\text{tar}}, \mathcal{D}_{\text{off}}$, behavioral readout b, preservation observable o and divergence d_{off}, differentiable surrogate ψ, preservation metric Ω, candidate library $\mathcal{V}$, fixed direction rules, costs $c_v \geq 1$, and prescribed candidate-exclusion and tie-breaking rules. Here $k_c \geq 0$ denotes the selection-cost budget written as k in the weighted-sparsity constraints of Sections 2–3; under unit costs it bounds the number of selected candidates. The formal refit uses $\mu \geq 0$, $\lambda > 0$, and $\kappa \geq 0$. The library and admissibility rules are assumed to give $W_{SS} > 0$ for every nonempty support passed to the positive-definite refit of Appendix B.4. Site-disjoint candidates with nonzero mean squared direction norms satisfy this condition. The ordinary-OMP specialization in Appendix B.6 uses unconstrained least-squares refitting.

A selected set S and its coefficient vector β_S define a full-library vector by setting coefficients outside S to zero. Its active support satisfies $S_\beta \subseteq S$. Support selection enforces the cost budget and specified exclusions; SharedRefit enforces the energy constraint. Algorithms C.3–C.4 return proposals. Algorithm C.5 accepts a proposal when all three conditions of Definition 1 are satisfied.

The numerical parameters are a target-normalization stabilizer $\xi > 0$, a cost stabilizer $\eta_0 > 0$, a coverage stabilizer $\delta_0 > 0$, a relative linear-residual tolerance $\varepsilon_{\text{lin}} \in [0, 1)$, an improvement threshold $\eta_{\text{tol}} \geq 0$, a coverage threshold $\tau \in (0, 1]$, and a normalized gain threshold $g_{\min} \geq 0$. SharedRefit uses a multiplier-interval tolerance $t_\nu > 0$ and a relative energy-deficit tolerance $t_E \in (0, 1)$; these are numerical tolerances, not the certificate thresholds $(\varepsilon, \delta, \kappa)$.

C.2 Building local dictionaries

The target dictionary G stores the first-order effect of every candidate on every target row, and H stores its first-order effect on the preservation surrogate. The Gram matrix W represents the energy of simultaneous patches. The evaluation set is $\mathcal{X}_{\text{eval}} = \{x_i : (x_i, \tau_i) \in \mathcal{D}_{\text{tar}}\} \cup \mathcal{D}_{\text{off}}$, with $n, n_{\text{off}} > 0$. Each direction rule is defined at every input in this set and remains fixed while the intervention coefficients vary. All derivatives are evaluated at the unpatched model, $\beta = 0$. The returned objects are $(\bar{q}, \bar{G}, \bar{H}, W, \Omega)$. For the energy matrix, inner products are evaluated only between directions at the same physical site; entries for different sites are set to zero, as in Appendix A.

Algorithm C.1 BuildLocalTables: target, preservation, and energy dictionaries

Require: Frozen model f_θ; sets $\mathcal{D}_{\text{tar}}, \mathcal{D}_{\text{off}}$ with $n, n_{\text{off}} > 0$; library $\mathcal{V}$ with fixed site/direction rules; readouts b, ψ; metric Ω.
Ensure: Normalized target $\bar{q}$, dictionaries $\bar{G}, \bar{H}$, energy matrix W, and metric Ω.

$\mathcal{X}_{\text{eval}} \leftarrow \{x_i : (x_i, \tau_i) \in \mathcal{D}_{\text{tar}}\} \cup \mathcal{D}_{\text{off}}$.
Evaluate $q_i \leftarrow \tau_i - b(f_\theta, x_i)$, stack q, and set $\bar{q} \leftarrow q/\sqrt{n}$.
for each $v \in \mathcal{V}$ **do**
  for each target input x_i **do**
    $g_{i,v} \leftarrow J_{i,s_v} u_v(x_i)$, with J_{i,s_v} evaluated at $\beta = 0$.
  end for
  for each preservation input $\tilde{x}_j$ **do**
    $h_{j,v} \leftarrow K_{j,s_v} u_v(\tilde{x}_j)$, with K_{j,s_v} evaluated at $\beta = 0$.
  end for
end for
Stack columns to form G, H; set $\bar{G} \leftarrow G/\sqrt{n}$, $\bar{H} \leftarrow H/\sqrt{n_{\text{off}}}$.
for each pair $v, w \in \mathcal{V}$ **do**
  if $s_v = s_w$ **then**
    $W_{vw} \leftarrow |\mathcal{X}_{\text{eval}}|^{-1} \sum_{x \in \mathcal{X}_{\text{eval}}} \langle u_v(x), u_w(x) \rangle$.
  else
    $W_{vw} \leftarrow 0$.
  end if
end for
return $(\bar{q}, \bar{G}, \bar{H}, W, \Omega)$.

In practice, the products $J_{i,s_v} u_v(x_i)$ and $K_{j,s_v} u_v(\tilde{x}_j)$ can be computed by automatic differentiation as Jacobian–vector products or vector–Jacobian products, depending on whether the readout dimension is small or large. For scalar margins, a single backward pass per example can score many candidate directions at a site. For vector-valued panels, rows are accumulated across readout coordinates.

C.3 Shared energy-constrained refit

On a fixed support, SharedRefit jointly solves the quadratic refit in Appendix B.4. Let

$$\hat{q} = \begin{bmatrix} \bar{q} \\ 0 \end{bmatrix}, \qquad \hat{G}_S = \begin{bmatrix} \bar{G}_S \\ \sqrt{\mu}\, \Omega^{1/2} \bar{H}_S \end{bmatrix}.$$

The refit solves

$$\min_{\beta_S} \|\hat{q} - \hat{G}_S \beta_S\|_2^2 + \lambda \beta_S^\top W_{SS} \beta_S \quad \text{s.t.} \quad \beta_S^\top W_{SS} \beta_S \le \kappa.$$

It returns an empty vector for an empty support and zero coefficients when $\kappa = 0$. For a nonempty support with $\kappa > 0$, it first tests the unconstrained ridge solution. If the energy budget is active, it brackets the multiplier and bisects while retaining a feasible upper endpoint. The Gram matrix and right-hand side are computed once for each support and reused in the linear solves. Bisection stops when either the multiplier interval is at most t_ν or the feasible endpoint's energy deficit is at most $t_E \kappa$. Formal reference values are $t_\nu = 10^{-8}$ and $t_E = 10^{-3}$ (Appendix F).

C.4 Driver-OMP: sparse residual fitting

Driver-OMP ranks candidates using their alignment with the remaining augmented residual. Define the augmented full-library column for candidate v as

$$\hat{g}_v = \begin{bmatrix} \bar{G}_v \\ \sqrt{\mu}\, \Omega^{1/2} \bar{H}_v \end{bmatrix}.$$

Algorithm C.2 SharedRefit: energy-constrained ridge refit

Require: S, $\hat{q}$, $\hat{G}_S$, $W_{SS} \succ 0$ for nonempty S; $\lambda > 0$; $\kappa \geq 0$; $t_\nu > 0$, $0 < t_E < 1$.
Ensure: Coefficients on S with energy at most κ.

if $S = \varnothing$ **then**
  return the empty coefficient vector.
end if
if $\kappa = 0$ **then**
  return $0 \in \mathbb{R}^{|S|}$.
end if
$B_S \leftarrow \hat{G}_S^\top \hat{G}_S$; $g_S \leftarrow \hat{G}_S^\top \hat{q}$.
function Solve(ν)
  Solve $[B_S + (\lambda + \nu) W_{SS}]\beta = g_S$.
  return $(\beta, \beta^\top W_{SS} \beta)$.
end function
$(\beta, e) \leftarrow$ Solve(0).
if $e \leq \kappa$ **then**
  return β.
end if
$\nu_{\mathrm{lo}} \leftarrow 0$; $\nu_{\mathrm{hi}} \leftarrow 1$; $(\beta_{\mathrm{hi}}, e_{\mathrm{hi}}) \leftarrow$ Solve(ν_{hi}).
while $e_{\mathrm{hi}} > \kappa$ **do**
  $\nu_{\mathrm{lo}} \leftarrow \nu_{\mathrm{hi}}$; $\nu_{\mathrm{hi}} \leftarrow 2\nu_{\mathrm{hi}}$.
  $(\beta_{\mathrm{hi}}, e_{\mathrm{hi}}) \leftarrow$ Solve(ν_{hi}).
end while
while $\nu_{\mathrm{hi}} - \nu_{\mathrm{lo}} > t_\nu$ **and** $\kappa - e_{\mathrm{hi}} > t_E \kappa$ **do**
  $\nu \leftarrow (\nu_{\mathrm{lo}} + \nu_{\mathrm{hi}})/2$; $(\beta, e) \leftarrow$ Solve(ν).
  if $e > \kappa$ **then**
    $\nu_{\mathrm{lo}} \leftarrow \nu$.
  else
    $(\nu_{\mathrm{hi}}, \beta_{\mathrm{hi}}, e_{\mathrm{hi}}) \leftarrow (\nu, \beta, e)$.
  end if
end while
return β_{hi}.

At iteration t, with residual $r_t = \hat{q} - \hat{G}_{S_t} \beta_{S_t}$, the candidate-ranking score is

$$\mathrm{score}(v \mid S_t) = \frac{|\langle \hat{g}_v, r_t \rangle|^2}{(\|\hat{g}_v\|_2^2 + \lambda W_{vv})(c_v + \eta_0)}. \tag{C.28}$$

Equation (C.28) is a cost-normalized, regularized correlation score for candidate ranking. The augmented columns include the linear preservation surrogate, and the selected coefficients are jointly recomputed by SharedRefit. The stopping statistic is the relative linear target residual. A trial support is retained when it reaches $\varepsilon_{\mathrm{lin}}$; otherwise, it must reduce that residual by more than η_{tol}.

C.5 Coverage-Driver: saturated row coverage

Coverage-Driver selects candidates using a saturated target-row coverage objective. Let $\mathcal{V}^{\pm}$ be the oriented candidate library and write $v^\sigma = (s_v, \sigma u_v, c_v)$ for $\sigma \in \{-1, +1\}$. All matrices passed to the oriented selector use the same indexing:

$$\bar{G}_{v^\sigma} = \sigma \bar{G}_v, \qquad \bar{H}_{v^\sigma} = \sigma \bar{H}_v, \qquad W_{v^\sigma, w^{\sigma'}} = \sigma \sigma' W_{vw}.$$

The costs and physical-site identifiers are inherited from the original candidates. After fitting oriented coefficients, the equivalent original coefficient is $\beta_v = \beta_{v^+} - \beta_{v^-}$, with an absent orientation contributing zero. Selection applies all specified site and orientation exclusions.

The index $j \in \{1, \ldots, nr\}$ refers to a target-readout coordinate. For $r = 1$ each row corresponds to a prompt. The row weights satisfy $w_j \geq 0$, and $W_{\mathrm{row}} = \sum_j w_j$ is a scalar weight sum, distinct from the energy matrix W. If $W_{\mathrm{row}} = 0$, the coverage objective is identically zero and the selector returns the empty support. The convention $\mathrm{sign}(0) = 0$ is used. For each row and oriented candidate, compute

$$a_{jv} = \mathrm{clip}\left(\frac{\mathrm{sign}(\bar{q}_j) \bar{G}_{jv}}{|\bar{q}_j| + \delta_0}, 0, 1 \right).$$

Algorithm C.3 DRIVER-OMP: residual-fitting driver selection

Require: $\bar{q}, \bar{G}, \bar{H}, W, \Omega$, library $\mathcal{V}$, costs c_v, exclusions, budget $k_c \geq 0$; $\lambda > 0$, $\mu \geq 0$, $\kappa \geq 0$; $\xi, \eta_0 > 0$, $0 \leq \varepsilon_{\text{lin}} < 1$, $\eta_{\text{tol}} \geq 0$; t_v, t_E.
Ensure: Proposal (S, β_S); nonlinear validation is a separate step.

Form $\hat{q} \leftarrow [\bar{q}; 0]$ and $\hat{G} \leftarrow [\bar{G};\ \sqrt{\mu}\,\Omega^{1/2}\bar{H}]$.
$S \leftarrow \varnothing$; $\beta_S \leftarrow [\,]$; $r \leftarrow \hat{q}$.
$\rho_{\text{prev}} \leftarrow \|\bar{q}\|_2 / (\|\bar{q}\|_2 + \xi)$.
if $\rho_{\text{prev}} \leq \varepsilon_{\text{lin}}$ **or** $\kappa = 0$ **then**
    return (S, β_S).
end if
while $\sum_{v \in S} c_v < k_c$ **do**
    $\mathcal{A} \leftarrow \{v \in \mathcal{V} \setminus S : \sum_{u \in S} c_u + c_v \leq k_c$ and exclusions hold$\}$.
    if $\mathcal{A} = \varnothing$ **then**
        break.
    end if
    Compute score$(v \mid S)$ from Eq. (C.28) for every $v \in \mathcal{A}$.
    Select a maximizer $v^\star$ using the prescribed tie rule.
    $S' \leftarrow S \cup \{v^\star\}$.
    $\beta' \leftarrow$ SHAREDREFIT$(S', \hat{q}, \hat{G}_{S'}, W_{S'S'}, \lambda, \kappa, t_v, t_E)$.
    $\rho' \leftarrow \|\bar{q} - \bar{G}_{S'}\beta'\|_2 / (\|\bar{q}\|_2 + \xi)$.
    if $\rho' \leq \varepsilon_{\text{lin}}$ **then**
        $(S, \beta_S) \leftarrow (S', \beta')$; **break**.
    end if
    if $\rho_{\text{prev}} - \rho' \leq \eta_{\text{tol}}$ **then**
        break. ▷ Return the previous support; reject the trial.
    end if
    $(S, \beta_S) \leftarrow (S', \beta')$; $r \leftarrow \hat{q} - \hat{G}_S\beta_S$; $\rho_{\text{prev}} \leftarrow \rho'$.
end while
return (S, β_S).

Let $\chi_j(S) = \min(1, \sum_{v \in S} a_{jv})$ be the current coverage of row j. The marginal gain of adding v is

$$\Delta(v \mid S) = \sum_j w_j \left[\min(1, \chi_j(S) + a_{jv}) - \chi_j(S)\right]. \tag{C.29}$$

The selected candidate maximizes $\Delta(v \mid S)/(c_v + \eta_0)$ among feasible candidates.

Saturation assigns marginal credit to target rows that remain uncovered. COVERAGE-DRIVER therefore allocates its support through row coverage, whereas DRIVER-OMP ranks candidates using the current residual. The coverage score defines the support-proposal objective; joint intervention strengths are obtained in the subsequent refit.

C.6 Nonlinear validation

The output of either selector is a proposal. Validation applies the additive patch in the original Transformer and recomputes the nonlinear quantities in the certificate. For the candidate library and evaluation sets supplied to validation, $R_{\text{tar}} = \text{Rel}_{\text{tar}}(1)$ denotes the target residual in Definition 1. The same direction rules define the patched model and its energy matrix.

Validation returns ACCEPT or REJECT together with R_{tar}, $L_{\text{off}}(\beta)$, and $E(\beta)$ for the specified intervention and evaluation sets. The selection-cost and exclusion constraints are enforced by the support-selection stage.

C.7 Selection recurrence and consensus drivers

Sparse dictionaries in Transformers can be redundant: multiple supports may produce similar target movement. Selection frequency summarizes a specified collection of configuration runs. For the reported IOI analysis, five seeds and five support budgets $k \in \{3, 5, 8, 12, 16\}$ give $B = 25$ configurations. The sets $S^{(b)}$ are the final supports returned by the IOI recurrence implementation, which includes target-readout evaluation, fallback selection, and pruning. In that implementation, the validation flag uses mean absolute error to the clean target, and the stored scores also include R_{clean}.

Algorithm C.4 COVERAGE-DRIVER: saturated row-coverage driver selection

Require: $\bar{q}$; oriented library $\mathcal{V}^{\pm}$ and consistently oriented $\bar{G}, \bar{H}, W$; Ω; $w_j \geq 0$; c_v, exclusions, $k_c \geq 0$; $\tau \in (0, 1]$, $g_{\min} \geq 0$, $\delta_0, \eta_0 > 0$; $\lambda > 0$, $\mu \geq 0$, $\kappa \geq 0$, t_v, t_E.
Ensure: Oriented proposal (S, β_S); map coefficients back before original-library validation.

$S \leftarrow \varnothing$; $\chi_j \leftarrow 0$ for all j; $W_{\text{row}} \leftarrow \sum_j w_j$.
if $W_{\text{row}} = 0$ **or** $\kappa = 0$ **then**
  return $(\varnothing, [\,])$.
end if
$a_{jv} \leftarrow \text{clip}(\text{sign}(\bar{q}_j)\bar{G}_{jv}/(|\bar{q}_j| + \delta_0), 0, 1)$, with $\text{sign}(0) = 0$.
while $\sum_{v \in S} c_v < k_c$ **do**
  $\text{cov} \leftarrow W_{\text{row}}^{-1} \sum_j w_j \chi_j$.
  if $\text{cov} \geq \tau$ **then**
    break.
  end if
  $\mathcal{A} \leftarrow \{v \in \mathcal{V}^{\pm} \setminus S : \sum_{u \in S} c_u + c_v \leq k_c$ and exclusions hold$\}$.
  if $\mathcal{A} = \varnothing$ **then**
    break.
  end if
  Compute $\Delta(v \mid S)$ from Eq. (C.29) for every $v \in \mathcal{A}$.
  Select $v^\star \in \arg\max_{v \in \mathcal{A}} \Delta(v \mid S)/(c_v + \eta_0)$ using the prescribed tie rule.
  if $\Delta(v^\star \mid S) \leq g_{\min} W_{\text{row}}$ **then**
    break.
  end if
  $S \leftarrow S \cup \{v^\star\}$; $\chi_j \leftarrow \min(1, \chi_j + a_{jv^\star})$ for every j.
end while
Form $\hat{q} \leftarrow [\bar{q}; 0]$, $\hat{G} \leftarrow [\bar{G};\ \sqrt{\mu}\,\Omega^{1/2}\bar{H}]$.
$\beta_S \leftarrow$ SHAREDREFIT$(S, \hat{q}, \hat{G}_S, W_{SS}, \lambda, \kappa, t_v, t_E)$.
return (S, β_S).

Algorithm C.5 VALIDATECERTIFICATE: nonlinear acceptance test

Require: f_θ; original candidate library with site/direction rules; (S, β_S); $\mathcal{D}_{\text{tar}} = \{(x_i, \tau_i)\}$ with $\tau_i \in \mathbb{R}^r$; $\mathcal{D}_{\text{off}}$; readouts b, o; divergence d_{off}; energy matrix W for these rules and sets; $\xi > 0$, tolerances $(\varepsilon, \delta, \kappa)$.
Ensure: Acceptance status and $(R_{\text{tar}}, L_{\text{off}}(\beta), E(\beta))$.

If necessary, map oriented coefficients to the original library as specified in Section C.5.
Form β by zero extension of the original-library coefficients.
Compute $q_i \leftarrow \tau_i - b(f_\theta, x_i)$ and stack q.
Evaluate the patched model and set $F(\beta) \leftarrow [b(f_\theta^\beta, x_i) - b(f_\theta, x_i)]_{i=1}^n$.
$R_{\text{tar}} \leftarrow \|q - F(\beta)\|_2/(\|q\|_2 + \xi) = \text{Rel}_{\text{tar}}(1)$.
$L_{\text{off}}(\beta) \leftarrow n_{\text{off}}^{-1} \sum_j d_{\text{off}}(o(f_\theta, \tilde{x}_j), o(f_\theta^\beta, \tilde{x}_j))$.
$E(\beta) \leftarrow \beta^\top W \beta$.
if $R_{\text{tar}} \leq \varepsilon$ **and** $L_{\text{off}}(\beta) \leq \delta$ **and** $E(\beta) \leq \kappa$ **then**
  return Accept, $(R_{\text{tar}}, L_{\text{off}}(\beta), E(\beta))$.
else
  return Reject, $(R_{\text{tar}}, L_{\text{off}}(\beta), E(\beta))$.
end if

These run-specific checks are distinct from ValidateCertificate in Algorithm C.5. The configurations are not bootstrap resamples. For candidate v, the frequency is

$$\pi(v) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}[v \in S^{(b)}].$$

For a threshold $\pi^\star$, the consensus core is

$$C_{\pi^\star} = \{v : \pi(v) \geq \pi^\star\}.$$

The frequency $\pi(v)$ measures recurrence across this configuration grid, and $\pi^\star$ defines the consensus set studied in Appendix G.4.1. The consensus core is not used to replace the nonlinear certificate. It is a descriptive recurrence diagnostic: candidates that appear across the specified configurations can be studied further, but the frequency alone is not a selection probability, an uncertainty interval, or a guarantee on unseen prompts.

C.8 Computational complexity

Let $d = nr + n_{\text{off}}p$ be the number of augmented rows, let $M = |\mathcal{V}|$, and let K bound the number of candidate additions permitted by the budget. For a nonempty candidate library, set $c_{\min} = \min_v c_v$; then

$$K \leq \min(M_{\text{sel}}, \lfloor k_c / c_{\min} \rfloor),$$

where M_{sel} is the number of columns scored by the selector, including any orientation expansion. For a fixed support of size s, precomputing $\hat{G}_S^\top \hat{G}_S$ and $\hat{G}_S^\top \hat{q}$ costs $O(ds^2)$. If the refit uses T_s dense linear solves, its cost is $O(ds^2 + T_s s^3)$; T_s includes the initial solve, multiplier bracketing, and bisection. For a bracket of width B_v, bisection to an absolute multiplier tolerance t_v takes at most

$$\max\{0, \lceil \log_2(B_v/t_v) \rceil\}$$

steps; the energy-deficit condition can terminate it earlier.

For $K \geq 1$, with $T = \max_{1 \leq s \leq K} T_s$, Driver-OMP has the dictionary-level bound

$$O(KMd + dK^3 + TK^4),$$

whereas Coverage-Driver has the bound

$$O(KM_{\pm}nr + dK^2 + TK^3),$$

with $M_{\pm} = |\mathcal{V}^{\pm}|$. These expressions assume constant-time candidate-exclusion checks and an incrementally maintained support cost, and count work after the augmented dictionaries are constructed. Dictionary differentiation, energy-matrix construction, nonlinear validation, and any experimental fallback or pruning calls are accounted for separately. Driver-OMP refits each trial support, including a final rejected trial if present; Coverage-Driver performs one final refit.

C.9 Algorithmic summary

BuildLocalTables constructs the target, preservation, and energy quantities. Driver-OMP or Coverage-Driver returns a candidate support with energy-constrained coefficients, and Validate-Certificate evaluates that intervention in the nonlinear model. The complete procedure returns the intervention, its three certificate metrics, and the acceptance status. Selection recurrence summarizes the supports of a separately specified collection of runs. Executed experimental variants and their recorded checks are described in Appendices F–G.

D Control evaluation protocol: complete definitions

This section gives the complete definitions of the six control axes introduced in Section 4: reachability, signed reversal, dose response, feedback controllability, off-target preservation, and held-out reuse. Experimental operating criteria are specified in Appendix F, and configurations are indexed in Table G.2.

An evaluated intervention specifies a support S, base coefficients β_S, and a direction rule $u_v(x)$. A dose $\gamma \in \mathbb{R}$ defines the patched model $f_\theta^{\gamma\beta}$, with $\gamma = 1$ denoting the nominal dose. Target displacement is $F(\gamma\beta) = [b(f_\theta^{\gamma\beta}, x_i) - b(f_\theta, x_i)]_{i=1}^n$. The six axes report properties of this intervention in the nonlinear model. Definition 1 separately specifies the joint target-recovery, off-target-preservation, and intervention-energy conditions.

D.1 Reachability

Reachability is measured by the relative residual between the desired displacement q and the displacement actually produced by the patch:

$$\mathrm{Rel}_{\mathrm{tar}}(\gamma) = \frac{\|q - F(\gamma\beta)\|_2}{\|q\|_2 + \xi}. \tag{D.30}$$

The stabilizer $\xi > 0$ regularizes normalization near zero requested displacement. At nominal dose, $\mathrm{Rel}_{\mathrm{tar}}(1)$ measures the normalized remaining target displacement and is compared with the specified tolerance ε. For scalar readouts, an auxiliary target-sign agreement rate compares the signs of the patched and target readouts. Here $b_{\mathrm{scalar}} = b$ and $\tau_{i,\mathrm{scalar}} = \tau_i$ when $r = 1$:

$$\mathrm{SignAgree}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[\mathrm{sign}(b_{\mathrm{scalar}}(f_\theta^\beta, x_i)) = \mathrm{sign}(\tau_{i,\mathrm{scalar}})]. \tag{D.31}$$

Agreement is measured at nominal dose, with $\mathrm{sign}(0) = 0$. For IOI, Sentiment, and Refusal, b_{scalar} is the specified task readout. The repeated MCQA controls use the correct-minus-mean-wrong margin M_{MCQA} defined in Section 5.1, with the reference answer index fixed across each pair. Nominal MCQA reconstruction uses the four option logits as a vector; it is a separate configuration and does not replace that scalar definition. Nominal reconstruction additionally reports R_{clean}, which normalizes by the clean-output norm (Appendix G.2.1).

D.2 Signed reversal

Signed reversal applies $-\beta$ to the target-side member of a prescribed pair while retaining the selected support and the original direction convention. Write x_i^r for the source member and x_i^c for the target member. In the paired-contrast condition, the original target-minus-source contrast is retained when the negative patch is applied. The subscript dis denotes the resulting target-side readout after that patch:

$$\begin{aligned}
\mathrm{Rel}_{\mathrm{dis}\to\mathrm{src}} &= \frac{\left\|[b(f_\theta^{-\beta}, x_i^c) - b(f_\theta, x_i^r)]_{i=1}^n\right\|_2}{D}, \\
\mathrm{Rel}_{\mathrm{dis}\to\mathrm{clean}} &= \frac{\left\|[b(f_\theta^{-\beta}, x_i^c) - b(f_\theta, x_i^c)]_{i=1}^n\right\|_2}{D}, \\
D &= \left\|[b(f_\theta, x_i^c) - b(f_\theta, x_i^r)]_{i=1}^n\right\|_2 + \xi, \\
\mathrm{Reversal}(\beta) &= \frac{\mathrm{Rel}_{\mathrm{dis}\to\mathrm{clean}}}{\mathrm{Rel}_{\mathrm{dis}\to\mathrm{src}}}.
\end{aligned} \tag{D.32}$$

A ratio greater than one places the reverse-patched readout closer to the source readout than to the unpatched target readout. If $\mathrm{Rel}_{\mathrm{dis}\to\mathrm{src}} = 0$ and $\mathrm{Rel}_{\mathrm{dis}\to\mathrm{clean}} > 0$, the ratio is $+\infty$; if both distances vanish, it is undefined. These cases are recorded explicitly. The common normalizer D cancels from the ratio. Operating criteria and per-seed results are given in Appendices F and G.2.2.

D.3 Dose response

Dose response measures changes in target progress across a specified dose grid Γ, with support, base coefficients, and the direction rule fixed. Negative doses reverse the patch sign on the source input; $\gamma = 0$ is the unpatched model and $\gamma = 1$ the nominal dose. This source-side scan is distinct from the target-side reversal test. The grid is specified for each experimental series. For vector readouts, projected target progress is

$$P(\gamma) = \frac{\langle q, F(\gamma\beta)\rangle}{\|q\|_2^2 + \xi}. \tag{D.33}$$

The signed Spearman coefficient ρ_{sp}, denoted ρ in the result figures and tables, summarizes the ordering of dose and target progress without requiring a linear response. Scalar mean-margin scans use P_{mean} from Section 4, distinct from the vector projection in Eq. (D.33). The sign of the coefficient records whether progress increases or decreases with dose. Fixed-grid scans and controller-trajectory diagnostics are identified separately. Task-specific grids and operating criteria are reported in Appendices F and G.2.3.

D.4 Feedback controllability

Feedback controllability is evaluated by tracking specified scalar targets with dose updates based on observed error. The selected support, base coefficients, and direction rule remain fixed. For an evaluation batch, the mean readout is

$$M(\gamma) = \frac{1}{n} \sum_i b_{\text{scalar}}(f_\theta^{\gamma\beta}, x_i) \tag{D.34}$$

and the tracking error is $e_t = m^\star - M(\gamma_t)$, following Section 4. A reference proportional controller is

$$\begin{aligned} e_t &= m^\star - M(\gamma_t), \\ \gamma_{t+1} &= \text{clip}(\gamma_t + \eta_t\, e_t, \gamma_{\min}, \gamma_{\max}), \\ \eta_{t+1} &= \begin{cases} \eta_t/2, & \text{sign}(e_{t+1}) \neq \text{sign}(e_t), \\ \eta_t, & \text{otherwise.} \end{cases} \end{aligned} \tag{D.35}$$

Here t indexes controller updates, $\eta_t > 0$ is the gain, and $[\gamma_{\min}, \gamma_{\max}]$ is the dose interval of this reference controller. The error e_{t+1} is evaluated at the updated dose before the next gain is selected.

Equation (D.35) specifies a reference controller. The executed MCQA and Refusal experiments use the fixed-gain update and stopping rules in Appendix F, without dose clipping or gain halving. A terminal hit satisfies $|m^\star - M(\gamma_{\text{end}})| \leq \tau_{\text{ctrl}}$. Endpoint errors, hit counts, and recorded readout-evaluation counts are reported by seed and target; evaluation counts are distinct from controller update counts. The recorded count includes the initial zero-dose readout and final endpoint readout; zero-dose calls may reuse the cached unpatched baseline. Initial zero-dose hits are identified separately. Targets within one seed share the fitted intervention and evaluation batch.

D.5 Off-target preservation

Off-target preservation measures changes in the predictive distribution on the specified preservation set $\mathcal{D}_{\text{off}}$. The nominal-dose loss is

$$L_{\text{off}}(\beta) = \frac{1}{n_{\text{off}}} \sum_{j=1}^{n_{\text{off}}} \text{KL}\Big(p_\theta(\cdot \mid \tilde{x}_j) \,\|\, p_\theta^\beta(\cdot \mid \tilde{x}_j)\Big). \tag{D.36}$$

Evaluation at dose γ uses $L_{\text{off}}(\gamma\beta)$. The unpatched distribution is the first KL argument, and the patched distribution is evaluated at the same readout position. KL divergence is computed using natural logarithms. For a specified tolerance δ, the seed-mean rule compares the average prompt KL with δ, whereas the all-prompt rule requires every prompt KL to meet δ. The panel size, operating threshold, and per-seed outcomes are given in Appendices F and G.2.5.

The IOI paired-position diagnostic instead measures $\text{KL}(p_{\text{patched}} \| p_{\text{clean}})$ at non-answer positions of paired prompts and reports the change relative to the unpatched-corrupt baseline. The unpaired Sentiment dose plot uses $\text{KL}(p_{\text{unpatched}} \| p_{\text{patched}})$ on its negative evaluation prompts. Appendix G.2.3 records these panel-specific definitions.

The auxiliary ratio

$$\text{Sel}(\gamma) = \frac{\|F(\gamma\beta)\|_2 / (\|q\|_2 + \xi)}{\sqrt{L_{\text{off}}(\gamma\beta)} + \xi}. \tag{D.37}$$

divides normalized displacement magnitude by $\sqrt{L_{\text{off}}(\gamma\beta)} + \xi$. Its numerator is a magnitude; $\text{Rel}_{\text{tar}}(\gamma)$ separately measures proximity to the requested target. The six-axis preservation measure remains $L_{\text{off}}(\gamma\beta)$.

D.6 Held-out reuse

Held-out reuse evaluates intervention components on prompts excluded from support selection. Four information conditions are distinguished.

Refit-β reuse. Selected sites remain fixed, while coefficients are refitted using the held-out target readouts and prescribed evaluation-pair directions.

Fixed-β reuse with evaluation-pair directions. Sites and training-fitted coefficients remain fixed, while directions use activation contrasts from the evaluation pairs.

Feedback with training-mean directions (F). Sites, base coefficients, and training-mean directions remain fixed, while dose is adjusted using evaluation-time feedback. This condition is evaluated under feedback controllability.

Fully frozen open-loop reuse. Sites, coefficients, training-mean directions, and dose are all fixed before evaluation. This is the complete-intervention condition called frozen-β reuse in Section 4. P/F identifies the direction source; it does not by itself specify whether the full intervention is frozen.

Each series retains its recorded split and statistic. Module-level MCQA uses four seeds per category, 32 questions split 16/16, and signed mean-margin recovery $(\bar{b}_{\text{patched}} - \bar{b}_{\text{source}})/(\bar{b}_{\text{target}} - \bar{b}_{\text{source}})$, with means over the evaluation batch and absolute denominators below 10^{-9} omitted. Sentiment's eight-seed paired fixed-dose producer uses the same numerator and $\max(\bar{b}_{\text{target}} - \bar{b}_{\text{source}}, 10^{-6})$ in the denominator. Three-seed IOI reuse instead reports one minus the mean pointwise clean-normalized absolute error, with evaluation-pair directions retained in its fixed-coefficient branch. These recovery statistics are distinct from $1 - \text{Rel}_{\text{tar}}^{\text{test}}(1)$. Both Sentiment conditions retain training-fitted coefficients: P uses evaluation-pair directions and F uses training-mean directions. Recorded recovery values are reported in Appendix G.2.6; series-specific statistical procedures are given in Appendix H.

E Task instantiations

This section specifies each task's model, prompt construction, behavioral readout, paired targets, direction rules, and candidate library. The notation follows Sections 2–4: b is the behavioral readout, τ_i the target readout, u_v a candidate direction rule, $\mathcal{V}$ the candidate library, and $\mathcal{D}_{\text{off}}$ the preservation panel. Fitting and evaluation samples are indexed in Table G.2; numerical settings are given in Appendix F.

Preservation readouts. The differentiable surrogate ψ determines the local dictionary H; its coordinate metric Ω is specified with the surrogate (Appendix B.1). The nonlinear observable for the independent preservation panels is the full next-token distribution, $o(f_\theta, x) = p_\theta(\cdot \mid x)$, with

$$d_{\text{off}}(p_\theta, p_\theta^{\gamma\beta}) = \text{KL}\Big(p_\theta \,\|\, p_\theta^{\gamma\beta}\Big).$$

Baseline and patched forwards use the same specified input and readout position, with the patch's direction rule defined for that input. IOI paired-position and unpaired Sentiment diagnostics are defined in Appendix G.2.3. Executed refits and penalty terms are specified in Appendix F.

Paired targets and feedback targets. For paired fitting and reconstruction, τ_i is the target member's readout under the specified experimental configuration. On IOI this is the clean counterpart of the corrupt source; MCQA and the paired Sentiment/Refusal controls use their configuration-specific target members. Feedback uses separately specified batch-mean targets $m^\star$, whose executed grids are listed in Appendix F. Those grids are not assignments to the paired-fitting τ_i.

For a prescribed source–target pair (x_i^r, x_i^c), the evaluation-pair direction and fitting-split mean direction are

$$u_v^{\text{P}}(x_i^r) = h_{s_v}(x_i^c) - h_{s_v}(x_i^r), \qquad u_v^{\text{F}} = \frac{1}{n_{\text{fit}}} \sum_{i \in \mathcal{I}_{\text{fit}}} \big[h_{s_v}(x_i^c) - h_{s_v}(x_i^r)\big],$$

where $n_{\text{fit}} = |\mathcal{I}_{\text{fit}}|$. P and F specify direction information; coefficient fitting and dose updates are stated separately for each series. The unpaired Sentiment dose series uses the class-mean construction described below.

E.1 Indirect Object Identification

The IOI experiments use GPT-2 small with twelve layers and twelve attention heads per layer. Prompts follow paired ABBA/BABA templates such as "Then, *A* and *B* went to the place. *B* gave an item to". Templates and vocabulary items are sampled using the recorded seed and constructor. The behavioral readout is the indirect-object-minus-subject logit margin at the answer position. The

reference IO and subject token identities are taken from the clean member and kept fixed when scoring both members. The source is the corrupt member, and its target readout is the clean member's margin.

END is the final valid prompt position; S1 and S2 are the first and second subject occurrences, and IO is the indirect-object occurrence in each member's role assignment. Position rules are resolved separately for the two members. The library spans twelve layers, twelve heads, and END, IO, S1, and S2, giving 576 candidates. The nominal reconstruction uses 64 pairs and ten seeds, with the same panel used for fitting and evaluation. The separate feedback, dose, reuse, and recurrence series retain the configurations in Table G.2.

The IOI feedback and dose producers use raw clean-minus-corrupt activation contrasts at head–position sites. Candidate-score column normalization is distinct from unit-normalizing activation differences. The IOI paired-position diagnostic measures full-vocabulary KL at non-answer positions of the same paired prompts: patched relative to clean, with the unpatched corrupt-relative-to-clean KL reported separately as a baseline. Baseline subtraction applies only to quantities explicitly defined as Δ_{off} or ΔKL. This diagnostic uses a different panel and KL reference from the independent preservation measurements.

The additional IOI reversal keeps the ten-seed nominal supports and coefficients, applies the opposite patch to the target member using each evaluation pair's contrast, and uses the same 64-pair panel. It adds no frozen-direction IOI condition.

E.2 Multiple-choice question answering

Nominal reconstruction. The nominal reconstruction in Table 2 uses Qwen3-4B (Qwen Team, 2025) on the Humanities and STEM subject categories of MMLU. The readout is the vector of four option logits. The library contains `resid_post`, `mlp_out`, and `attn_out` at each of 36 layers, giving 108 module-level candidates. Each category uses eight questions on the same fitting and evaluation panel over three seeds. Table 2 reports $\text{Rel}_{\text{tar}}(1)$; Table G.5 provides the clean-output-normalized reconstruction statistic R_{clean} (Appendix G.2.1).

Head-level control. The repeated scalar-control series uses 1,152 answer-position attention-head candidates: 36 layers times 32 heads. Each seed draws 24 category-filtered MMLU test questions and divides them into twelve fitting and twelve evaluation examples. The source member exchanges the correct option with a seeded incorrect option; the target member retains the original order. The original answer index $a^\star$ is used for both members in the correct-minus-mean-wrong readout M_{MCQA} from Section 5.1. Supports and base coefficients are fitted on the fitting half; P and F use evaluation-pair and training-mean directions, respectively. Feedback, signed reversal, and independent dose response use eight seeds per category.

Module-level held-out reuse. The separate reuse series uses the scalar correct-minus-mean-wrong readout M_{MCQA} with the module-level library. It uses 32 questions per category, a 16/16 fitting/evaluation split, and four seeds. Refit-β retains the selected support with evaluation-side coefficient fitting and paired directions. Fully frozen open-loop reuse applies the fitting-side support, coefficients, mean directions, and nominal dose. Its recovery statistic and fitting normalization are specified in Appendices D.6 and F.

The additional STEM dose scan reuses the saved head-level supports and coefficients in both P and F conditions. The independent preservation panel contains sixteen Humanities and sixteen STEM questions. Both baseline and patched forwards use the original unswapped panel prompt; its option-swapped companion supplies the clean-minus-swapped activation contrast. Each prompt is processed without padding, with the patch and full-vocabulary KL readout at the final valid input position. The native support and coefficients remain fixed at nominal dose. Thus this panel uses evaluation-pair directions, separately from fully frozen reuse.

E.3 Sentiment steering

Sentiment uses GPT-2 small. Its positive list is {good, great, wonderful, amazing, happy, excellent, fantastic, lovely}; the negative list is {bad, terrible, awful, horrible, miserable, disappointing, boring, depressing}. Each word is encoded with a leading space and `add_special_tokens=False`; only

its first subtoken is used. The readout is the mean next-token log-probability of the positive IDs minus that of the negative IDs, not a whole-word sequence probability.

The unpaired-dose producer constructs pools of sixteen positive and sixteen negative templates, with eight fitting and eight evaluation examples per class. Selection and directions use fitting-side population means. Its progress divides the patched-minus-negative evaluation mean by the positive-minus-negative mean gap, floored at 10^{-6}. Its KL uses the same negative evaluation prompts (unpatched relative to patched), rather than an independent panel. The fitting/evaluation split and producer settings are specified in `configs/sentiment_unpaired_dose.json`, with the implementation registered in `experiments/registry.json`. The three-seed curves in Figure 2 summarize the saved unpaired-dose records identified in the accompanying source index.

The paired fixed-dose comparison contains sixteen template pairs per seed, shuffled into eight fitting and eight evaluation pairs. The negative template is the source and the positive template is the target. Five of the sixteen template pairs have identical source and target strings and therefore zero requested displacement; these pairs are retained in both fitting and evaluation according to the seeded split. The library has 144 last-position attention-head candidates. Support selection and ridge fitting use the fitting half. At nominal dose $\gamma = 1$, P uses evaluation-pair activation contrasts and F uses fitting-split mean contrasts; both retain training-fitted coefficients. Eight seeds and support budgets $k \in \{3, 5, 8\}$ are used for this comparison. Its mean-margin recovery retains the denominator specified in Appendix D.6. Excluding the zero-displacement pairs from the saved evaluation vectors leaves the $\varepsilon = 0.4$ pass counts unchanged for all three selectors and both direction conditions.

Additional target-side reversal and independent preservation reuse only the saved $k = 8$ interventions. Reversal tests both P and F on the eight evaluation pairs per seed, applying $-\beta$ to the positive target member. Preservation broadcasts fitting-split mean directions on the independent 32-prompt MMLU panel at nominal dose. Sentiment dose response and fixed-dose reuse are evaluated in separate unpaired and paired experiments, respectively. The reported Sentiment series contain no multi-target feedback experiment, so feedback controllability remains unassessed for this task.

E.4 Refusal steering

The Refusal experiments use Qwen3-1.7B and Arditi-inspired harmful/harmless instruction pairs (Arditi et al., 2024). The refusal list is {I, Sorry, Unfortunately, Cannot, Can}; the compliance list is {Sure, Of, Here, Let, Certainly}. Each word is encoded with a leading space and `add_special_tokens=False`, retaining its first subtoken. The readout is the mean refusal-ID minus mean compliance-ID next-token log-probability. This lexical readout is evaluated before generating a token, at the input position specified by the producer's ANSWER rule (Appendix F). It measures a lexical margin rather than completed-answer semantic safety.

In the paired controls, the harmless instruction is the source x_i^r and the harmful instruction is the target x_i^c. Fitting uses harmful-minus-harmless activation contrasts. Forward control applies β to the harmless member; signed reversal applies $-\beta$ to the harmful member using the same direction convention. Feedback targets are specified separately from the paired target readouts.

The library spans 28 layers and sixteen heads at ANSWER, giving 448 candidates. Twenty instruction pairs are shuffled by seed and split into ten fitting and ten evaluation pairs. P and F are both evaluated, with support and base coefficients fitted on the fitting half. The nominal-dose preservation test uses the independent 32-prompt MMLU panel with fitting-split mean directions.

The separate short-generation probe evaluates ten harmful prompts per seed across eight seeds under four conditions: the unpatched model and the three selectors. It generates exactly sixteen greedy tokens per continuation, giving 320 continuations on twenty prompt identities reused across seeds. Decoding and patch-position rules are specified in Appendix F; outcomes are reported in Appendix G.4.3.

F Hyperparameters and implementation details

Refit objectives and normalization. The formal refit in Appendices B–C uses sample-normalized target and preservation terms with an activation-energy penalty and constraint. The repeated scalar

MCQA and Refusal runs fit the unnormalized training dictionary by coefficient ridge:

$$\min_{\beta_S} \|q_{\text{fit}} - G_{\text{fit},S}\beta_S\|_2^2 + \lambda\|\beta_S\|_2^2, \qquad \lambda = 1,$$

with solution

$$\widehat{\beta}_S = (G_{\text{fit},S}^\top G_{\text{fit},S} + \lambda I)^{-1} G_{\text{fit},S}^\top q_{\text{fit}}.$$

These runs use raw activation contrasts and omit the preservation block and independent energy-ball constraint. The directions are not unit-normalized, so coefficient ridge does not generally penalize the activation-space energy $E(\beta) = \beta^\top W\beta$. The conditions for equality with $\|\beta\|_2^2$ are stated in Appendix A. For fixed directions, $E(\gamma\beta) = \gamma^2 E(\beta)$; its evaluation requires W on the specified evaluation set. The numerical stabilizer is $\xi = 10^{-6}$.

Nominal module-level MCQA uses $\lambda = 10^{-5}$, while nominal IOI uses $\lambda = 1$, with their raw flattened dictionaries and displacements. The retained paired and unpaired Sentiment producers likewise fit raw dictionaries; their recorded reconstruction recipes specify `--ridge 1`. Module-level MCQA reuse uses $\lambda = 1$ after dividing both its dictionary and displacement by the square root of the flattened target-readout row count n_{row}. The relative penalty scale therefore differs:

$$\frac{1}{n_{\text{row}}}\|q - G\beta\|_2^2 + \lambda\|\beta\|_2^2 \quad \Longleftrightarrow \quad \|q - G\beta\|_2^2 + n_{\text{row}}\lambda\|\beta\|_2^2.$$

Each series retains its own scaling and refit parameter. The paired and unpaired Sentiment configurations are provided in `configs/sentiment_paired.json` and `configs/sentiment_unpaired_dose.json`. Module-level MCQA reuse is configured in `configs/mcqa_humanities_reuse.json` and `configs/mcqa_stem_reuse.json`. Their execution modules are registered in `experiments/registry.json`.

Intervention coordinates. The retained TransformerLens head-position adapter patches the selected head slice at `hook_z`, before the attention output projection. Its position rule selects the specified token position. The module-level adapter defines `resid_post` as the post-block residual, `mlp_out` as the MLP output, and `attn_out` as the attention-module output. Tokenizer and position rules remain specific to each experimental series.

Executed support selection. For the repeated scalar MCQA and Refusal controls, the executed OMP score is

$$\text{score}_{\text{exec}}(v \mid S_t) = \frac{|\langle G_v, r_t\rangle|^2}{\max(\|G_v\|_2^2, 10^{-12})\, c_v}, \qquad c_v = 1.$$

There is no λW_{vv} term in this selection denominator; ridge enters the subsequent strength refit. The OMP stopping defaults are residual tolerance 10^{-3} and improvement threshold 10^{-4}. The executed OMP implementation evaluates $\|q - G_S\widehat{\beta}_S\|_2 / \max(\|q\|_2, 10^{-12})$ on the current support at the start of each iteration; a stagnation exit retains that already selected support. This differs from the trial acceptance and rollback rule in Algorithm C.3. Coverage uses `signed_split=False`, `cover_mode=soft`, `weight_mode=absq`, `eps_norm=` 10^{-6}, `coverage_target=` 0.95, and `min_marginal_gain=` 10^{-4}. Its executed row weights and contributions are

$$d_j = \max(|q_j|, 10^{-6}), \qquad w_j = d_j, \qquad a_{jv} = \text{clip}\left(\frac{\text{sign}(q_j)G_{jv}}{d_j}, 0, 1\right).$$

The implementation thus uses a maximum floor, rather than the additive floor of the formal contribution in Eq. (B.26). Negative alignment is clipped during proposal; final ridge coefficients may have either sign. Coverage is accumulated as $C(S) = \sum_j w_j \min(1, \sum_{v\in S} a_{jv})$, with stopping based on $C(S)/\sum_j w_j$ and marginal gain normalized by $\sum_j w_j$. The formal settings in Table F.1 do not override these executed values.

Module-level MCQA reuse implementation. The module-level held-out reuse records use a separate selector implementation and explicit `refit_mode=ridge`. Its target dictionary and displacement are normalized by the square root of the flattened target-readout row count. The OMP defaults are residual tolerance 0.05, minimum marginal gain 10^{-3}, and cost stabilizer $\eta_0 = 10^{-4}$; these differ from the scalar-control defaults above. The selected training and evaluation dictionaries are refitted by plain ridge for their respective reuse conditions. The optional constrained-refit and certificate helpers are not called for these recorded runs.

Executed feedback and fixed-dose settings. For the repeated MCQA and Refusal controls, the support budget is at most eight, ridge is $\lambda = 1$, and the three selectors are Driver-OMP, Coverage-Driver, and Top-k. Top-k ranks $|G_{\text{train}}^{\top} q_{\text{train}}|$; all three fit the selected coefficients on the fitting split. Feedback observes the evaluation batch's mean margin and updates $\gamma_{t+1} = \gamma_t + \eta(m^\star - M(\gamma_t))$, starting from $\gamma_0 = 0$, with no dose clipping or gain halving. The iteration cap is twelve, followed by the producer's final endpoint evaluation. MCQA uses $\eta = 0.15$, endpoint tolerance 0.3, and targets $\{-2, -1, 0, 1, 2, 3\}$ for both categories. Refusal uses $\eta = 0.5$, tolerance 0.1, and targets $\{-1.55, -1.45, -1.35, -1.25, -1.15\}$. A hit is determined from the final realized batch-mean error, including any target already within tolerance before a nonzero intervention. These executed settings are distinct from the clipped reference controller in Appendix D.4.

The separate IOI feedback record contains three seeds and eight batch-mean targets $\{-3, -1, 0, 1, 2, 3, 4, 5\}$, with endpoint tolerance $\tau_{\text{ctrl}} = 0.3$ and the same 64-prompt fitting/evaluation panel. Its executable settings are specified in `configs/ioi_feedback.json`.

Dose-response grids and additional measurements. The independent Humanities, STEM, and Refusal scans use

$$\Gamma = \{-1, -0.5, 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 2\}.$$

The IOI dose record uses the same ten points. The unpaired Sentiment dose record uses

$$\Gamma_{\text{Sentiment}} = \{0, 0.25, 0.5, 1, 1.5, 2, 3\}.$$

Support, base coefficients, and direction construction remain fixed within each scan; the series retain their own sample sizes, seeds, and support budgets. Additional IOI/Sentiment reversal applies target-side $-\beta$ without reselection or refitting. The added Sentiment reversal and preservation use the saved $k = 8$ controls, rather than the full $k \in \{3, 5, 8\}$ fixed-dose comparison. The Refusal preservation panel uses training-mean directions at $\gamma = 1$ on 32 prompts reused across eight seeds. MCQA uses the panel-specific pairing and input member in Appendix E.2.

Refusal formatting and generation. The retained batched Refusal producer applies the tokenizer's chat template to a user message with `tokenize=False` and `add_generation_prompt=True`, then tokenizes with padding, truncation, and maximum input length 64. Its ANSWER index is computed as $\sum_t \text{attention_mask}_{it} - 1$. The template call does not pass an explicit `enable_thinking` argument. The fixed nominal-dose readout measurements record right padding and right truncation. The short-generation probe uses the same template-call arguments and maximum input length 64 with truncation enabled, but processes one harmful prompt at a time without padding. At each step it applies the saved nominal-dose patch at the current final valid input position, with fixed support, coefficients, and fitting-split mean directions. It generates exactly sixteen argmax tokens, without EOS-based early termination. The generation configuration is provided in `configs/refusal_generation.json`, with execution instructions in `docs/experiments.md`. The batched lexical metrics use the implemented ANSWER index; the unpadded generation probe directly uses the final valid position.

Operating criteria and statistical units. Signed reversal uses a distance ratio of at least two. Dose response uses signed Spearman correlation $\rho \geq 0.95$, with strong negative ordering $\rho \leq -0.95$ reported separately. Feedback uses the task-specific endpoint tolerances above. Each feedback observation is a seed–target batch-mean endpoint; initial zero-dose hits are included in the total and reported separately.

The independent preservation panels report both a seed mean over 32 prompts and an all-prompt rule, each requiring KL at most 0.001. The same 32 prompts are reused across eight seeds, giving 256 prompt–seed measurements per selector. For sorted pooled values $z_{(1)} \leq \cdots \leq z_{(256)}$, the nearest-rank percentile is

$$P_{95} = z_{(\lceil 0.95 \times 256 \rceil)} = z_{(244)}.$$

The panel measurements use nominal dose and the saved native supports, coefficients, and direction rules in Appendix G.2.5. These axis-specific operating criteria are separate from the joint nonlinear certificate in Definition 1.

The reference linear algebra uses PyTorch `float32`. The repeated control runs use MPS or CUDA execution profiles. Available runtime metadata are included with the corresponding measurement records. The active-gain step in Driver-OMP was checked against `float64`; the only observed

Table F.1: Formal reference settings and recorded IOI recurrence. Setting type distinguishes the formal reference from an executed analysis. Repeated MCQA/Refusal feedback parameters are specified only in the executed-settings paragraph above; the clipped controller of Eq. (D.35) is a reference rule.

Setting type	Component	Setting
Reference	Refit and certificate	Ridge and preservation weights $\lambda = \mu = 1$; activation-energy bound $E \le \kappa = 2$; target criterion $\mathrm{Rel}_{\mathrm{tar}}(1) \le 0.4$; preservation bound $L_{\mathrm{off}} \le 0.001$.
Reference	Sparse selection	DRIVER-OMP residual tolerance $\varepsilon_{\mathrm{lin}} = 0.05$ and improvement tolerance $\eta_{\mathrm{tol}} = 10^{-3}$; COVERAGE-DRIVER coverage $\tau = 0.95$. Denominator stabilizers $\delta_0 = \eta_0 = 10^{-3}$; normalized marginal-gain threshold $g_{\min} = 10^{-3}$. Uniform row weights $w_j = 1/n$ for $j = 1, \ldots, nr$ give $W_{\mathrm{row}} = \sum_j w_j = r$; Algorithm C.4 normalizes coverage by W_{row}. Target normalization guard $\xi = 10^{-6}$.
Executed	IOI recurrence	Five seeds and five budgets $k \in \{3, 5, 8, 12, 16\}$ give 25 seed–budget configurations; consensus threshold $\pi^\star = 0.6$.

differences were isolated tie breaks between candidates with nearly identical scores. The formal reference procedure in Algorithm C.2 uses $t_E = 10^{-3}$ and $t_\nu = 10^{-8}$. It returns the feasible upper-endpoint coefficients when $\kappa - e_{\mathrm{hi}} \le t_E \kappa$ or $\nu_{\mathrm{hi}} - \nu_{\mathrm{lo}} \le t_\nu$, solving the cached positive-definite ridge system at each multiplier. These tolerances specify the formal procedure and do not describe the plain-ridge refits used in the recorded scalar-control and module-level reuse runs.

G Empirical analysis

This appendix organizes the quantitative evidence in the same six-axis order as Sections 4 and 5.2. The comparisons retain each experiment's readout, direction information, and reporting unit.

G.1 Experimental configurations and reporting units

Table G.2 distinguishes the experimental series. IOI and Sentiment use GPT-2 small; the main MCQA studies use Qwen3-4B, and Refusal uses Qwen3-1.7B. P denotes evaluation-pair activation contrasts; F denotes fitting-split mean directions. IOI/Sentiment reversal, STEM dose, and MCQA/Sentiment preservation measurements use fixed supports and coefficients without reselection or refitting. Task definitions and executed settings are given in Appendices E and F.

Table G.2: Experimental configurations and reporting units. Panel sizes are per seed; a fitting/evaluation split is written as fit/eval. P and F specify direction information, not different sample sizes.

Series	Readout and panel	Repetitions and reporting unit
IOI nominal	IO–S margin; 64 pairs, same panel	10 seeds; stacked R_{clean} and $\mathrm{Rel}_{\mathrm{tar}}(1)$
MCQA nominal	Four option logits; 8 questions/category, same panel	3 seeds/category; stacked R_{clean} and $\mathrm{Rel}_{\mathrm{tar}}(1)$
MCQA control	Scalar mean-wrong margin; 24 questions/category, 12/12; head library; P/F	8 seeds; $8 \times 6 = 48$ feedback endpoints/selector/mode; 8 reversal ratios; 8 nominal-dose $\mathrm{Rel}_{\mathrm{tar}}(1)$ values/selector/mode
Refusal control	Lexical margin; 20 pairs, 10/10; head library; P/F	8 seeds; $8 \times 5 = 40$ feedback endpoints/selector/mode; 8 reversal ratios; 8 nominal-dose $\mathrm{Rel}_{\mathrm{tar}}(1)$ values/selector/mode
IOI reversal	Saved nominal interventions; same 64 pairs; P	10 seed-level ratios/selector
Sentiment reversal	Saved $k = 8$ paired interventions; 16 pairs, 8/8; P/F	8 seed-level ratios/selector/mode; 8 nominal-dose $\mathrm{Rel}_{\mathrm{tar}}(1)$ values/selector/mode
Independent dose	Humanities, STEM, Refusal; control evaluation splits; P/F	8 seeds, 10 doses; 48 curves and 480 points/task
Preservation	Fixed 32-prompt MMLU panel; MCQA P, Sentiment/Refusal F; nominal dose	8 seeds; 256 prompt–seed observations/selector on 32 unique prompts

Series	Readout and panel	Repetitions and reporting unit
MCQA module reuse	Signed margin recovery; 32 questions/category, 16/16	4 seeds/category; evaluation-side coefficient refit or fully frozen
Sentiment reuse	Signed margin recovery; 16 pairs, 8/8; P/F; $k = 3, 5, 8$	8 seeds; fixed-dose recovery
IOI reuse	Pointwise clean-normalized recovery; 64 fitting/64 test pairs; $k \leq 8$	3 seeds; refit or fixed coefficients, both with evaluation-pair directions
IOI dose	64 pairs, same panel; mean-margin progress and non-answer-position ΔKL	Driver-OMP: 2 seeds, $k = 4$; Random: 3 seeds, $k = 8$
Sentiment unpaired dose	Word-set margin; unpaired prompt pools; seven doses	3 seeds/method; progress and same-prompt KL
IOI feedback	Batch-mean IO–S margin; same 64-prompt panel	3 seeds, 8 targets; 24 endpoints
IOI supporting analyses	Recurrence: $k = 3, 5, 8, 12, 16$. Backup: separate 64-prompt panel	Recurrence: 5 seeds, 25 configurations. Backup: one to seven sites
Model comparison	Four Humanities questions, same panel; four option logits	Qwen3-8B: 5 seeds, residual-only; Mistral-7B: 3 seeds, three module types
Refusal continuations	10 harmful evaluation prompts/seed; 16 greedy tokens; unpatched + 3 selectors	8 seeds; 320 continuations on 20 reused prompt identities

Nominal reconstruction, formal reachability, and reuse report means and sample SD across seeds. Feedback observations are seed–target batch-mean endpoints; reversal ratios and dose correlations are seed-level quantities. Preservation retains both prompt and seed identities: 256 measurements on a reused 32-prompt panel are not 256 different prompts. Continuation counts likewise retain seed, method, and prompt identities.

G.2 Results across the six control axes

G.2.1 Reachability: same-panel reconstruction and held-out nominal-dose control

Table 2 compares five methods on the same fitting and evaluation panels, using ten IOI seeds and three seeds per MCQA category. The stacked readouts are IO–S margins for IOI and four option logits for module-level MCQA. Target recovery is evaluated using the displacement-normalized residual

$$\mathrm{Rel}_{\mathrm{tar}}(1) = \frac{\|B_{\mathrm{target}} - B_{\mathrm{patched}}\|_2}{\|B_{\mathrm{target}} - B_{\mathrm{source}}\|_2 + 10^{-6}}.$$

The complete stacked readout vectors are evaluated within each seed before computing the across-seed mean, sample standard deviation, and pass count at $\varepsilon = 0.4$. The maximum support budget is eight. Supports, coefficients, and direction rules are held fixed at nominal dose $\gamma = 1$; the results do not use feedback-selected doses. Scalar head-level MCQA control uses a separate correct-minus-mean-wrong margin with the original reference answer index. Nonlinear greedy obtains the lowest mean residual in all three same-panel tasks, at additional model-evaluation cost.

Table G.3 gives the same-panel residuals and threshold pass counts. For IOI Driver-OMP, the residual is 0.201 ± 0.058 and passes in 10/10 seeds. Coverage-Driver passes in 3/3 module-level seeds for both MCQA categories; Humanities Top-k passes in 2/3 seeds. These results use the fitting panels and do not by themselves establish held-out generalization.

Table G.5 reports the complementary clean-output-normalized reconstruction error

$$R_{\mathrm{clean}} = \frac{\|B_{\mathrm{clean}} - B_{\mathrm{patched}}\|_2}{\max(\|B_{\mathrm{clean}}\|_2, 10^{-12})}.$$

Here $B_{\mathrm{clean}} = B_{\mathrm{target}}$: both statistics measure the same reconstruction error, but use different normalization denominators. The clean-output-normalized scores do not directly determine the target-residual reachability criterion.

Table G.4 reports the separate held-out fixed-dose experiments, with evaluation-pair directions (P) and training-mean directions (F). On Humanities and STEM, Coverage-Driver and Top-k pass in 8/8 P seeds per category, but in 0/8 F seeds. Driver-OMP passes in 6/8 P seeds in each category and 0/8 F seeds. For Sentiment, Coverage-Driver and Top-k pass in 8/8 P seeds and 0/8 F seeds; Driver-OMP

passes in 3/8 P seeds and 0/8 F seeds. Refusal passes in 0/8 seeds for all selectors in either mode. Evaluation-pair directions depend on paired evaluation inputs and are not fully frozen transferable interventions. All counts concern only the target-residual component of the control specification, not the joint nonlinear certificate.

Table G.3: Formal $\mathrm{Rel}_{\mathrm{tar}}(1)$ on the same fitting/evaluation panel at nominal dose $\gamma = 1$ and maximum support budget eight. Cells give mean ± sample SD (number of seeds satisfying $\varepsilon = 0.4$/total). IOI uses 64 paired logit-margin readouts (10 seeds); MCQA uses four option logits on eight questions (3 seeds/category). All directions are evaluation-pair (P); lower residuals are better.

Method	IOI (10 seeds)	Humanities (3 seeds)	STEM (3 seeds)
Driver-OMP	0.201 ± 0.058 (10/10)	0.279 ± 0.032 (3/3)	0.175 ± 0.035 (3/3)
Coverage-Driver	0.294 ± 0.029 (10/10)	0.152 ± 0.154 (3/3)	0.061 ± 0.033 (3/3)
Top-k	0.240 ± 0.035 (10/10)	0.333 ± 0.117 (2/3)	0.221 ± 0.046 (3/3)
Random	0.747 ± 0.240 (1/10)	1.209 ± 0.949 (0/3)	1.167 ± 1.141 (0/3)
Nonlinear greedy	0.151 ± 0.037 (10/10)	0.073 ± 0.093 (3/3)	0.020 ± 0.012 (3/3)

Table G.4: Formal $\mathrm{Rel}_{\mathrm{tar}}(1)$ on held-out evaluation prompts at fixed dose $\gamma = 1$ and maximum support budget eight. The unit is one seed-level stacked vector; entries report mean ± sample SD and number of seeds satisfying $\varepsilon = 0.4$ out of eight. Head-level MCQA, paired Sentiment, and Refusal use distinct task-specific readouts. P uses evaluation-pair directions; F uses training-mean directions.

Task	Method	P: mean ± SD (pass)	F: mean ± SD (pass)
Humanities	Driver-OMP	0.495 ± 0.428 (6/8)	1.009 ± 0.048 (0/8)
	Coverage-Driver	0.188 ± 0.072 (8/8)	1.012 ± 0.058 (0/8)
	Top-k	0.192 ± 0.071 (8/8)	1.010 ± 0.055 (0/8)
STEM	Driver-OMP	0.497 ± 0.439 (6/8)	1.016 ± 0.048 (0/8)
	Coverage-Driver	0.169 ± 0.067 (8/8)	1.031 ± 0.053 (0/8)
	Top-k	0.170 ± 0.046 (8/8)	1.030 ± 0.051 (0/8)
Sentiment	Driver-OMP	0.445 ± 0.088 (3/8)	0.766 ± 0.065 (0/8)
	Coverage-Driver	0.171 ± 0.084 (8/8)	0.746 ± 0.056 (0/8)
	Top-k	0.167 ± 0.084 (8/8)	0.742 ± 0.052 (0/8)
Refusal	Driver-OMP	0.932 ± 0.045 (0/8)	0.928 ± 0.053 (0/8)
	Coverage-Driver	0.799 ± 0.037 (0/8)	0.801 ± 0.039 (0/8)
	Top-k	0.774 ± 0.043 (0/8)	0.779 ± 0.044 (0/8)

Table G.5: Clean-output-normalized reconstruction. Mean ± sample SD of R_{clean}, lower is better. Same-panel fitting and evaluation use 64 IOI pairs or eight MCQA questions per category. Evaluation-pair directions and nominal dose $\gamma = 1$ are used. Maximum budget is eight; early stopping can use fewer candidates. Nonlinear greedy uses additional model evaluations.

Method	IOI (10 seeds)	Humanities (3 seeds)	STEM (3 seeds)
Driver-OMP	0.386 ± 0.114	0.045 ± 0.013	0.030 ± 0.008
Coverage-Driver	0.567 ± 0.059	0.027 ± 0.031	0.010 ± 0.005
Top-k	0.464 ± 0.078	0.054 ± 0.024	0.038 ± 0.012
Random	1.430 ± 0.442	0.185 ± 0.132	0.196 ± 0.185
Nonlinear greedy	0.289 ± 0.068	0.013 ± 0.018	0.003 ± 0.002

G.2.2 Signed reversal

Table G.6 combines the extended MCQA/Refusal experiments with the IOI/Sentiment measurements. The patch $-\beta$ is applied to the target member. A ratio of target-distance to source-distance of at least two defines the operating pass. Each ratio uses a full evaluation vector; multiple feedback targets do not multiply the number of reversal replicates. IOI has ten P seeds and no added F test. Sentiment has eight P and F seeds, using eight evaluation pairs per seed. The IOI/Sentiment tests keep the fitted supports and coefficients fixed.

Table G.6: Formal target-side reversal. Ratio at least two is the operating criterion; mean and sample SD are over seeds, not feedback targets. IOI is P-only; P and F are defined in Table G.2.

Task	Mode	Selector	Pass/seeds	Ratio mean ± SD
IOI	P	DRIVER-OMP	10/10	4.7425 ± 1.1143
IOI	P	COVERAGE-DRIVER	10/10	2.5702 ± 0.3689
IOI	P	Top-k	10/10	3.4315 ± 0.8456
Sentiment	P	DRIVER-OMP	4/8	2.3722 ± 1.0708
Sentiment	F	DRIVER-OMP	0/8	0.6547 ± 0.2871
Sentiment	P	COVERAGE-DRIVER	8/8	8.5143 ± 2.9573
Sentiment	F	COVERAGE-DRIVER	0/8	0.9969 ± 0.4172
Sentiment	P	Top-k	8/8	8.7627 ± 3.4087
Sentiment	F	Top-k	0/8	1.0000 ± 0.4175
Humanities	P	DRIVER-OMP	5/8	3.0596 ± 1.7541
Humanities	F	DRIVER-OMP	0/8	0.1304 ± 0.0733
Humanities	P	COVERAGE-DRIVER	8/8	3.7768 ± 0.8716
Humanities	F	COVERAGE-DRIVER	0/8	0.1428 ± 0.0905
Humanities	P	Top-k	8/8	4.2961 ± 2.2079
Humanities	F	Top-k	0/8	0.1424 ± 0.0880
STEM	P	DRIVER-OMP	7/8	3.6450 ± 2.0158
STEM	F	DRIVER-OMP	0/8	0.1312 ± 0.0658
STEM	P	COVERAGE-DRIVER	8/8	4.8023 ± 1.4251
STEM	F	COVERAGE-DRIVER	0/8	0.1520 ± 0.0533
STEM	P	Top-k	8/8	5.0921 ± 1.5127
STEM	F	Top-k	0/8	0.1441 ± 0.0516
Refusal	P	DRIVER-OMP	0/8	0.0770 ± 0.0513
Refusal	F	DRIVER-OMP	0/8	0.0713 ± 0.0519
Refusal	P	COVERAGE-DRIVER	0/8	0.1123 ± 0.0452
Refusal	F	COVERAGE-DRIVER	0/8	0.1252 ± 0.0573
Refusal	P	Top-k	0/8	0.1356 ± 0.0527
Refusal	F	Top-k	0/8	0.1452 ± 0.0606

Across selectors, Humanities and STEM have 21/24 and 23/24 passing P method–seed cells, respectively, and no passing F cells. Refusal has 0/24 in both modes, despite its ordered dose responses below. The per-seed behavioral vectors and ratios remain in the saved records. Source-side negative-dose scans start from a different input member and are not substituted for target-side reversal.

G.2.3 Dose response

Figure 3 uses the independent grid $\{-1, -0.5, 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 2\}$. STEM, Humanities, and Refusal use the same dose grid. Each task has three selectors, two direction conditions, and eight seeds: 48 curves and 480 measured points. Ordering uses the signed Spearman coefficient, with $\rho \geq 0.95$ as the positive-response criterion. Strong negative ordering $\rho \leq -0.95$ is reported separately. Supports, coefficients, and direction construction remain fixed within each scan; no controller chooses the visited doses.

Table G.7: Signed Spearman ordering over ten independent doses per seed. Positive and strong negative counts are separate; absolute correlations are not a pass criterion.

Task	Mode	Selector	$\rho \geq .95$	$\rho \leq -.95$	Mean	SD
Humanities	P	DRIVER-OMP	3/8	0/8	0.5970	0.6005
Humanities	P	COVERAGE-DRIVER	5/8	0/8	0.9394	0.0987

Table G.7 (continued)

Task	Mode	Selector	$\rho \geq .95$	$\rho \leq -.95$	Mean	SD
Humanities	P	Top-k	2/8	0/8	0.8939	0.0727
Humanities	F	DRIVER-OMP	4/8	3/8	0.2303	1.0202
Humanities	F	COVERAGE-DRIVER	5/8	3/8	0.2500	1.0351
Humanities	F	Top-k	5/8	3/8	0.2500	1.0351
STEM	P	DRIVER-OMP	4/8	0/8	0.5076	0.6972
STEM	P	COVERAGE-DRIVER	6/8	0/8	0.9742	0.0352
STEM	P	Top-k	5/8	0/8	0.9621	0.0495
STEM	F	DRIVER-OMP	4/8	4/8	-0.0015	1.0674
STEM	F	COVERAGE-DRIVER	4/8	4/8	0.0000	1.0690
STEM	F	Top-k	4/8	4/8	0.0000	1.0690
Refusal	P	DRIVER-OMP	8/8	0/8	1.0000	0.0000
Refusal	P	COVERAGE-DRIVER	8/8	0/8	1.0000	0.0000
Refusal	P	Top-k	8/8	0/8	1.0000	0.0000
Refusal	F	DRIVER-OMP	8/8	0/8	1.0000	0.0000
Refusal	F	COVERAGE-DRIVER	8/8	0/8	1.0000	0.0000
Refusal	F	Top-k	8/8	0/8	1.0000	0.0000

Each selector has three strongly negative Humanities F seeds and four strongly negative STEM F seeds. All Refusal curves have $\rho = 1$, although formal reversal fails in both modes. Positive ordering of forward doses and target-side reversibility are empirically distinct.

IOI and Sentiment dose diagnostics. The IOI panels in Figure 2 compare Driver-OMP with four selected sites and seeds 0/1 against random supports of size eight and seeds 0/1/2. The full IOI dose series also contains a third Driver-OMP seed with seven sites. The illustration therefore compares unequal support sizes and seed counts. Progress is computed from batch-mean margins. The non-answer-position diagnostic subtracts KL(corrupt || clean) from KL(patched || clean) on the same paired prompts; it is distinct from the independent 32-prompt preservation panel in Appendix G.2.5.

The Sentiment panels summarize three seeds per method at seven doses, $\{0, 0.25, 0.5, 1, 1.5, 2, 3\}$, from the saved unpaired-dose records. They report normalized positive-minus-negative mean-margin progress and KL on the same negative evaluation prompts, separately from the paired eight-seed fixed-dose comparison and the independent 32-prompt preservation panel. The record files and their producer configuration are identified in the accompanying experiment archive.

G.2.4 Feedback controllability

Table 3 reports terminal target hits, while Figure 4 illustrates recorded batch-mean tracking trajectories. Table G.8 additionally reports terminal mean absolute errors, recorded readout-evaluation counts, and initial zero-dose hits.

Within each seed, the selected support, base coefficients, and direction rule remain fixed across requested targets; only the dose is adjusted. The operating tolerances are 0.3 for IOI and MCQA and 0.1 for Refusal. The MCQA and Refusal controllers use up to twelve update iterations followed by a final endpoint evaluation, without dose clipping or gain halving.

Table G.8: Repeated scalar feedback. Terminal target hits, mean absolute tracking error (MAE), mean recorded readout-evaluation count, and initial zero-dose hits. Zero-dose readouts may reuse the cached unpatched baseline. Errors and evaluation counts are averaged over all seed–target endpoints, including unsuccessful trials. Initial hits are included in the total hit count and reported separately.

Task	Selector	Mode	Hits	MAE	Evals	Initial
IOI	DRIVER-OMP	P	24/24	0.1351	3.958	2/24
Humanities	DRIVER-OMP	P	38/48	1.0026	6.125	6/48
Humanities	DRIVER-OMP	F	8/48	2.2641	11.479	6/48
Humanities	COVERAGE-DRIVER	P	46/48	0.2378	4.417	6/48
Humanities	COVERAGE-DRIVER	F	8/48	1.9239	11.479	6/48
Humanities	Top-k	P	45/48	0.2679	4.583	6/48
Humanities	Top-k	F	8/48	1.8825	11.438	6/48
STEM	DRIVER-OMP	P	39/48	1.0743	5.354	3/48

Task	Selector	Mode	Hits	MAE	Evals	Initial
STEM	DRIVER-OMP	F	8/48	2.1875	11.896	3/48
STEM	COVERAGE-DRIVER	P	47/48	0.1321	3.854	3/48
STEM	COVERAGE-DRIVER	F	7/48	2.4494	11.917	3/48
STEM	Top-k	P	47/48	0.1364	3.771	3/48
STEM	Top-k	F	7/48	2.4310	11.938	3/48
Refusal	DRIVER-OMP	P	12/40	0.1885	10.100	10/40
Refusal	DRIVER-OMP	F	12/40	0.1870	10.100	10/40
Refusal	COVERAGE-DRIVER	P	21/40	0.1167	9.450	10/40
Refusal	COVERAGE-DRIVER	F	21/40	0.1174	9.375	10/40
Refusal	Top-k	P	21/40	0.1084	9.225	10/40
Refusal	Top-k	F	22/40	0.1098	9.225	10/40

Figure 4(a) shows all three IOI seeds, with eight requested targets per seed on the same 64-prompt fitting and evaluation panel. All $24/24$ endpoints are within tolerance, including two initial zero-dose hits.

Panels (b) and (c) compare Humanities seed 0 and the same six target margins under P and F directions, with $6/6$ and $1/6$ terminal hits, respectively. Panel (d) illustrates Refusal under F directions for seed 0, with $2/5$ hits. These seed-specific trajectories are included in the aggregate results of Table 3 and do not represent additional independent trials.

G.2.5 Off-target preservation

The fixed panel has sixteen Humanities and sixteen STEM questions. Its 32 prompt identities are reused across seeds and methods. At $\gamma = 1$, the MCQA measurements use each off-target prompt's own clean-minus-option-exchanged activation contrast. Both baseline and patched forwards use the original unswapped prompt; the companion supplies the direction only. Inputs are processed without padding, and the patch and full-vocabulary KL readout use the final valid input position. Saved task-native supports and coefficients remain fixed. Sentiment and Refusal broadcast their fitting-split mean directions; the added Sentiment measurements use the saved $k = 8$ interventions. Consequently, the MCQA measurement is not a fully frozen-direction intervention on unseen inputs. These measurements are distinct from the paired-position diagnostics in Figure 2.

For Refusal, 196/256, 82/256, and 93/256 prompt–seed measurements satisfy the KL criterion of at most 0.001 for DRIVER-OMP, COVERAGE-DRIVER, and Top-k, respectively.

Table G.9: Full-vocabulary KL(unpatched || patched) at nominal dose, using natural logarithms. Each row pools 32 prompts × eight seeds (256 measurements on 32 unique prompts). P95 is nearest rank. Last columns count seeds meeting mean-KL or all-32-prompt KL ≤ 0.001. MCQA uses each off-prompt pair; Sentiment/Refusal use training means.

Task	Selector	Mean KL	Median	P95	Max	Mean-pass	All-pass
Sentiment	DRIVER-OMP	0.005527	0.004959	0.011119	0.0286	0/8	0/8
Sentiment	COVERAGE-DRIVER	0.011015	0.009627	0.024779	0.0381	0/8	0/8
Sentiment	Top-k	0.011567	0.010322	0.024746	0.0387	0/8	0/8
Humanities	DRIVER-OMP	0.496530	0.050394	2.349920	15.5523	0/8	0/8
Humanities	COVERAGE-DRIVER	0.109692	0.050405	0.392607	1.0154	0/8	0/8
Humanities	Top-k	0.115396	0.056724	0.478003	0.6963	0/8	0/8
STEM	DRIVER-OMP	0.581701	0.051584	3.088410	17.7225	0/8	0/8
STEM	COVERAGE-DRIVER	0.128467	0.060374	0.465265	1.0098	0/8	0/8
STEM	Top-k	0.118910	0.055203	0.511026	0.7559	0/8	0/8
Refusal	DRIVER-OMP	0.001080	0.000174	0.005086	0.0410	4/8	2/8
Refusal	COVERAGE-DRIVER	0.008658	0.002119	0.035977	0.1698	0/8	0/8
Refusal	Top-k	0.007986	0.001948	0.036849	0.1572	0/8	0/8

For Humanities, STEM, and Sentiment, none of the eight seeds satisfies either preservation criterion under any selector. On the Refusal panel, Driver-OMP passes in four seeds by the mean rule and two by the all-prompt rule; Coverage-Driver and Top-k pass in none. Large predictive-distribution

changes are not confined to one extreme observation: the MCQA medians are approximately 0.05–0.06. The normalized question-string check found no exact duplicates between the fixed panel and the native MCQA task pools, but that does not establish semantic independence. These KL results measure predictive-distribution change, not completed-answer safety or semantic correctness.

G.2.6 Held-out reuse

The module-level MCQA study uses four seeds per category, 32 questions split 16/16, and signed mean-margin recovery. Table G.10 reports this four-seed series separately from the head-level controls. Refit uses evaluation targets and evaluation-pair directions; fully frozen open-loop reuses the fitting-side support, coefficients, mean directions, and nominal dose. Recovery can exceed one because it is a signed margin ratio, not an accuracy or clipped probability.

Table G.10: Module-level MCQA reuse. Seed means and sample SD. Refit retains evaluation-target access; fully frozen means the full open-loop intervention is fixed. Each category has four seeds.

Task	Method	Seeds	Refit-β	Fully frozen
Humanities	Driver-OMP	4	0.992 ± 0.009	0.021 ± 0.026
Humanities	Coverage-Driver	4	0.995 ± 0.053	0.033 ± 0.029
Humanities	Top-k	4	1.078 ± 0.017	0.028 ± 0.024
Humanities	Random	4	0.758 ± 0.443	0.022 ± 0.023
STEM	Driver-OMP	4	0.987 ± 0.034	0.024 ± 0.066
STEM	Coverage-Driver	4	0.944 ± 0.056	0.038 ± 0.108
STEM	Top-k	4	1.075 ± 0.029	0.031 ± 0.093
STEM	Random	4	0.744 ± 0.424	0.041 ± 0.108

The IOI study in Table G.11 uses seeds $s \in \{0, 1, 2\}$, a maximum support budget of eight, and 64 fitting pairs plus 64 test pairs per seed. The two panels are sampled separately from the same prompt distribution, using seeds s and $s + 10{,}000$, respectively. For test pair i, let b_i^{clean} and b_i^{patched} be the scalar IO–S margins. The pointwise error and seed-level recovery are

$$e_i = \frac{|b_i^{\text{patched}} - b_i^{\text{clean}}|}{\max(|b_i^{\text{clean}}|, 10^{-6})}, \qquad \text{Rec} = 1 - \frac{1}{64}\sum_{i=1}^{64} e_i.$$

A test pair fails when $e_i > 0.3$. Each failure fraction is $64^{-1}\sum_{i=1}^{64} \mathbf{1}\{e_i > 0.3\}$; the table reports its arithmetic mean across the three seeds. Recovery is summarized by the mean and sample SD across seeds. Both coefficient-refit and fixed-coefficient branches use evaluation-pair clean–corrupt activation differences. The fixed-coefficient branch is therefore not a fully frozen intervention, and this pointwise statistic is distinct from MCQA mean-margin recovery.

Table G.11: IOI reuse on the test split. Recovery is reported as the mean ± sample SD across three seeds. Fixed coefficients retain access to evaluation-pair directions. Failure fractions use the pointwise error and threshold specified above.

Method	Refit rec.	Fixed-β rec.	Refit fail frac.	Fixed-β fail frac.
Driver-OMP	0.461 ± 0.144	0.391 ± 0.111	0.583	0.661
Coverage-Driver	0.449 ± 0.083	0.298 ± 0.143	0.667	0.833
Top-k	0.632 ± 0.012	0.434 ± 0.160	0.464	0.734
Random	-0.670 ± 0.612	-0.649 ± 0.637	0.984	0.984

Table G.12: Sentiment paired fixed-dose recovery: seed mean and sample SD across eight seeds, by support budget and direction condition. Forward mean-margin recovery is distinct from formal vector reversal.

Selector	k	Mode	Mean	SD	Seeds
Driver-OMP	3	P	0.3602	0.1887	8
Driver-OMP	3	F	0.3536	0.1723	8
Coverage-Driver	3	P	0.7687	0.1038	8
Coverage-Driver	3	F	0.8137	0.4851	8
Top-k	3	P	0.7804	0.1173	8
Top-k	3	F	0.8127	0.4852	8

Table G.12 (continued)

Selector	k	Mode	Mean	SD	Seeds
Random	3	P	0.0745	0.1221	8
Random	3	F	0.0743	0.1072	8
Driver-OMP	5	P	0.4949	0.1142	8
Driver-OMP	5	F	0.4538	0.1586	8
Coverage-Driver	5	P	0.9542	0.0969	8
Coverage-Driver	5	F	0.8891	0.5155	8
Top-k	5	P	0.9328	0.1041	8
Top-k	5	F	0.8880	0.5157	8
Random	5	P	0.1065	0.1376	8
Random	5	F	0.0845	0.1030	8
Driver-OMP	8	P	0.5928	0.1053	8
Driver-OMP	8	F	0.5291	0.2137	8
Coverage-Driver	8	P	1.0311	0.1120	8
Coverage-Driver	8	F	0.9420	0.5354	8
Top-k	8	P	1.0325	0.1085	8
Top-k	8	F	0.9370	0.5262	8
Random	8	P	0.1532	0.1601	8
Random	8	F	0.1122	0.1246	8

Interpretation of reuse. High refit recovery demonstrates reuse of selected sites with evaluation-side coefficient refitting. High frozen forward recovery does not establish backward reversal, preservation, or feedback coverage. The results therefore retain continuous recovery and their information conditions instead of converting heterogeneous series into one cross-task pass flag.

The paired Sentiment series can show high mean forward recovery with training-mean directions, particularly for Coverage-Driver and Top-k, while formal F reversal fails for every method and seed. In the MCQA and Refusal feedback experiments, F directions remain fixed while dose adapts to observed target errors; fully frozen open-loop reuse also fixes the dose. No common cross-task recovery threshold or correction family is imposed on these different statistics.

The strong Sentiment F mean-margin recovery reported in the fixed-dose reuse table does not imply low formal target residuals: mean-margin recovery aggregates across prompts, whereas $\mathrm{Rel}_{\mathrm{tar}}$ measures the remaining error in the entire vector of held-out prompt-level readouts. Under the maximum support budget of eight, none of the three evaluated selectors satisfies $\varepsilon = 0.4$ in any of the eight Sentiment F seeds.

G.3 Comparison methods and sensitivity

G.3.1 Comparison methods

The nominal table compares both sparse selectors, Top-k, random supports, and higher-cost nonlinear greedy using the task-specific dictionaries and refits. Equal maximum support budgets do not imply equal realized support sizes after early stopping or equal model-evaluation cost. In the repeated scalar controls, Top-k ranks $|G_{\mathrm{train}}^{\top} q_{\mathrm{train}}|$, and all three selectors fit on the same fitting split.

Random and nonlinear-greedy references. Random supports are sampled from the task-specific candidate library and evaluated using the task-specific direction and refitting rules. Budget matching holds the maximum selection-cost or cardinality budget fixed; support-size matching holds the realized number of selected candidates fixed. The matching condition is specified for each experimental series. In nominal reconstruction, random supports contain $\min(k, |\mathcal{V}|)$ candidates at the prescribed cardinality budget k; the other selectors may stop earlier.

Nonlinear greedy provides a higher-cost reconstruction reference. At each iteration, it shortlists up to 32 remaining candidates by absolute correlation with the current local residual. Each trial support is refitted by ridge regression and evaluated in the nonlinear model; the candidate with the lowest R_{clean} is added. The procedure stops at its support budget or the implemented mean-absolute-error criterion. It is a greedy heuristic rather than a global optimization bound.

The Refusal single-site reference fixes answer head L15.H0. Its P and F feedback both give 10/40 hits, equal to the ten initial zero-dose hits; it adds no newly hit endpoints. Its formal reversal is 0/8

in each mode. At nominal dose, its mean preservation KL is 5.29×10^{-7} over 256 prompt–seed measurements: very small distributional change accompanies little target effect. This head-level proxy is not canonical residual-stream ActAdd (Turner et al., 2023), CAA (Rimsky et al., 2024), or RepE (Zou et al., 2023). Systematic comparisons with those steering families require matched readouts, information access, budgets, and intervention definitions.

G.3.2 Sensitivity

Formal reachability was separately reclassified at $\varepsilon \in \{0.30, 0.35, 0.40, 0.45, 0.50\}$, distinct from the absolute feedback-endpoint tolerances in the existing feedback sensitivity table. IOI Driver-OMP passes in 9/10 seeds at $\varepsilon = 0.30$ and 10/10 at $\varepsilon \geq 0.35$. On the held-out MCQA and paired Sentiment panels, all training-mean-direction (F) configurations remain at 0/8 seeds throughout this range. Refusal similarly fails in all P and F conditions even at $\varepsilon = 0.50$. For Coverage-Driver in P mode, Humanities passes in 8/8 seeds throughout the range, STEM reaches 8/8 at $\varepsilon \geq 0.35$, and Sentiment reaches 8/8 at $\varepsilon \geq 0.40$. Some individual classifications are sensitive: nominal module-level Humanities Top-k passes in 2/3 seeds at $\varepsilon = 0.40$ and 3/3 at $\varepsilon = 0.45$. Complete per-condition sensitivity counts are included in the accompanying reachability records.

The sensitivity analyses vary support budget, target-residual threshold, endpoint tolerance, signed ordering threshold, and preservation aggregation within each experimental series.

Sentiment's support-budget dependence is retained in Table G.12; signed dose tails are in Table G.7; mean versus all-prompt preservation rules are in Table G.9. Table G.13 reclassifies the same measured feedback endpoints at five absolute-error tolerances.

Table G.13: Feedback endpoint hits reclassified at five absolute-error tolerances. These are the same endpoints, not new runs or independent samples.

Task	Selector	Mode	0.10	0.20	0.30	0.40	0.50
Humanities	Coverage-Driver	F	1/48	2/48	8/48	9/48	10/48
Humanities	Coverage-Driver	P	11/48	24/48	46/48	46/48	46/48
Humanities	Driver-OMP	F	1/48	2/48	8/48	8/48	10/48
Humanities	Driver-OMP	P	6/48	17/48	38/48	38/48	38/48
Humanities	Top-k	F	1/48	2/48	8/48	10/48	10/48
Humanities	Top-k	P	10/48	29/48	45/48	45/48	45/48
STEM	Coverage-Driver	F	1/48	2/48	7/48	10/48	13/48
STEM	Coverage-Driver	P	19/48	39/48	47/48	48/48	48/48
STEM	Driver-OMP	F	1/48	2/48	8/48	9/48	11/48
STEM	Driver-OMP	P	15/48	29/48	39/48	39/48	39/48
STEM	Top-k	F	1/48	2/48	7/48	10/48	12/48
STEM	Top-k	P	21/48	35/48	47/48	47/48	47/48
Refusal	Coverage-Driver	F	21/40	37/40	40/40	40/40	40/40
Refusal	Coverage-Driver	P	21/40	37/40	40/40	40/40	40/40
Refusal	Driver-OMP	F	12/40	22/40	33/40	38/40	39/40
Refusal	Driver-OMP	P	12/40	22/40	32/40	38/40	39/40
Refusal	Top-k	F	22/40	39/40	40/40	40/40	40/40
Refusal	Top-k	P	21/40	39/40	40/40	40/40	40/40

At tolerance 0.50, the three Humanities F methods reach only 10/48 endpoints each, and the STEM F methods reach 11/48, 13/48, and 12/48 for Driver-OMP, Coverage-Driver, and Top-k, respectively. The MCQA P/F difference therefore persists across the reported 0.10–0.50 range. Refusal's 48 independent dose curves all have $\rho = 1$, so its positive-ordering counts remain unchanged for signed thresholds from 0.80 to 0.99. Strongly negative MCQA curves are not counted as positive responses at any of these thresholds.

G.4 Supporting analyses

G.4.1 IOI selection recurrence and backup augmentation

The recurrence analysis crosses five random seeds with five budgets $k \in \{3, 5, 8, 12, 16\}$, giving 25 configurations. Two late head–position candidates, L9.END.H9 and L10.END.H7, each appear in

15/25 configurations. The configured recurrence threshold is 0.6. These frequencies describe this seed–budget grid; they are not bootstrap probabilities or evidence that either site is necessary. Their locations can be compared with the manual IOI circuit of Wang et al. (2023).

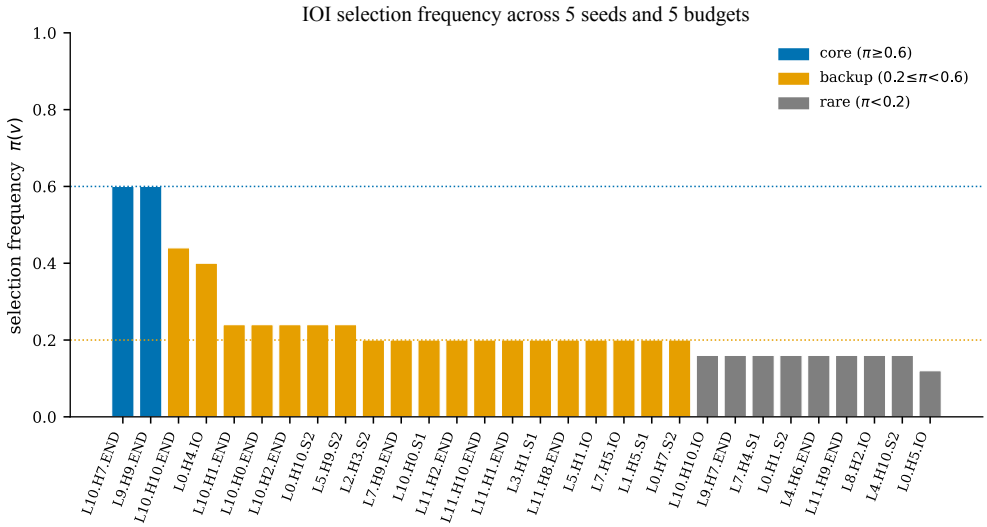


Figure G.1: IOI selection recurrence. Frequencies across five seeds and five budgets. Two candidates occur in 15/25 configurations; the labels "core", "backup", and "rare" denote recurrence bands in this analysis, not proofs of necessity or universal substitutability.

The separate backup experiment starts from L9.END.H9 and adds up to six candidates. Across 64 prompts, mean pointwise absolute error normalized by $\max(|b_{\text{clean}}|, 10^{-6})$ changes from 0.5894 with one site to 0.3503 with seven. The first addition worsens the mean to 0.698. At threshold 0.3, 26 prompts cross toward smaller error, six cross in the opposite direction, and 29/64 still fail at the final support. This statistic differs from stacked-vector R_{clean}, and the six-site random reference is not size-matched to the final seven-site support. The complete support-by-prompt record is retained with the experimental data.

G.4.2 Model scope

Table G.14 compares nominal reconstruction across three model configurations. All use four option logits and clean-output vector normalization, but Qwen3-4B uses eight questions per seed and three module types; Qwen3-8B uses four questions and residual-only candidates; Mistral uses four questions and three module types. They have three, five, and three seeds, respectively. Their differences cannot isolate the causal effect of model size or architecture.

Table G.14: Configuration-specific model comparisons. Mean ± sample SD of nominal R_{clean}. Panels and libraries differ as specified in the text; these are configuration-specific results.

Model	Seeds	Driver-OMP	Coverage-Driver	Top-k	Random
Qwen3-4B	3	0.045 ± 0.013	0.027 ± 0.031	0.054 ± 0.024	0.185 ± 0.132
Qwen3-8B	5	0.016 ± 0.008	0.018 ± 0.009	0.027 ± 0.012	0.077 ± 0.032
Mistral-7B	3	0.044 ± 0.005	0.015 ± 0.003	0.008 ± 0.006	0.050 ± 0.047

G.4.3 Refusal short continuations

The Refusal probe has eight seeds, ten harmful evaluation prompts per seed, and four conditions (unpatched plus three selectors), giving 320 continuations on twenty distinct prompt identities reused across seeds. All continuations have exactly sixteen greedy tokens, begin with <think>, and contain no </think> marker. The observed window therefore does not expose completed final answers. It cannot determine final semantic refusal or compliance rates. Patches use fixed training-mean directions at each current last valid position. This unpadded generation procedure is separate from the batched lexical-readout configuration in Appendix F.

G.5 Limitations

The experiments cover a finite set of models, task panels, readouts, and candidate libraries. Evaluation-pair directions require activation contrasts from paired evaluation inputs, so their strong MCQA results do not establish control without this information. Training-mean feedback still adapts dose at evaluation time, whereas fully frozen open-loop reuse fixes the entire intervention. The executed MCQA and Refusal plain-ridge fits impose neither an independent physical-energy bound nor an off-target penalty, and separate axis results do not establish the joint nonlinear certificate. Preservation is measured on 32 reused MMLU prompts; it quantifies predictive-distribution change rather than semantic correctness or safety. Refusal's lexical readout and sixteen-token continuations leave completed-answer behavior unresolved. Sentiment feedback controllability remains unassessed because the reported Sentiment series contain no multi-target feedback trajectories. Different model panels and candidate libraries preclude an isolated scaling comparison, while the single-site proxy does not replace matched implementations of established steering methods. Broader control and transfer claims require corresponding evaluations beyond these settings.

H Reproducibility and Statistical Reporting

Code and experimental records. Code and experimental records are publicly available at `https://github.com/zxzok/sparse-internal-control`. The repository contains the implementation, experiment configurations, evaluation inputs, saved interventions, task-specific readout vectors, prompt-level KL measurements, and analysis scripts for reproducing the reported tables and statistics. Experiment identities and sample configurations are indexed in Table G.2.

Models and execution environments. The experiments use GPT-2 small, Qwen3-1.7B, Qwen3-4B, and the additional models specified in Appendix G.4.2. MCQA nominal reconstruction and the documented model-scale runs used an Apple M2 Ultra system with the MPS backend. The IOI/Sentiment reversal experiments, Sentiment and MCQA preservation measurements, and independent STEM dose scan used NVIDIA RTX A6000 GPUs with CUDA. Series-specific execution profiles and available runtime metadata are documented in `experiments/execution_profiles.json`, `docs/compute_resources.md`, and the corresponding measurement records. Batched Refusal lexical scores use the ANSWER index $\sum_t \text{attention_mask}_{it} - 1$; this index coincides with the last valid input token under right padding. The short-generation probe uses unpadded inputs and the actual final valid token. Available model repository identifiers and checkpoint revisions are listed in `experiments/model_assets.json`. Software dependencies and execution profiles are documented in `requirements/model-*.lock.txt` and `experiments/execution_profiles.json`, respectively.

Statistical analysis. Nominal reconstruction, formal reachability, and held-out reuse report means and sample standard deviations across seeds. Feedback results use seed–target endpoints, whereas signed reversal and dose response use seed-level ratios and curves. Preservation statistics retain both prompt and seed identities. No common multiplicity correction is applied across the six control axes.