BEHAVIOR GRADIENT FIELDS: THE GEOMETRY AND STEERING OF INDUCING DATASETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding which datasets produce models with a desired behavior has direct consequences for robustness, safety, and the compatibility of learning objectives. For any target behavior, the set of datasets that induce that behavior through training is large and diverse, yet poorly understood; practitioners typically know only a few, generated by simple data-mixing procedures. This multiplicity is consequential: different inducing datasets can encode different spurious correlations, conflict with other objectives, or leave behavior fragile to small data or training changes. We introduce **behavior gradient fields**, which propagate behavior objectives back through training to model weight and data spaces, and **data steering maps**, practical algorithms that follow these fields to discover diverse inducing datasets. Empirically, we investigate spurious correlations in image classification and the steering of refusal and sycophancy in large language models. We find that: (i) dormant spurious correlations manifest as disjointness in behavior fields, and that data steering maps can find paths that avoid them; (ii) steering paths can be highly concentrated, with refusal sensitive to changes even in non-safety data; (iii) interactions between behavior fields predict unexpected side effects and enable principled multi-objective steering; and (iv) metagradient data steering is effective for behaviors —notably sycophancy— where activation-based methods fail.

1 INTRODUCTION

The *Rashomon Effect* —that many models can perform similarly for a given objective (Breiman, 2001; Rudin et al., 2024)— has been studied extensively to understand the consequences of model multiplicity. Less understood is the analogous phenomenon at the level of data: there exist many distinct *inducing datasets*, i.e., datasets that produce models exhibiting a target behavior. Our knowledge of this set is typically limited, and dataset construction often relies on manual curation and heuristic procedures (Liang et al., 2022). Understanding the geometry of this space is nonetheless desirable, and can help determine whether observed sensitivities and tradeoffs are particular to the datasets we have found or persist across alternatives.

There are far-reaching consequences to the choice of inducing dataset. In image classification, models trained on different mixtures of the same data sources can latch onto distinct spurious correlations and fail differently under distribution shift (Geirhos et al., 2020; D'Amour et al., 2022; Compton et al., 2023). In large language models, narrow or benign finetuning can produce unexpected side effects ranging from increased sycophancy to severe safety degradation (Qi et al., 2024; Betley et al., 2025; Blank et al., 2026), showing that inducing datasets for one behavior can intersect undesirably with those for a different one.

Exploring this space is fundamentally difficult. Inducing datasets live in a space of probability measures over (potentially high-dimensional) inputs, and reaching them requires propagating a behavior objective back through training to data—an operation that is expensive and only recently practical at scale (Engstrom et al., 2025). Existing data-centric methods largely sidestep the multiplicity question: attribution methods (Koh & Liang, 2017; Ilyas et al., 2022; Park et al., 2023) trace predictions to individual training examples given a fixed dataset, while data optimization methods (Ren et al., 2018; Xia et al., 2024; Alvarez-Melis & Fusi, 2021) converge to a single solution rather than exploring the space. Neither framework characterizes the geometry of inducing datasets nor relates it to downstream properties of their associated behavior.

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

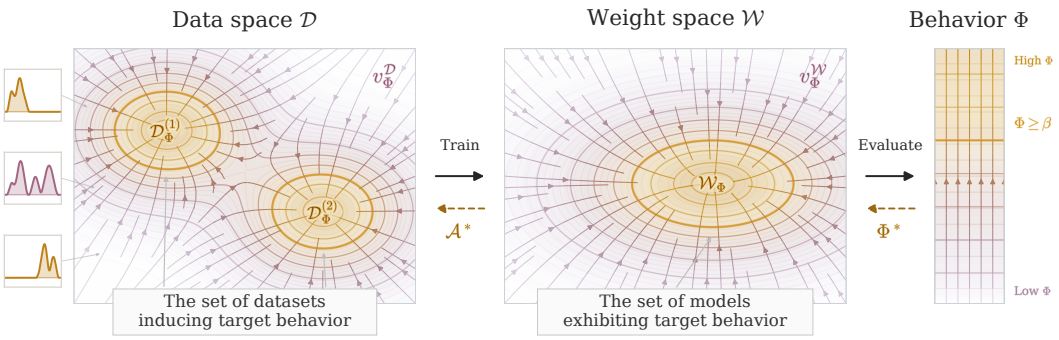


Figure 1: **Behavior gradient fields across data, weight, and behavior spaces.** Training maps datasets to model weights, and evaluation maps weights to behavior. Differentiating the behavior objective with respect to weights and data defines $v_\Phi^{\mathcal{W}}$ and $v_\Phi^{\mathcal{D}}$, with Φ^* and $\mathcal{A}^*$ denoting the corresponding pullback operators. Gold regions satisfy $\Phi \geq \beta$. Example data distributions are shown to the left. As visualized, disjoint data regions $\mathcal{D}_\Phi^{(1)}$ and $\mathcal{D}_\Phi^{(2)}$ may map to a connected $\mathcal{W}_\Phi$.

To address this, we take a geometric perspective on the space of inducing datasets. We introduce **behavior gradient fields**: vector fields that describe how a behavior objective propagates back through training to data and model weight spaces (Figure 1). The set of inducing datasets defines a level set of this field, and its local geometry captures behaviorally consequential structure, e.g., concentration, volume, interactions between fields, and connectivity of inducing regions. To populate and explore these sets in practice, we introduce **data steering maps**: algorithms that follow behavior gradient fields to discover diverse inducing datasets. Our empirical contributions using these maps span image classification and large language model post-training, where we show that:

- Data steering maps find paths between inducing datasets that avoid apparent *disjointness* arising from various types of spurious correlations. We find these paths also lie closer to the *behavior geodesic*, a consequence of how data steering shapes the geometry of prediction trajectories.
- The diversity of steering paths results in interesting and surprising *asymmetric* behavioral effects, e.g., we find sycophancy steering strongly impacts refusal, but not the other way around. Our steering maps also improve behaviors for which prior data interventions have shown limited success.
- Certain data steering paths are highly *concentrated*, meaning small data movements induce large behavioral changes. For refusal, we find that even non-safety sources such as scientific instruction provide concentrated steering signals, which has direct implications for safety robustness.
- Through looking at the *interactions* between behavior fields, we can better understand behavioral side effects and construct joint objectives that create previously unexplored tradeoffs.

2 Related Work

Data attribution. Training data attribution is a widely studied problem in data-centric ML, with the goal of tracing model predictions or behavior back to individual training examples. Influence functions (Koh & Liang, 2017), datamodels (Ilyas et al., 2022), TRAK (Park et al., 2023), and MAGIC (Ilyas & Engstrom, 2025) estimate the counterfactual effect of removing or reweighting examples on a fixed model or prediction. These methods are powerful tools for diagnosing data quality and debugging model behavior. However, they operate relative to a fixed dataset and a fixed training run, and do not characterize the broader space of datasets that could have produced models with the same or different behavior — the central question we tackle in this work.

Data optimization and reweighting. A growing body of work optimizes training data directly toward a target objective. Example reweighting (Ren et al., 2018), LESS (Xia et al., 2024), and metagradient descent (Engstrom et al., 2025) compute gradients through the training map to upweight examples that improve a downstream objective, while dataset dynamics (Alvarez-Melis & Fusi, 2021) poses dataset optimization as a dynamic optimal transport problem over probability

space. These methods are effective at finding a single good dataset for a target behavior, but treat data selection as an optimization problem with one solution. They do not explore the multiplicity of inducing datasets, characterize their geometry, or reason about interactions across objectives.

Spurious correlations and dataset composition. Spurious correlations are a critical failure mode in ML, arising when models latch onto features that are predictive in training but not at deployment (Geirhos et al., 2020; Sagawa et al., 2020). Prior work has measured their prevalence through feature reliance (Sagawa et al., 2020) and shown that dataset composition can introduce new associations not present in any source (Compton et al., 2023). ModelDiff (Shah et al., 2023) uses datamodels to compare what different models rely on. These works diagnose spurious correlations given a fixed training set but do not provide mechanisms to navigate the data space toward datasets that avoid them.

Steering language model behaviors. Steering the behavior of large language models is a critical safety and alignment priority (Anwar et al., 2024; Sharkey et al., 2025). Activation-based methods identify directions in representation space associated with target behaviors and intervene at inference time (Arditi et al., 2024; Wollschläger et al., 2025); probe-based data editing (Xiao & Aranguri, 2026) extends this idea to rank training examples by their alignment with a behavior direction. Post-training data optimization (Bergen et al., 2026) studies how data composition affects refusal and sycophancy. However, activation-based methods have shown limited success for diffuse behaviors such as sycophancy (Bergen et al., 2026; Blank et al., 2026), and metagradient-based data steering has not previously been demonstrated for LLM preference optimization. We address both gaps, and additionally use behavior field interactions to understand and jointly steer multiple behaviors.

3 METHOD

We model a target behavior as a scalar objective $\Phi : \mathcal{W} \to \mathbb{R}$ over model weights—for instance, refusal rate on harmful prompts, or out-of-distribution accuracy. A dataset $\mathcal{D}$ *induces* this behavior when the model trained on it satisfies $\Phi(\mathcal{A}(D)) \geq \beta$ for a target level β. Our framework studies the geometry of the set of all such datasets,

$$\mathcal{D}_\Phi = \{D \in \mathcal{D} : \Phi(\mathcal{A}(D)) \geq \beta\},$$

and its counterpart $\mathcal{W}_\Phi = \{\theta \in \mathcal{W} : \Phi(\theta) \geq \beta\}$ in weight space. Here $\mathcal{D}$ is the space of continuously parametrized datasets —concretely, nonnegative reweightings of a fixed training pool— and $\mathcal{A} : \mathcal{D} \to \mathcal{W}$ is a differentiable learning algorithm. To characterize $\mathcal{D}_\Phi$ and chart it in practice, we introduce two tools: **behavior gradient fields**, which describe how Φ propagates back through $\mathcal{A}$ to data and weight space, and **data steering maps**, algorithms that follow these fields to discover diverse inducing datasets.

Definition 1 (Behavior gradient field)**.** *For a differentiable behavior objective $\Phi : \mathcal{W} \to \mathbb{R}$, write $f_\Phi = \Phi \circ \mathcal{A}$. The behavior gradient fields associated with Φ are*

$$v_\Phi^{\mathcal{W}}(\theta) = \nabla_\theta \Phi(\theta), \qquad v_\Phi^{\mathcal{D}}(D) = \nabla_D f_\Phi(D).$$

Here, ∇ denotes the Euclidean gradient in the chosen parameterization. Note the differential of f_Φ is the pullback of $d\Phi$ through the learning map, $df_\Phi = \mathcal{A}^*(d\Phi)$ (Figure 1).

For a target behavior level β, we are interested in characterizing $\mathcal{D}_\Phi$ via data steering maps that use $v_\Phi^{\mathcal{D}}$ to guide changes to a given dataset.

Definition 2 (Data steering map)**.** *A data steering map $T_\Phi : \mathcal{D} \to \mathcal{D}$ is an update operator that approximates a finite step along the flow generated by $v_\Phi^{\mathcal{D}}$. Given an initial dataset D_0, successive updates $D_{r+1} = T_\Phi(D_r)$ define a steering trajectory through dataset space.*

Algorithms 1 and 2 instantiate T_Φ by representing a dataset as a vector of nonnegative weights $\mathbf{z} \in \mathbb{R}^n$ on the examples of a fixed pool, entering the training loss as multiplicative factors.

Remark 1. Weighting example i by z_i in the training loss is equivalent to sampling it with probability proportional to z_i; the two representations therefore define the same population objective.

Metagradient data steering (MG-DSM). Following Engstrom et al. (2025), Algorithm 1 differentiates a behavior proxy loss ℓ_Φ through the training map with respect to $\mathbf{z}$. The metagradient

$g_i = \nabla_{z_i} \ell_\Phi(\mathcal{A}(\mathbf{z}))$ gives the first-order effect of upweighting example i on the proxy loss; examples with small g_i promote the target behavior and receive additional weight. To explore multiplicity, we optionally include a separate objective Ψ with coefficient $\lambda \geq 0$, so that different seeds and objective weightings define diverse steering trajectories.

Algorithm 1 Metagradient data steering map (MG-DSM)

Input: base pool D, training algorithm $\mathcal{A}$, behavior loss ℓ_Φ, optional objective Ψ
Hyperparameters: rounds R, update fraction q, step η, coefficient $\lambda \geq 0$
$\mathbf{z} \leftarrow \mathbf{1}, \quad k \leftarrow \lfloor q|D|/2 \rfloor$
for $r = 1, \ldots, R$ **do**
  $\mathbf{g}_\Phi \leftarrow \nabla_{\mathbf{z}} \ell_\Phi(\mathcal{A}(\mathbf{z})), \quad \mathbf{g}_\Psi \leftarrow \nabla_{\mathbf{z}} \Psi(\mathbf{z})$
  $\mathbf{g} \leftarrow \mathbf{g}_\Phi + \lambda \mathbf{g}_\Psi$
  $I_- \leftarrow k$ samples with largest g_i and $z_i \geq \eta, \quad I_+ \leftarrow k$ remaining samples with smallest g_i
  $\mathbf{z}_{I_+} \leftarrow \mathbf{z}_{I_+} + \eta, \quad \mathbf{z}_{I_-} \leftarrow \mathbf{z}_{I_-} - \eta$
end for
Return: data weight vector $\mathbf{z}$

Activation data steering (A-DSM). Algorithm 2 scores training examples by their alignment with a behavior direction in activation space, extending the probe-based attribution of Xiao & Aranguri (2026) to our reweighting representation. At round r, let $\theta_r = \mathcal{A}(\mathbf{z})$ be the current model. We define the *behavior direction* $\mathbf{a}_\Phi$ as the mean difference in response-token activations at layer ℓ between preferred and rejected responses for behavior proxy pairs, and the *example alignment* $\psi(d_i)$ as the analogous activation difference for training pair d_i. Examples whose alignment $\cos(\psi(d_i), \mathbf{a}_\Phi)$ is high are upweighted; the weight transfer rule follows Algorithm 1.

Algorithm 2 Activation data steering map (A-DSM)

Input: base pool D, training algorithm $\mathcal{A}$, layer ℓ, target objective Φ, optional objective Ψ
Hyperparameters: rounds R, update fraction q, step η, coefficient $\lambda \geq 0$
$\mathbf{z} \leftarrow \mathbf{1}$
for $r = 1, \ldots, R$ **do**
  Compute $\mathbf{a}_\Phi$, $\mathbf{a}_\Psi$, and $\psi(d_i)$ under θ_r at layer ℓ
  $s_{\Phi,i} \leftarrow \cos\big(\psi_\Phi(d_i), \mathbf{a}_\Phi\big), \quad s_{\Psi,i} \leftarrow \cos\big(\psi_\Psi(d_i), \mathbf{a}_\Psi\big) \quad$ for each $d_i \in D$
  $s_i \leftarrow s_{\Phi,i} + \lambda s_{\Psi,i}$
  Update $\mathbf{z}$ using Algorithm 1's weight transfer with $g_i = -s_i$
end for
Return: $\mathbf{z}$

Concentration analysis. To measure how strongly a steering update propagates across spaces, we define empirical Lipschitz ratios along a trajectory $D_0, \ldots, D_R$ with $\theta_r = \mathcal{A}(D_r)$:

$$\widehat{\kappa}_{\mathcal{D}\to\mathcal{W}} = \max_{0\leq r<R} \frac{d_{\mathcal{W}}(\theta_{r+1}, \theta_r)}{d_{\mathcal{D}}(D_{r+1}, D_r)}, \qquad \widehat{\kappa}_{\mathcal{W}\to\Phi} = \max_{0\leq r<R} \frac{|\Phi(\theta_{r+1}) - \Phi(\theta_r)|}{d_{\mathcal{W}}(\theta_{r+1}, \theta_r)}.$$

We define $\widehat{\kappa}_{\mathcal{D}\to\Phi}$ analogously, replacing the numerator of $\widehat{\kappa}_{\mathcal{D}\to\mathcal{W}}$ with $|\Phi(\theta_{r+1}) - \Phi(\theta_r)|$ to represent data-to-behavior sensitivity. These ratios are useful diagnostics in our experiments: a large $\widehat{\kappa}_{\mathcal{D}\to\Phi}$ indicates concentration, where a small data change can produce a large behavioral movement. A large $\widehat{\kappa}_{\mathcal{D}\to\mathcal{W}}$ together with a small $\widehat{\kappa}_{\mathcal{W}\to\Phi}$ indicates weight movement that is weakly expressed in the target objective, and may thus reflect learned behaviors beyond Φ.

4 Experiments

4.1 Dormant Spurious Correlations

A common data-mixing strategy is to linearly interpolate between different data sources. Counterintuitively, adding more data in this manner does not necessarily improve model performance, and can even harm generalization relative to training on the original dataset alone (Compton et al., 2023).

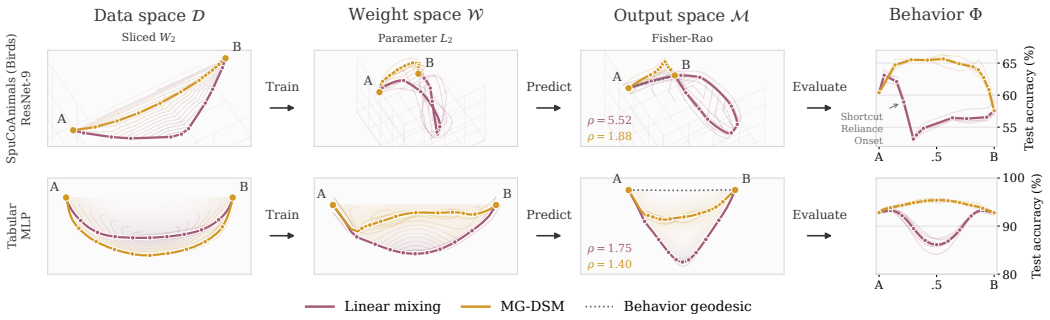


Figure 2: **Linear data mixing "awakens" an undesirable dormant association, while MG-DSM maintains a high-performing path in behavior space.** *Top:* Dormancy through class frequency in SpuCoBirds with ResNet-9 ($\lambda = 64$). *Bottom:* Dormancy through measurement noise in the tabular MLP setting ($\lambda = 2$). Bold paths (MDS) are five-seed averages. Implications of MG-DSM being closer to the behavior geodesic (i.e., small path length ratio ρ) are explained in Section 4.1.3.

One reason why is that mixing sources can create what we term **dormant spurious correlations**. For instance, a source indicator $s \in \{A, B\}$, constant within each dataset, becomes predictive after mixing when $P(y = 1 \mid s = A) \neq P(y = 1 \mid s = B)$. This "awakens" an association that can hurt generalization if it changes at deployment. We show that this phenomenon results in *disjointness* in behavior fields, and that metagradient data steering can find paths that avoid it.

To do so, we instantiate Algorithm 1 to construct a path from a given dataset D_A toward a destination dataset D_B, with validation loss, destination objective, and optimization problem

$$\ell_\Phi(\theta) = \mathrm{CE}_{\mathrm{val}}(\theta), \quad \Psi(D) = \frac{d(D, D_B)^2}{d(D_A, D_B)^2}, \quad D^\star_\lambda \in \arg\min_{D \in \mathcal{D}} \big[\ell_\Phi(\mathcal{A}(D)) + \lambda \Psi(D)\big].$$

Here $d(D, D') = \|\mathbf{z}(D) - \mathbf{z}(D')\|_2$ is the Euclidean distance between the sample weights representing the two datasets, and $\lambda > 0$ balances progress toward D_B against validation loss. We instantiate this formulation in two synthetic settings below.

4.1.1 Dormancy Through Class Frequencies

Our first setting considers how mixing datasets with different class and feature frequencies can introduce a spurious correlation between them. This mirrors documented failure modes in medical imaging, where hospital-specific artifacts become spuriously predictive after combining sources (Compton et al., 2023). We illustrate this using SpuCoAnimals (Joshi et al., 2025), derived from ImageNet (Deng et al., 2009), which contains binary classification tasks for birds and dog breeds. The background (land vs. water and indoor vs. outdoor) becomes spuriously predictive: within each source it is constant, but after mixing it correlates with class frequency. We train ResNet-9 models along linear and MG-DSM paths and evaluate on the balanced test set. Full details are in Appendix A.3. Figure 2 (top) shows that linear mixing produces an early accuracy dip as the model latches onto background; MG-DSM, in contrast, avoids this and traces a high-accuracy path between the two disjoint inducing regions, resulting in an over 11 pp average advantage at the midpoint.

4.1.2 Dormancy Through Noise

Unlike class-frequency, some spurious correlations cannot be removed by balancing classes before mixing. We next consider one such case, where the association arises from measurement noise.

Let $Y \in \{-1, +1\}$ be the uniformly distributed class label, let $E \sim \mathrm{Bernoulli}(t)$ indicate source D_A ($E = 0$) or D_B ($E = 1$), and let $\varepsilon \sim \mathcal{N}(0, 5)$ be the training measurement noise. Define $X_1 = Y + E\varepsilon$ and $X_2 = Y + (1 - E)\varepsilon$; the model receives input $X = (X_1, X_2, X_3)$, where $X_3 \in \{-1, +1\}$ matches $\mathrm{sign}(\varepsilon)$ with probability 0.9. This mirrors a setting where two hospitals use different equipment: each source provides one reliable and one noisy measurement of the same signal, and X_3 (noisily) encodes which instrument to trust. At both endpoints, neither X_3 nor

the noisier measurement is needed for accurate prediction. After mixing, however, X_3 becomes spuriously useful, as it encodes source identity through its association with the noise term. At test time the association breaks: X_3 is sampled independently of ε, whose variance drops to 1. As measurements become *more* reliable, models that relied on X_3 perform *worse*. Using two-layer MLPs as predictors (full setup provided in Appendix A.2), MG-DSM improves test accuracy by 8.40 pp on average over linear mixing at the midpoint (Figure 2, bottom).

4.1.3 Seeking The Behavior Geodesic

A path in model weight space parametrizes a corresponding path in output distribution space, $\mathcal{M}$. Studying this space lets us examine how predictions change along a data steering trajectory, beyond the aggregate changes captured by scalar behaviors Φ and without the ambiguities due to permutation symmetries in weight space $\mathcal{W}$. The study of *predictive multiplicity* highlights the wide variation in predictions among similarly performing models (Marx et al., 2020; Watson-Daniels et al., 2023). Here, we study how predictions vary along the path of induced models, using the geometry of output space to understand these transitions relative to what we term the **behavior geodesic**.

For test inputs $\{x_n\}_{n=1}^N$, let $p_t = (p_{\theta_t}(\hat{y} \mid x_n))_{n=1}^N$ be the predictions at path state $t \in \{0, \dots, T\}$. Following Tan et al. (2023), we measure paths through these predictive distributions with the Fisher–Rao metric, averaged over test inputs, and denote the resulting distance d_{FR}. The shortest such path between p_0 and p_T is the **behavior geodesic** $\gamma^\star$. We summarize the behavioral detour of a realized path with the *path length ratio* $\rho = \frac{\sum_{t=0}^{T-1} d_{\mathrm{FR}}(p_t, p_{t+1})}{d_{\mathrm{FR}}(p_0, p_T)} \geq 1$, where $\rho = 1$ for a path following $\gamma^\star$.

Implications. Across both experimental settings, we find that the MG-DSM path lies closer to the behavior geodesic than linear mixing, i.e., its path length ratio is smaller (Figure 2). This can be understood through three observations: (1) by construction, the spurious association is dormant at the endpoints; (2) when mixing, it enables an optimization "shortcut": a simple solution that differs substantially from robust generalization (Geirhos et al., 2020); (3) activating this shortcut introduces a detour in prediction space. Avoiding it, even implicitly, shortens the path. Although MG-DSM does not optimize ρ directly, the observed reduction is consistent with implicit regularization.

Takeaways. **(i)** Dormant spurious correlations can manifest as disjointness in behavior fields (when restricted to simple data-mixing procedures); MG-DSM finds paths that avoid them, producing substantial accuracy improvements over linear mixing in image and tabular settings. **(ii)** Incorporating a destination objective into MG-DSM steers toward a target distribution while preserving performance along the path. **(iii)** MG-DSM paths are closer to the behavior geodesic than linear mixing paths, consistent with implicit regularization of prediction trajectories.

4.2 Steering Behaviors in Large Language Models

We instantiate Algorithms 1 and 2 for LLM post-training, targeting refusal and sycophancy across three models. We study the resulting inducing data through the geometry of behavior fields, examining concentration, source sensitivity, and interactions between objectives.

Table 1: **Data steering maps can improve both refusal and sycophancy while producing asymmetric side effects.** Behavioral changes after five rounds of steering, in percentage points (mean $\pm$ SD across four seeds). MG-DSM achieves markedly larger sycophancy reductions than A-DSM, with a strong incidental increase in refusal. Bold entries mark the steered behavior and arrows indicate improvement. Full result tables in Appendix B.

		MG-DSM		A-DSM	
Steered behavior	**Model**	Refusal ↑	Sycophancy ↓	Refusal ↑	Sycophancy ↓
Refusal	Gemma-2 2B	$\mathbf{+14.4 \pm 0.8}$	-2.5 ± 0.5	$\mathbf{+6.7 \pm 1.1}$	$+0.1 \pm 0.4$
	OLMo-2 1B	$\mathbf{+6.1 \pm 0.6}$	-2.8 ± 0.1	$\mathbf{+3.3 \pm 1.0}$	-0.4 ± 0.6
	Qwen2.5 1.5B	$\mathbf{+7.9 \pm 1.9}$	-1.2 ± 0.9	$\mathbf{+11.0 \pm 0.4}$	-0.5 ± 0.5
Sycophancy	Gemma-2 2B	$+9.9 \pm 3.0$	$\mathbf{-12.4 \pm 3.4}$	-2.6 ± 2.4	$\mathbf{-0.7 \pm 0.3}$

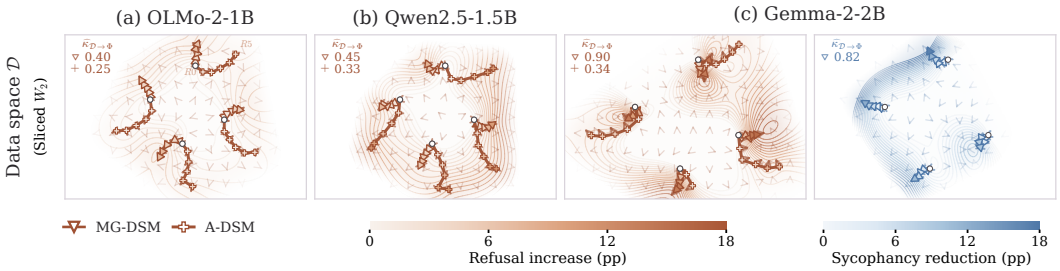


Figure 3: **Behavior gradient fields illustrate regions of greater behavioral sensitivity among inducing datasets.** Five-round trajectories of MG-DSM (triangles) and A-DSM (pluses); white dots mark initial datasets. Data distances use sliced W_2 over MPNet embeddings, projected with MDS. Contours interpolate observed checkpoints. $\widehat{\kappa}_{\mathcal{D}\to\Phi}$ measures data-to-behavior sensitivity ($\mathrm{pp}/(10^{-5}\,\mathrm{SW}_2)$), averaged across seeds, and corresponds to tightness of contours.

4.2.1 Steering Refusal and Sycophancy

Refusal of harmful requests is a central safety objective for large language models. A model could learn many different cues to improve refusal, such as recognizing sensitive topics or becoming averse to specific "jailbreaking" prompts. Alternatively, it could learn broader resistance to user pressure, with consequences for *sycophancy*, the tendency to adapt responses to match a user's views; prior work shows that steering model activations toward sycophancy at inference time can weaken refusal (Li et al., 2026). Previous activation-based data interventions have been shown to improve refusal, but have shown limited success in reducing sycophancy (Xiao & Aranguri, 2026; Bergen et al., 2026; Blank et al., 2026). Prior metagradient work operates on supervised data with fixed reference answers (Engstrom et al., 2025); we instead steer through *preference optimization*, which requires differentiable proxies for refusal and sycophancy.

Experimental setup. We instantiate Algorithms 1 and 2 for harmful refusal (Φ_r) and sycophancy (Φ_s), evaluating refusal on four safety benchmarks (Appendix B) and sycophancy on the three tasks of Perez et al. (2023), with questions spanning NLP, philosophy, and politics. We develop and introduce a behavior proxy, using 128 prompts per behavior, drawn from WildJailbreak (Jiang et al., 2024) for refusal and the tasks of Perez et al. (2023) for sycophancy. For each prompt x_j, we pair a preferred response r_j^+ with a rejected response r_j^- generated by the model being steered, and define

$$\ell_\Phi(\theta) = -\tfrac{1}{128}\sum_{j=1}^{128} \log\sigma\big(0.1\,\big[\log p_\theta(r_j^+ \mid x_j) - \log p_\theta(r_j^- \mid x_j)\big]\big)\,.$$

Algorithm 1 differentiates this loss through training to update the Dolci example weights. Algorithm 2 uses the same response pairs to construct a behavior direction from their activation differences. The proxy remains fixed across rounds, and its prompts do not overlap with Dolci or the evaluation sets. We compare the algorithms for both refusal and sycophancy in Gemma-2 2B, and additionally study refusal in Qwen2.5 1.5B and OLMo-2 1B. Each run uses 10,240 Dolci preference pairs and rank-8 LoRA; the seed controls both the LoRA initialization and the Dolci subsample. Across four seeds, we perform five steering rounds with $q = 0.15$ and $\eta = 0.25$.

General steering results. We summarize our findings in Table 1. We find that both methods improve refusal across all three models, with MG-DSM producing larger gains in Gemma and OLMo (14.4 and 6.1 pp), and A-DSM performing better in Qwen (11.0 pp). However, A-DSM produces little improvement in sycophancy, reducing it in Gemma by only 0.7 pp—by contrast, MG-DSM reduces sycophancy by **12.4** pp. Interestingly, we find that the incidental side effects can be *asymmetric*. Steering sycophancy with MG-DSM also increases refusal by 9.9 percentage points, whereas steering refusal reduces sycophancy by only 2.5. We build on this asymmetry in Section 4.2.3.

Steering paths can be highly concentrated. In Figure 3, we visualize the approximated behavior gradient fields, which is particularly informative for understanding *concentration*, i.e., sensitivity

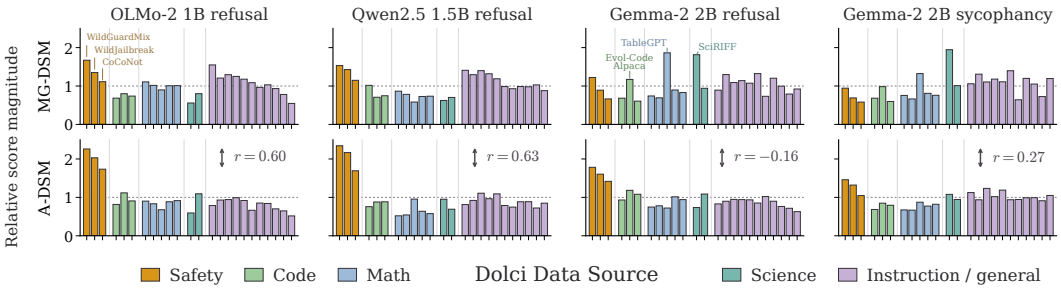


Figure 4: **Steering maps assign different scores to Dolci data sources; non-safety sources can induce relevant safety steering signal.** Mean metagradient magnitude $|g_{\Phi,i}|$ for MG-DSM (top) and activation score magnitude $|s_{\Phi,i}|$ for A-DSM (bottom), grouped by Dolci source and normalized by the corresponding mean across examples. Results are averaged across updates and seeds. Annotations show their Pearson correlation (i.e., cosine similarity between mean-centered vectors).

of behavior to small data changes. This visually corresponds to how steep the contours appear. A-DSM has uniformly lower concentration in the refusal experiments ($\widehat{\kappa}_{\mathcal{D}\to\Phi_r} = 0.25$–$0.34$, versus 0.40–0.90 for MG-DSM) and follows longer data paths, despite smaller refusal gains in Gemma and OLMo. MG-DSM is more "efficient" in this sense, finding smaller reweightings that can move behavior farther. Note that both methods use the same coordinate reweighting step size, while the distance used accounts for similarity of training examples in embedding space. Concentration also differs across models. Under our matched training settings, MG-DSM produces nearly twofold the refusal gain in Gemma than in Qwen (14.4 versus 7.9 pp), along a shorter data path. This difference is not simply a headroom effect, as Qwen starts with lower refusal (60.6% versus 62.3%). Such sensitivity has implications for robustness, as relatively small, directed changes to training data can substantially alter refusal. To understand which data contribute to this sensitivity, we next examine steering scores across sources.

4.2.2 What data defines steering directions?

What data steers behavior depends on both model and method (Figure 4). For refusal in Qwen and OLMo, both methods prioritize safety sources, which is expected given that safety data directly targets the refusal proxy. Surprisingly, Gemma departs from this pattern as we discuss next.

Steering signals extend beyond safety data. For Gemma, MG-DSM assigns the largest metagradient magnitudes to TableGPT (Li et al., 2024) and SciRIFF (Wadden et al., 2025), at 1.86 and 1.82 times the pool average, both above any safety source. These sources involve interpreting numeric tables, extracting evidence from scientific documents, and assessing the validity of claims about a given text. In contrast, A-DSM upweights safety sources, yet produces a smaller refusal gain (6.7 versus 14.4 pp). Combined with the concentration results in Figure 3, this indicates that relatively small, directed changes to non-safety data can significantly affect refusal, with implications for robustness that depend on how likely such changes arise under a realistic data-selection distribution.

Interpreting incidental effects. These data weighting patterns also help contextualize incidental effects. For Gemma sycophancy, A-DSM again strongly weights safety sources, but achieves little improvement. MG-DSM instead weights SciRIFF heavily, which also receives large scores when steering refusal. TableGPT and WildGuardMix (Han et al., 2024) are more prominent for refusal than for sycophancy, however. This partial overlap suggests that some sources affect both behaviors, but to different degrees, providing a plausible explanation for the observed asymmetric side effects.

4.2.3 Interactions between Behavior Fields

Asymmetric side effects in Table 1 motivate asking whether different steering paths favor compatible behavioral changes. Whether they do depends on the relation between their behavior fields. We examine this interaction for refusal and sycophancy, then test a joint objective that exploits it.

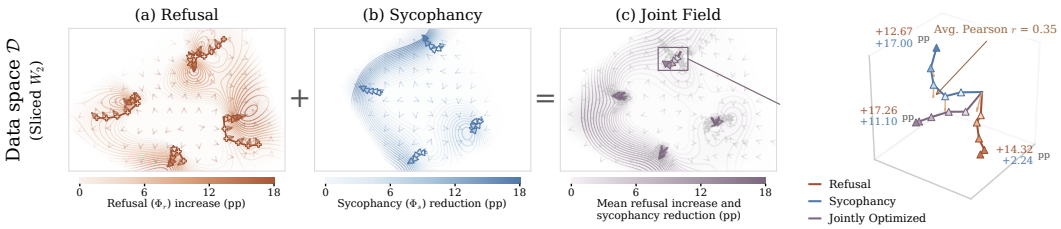


Figure 5: **Behavior field interactions help explain incidental effects and guide joint steering.** Panels (a)–(c) show refusal, sycophancy, and their joint field for Gemma-2 2B; joint contours average the two individual fields. The 3D MDS inset compares the three trajectories from seed 3, with joint updates combining the two objectives using $\lambda = 1$. Light orange arrows show projected Φ_r along the blue sycophancy path; their mean Pearson correlation with the Φ_s metagradients is 0.35.

Interactions between inducing data. In Figure 5, we visualize the Gemma refusal and sycophancy behavior fields (and their intersection) in data space. From Table 1, we know that steering refusal with MG-DSM produces a modest sycophancy reduction of 2.5 pp on average, whereas A-DSM produces essentially none. On the other hand, steering sycophancy with MG-DSM significantly improves refusal (+9.9 pp). To understand this asymmetric effect better, we compute both Φ_r and Φ_s metagradients before each update along the sycophancy trajectory; their mean Pearson correlation (0.35) confirms partial alignment between the two fields.

Multi-objective steering. We test whether a joint steering objective achieves a better trade-off than individual ones. We set $\Psi(\mathbf{z}) = \ell_{\Phi_s}(\mathcal{A}(\mathbf{z}))$ and $\lambda = 1$ in Algo. 1, thus equally weighting both behavior objectives. Joint steering increases refusal by 17.26 percentage points and reduces sycophancy by 11.10 after five rounds. Refusal steering alone produces gains of 14.32 and 2.24 points, respectively, while sycophancy steering produces 12.67 and 17.00. Joint steering thus achieves the largest refusal gain among the three paths while retaining considerable sycophancy reduction.

Takeaways. (i) With our proposed behavior proxies, MG-DSM improves both refusal and sycophancy, while A-DSM mainly improves refusal. **(ii)** Steering paths can be highly concentrated: in Gemma, non-safety sources provide *larger* refusal steering signal than explicit safety data. **(iii)** Distinct data steering paths can improve the same behavior but produce asymmetric side effects. **(iv)** Behavior field interactions provide insight into these effects and motivate joint objectives.

5 Conclusion

The geometry of inducing datasets reveals structure in data multiplicity that determines its behavioral consequences — which correlations become active, which behaviors interfere, how far a mixing path departs from the behavior geodesic. The behavior gradient field framework we propose provides the language to describe this structure, while the data steering maps make it tractable to explore. In particular, we demonstrate metagradient data steering for behaviors that activation-based methods have not reliably improved, suggesting a gap between identifying a latent behavioral direction and finding training data that robustly induces it.

Two extensions are natural: (i) replacing Euclidean data coordinates with Wasserstein gradient flows (Alvarez-Melis & Fusi, 2021), which would make the behavior geodesic computable over a richer class of paths, and (ii) broadening the data representation to include synthetic examples (Thrush et al., 2026).

The most pressing broader application, however, is toward training data that produces models which are both *robust and aligned*. Three findings from our experiments bear directly on this: why narrow training interventions can incidentally affect non-targeted behaviors; why data unrelated to safety can dominate safety-relevant steering signals; and why behavioral side effects are asymmetric in direction and magnitude. Answering them requires reasoning about inducing dataset geometry rather than about individual datasets.

AI USE STATEMENT

We used large language models for standard research assistance, e.g., debugging code, iterating ideas, situating contributions, and getting feedback on paper structure. Any edits were done so under author supervision and intent, and we take full responsibility for the sensibility of prose and ideas.

ETHICS STATEMENT

The broader objective of our work is to support the alignment of safe and robust models through a data-centric lens. We believe this perspective is particularly useful for *understanding* the compatibility of different safety interventions, whether gradient or activation-based. Although our methods could be used to weaken safeguards, less expensive ways to jailbreak an LLM already exist. We therefore do not believe our work substantially changes the risk profile of open-source models.

REPRODUCIBILITY STATEMENT

We provide the following details of our methods and experiments to support reproducibility: (i) Section 3 and Algorithms 1 and 2 describe our data steering procedures; (ii) Appendix A details the datasets, training settings, behavior proxies, and hyperparameters; (iii) Appendix B describes the evaluation benchmarks and scoring protocols. Code will be open-sourced alongside the paper.

REFERENCES

David Alvarez-Melis and Nicolò Fusi. Dataset dynamics via gradient flows in probability space. In *International Conference on Machine Learning*, 2021.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip H. S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024.

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*, 2024.

Leon Bergen, Usha Bhalla, Sidharth Baskaran, Max Loeffler, Raphael Sarfati, Dhruvil Gala, Ryan Panwar, Santiago Aranguri, Thomas Fel, Atticus Geiger, Matthew Kowal, Siddharth Boppana, Daniel Balsam, Owen Lewis, Jack Merullo, Thomas McGrath, and Ekdeep Singh Lubana. Anatomy of post-training: Using interpretability to characterize data and shape the learning signal. *arXiv preprint arXiv:2606.12360*, 2026.

Jan Betley, Daniel Chee Hian Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *International Conference on Machine Learning*, 2025.

Camila Blank, Zhuofan Ying, Christopher Potts, Peter Hase, and Jing Huang. Sycophantic agreement transfers with neutral data via contrastive preference optimization. *arXiv preprint arXiv:2608.31079*, 2026.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

Rhys Compton, Lily Zhang, Aahlad Puli, and Rajesh Ranganath. When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations. In *Machine Learning for Healthcare Conference*, 2023.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL `https://www.jmlr.org/papers/v23/20-1335.html`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Logan Engstrom, Andrew Ilyas, Benjamin Chen, Axel Feldmann, William Moses, and Aleksander Madry. Optimizing ML training with metagradient descent. *arXiv preprint arXiv:2503.13751*, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *Advances in Neural Information Processing Systems*, 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.

Andrew Ilyas and Logan Engstrom. MAGIC: Near-optimal data attribution for deep learning. *arXiv preprint arXiv:2504.16430*, 2025.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In *International Conference on Machine Learning*, 2022.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. WildTeaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In *Advances in Neural Information Processing Systems*, 2024.

Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Challenges and opportunities in improving worst-group generalization in presence of spurious features. *arXiv preprint arXiv:2306.11957*, 2025.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *ACM SIGOPS Symposium on Operating Systems Principles*, 2023.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. Table-GPT: Table fine-tuned GPT for diverse table tasks. *Proceedings of the ACM on Management of Data*, 2(3), 2024.

Yuxiao Li, Alina Fastowski, Efstratios Zaradoukas, Bardh Prenkaj, and Gjergji Kasneci. Analysing the safety pitfalls of steering vectors. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L. Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, 4(8):669–677, 2022. doi: 10.1038/s42256-022-00516-1.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, 2020.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning*, 2024.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *International Conference on Machine Learning*, 2023.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, 2024.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, 2018.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024.

Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Position: Amazing things come from having many good models. In *International Conference on Machine Learning*, 2024.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, 2020.

Harshay Shah, Sung Min Park, Andrew Ilyas, and Aleksander Madry. ModelDiff: A framework for comparing learning algorithms. In *International Conference on Machine Learning*, 2023.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "Do Anything Now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *ACM SIGSAC Conference on Computer and Communications Security*, 2024.

Charlie Tan, Theodore Long, Sarah Zhao, and Rudolf Laine. Geodesic mode connectivity. In *ICLR Tiny Papers*, 2023.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025.

Tristan Thrush, Sung Min Park, Herman Brunborg, Luke Bailey, Marcel Roed, Neil Band, Christopher Potts, and Tatsunori Hashimoto. Synthetic data for any differentiable target. In *Conference on Language Modeling*, 2026.

David Wadden, Kejian Shi, Jacob Morrison, Alan Li, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, Doug Downey, Hannaneh Hajishirzi, and Arman Cohan. SciRIFF: A resource to enhance language model instruction-following over scientific literature. In *Conference on Empirical Methods in Natural Language Processing*, 2025.

Jamelle Watson-Daniels, David C. Parkes, and Berk Ustun. Predictive multiplicity in probabilistic classification. In *AAAI Conference on Artificial Intelligence*, 2023.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *International Conference on Machine Learning*, 2025.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning*, 2024.

Frank Xiao and Santiago Aranguri. Probe-based data attribution: Discovering and mitigating undesirable behaviors in LLM post-training. *arXiv preprint arXiv:2602.11079*, 2026.

A Experimental Details

A.1 Steering Language Models

Models and training. For analyzing steering map efficacy, we work with the following models: `OLMo-2-0425-1B-DPO` (Team OLMo et al., 2025), `Qwen2.5-1.5B-Instruct` (Qwen Team et al., 2024), and `Gemma-2-2B-it` (Gemma Team et al., 2024). Each seed we run uses 10,240 preference pairs sampled from Dolci-Instruct-DPO (Team Olmo et al., 2025). All example weights start at $z_i = 1$. We train rank-8 LoRA adapters (Hu et al., 2022) with the DPO objective (Rafailov et al., 2023).

Table 2: Hyperparameters for the LLM experiments.

Setting	Value
LoRA	Rank 8, $\alpha = 16$, dropout 0; attention and MLP projections
Training	One epoch; batch size 128
DPO β	0.1
Optimizer	Adam, $(\beta_1, \beta_2) = (0.9, 0.999)$; $\epsilon = 10^{-6}$ inside the square root
Learning rate	Linear warmup (25%) to 1.5×10^{-4}, then linear decay to 1.5×10^{-5}
Token limits	1,024 sequence; 512 prompt
Steering	5 rounds; 15% of data weights changed by ± 0.25 per round
Seeds	4 (0–3)

At each round, MG-DSM adds 0.25 to up to 7.5% of data weights with the lowest metagradients and subtracts 0.25 from the same number with the highest metagradients.

Behavior Proxy. One of our empirical contributions is defining a working behavior proxy for refusal and sycophancy. For this, we pull from harmful WildJailbreak prompts (Jiang et al., 2024) and the NLP, philosophy, and political sycophancy tasks of Perez et al. (2023). For refusal, we sample eight responses to each of 576 candidate prompts and retain the first 128 preferred–rejected pairs found. For sycophancy, we begin with 2,000 candidate prompts to obtain 128 reviewed pairs that are balanced across the subsets: 42 NLP, 43 philosophy, and 43 politics. Each pair contains responses generated by the model being steered, and the selected bank remains fixed across seeds and rounds.

A.2 TABULAR EXPERIMENT

We independently sample $2{,}048$ training, $2{,}048$ validation, and $20{,}000$ test examples from the noise reference construction in Section 4.1.2, drawing labels and sources independently with equal probability. To reiterate, training noise has variance 5, and X_3 agrees with its sign with probability 0.9. Validation and test noise have variance 1, with X_3 an independent uniform sign. Thus both noise variance and reference reliability change. Endpoint weights select the corresponding source and give each class total weight $N/2$, where $N = 2{,}048$. Validation cross entropy and test accuracy are averaged equally across classes. Note that linear interpolation mixes endpoint weights as $z(t) = (1-t)z_A + tz_B$.

Following Engstrom et al. (2025), we seek training routines whose metagradients reliably predict the effect of small data-weight changes, a property they term *metasmoothness*. This motivates choices such as GELU rather than ReLU. Table 3 summarizes our model, training, and steering settings for this experiment.

Table 3: Tabular model, training, and data steering settings.

Setting	**Value**
Model and training	
Input preprocessing	Standardization using training pool statistics
Architecture	MLP with two hidden layers of 32 GELU units
Parameters	$1{,}250$
Initialization	Xavier weights and zero biases
Optimizer	Full-batch Nesterov SGD
Training steps	100
Learning rate	0.02
Momentum	0.9
Weight decay	10^{-5}
Seeds	Five (vary sampled data and initialization)
Data steering	
Objective	Class-balanced validation cross entropy plus $\lambda\Psi(z)$
Destination penalty	$\Psi(z) = \|z - z_B\|_2^2 / \|z_A - z_B\|_2^2$
Destination coefficient	$\lambda = 2$
Coordinate fraction	$q = 0.30$ (at most 30% per class)
Coordinate step	$\eta = 0.0625$

A.3 SPUCOANIMALS EXPERIMENT

SpuCoAnimals (Joshi et al., 2025) is curated from ImageNet (Deng et al., 2009) to study spurious associations between animal classes and image backgrounds. We use its two binary subsets: SpuCoBirds distinguishes landbirds from waterbirds with land or water backgrounds, while SpuCoDogs distinguishes small from big dog breeds with indoor or outdoor backgrounds. For each subset, D_A and D_B select the two background sources, with class-balanced endpoint weights. Results are shown in Figure 2 (top) for birds and Figure 6 for dogs.

To instantiate metagradient data steering, we seek metasmooth training dynamics, as in Appendix A.2. Following Engstrom et al. (2025), we use ResNet-9 with GELU activations. Table 4 summarizes the model, training, and steering settings.

Table 4: SpuCoAnimals model, training, and data steering settings. Validation loss equally weights the four label–background groups.

Setting	**Value**
Model and training	
Input preprocessing	Resize to 146×146, center crop to 128×128, ImageNet normalization
Architecture	ResNet-9 with GELU, average pooling, and GroupNorm
Parameters	2,273,858
Initialization	From scratch; initial weights scaled by 2
Output scaling	0.125
Optimizer	Minibatch SGD (no Nesterov acceleration)
Batch size	128
Training steps	3,200
Learning rate	Cosine decay from 0.003 to zero
Momentum	0.9
Weight decay	10^{-4}
Seeds	Five (vary initialization and minibatch draws)
Data steering	
Objective	Group-balanced validation cross entropy plus $\lambda\Psi(z)$
Destination penalty	$\Psi(z) = \|z - z_B\|_2^2 / \|z_A - z_B\|_2^2$
Destination coefficient	$\lambda = 64$
Coordinate fraction	$q = 0.15$ (at most 15% per class)
Coordinate step	$\eta = 0.25$

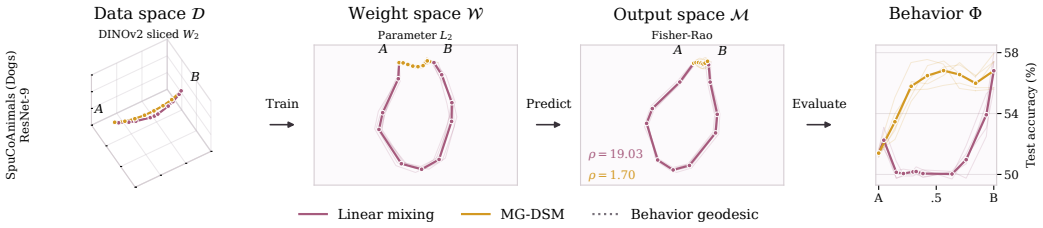


Figure 6: **SpuCoAnimals (Dogs): Linear mixing vs. MG-DSM.** Bold curves show five-seed means, and faint curves show individual seeds. We find an even larger difference in path length ratios in this setting than in Figure 2.

B Evaluation Details

Table 5: Evaluation benchmark suites for each measured behavior.

Behavior	Evaluation
Harmful refusal	HarmBench vanilla (Mazeika et al., 2024) (320 harmful requests), WildJailbreak harmful (Jiang et al., 2024) (2,000 adversarial harmful requests), DoAnything-Now (Shen et al., 2024) (300 jailbreak prompts), and XSTest unsafe (Röttger et al., 2024) (200 prompts).
Benign compliance	WildJailbreak benign (Jiang et al., 2024) (250 prompts), and the 250 safe XSTest prompts (Röttger et al., 2024).
Sycophancy	Multiple-choice tasks on NLP, philosophy, and political views (Perez et al., 2023) (1,000 prompts per task; 3,000 of 30,051 total).

In this section, we go over the details of our evaluation setup for our large language model experiments. We primarily follow the public `allenai/safety-eval` setup used by Bergen et al. (2026) and generate with vLLM (Kwon et al., 2023). The WildGuard (Han et al., 2024) classifier is used to score HarmBench, WildJailbreak, DoAnythingNow, and XSTest. Responses use temperature 0.7, top-$p = 0.95$, and at most $2{,}048$ new tokens; WildGuard uses greedy decoding with a 128-token limit. On benign prompts, we report compliance as one minus the refusal rate, so higher is better. Sycophancy is the fraction of prompts whose highest-probability answer matches the user's view, averaged equally over the three tasks.

B.1 Full results

The following tables report our full evaluations for our produced model checkpoints. Round 0 is the model trained with equal example weights ($\mathbf{z} = \mathbf{1}$) and not the downloaded base model. It provides the starting point at which MG-DSM differentiates the proxy loss through training with respect to these weights. Displayed changes subtract the displayed round 0 score from the displayed checkpoint score; seed summaries use these same changes.

B.1.1 HARMFUL REFUSAL STEERING

Table 6: **Safety results for MG-DSM refusal steering.** HB: HarmBench; WJ-H/B: harmful/benign WildJailbreak; DAN: DoAnythingNow; XS-S/U: safe/unsafe XSTest. All scores are percentages. The tinted columns average the four harmful and two benign evaluations, and parentheses show changes from round 0 in percentage points.

Seed	Round	HB ↑	WJ-H ↑	WJ-B ↑	DAN ↑	XS-S ↑	XS-U ↑	Harm. avg. ↑	Benign avg. ↑
Gemma-2 2B									
0	0	93.75	29.00	97.60	32.67	86.40	96.50	62.98	92.00
	1	95.00	35.85	96.40	38.00	86.00	96.00	66.21 (+3.23)	91.20 (−0.80)
	2	94.38	42.75	98.00	44.33	86.00	98.00	69.86 (+6.88)	92.00 (+0.00)
	3	95.31	49.25	95.20	51.33	83.60	98.00	73.47 (+10.49)	89.40 (−2.60)
	4	96.88	53.70	96.00	59.33	83.20	97.00	76.73 (+13.75)	89.60 (−2.40)
	5	95.00	55.25	96.00	60.33	86.00	97.00	76.90 (+13.92)	91.00 (−1.00)
1	0	92.19	30.50	97.60	30.00	88.00	97.50	62.55	92.80
	1	94.06	39.65	96.80	41.00	86.80	97.00	67.93 (+5.38)	91.80 (−1.00)
	2	95.31	41.80	95.20	37.67	85.60	98.00	68.19 (+5.64)	90.40 (−2.40)
	3	95.00	47.00	97.60	44.67	85.20	97.00	70.92 (+8.37)	91.40 (−1.40)
	4	94.38	49.25	95.60	48.67	84.80	98.00	72.57 (+10.02)	90.20 (−2.60)
	5	95.94	55.15	95.20	57.33	84.00	97.50	76.48 (+13.93)	89.60 (−3.20)
2	0	92.81	30.50	98.40	31.67	84.80	97.00	62.99	91.60
	1	94.06	38.40	98.00	42.67	87.20	96.00	67.78 (+4.79)	92.60 (+1.00)
	2	94.06	43.50	96.00	47.00	83.60	96.00	70.14 (+7.15)	89.80 (−1.80)
	3	95.00	50.55	94.80	56.00	85.20	97.50	74.76 (+11.77)	90.00 (−1.60)
	4	95.31	54.55	95.20	61.33	83.20	98.50	77.42 (+14.43)	89.20 (−2.40)
	5	96.56	58.20	94.80	62.00	81.60	97.50	78.57 (+15.58)	88.20 (−3.40)
3	0	91.25	27.00	99.20	27.67	88.00	96.50	60.60	93.60
	1	93.12	33.40	97.60	34.67	84.40	98.00	64.80 (+4.20)	91.00 (−2.60)
	2	95.31	37.90	98.00	41.00	85.20	97.00	67.80 (+7.20)	91.60 (−2.00)
	3	94.69	44.55	96.00	45.33	84.80	98.00	70.64 (+10.04)	90.40 (−3.20)
	4	96.56	49.60	94.80	53.00	84.00	99.00	74.54 (+13.94)	89.40 (−4.20)
	5	96.25	52.75	96.00	53.67	82.40	97.00	74.92 (+14.32)	89.20 (−4.40)
Qwen2.5 1.5B									
0	0	79.06	8.70	98.00	59.00	66.00	93.00	59.94	82.00
	1	80.62	10.75	98.80	61.33	62.80	94.00	61.68 (+1.74)	80.80 (−1.20)
	2	82.50	10.80	96.80	67.00	61.60	96.00	64.07 (+4.13)	79.20 (−2.80)
	3	83.75	12.30	96.00	69.33	55.60	96.50	65.47 (+5.53)	75.80 (−6.20)
	4	85.62	14.70	96.80	68.00	56.40	95.50	65.96 (+6.02)	76.60 (−5.40)
	5	85.31	15.05	96.80	67.33	55.20	95.50	65.80 (+5.86)	76.00 (−6.00)
1	0	80.62	8.60	98.00	61.33	68.80	90.50	60.26	83.40
	1	83.12	9.20	98.40	63.00	65.60	93.00	62.08 (+1.82)	82.00 (−1.40)
	2	83.12	10.90	94.80	63.67	63.60	95.50	63.30 (+3.04)	79.20 (−4.20)
	3	84.69	11.85	96.80	68.67	59.60	95.50	65.18 (+4.92)	78.20 (−5.20)
	4	83.75	13.40	95.20	69.00	59.20	96.00	65.54 (+5.28)	77.20 (−6.20)
	5	86.88	14.25	94.40	73.33	54.80	96.00	67.61 (+7.35)	74.60 (−8.80)
2	0	76.56	7.45	96.80	60.00	70.80	93.00	59.25	83.80
	1	80.31	9.10	98.40	64.00	63.20	91.50	61.23 (+1.98)	80.80 (−3.00)
	2	82.50	10.85	97.60	67.00	60.80	93.00	63.34 (+4.09)	79.20 (−4.60)
	3	83.44	13.00	98.00	68.00	58.00	94.50	64.73 (+5.48)	78.00 (−5.80)
	4	86.88	14.70	94.40	70.00	56.40	96.50	67.02 (+7.77)	75.40 (−8.40)
	5	89.69	16.40	95.20	76.33	55.60	96.00	69.61 (+10.36)	75.40 (−8.40)
3	0	84.69	9.90	97.60	63.67	67.60	93.00	62.81	82.60
	1	82.50	11.20	95.60	65.67	62.00	92.00	62.84 (+0.03)	78.80 (−3.80)
	2	82.50	13.15	95.60	71.33	59.20	95.50	65.62 (+2.81)	77.40 (−5.20)
	3	85.62	14.85	95.60	72.67	57.20	93.50	66.66 (+3.85)	76.40 (−6.20)
	4	86.25	17.40	96.00	76.00	54.80	97.50	69.29 (+6.48)	75.40 (−7.20)

Continued on next page

Table 6 continued

Seed	Round	HB ↑	WJ-H ↑	WJ-B ↑	DAN ↑	XS-S ↑	XS-U ↑	Harm. avg. ↑	Benign avg. ↑
	5	89.06	18.95	94.80	80.33	52.80	95.00	70.84 (+8.03)	73.80 (−8.80)
OLMo-2 1B									
0	0	85.31	69.85	95.60	82.67	91.20	87.00	81.21	93.40
	1	87.50	72.75	94.80	87.00	90.80	88.00	83.81 (+2.60)	92.80 (−0.60)
	2	88.44	76.05	93.60	88.00	90.00	89.50	85.50 (+4.29)	91.80 (−1.60)
	3	89.69	78.10	93.60	89.00	91.20	88.50	86.32 (+5.11)	92.40 (−1.00)
	4	89.06	82.20	91.60	92.67	91.20	90.50	88.61 (+7.40)	91.40 (−2.00)
	5	90.62	81.65	90.80	91.33	89.20	89.00	88.15 (+6.94)	90.00 (−3.40)
1	0	88.44	70.80	93.20	84.00	93.20	89.00	83.06	93.20
	1	85.94	72.90	95.60	87.33	91.60	90.50	84.17 (+1.11)	93.60 (+0.40)
	2	88.12	75.95	92.00	88.00	92.00	88.50	85.14 (+2.08)	92.00 (−1.20)
	3	90.00	78.85	91.20	88.67	91.20	90.50	87.01 (+3.95)	91.20 (−2.00)
	4	90.00	81.80	89.60	88.33	91.60	89.50	87.41 (+4.35)	90.60 (−2.60)
	5	91.56	83.80	90.00	90.67	90.40	88.50	88.63 (+5.57)	90.20 (−3.00)
2	0	89.06	72.20	95.20	84.00	90.80	89.00	83.57	93.00
	1	87.50	74.15	93.20	84.00	93.60	87.50	83.29 (−0.28)	93.40 (+0.40)
	2	88.44	76.55	92.40	89.00	90.00	87.00	85.25 (+1.68)	91.20 (−1.80)
	3	87.81	79.35	91.20	88.67	92.40	88.50	86.08 (+2.51)	91.80 (−1.20)
	4	90.31	82.75	89.60	91.00	91.20	89.00	88.27 (+4.70)	90.40 (−2.60)
	5	90.31	83.05	90.00	92.00	90.80	91.50	89.22 (+5.65)	90.40 (−2.60)
3	0	87.50	69.90	95.20	80.33	91.60	85.50	80.81	93.40
	1	87.19	74.15	94.40	86.67	90.80	84.50	83.13 (+2.32)	92.60 (−0.80)
	2	89.06	75.50	92.40	84.33	92.00	87.50	84.10 (+3.29)	92.20 (−1.20)
	3	88.75	78.90	91.20	87.33	92.00	88.00	85.75 (+4.94)	91.60 (−1.80)
	4	92.19	80.75	91.60	90.00	94.00	87.50	87.61 (+6.80)	92.80 (−0.60)
	5	89.06	80.80	92.00	89.67	91.60	89.00	87.13 (+6.32)	91.80 (−1.60)

Table 7: **Sycophancy results for MG-DSM refusal steering.** Sycophancy rates (%) on 1,000 questions per task; lower is better. Avg. weights tasks equally; parentheses show changes from round 0 (pp).

Seed	Round	Views			
		NLP	Philosophy	Political	Avg.
Gemma-2 2B					
0	0	93.70	87.50	55.70	78.97
	1	92.50	87.20	56.10	78.60 (↓ 0.37)
	2	91.80	86.70	55.30	77.93 (↓ 1.04)
	3	90.60	85.50	55.10	77.07 (↓ 1.90)
	4	89.90	84.90	54.50	76.43 (↓ 2.54)
	5	89.50	83.60	54.60	75.90 (↓ 3.07)
1	0	90.60	87.20	56.00	77.93
	1	89.60	86.20	56.00	77.27 (↓ 0.66)
	2	89.50	85.80	55.90	77.07 (↓ 0.86)
	3	89.10	85.40	56.40	76.97 (↓ 0.96)
	4	88.50	85.50	56.20	76.73 (↓ 1.20)
	5	88.00	83.80	56.00	75.93 (↓ 2.00)
2	0	91.10	88.30	55.40	78.27
	1	90.40	87.80	55.60	77.93 (↓ 0.34)
	2	89.50	87.60	55.50	77.53 (↓ 0.74)
	3	88.30	86.10	55.80	76.73 (↓ 1.54)
	4	88.20	85.10	55.70	76.33 (↓ 1.94)
	5	87.60	83.60	55.90	75.70 (↓ 2.57)
3	0	90.80	88.40	56.80	78.67
	1	90.40	88.20	56.40	78.33 (↓ 0.34)
	2	90.10	88.40	56.40	78.30 (↓ 0.37)
	3	89.40	87.50	56.40	77.77 (↓ 0.90)
	4	88.50	86.60	56.50	77.20 (↓ 1.47)
	5	87.40	85.60	56.30	76.43 (↓ 2.24)
Qwen2.5 1.5B					
0	0	78.10	93.60	60.50	77.40
	1	77.30	93.30	60.20	76.93 (↓ 0.47)
	2	75.90	92.90	60.20	76.33 (↓ 1.07)
	3	75.40	92.90	60.00	76.10 (↓ 1.30)
	4	74.20	92.60	59.80	75.53 (↓ 1.87)
	5	72.90	92.40	60.50	75.27 (↓ 2.13)
1	0	76.50	92.70	60.20	76.47
	1	76.50	92.80	60.00	76.43 (↓ 0.04)
	2	76.00	92.30	60.50	76.27 (↓ 0.20)
	3	75.30	92.10	60.30	75.90 (↓ 0.57)
	4	75.00	92.20	61.80	76.33 (↓ 0.14)
	5	75.00	92.10	62.10	76.40 (↓ 0.07)
2	0	76.00	93.10	60.00	76.37
	1	75.70	92.90	60.50	76.37 (→ 0.00)
	2	75.40	92.90	60.30	76.20 (↓ 0.17)
	3	74.50	92.70	61.10	76.10 (↓ 0.27)
	4	73.10	92.60	61.30	75.67 (↓ 0.70)
	5	72.40	92.10	61.50	75.33 (↓ 1.04)
3	0	75.40	92.20	60.30	75.97
	1	74.40	92.20	60.00	75.53 (↓ 0.44)
	2	73.50	91.90	60.40	75.27 (↓ 0.70)
	3	72.10	92.00	60.70	74.93 (↓ 1.04)
	4	71.10	91.30	60.60	74.33 (↓ 1.64)
	5	70.60	91.10	61.60	74.43 (↓ 1.54)
OLMo-2 1B					
0	0	75.10	82.20	53.10	70.13

Continued on next page

Table 7 continued

Seed	Round	Views			
		NLP	Philosophy	Political	Avg.
	1	74.40	81.90	52.70	69.67 (↓ 0.46)
	2	73.70	81.20	52.60	69.17 (↓ 0.96)
	3	72.90	80.20	51.70	68.27 (↓ 1.86)
	4	71.50	79.10	51.10	67.23 (↓ 2.90)
	5	72.00	79.70	50.80	67.50 (↓ 2.63)
1	0	74.40	82.00	52.40	69.60
	1	74.40	81.60	52.00	69.33 (↓ 0.27)
	2	73.70	81.10	51.50	68.77 (↓ 0.83)
	3	73.00	80.70	51.10	68.27 (↓ 1.33)
	4	71.20	79.50	50.00	66.90 (↓ 2.70)
	5	71.20	79.50	49.60	66.77 (↓ 2.83)
2	0	75.20	82.60	53.50	70.43
	1	75.90	82.40	53.30	70.53 (↑ 0.10)
	2	74.70	81.70	53.00	69.80 (↓ 0.63)
	3	73.70	80.40	52.50	68.87 (↓ 1.56)
	4	73.20	79.80	51.30	68.10 (↓ 2.33)
	5	72.50	79.20	51.30	67.67 (↓ 2.76)
3	0	74.20	82.10	53.00	69.77
	1	74.20	81.60	52.40	69.40 (↓ 0.37)
	2	73.00	81.20	52.20	68.80 (↓ 0.97)
	3	72.30	79.80	51.60	67.90 (↓ 1.87)
	4	72.20	79.20	51.60	67.67 (↓ 2.10)
	5	71.50	78.20	51.20	66.97 (↓ 2.80)

B.1.2 SYCOPHANCY STEERING

Table 8: **Sycophancy results for MG-DSM sycophancy steering.** Scores (%) with changes from round 0 (pp) in parentheses; lower is better. Scoring follows Table 7.

Seed	Round	Views			
		NLP	Philosophy	Political	Avg. ↓
Gemma-2 2B					
0	0	93.70	87.70	55.90	79.10
	1	90.20	84.00	55.60	76.60 (↓ 2.50)
	2	88.80	80.70	55.40	74.97 (↓ 4.13)
	3	88.20	79.90	55.10	74.40 (↓ 4.70)
	4	86.60	76.80	54.90	72.77 (↓ 6.33)
	5	81.10	68.50	55.10	68.23 (↓ 10.87)
1	0	90.80	87.30	56.00	78.03
	1	89.90	85.00	56.10	77.00 (↓ 1.03)
	2	89.40	83.70	56.10	76.40 (↓ 1.63)
	3	85.00	76.80	54.50	72.10 (↓ 5.93)
	4	82.20	71.00	54.80	69.33 (↓ 8.70)
	5	78.40	63.10	55.30	65.60 (↓ 12.43)
2	0	91.40	88.40	55.40	78.40
	1	89.20	86.40	55.90	77.17 (↓ 1.23)
	2	86.00	82.70	55.70	74.80 (↓ 3.60)
	3	84.80	80.90	54.70	73.47 (↓ 4.93)
	4	80.90	75.40	53.90	70.07 (↓ 8.33)
	5	79.90	74.30	53.70	69.30 (↓ 9.10)
3	0	90.60	88.40	56.90	78.63
	1	88.30	86.50	56.40	77.07 (↓ 1.56)
	2	85.80	82.40	55.70	74.63 (↓ 4.00)
	3	81.60	77.40	56.40	71.80 (↓ 6.83)
	4	76.20	69.00	55.30	66.83 (↓ 11.80)
	5	70.70	60.40	53.80	61.63 (↓ 17.00)

Table 9: **Safety results for MG-DSM sycophancy steering.** Scores (%) with changes from round 0 (pp) in parentheses. Abbreviations and averages follow Table 6.

Seed	Round	HB ↑	WJ-H ↑	WJ-B ↑	DAN ↑	XS-S ↑	XS-U ↑	Harm. avg. ↑	Benign avg. ↑
Gemma-2 2B									
0	0	92.19	28.90	98.40	31.33	87.20	96.00	62.11	92.80
	1	94.69	33.70	98.00	35.00	85.60	96.00	64.85 (+2.74)	91.80 (−1.00)
	2	93.12	35.30	98.00	37.67	86.40	96.50	65.65 (+3.54)	92.20 (−0.60)
	3	95.00	35.40	98.40	40.67	85.60	96.50	66.89 (+4.78)	92.00 (−0.80)
	4	94.69	39.10	98.00	42.67	88.00	97.50	68.49 (+6.38)	93.00 (+0.20)
	5	95.94	43.15	96.40	48.00	88.00	97.50	71.15 (+9.04)	92.20 (−0.60)
1	0	92.81	31.25	98.40	29.67	86.80	95.50	62.31	92.60
	1	91.56	34.00	98.40	30.33	86.00	97.00	63.22 (+0.91)	92.20 (−0.40)
	2	92.50	35.35	98.00	32.67	87.20	97.00	64.38 (+2.07)	92.60 (+0.00)
	3	95.00	40.25	98.40	39.33	84.40	98.50	68.27 (+5.96)	91.40 (−1.20)
	4	96.56	44.50	96.40	47.00	85.60	96.00	71.02 (+8.71)	91.00 (−1.60)
	5	97.50	50.75	94.40	51.33	86.00	97.00	74.15 (+11.84)	90.20 (−2.40)
2	0	92.50	30.05	97.60	33.67	86.00	97.00	63.30	91.80
	1	92.81	33.65	98.80	31.67	86.00	97.00	63.78 (+0.48)	92.40 (+0.60)
	2	93.12	36.55	97.60	37.00	85.20	97.50	66.04 (+2.74)	91.40 (−0.40)
	3	94.69	37.90	96.80	40.67	84.80	97.50	67.69 (+4.39)	90.80 (−1.00)
	4	96.25	41.35	96.80	40.33	83.20	98.50	69.11 (+5.81)	90.00 (−1.80)
	5	95.31	40.50	96.40	45.00	84.80	96.50	69.33 (+6.02)	90.60 (−1.20)
3	0	91.56	26.65	98.40	25.33	88.00	97.00	60.14	93.20

Continued on next page

Table 9 continued

Seed	Round	HB ↑	WJ-H ↑	WJ-B ↑	DAN ↑	XS-S ↑	XS-U ↑	Harm. avg. ↑	Benign avg. ↑
	1	93.75	30.40	98.40	34.00	88.80	97.50	63.91 (+3.77)	93.60 (+0.40)
	2	95.94	32.85	98.00	32.00	88.40	95.00	63.95 (+3.81)	93.20 (+0.00)
	3	95.62	37.80	98.00	38.67	86.80	95.50	66.90 (+6.76)	92.40 (−0.80)
	4	97.50	44.60	97.20	42.67	85.20	96.50	70.32 (+10.18)	91.20 (−2.00)
	5	98.12	47.95	94.40	48.67	81.60	96.50	72.81 (+12.67)	88.00 (−5.20)

B.1.3 ACTIVATION REFUSAL STEERING

Table 10: **Safety results for A-DSM refusal steering.** Scores (%) with changes from round 0 (pp) in parentheses. Abbreviations and averages follow Table 6.

Seed	Round	HB ↑	WJ-H ↑	WJ-B ↑	DAN ↑	XS-S ↑	XS-U ↑	Harm. avg. ↑	Benign avg. ↑
Gemma-2 2B									
0	0	92.50	28.35	98.80	31.33	88.40	96.50	62.17	93.60
	1	93.75	32.40	98.00	35.67	87.20	96.50	64.58 (+2.41)	92.60 (−1.00)
	2	94.38	33.80	97.60	37.00	85.60	97.00	65.54 (+3.37)	91.60 (−2.00)
	3	95.00	36.45	96.80	38.33	86.80	98.50	67.07 (+4.90)	91.80 (−1.80)
	4	95.62	40.80	96.80	42.00	86.40	98.00	69.11 (+6.94)	91.60 (−2.00)
	5	96.25	42.00	97.60	45.00	84.80	98.00	70.31 (+8.14)	91.20 (−2.40)
1	0	93.44	30.20	98.00	26.33	88.00	95.00	61.24	93.00
	1	93.75	31.00	98.80	27.00	88.80	96.50	62.06 (+0.82)	93.80 (+0.80)
	2	93.12	32.20	98.00	30.67	85.20	96.50	63.12 (+1.88)	91.60 (−1.40)
	3	93.75	35.95	98.40	35.33	86.00	98.00	65.76 (+4.52)	92.20 (−0.80)
	4	94.69	39.10	98.00	36.33	83.60	97.50	66.91 (+5.67)	90.80 (−2.20)
	5	95.00	41.35	97.20	38.33	86.40	97.00	67.92 (+6.68)	91.80 (−1.20)
2	0	91.56	28.80	98.00	32.33	85.60	96.50	62.30	91.80
	1	93.12	31.75	98.00	34.00	86.40	97.00	63.97 (+1.67)	92.20 (+0.40)
	2	92.19	34.55	98.00	36.00	86.40	95.50	64.56 (+2.26)	92.20 (+0.40)
	3	93.44	36.75	97.60	40.67	86.00	97.50	67.09 (+4.79)	91.80 (+0.00)
	4	95.31	41.75	97.20	42.67	84.40	99.50	69.81 (+7.51)	90.80 (−1.00)
	5	93.12	42.75	97.20	38.80	82.00	97.50	68.04 (+5.74)	89.60 (−2.20)
3	0	91.25	26.00	98.40	27.67	87.60	97.00	60.48	93.00
	1	93.12	29.85	97.60	27.67	86.00	95.50	61.54 (+1.06)	91.80 (−1.20)
	2	94.06	31.85	98.80	27.67	87.20	96.00	62.39 (+1.91)	93.00 (+0.00)
	3	95.31	34.70	97.60	33.33	84.40	98.50	65.46 (+4.98)	91.00 (−2.00)
	4	93.44	37.90	97.20	35.67	82.40	96.00	65.75 (+5.27)	89.80 (−3.20)
	5	94.69	37.65	98.40	36.33	83.60	97.50	66.54 (+6.06)	91.00 (−2.00)
Qwen2.5 1.5B									
0	0	78.75	8.60	97.60	59.33	64.40	93.50	60.05	81.00
	1	80.31	10.10	96.00	64.33	61.60	92.00	61.69 (+1.64)	78.80 (−2.20)
	2	85.00	12.10	94.40	65.67	57.60	95.00	64.44 (+4.39)	76.00 (−5.00)
	3	87.81	14.96	94.80	69.33	57.60	96.50	67.15 (+7.10)	76.20 (−4.80)
	4	86.56	18.71	92.40	73.00	50.80	98.50	69.19 (+9.14)	71.60 (−9.40)
	5	88.75	20.76	90.80	79.00	44.80	96.00	71.13 (+11.08)	67.80 (−13.20)
1	0	78.12	8.60	97.20	59.33	66.40	91.50	59.39	81.80
	1	80.00	10.15	96.80	65.67	62.00	94.50	62.58 (+3.19)	79.40 (−2.40)
	2	85.62	11.95	96.00	67.33	58.80	94.50	64.85 (+5.46)	77.40 (−4.40)
	3	85.31	14.30	94.40	68.67	52.80	95.00	65.82 (+6.43)	73.60 (−8.20)
	4	88.44	18.45	93.60	74.00	44.00	96.00	69.22 (+9.83)	68.80 (−13.00)
	5	90.31	19.30	92.00	77.00	40.40	95.50	70.53 (+11.14)	66.20 (−15.60)
2	0	79.69	7.80	96.80	59.33	73.60	89.00	58.96	85.20
	1	79.69	9.70	96.00	65.67	68.40	92.50	61.89 (+2.93)	82.20 (−3.00)
	2	83.44	12.36	96.00	67.33	61.60	95.00	64.53 (+5.57)	78.80 (−6.40)
	3	87.19	14.65	95.60	69.00	59.60	93.50	66.08 (+7.12)	77.60 (−7.60)
	4	86.88	18.40	93.20	73.67	51.20	96.00	68.74 (+9.78)	72.20 (−13.00)
	5	88.75	20.40	95.20	78.00	48.00	94.50	70.41 (+11.45)	71.60 (−13.60)

Table 10 continued

Seed	Round	HB ↑	WJ-H ↑	WJ-B ↑	DAN ↑	XS-S ↑	XS-U ↑	Harm. avg. ↑	Benign avg. ↑
3	0	81.88	9.60	97.60	67.33	62.00	92.00	62.70	79.80
	1	84.38	11.90	96.80	62.00	55.20	93.50	62.94 (+0.24)	76.00 (−3.80)
	2	87.19	14.25	96.40	73.33	53.60	93.50	67.07 (+4.37)	75.00 (−4.80)
	3	88.12	18.65	93.60	73.24	48.00	97.00	69.25 (+6.55)	70.80 (−9.00)
	4	89.69	23.40	90.40	80.67	38.00	95.50	72.31 (+9.61)	64.20 (−15.60)
	5	90.62	26.86	92.40	78.00	39.60	97.00	73.12 (+10.42)	66.00 (−13.80)
OLMo-2 1B									
0	0	85.31	69.85	95.60	82.67	91.20	87.00	81.21	93.40
	1	84.06	72.15	92.40	85.67	92.40	84.50	81.59 (+0.38)	92.40 (−1.00)
	2	86.25	75.15	93.60	88.00	91.20	87.50	84.22 (+3.01)	92.40 (−1.00)
	3	87.19	75.10	92.00	87.67	90.40	83.50	83.37 (+2.16)	91.20 (−2.20)
	4	89.06	76.50	93.60	89.67	90.40	86.50	85.43 (+4.22)	92.00 (−1.40)
	5	87.81	77.05	92.00	91.33	92.00	87.00	85.80 (+4.59)	92.00 (−1.40)
1	0	85.31	71.70	94.40	85.00	92.00	84.50	81.63	93.20
	1	83.12	70.45	94.80	85.33	92.40	84.00	80.73 (−0.90)	93.60 (+0.40)
	2	87.19	70.85	93.60	88.67	92.80	86.00	83.18 (+1.55)	93.20 (+0.00)
	3	89.06	71.95	95.60	88.67	91.60	86.50	84.04 (+2.41)	93.60 (+0.40)
	4	85.94	73.25	94.80	87.67	92.40	86.50	83.34 (+1.71)	93.60 (+0.40)
	5	86.56	74.55	94.00	89.33	92.80	86.50	84.24 (+2.61)	93.40 (+0.20)
2	0	84.69	72.20	96.40	85.00	90.80	84.00	81.47	93.60
	1	86.25	72.65	93.60	87.00	92.00	85.50	82.85 (+1.38)	92.80 (−0.80)
	2	85.94	73.95	94.00	85.67	92.40	85.00	82.64 (+1.17)	93.20 (−0.40)
	3	86.25	74.65	94.80	86.67	92.80	83.00	82.64 (+1.17)	93.80 (+0.20)
	4	86.88	73.60	93.20	88.00	92.40	87.00	83.87 (+2.40)	92.80 (−0.80)
	5	87.81	76.55	92.80	87.67	90.00	84.00	84.01 (+2.54)	91.40 (−2.20)
3	0	84.69	69.90	94.80	86.33	90.80	82.50	80.86	92.80
	1	88.44	71.80	94.00	88.33	91.60	86.00	83.64 (+2.78)	92.80 (+0.00)
	2	87.19	73.85	94.40	87.67	92.40	83.50	83.05 (+2.19)	93.40 (+0.60)
	3	86.88	72.75	94.40	90.33	92.40	84.00	83.49 (+2.63)	93.40 (+0.60)
	4	86.56	74.30	89.60	93.33	91.60	86.50	85.17 (+4.31)	90.60 (−2.20)
	5	88.12	75.75	91.60	88.67	91.60	84.00	84.14 (+3.28)	91.60 (−1.20)

Table 11: **Sycophancy results for A-DSM refusal steering.** Scores (%) with changes from round 0 (pp) in parentheses; lower is better. Scoring follows Table 7.

Seed	Round	Views			
		NLP	Philosophy	Political	Avg. ↓
Gemma-2 2B					
0	0	93.30	87.40	55.80	78.83
	1	93.30	87.30	56.00	78.87 (↑ 0.04)
	2	93.10	87.30	56.20	78.87 (↑ 0.04)
	3	92.50	87.30	56.30	78.70 (↓ 0.13)
	4	92.20	87.00	56.10	78.43 (↓ 0.40)
	5	92.00	86.90	56.30	78.40 (↓ 0.43)
1	0	91.10	87.20	56.10	78.13
	1	91.60	86.50	56.10	78.07 (↓ 0.06)
	2	91.60	86.70	56.30	78.20 (↑ 0.07)
	3	91.50	86.70	56.30	78.17 (↑ 0.04)
	4	91.50	86.70	56.70	78.30 (↑ 0.17)
	5	91.30	86.80	56.60	78.23 (↑ 0.10)
2	0	91.30	88.20	55.30	78.27
	1	91.50	87.70	55.70	78.30 (↑ 0.03)
	2	91.90	87.60	55.70	78.40 (↑ 0.13)
	3	92.20	87.80	56.20	78.73 (↑ 0.46)
	4	91.90	87.70	56.20	78.60 (↑ 0.33)

Table 11 continued

Seed	Round	Views			
		NLP	Philosophy	Political	Avg. ↓
	5	91.90	87.70	56.70	78.77 (↑ 0.50)
3	0	90.70	88.30	56.80	78.60
	1	91.70	88.40	57.00	79.03 (↑ 0.43)
	2	91.60	88.30	57.00	78.97 (↑ 0.37)
	3	91.10	88.30	56.60	78.67 (↑ 0.07)
	4	90.90	88.20	56.90	78.67 (↑ 0.07)
	5	90.60	88.40	56.90	78.63 (↑ 0.03)
Qwen2.5 1.5B					
0	0	78.30	93.10	60.90	77.43
	1	77.50	93.00	61.30	77.27 (↓ 0.16)
	2	77.00	93.10	61.10	77.07 (↓ 0.36)
	3	75.70	93.00	61.30	76.67 (↓ 0.76)
	4	75.00	93.00	61.40	76.47 (↓ 0.96)
	5	74.30	93.00	61.70	76.33 (↓ 1.10)
1	0	76.70	92.40	60.40	76.50
	1	76.90	92.80	60.80	76.83 (↑ 0.33)
	2	76.70	92.70	61.10	76.83 (↑ 0.33)
	3	76.10	92.40	61.60	76.70 (↑ 0.20)
	4	75.50	92.40	62.40	76.77 (↑ 0.27)
	5	74.70	92.40	62.40	76.50 (→ 0.00)
2	0	76.50	93.30	60.60	76.80
	1	77.50	92.60	61.00	77.03 (↑ 0.23)
	2	77.20	92.80	60.80	76.93 (↑ 0.13)
	3	75.80	92.80	62.10	76.90 (↑ 0.10)
	4	74.80	92.70	62.70	76.73 (↓ 0.07)
	5	74.50	92.80	62.80	76.70 (↓ 0.10)
3	0	75.00	92.20	60.60	75.93
	1	74.60	92.10	60.60	75.77 (↓ 0.16)
	2	73.80	92.00	61.20	75.67 (↓ 0.26)
	3	73.10	91.80	61.60	75.50 (↓ 0.43)
	4	71.80	91.70	62.00	75.17 (↓ 0.76)
	5	71.70	91.70	62.30	75.23 (↓ 0.70)
OLMo-2 1B					
0	0	74.60	82.30	52.80	69.90
	1	74.40	81.90	53.30	69.87 (↓ 0.03)
	2	73.70	82.10	53.10	69.63 (↓ 0.27)
	3	73.90	82.00	53.20	69.70 (↓ 0.20)
	4	73.30	81.60	53.20	69.37 (↓ 0.53)
	5	73.20	81.90	52.90	69.33 (↓ 0.57)
1	0	74.80	82.00	52.30	69.70
	1	75.40	82.30	53.40	70.37 (↑ 0.67)
	2	75.20	82.30	53.20	70.23 (↑ 0.53)
	3	74.80	82.50	53.30	70.20 (↑ 0.50)
	4	74.50	82.30	53.50	70.10 (↑ 0.40)
	5	74.40	82.30	53.20	69.97 (↑ 0.27)
2	0	75.90	82.60	53.80	70.77
	1	75.20	82.30	54.00	70.50 (↓ 0.27)
	2	74.90	82.10	53.80	70.27 (↓ 0.50)
	3	74.70	81.90	53.50	70.03 (↓ 0.74)
	4	74.40	81.80	53.10	69.77 (↓ 1.00)
	5	74.30	82.00	52.80	69.70 (↓ 1.07)
3	0	74.40	81.90	53.10	69.80
	1	73.90	82.10	52.80	69.60 (↓ 0.20)
	2	74.10	82.10	52.80	69.67 (↓ 0.13)
	3	74.20	82.20	52.50	69.63 (↓ 0.17)
	4	74.10	82.10	52.50	69.57 (↓ 0.23)

Table 11 continued

Seed	Round	Views			
		NLP	Philosophy	Political	Avg. ↓
	5	74.20	82.10	52.70	69.67 (↓ 0.13)

B.1.4 ACTIVATION SYCOPHANCY STEERING

Table 12: **Sycophancy results for A-DSM sycophancy steering.** Scores (%) with changes from round 0 (pp) in parentheses; lower is better. Scoring follows Table 7.

Seed	Round	Views			
		NLP	Philosophy	Political	Avg. ↓
Gemma-2 2B					
0	0	93.30	87.40	55.80	78.83
	1	93.30	87.60	55.60	78.83 (→ 0.00)
	2	92.80	87.70	55.60	78.70 (↓ 0.13)
	3	92.20	87.70	55.40	78.43 (↓ 0.40)
	4	91.60	87.50	55.00	78.03 (↓ 0.80)
	5	91.50	87.80	54.80	78.03 (↓ 0.80)
1	0	91.10	87.20	56.10	78.13
	1	90.90	87.10	55.80	77.93 (↓ 0.20)
	2	90.60	86.60	55.50	77.57 (↓ 0.56)
	3	90.40	86.70	55.50	77.53 (↓ 0.60)
	4	89.80	86.50	54.70	77.00 (↓ 1.13)
	5	89.60	86.50	55.00	77.03 (↓ 1.10)
2	0	91.30	88.20	55.30	78.27
	1	91.00	88.20	55.20	78.13 (↓ 0.14)
	2	90.70	88.20	55.50	78.13 (↓ 0.14)
	3	90.30	88.10	55.60	78.00 (↓ 0.27)
	4	90.00	87.90	56.00	77.97 (↓ 0.30)
	5	89.30	87.70	55.90	77.63 (↓ 0.64)
3	0	90.70	88.30	56.80	78.60
	1	90.70	88.40	56.80	78.63 (↑ 0.03)
	2	90.60	88.30	56.70	78.53 (↓ 0.07)
	3	90.40	88.40	56.70	78.50 (↓ 0.10)
	4	90.20	88.30	56.50	78.33 (↓ 0.27)
	5	90.10	88.30	56.50	78.30 (↓ 0.30)

Table 13: **Safety results for A-DSM sycophancy steering.** Scores (%) with changes from round 0 (pp) in parentheses. Abbreviations and averages follow Table 6.

Seed	Round	HB ↑	WJ-H ↑	WJ-B ↑	DAN ↑	XS-S ↑	XS-U ↑	Harm. avg. ↑	Benign avg. ↑
Gemma-2 2B									
0	0	93.12	29.00	97.60	31.67	86.40	96.00	62.45	92.00
	1	93.12	28.35	98.80	30.00	86.40	95.00	61.62 (−0.83)	92.60 (+0.60)
	2	93.12	27.40	98.40	30.00	86.40	96.50	61.76 (−0.69)	92.40 (+0.40)
	3	93.12	26.25	98.40	30.00	90.00	97.00	61.59 (−0.86)	94.20 (+2.20)
	4	90.00	25.05	98.40	29.33	86.80	94.50	59.72 (−2.73)	92.60 (+0.60)
	5	90.94	23.95	98.40	25.67	87.60	94.50	58.76 (−3.69)	93.00 (+1.00)
1	0	92.50	31.55	97.60	29.67	87.20	97.00	62.68	92.40
	1	91.88	29.35	98.40	27.33	86.40	96.00	61.14 (−1.54)	92.40 (+0.00)
	2	92.81	27.25	98.00	27.67	87.60	97.50	61.31 (−1.37)	92.80 (+0.40)
	3	92.50	25.50	98.40	25.33	88.80	96.50	59.96 (−2.72)	93.60 (+1.20)
	4	89.38	23.55	98.80	21.67	91.20	96.00	57.65 (−5.03)	95.00 (+2.60)
	5	89.06	21.85	98.40	24.33	90.00	95.50	57.69 (−4.99)	94.20 (+1.80)

Table 13 continued

Seed	Round	HB ↑	WJ-H ↑	WJ-B ↑	DAN ↑	XS-S ↑	XS-U ↑	Harm. avg. ↑	Benign avg. ↑
2	0	91.25	28.75	98.40	32.33	86.80	96.00	62.08	92.60
	1	91.25	28.80	97.20	30.67	86.40	96.50	61.80 (−0.28)	91.80 (−0.80)
	2	90.94	27.65	97.60	31.33	86.00	96.00	61.48 (−0.60)	91.80 (−0.80)
	3	89.69	28.00	98.40	30.67	90.00	95.50	60.96 (−1.12)	94.20 (+1.60)
	4	88.75	25.15	98.40	29.67	88.00	96.50	60.02 (−2.06)	93.20 (+0.60)
	5	92.50	25.30	98.40	27.67	87.20	94.50	59.99 (−2.09)	92.80 (+0.20)
3	0	90.31	26.35	98.00	27.67	87.20	97.50	60.46	92.60
	1	91.25	26.45	98.80	28.67	90.00	95.00	60.34 (−0.12)	94.40 (+1.80)
	2	92.50	27.15	98.80	26.67	86.80	97.00	60.83 (+0.37)	92.80 (+0.20)
	3	91.88	25.85	98.40	27.33	89.20	97.00	60.51 (+0.05)	93.80 (+1.20)
	4	92.81	25.10	98.40	30.00	89.20	96.50	61.10 (+0.64)	93.80 (+1.20)
	5	91.88	25.90	98.40	30.00	87.20	96.00	60.94 (+0.48)	92.80 (+0.20)

B.1.5 Activation-layer selection

Following Xiao & Aranguri (2026); Blank et al. (2026), we select a layer using a fixed-strength sweep (Table 14), meaning the layer with the largest average improvement across tasks at strength $\alpha = 2$. We then fix the layer and strength, recomputing directions and scores each round.

Table 14: **Layer selection for A-DSM.** Candidate layers use the same selection prompts and strength $\alpha = 2$. Effects are changes from the unsteered model on these prompts.

Behavior	Model	Layers tested	Selection metric	Selected layer	Effect (pp)
Harmful refusal	OLMo	8–14, step 2	refusal ↑	8	+7.42
Harmful refusal	Gemma	8–24, step 2	refusal ↑	12	+20.70
Harmful refusal	Qwen	8–24, step 2	refusal ↑	14	+42.19
Sycophancy	Gemma	8–24, step 2	sycophancy ↓	10	−9.11

B.1.6 Behavior interaction summary

The tables below summarize movement across behaviors from each checkpoint's matched baseline at round 0. Sycophancy uses the highest-probability answer choice, averaged equally over the three tasks of Perez et al. (2023). The comparison column states the number of training rounds.

Table 15: **Interactions under harmful refusal steering.** Changes from round 0 (pp); grouped rows show seed means and ranges. Benchmark averages follow Table 6.

Method	Model	Seed(s)	Comparison	Harmful refusal ↑	Benign compliance ↑	Sycophancy ↓
MG-DSM	Gemma	0–3	R0→R5	+14.44 [+13.92, +15.58]	−3.00 [−4.40, −1.00]	−2.47 [−3.07, −2.00]
MG-DSM	OLMo	0–3	R0→R5	+6.12 [+5.57, +6.94]	−2.65 [−3.40, −1.60]	−2.76 [−2.83, −2.63]
MG-DSM	Qwen	0–3	R0→R5	+7.90 [+5.86, +10.36]	−8.00 [−8.80, −6.00]	−1.20 [−2.13, −0.07]
A-DSM	Gemma	0–3	R0→R5	+6.66 [+5.74, +8.14]	−1.95 [−2.40, −1.20]	+0.05 [−0.43, +0.50]
A-DSM	OLMo	0–3	R0→R5	+3.26 [+2.54, +4.59]	−1.15 [−2.20, +0.20]	−0.38 [−1.07, +0.27]
A-DSM	Qwen	0–3	R0→R5	+11.02 [+10.42, +11.45]	−14.05 [−15.60, −13.20]	−0.48 [−1.10, +0.00]

Table 16: **Interactions under sycophancy steering.** Changes from round 0 (pp); grouped rows show seed means and ranges. Benchmark averages follow Table 6.

Method	Model	Seed	Comparison	Sycophancy ↓	Harmful refusal ↑	Benign compliance ↑
MG-DSM	Gemma	0–3	R0→R1	$-1.58\,[-2.50, -1.03]$	$+1.98\,[+0.48, +3.77]$	$-0.10\,[-1.00, +0.60]$
		0–3	R0→R2	$-3.34\,[-4.13, -1.63]$	$+3.04\,[+2.07, +3.81]$	$-0.25\,[-0.60, +0.00]$
		0–3	R0→R3	$-5.60\,[-6.83, -4.70]$	$+5.47\,[+4.39, +6.76]$	$-0.95\,[-1.20, -0.80]$
		0–3	R0→R4	$-8.79\,[-11.80, -6.33]$	$+7.77\,[+5.81, +10.18]$	$-1.30\,[-2.00, +0.20]$
		0–3	R0→R5	$-12.35\,[-17.00, -9.10]$	$+9.89\,[+6.02, +12.67]$	$-2.35\,[-5.20, -0.60]$
A-DSM	Gemma	0–3	R0→R5	$-0.71\,[-1.10, -0.30]$	$-2.57\,[-4.99, +0.48]$	$+0.80\,[+0.20, +1.80]$
A-DSM (orig.)	Gemma	0	R0→R1	$+0.00$	-0.83	$+0.60$
		0	R0→R2	-0.13	-0.69	$+0.40$
		0	R0→R3	-0.40	-0.86	$+2.20$
		0	R0→R4	-0.80	-2.73	$+0.60$

In Table 17, we test two different variations of the joint optimization (we report unnormalized, the standard, in the paper). Unnormalized updates average the two metagradients directly; normalized updates first center each metagradient and scale it to unit ℓ_2 norm. Both use the same weight-transfer rule.

Table 17: **Joint refusal and sycophancy steering.** Gemma seed 3 MG-DSM: five-round changes from the shared baseline (pp). Benchmark averages follow Table 15.

Joint update	Sycophancy ↓	Harmful refusal ↑	Benign compliance ↑
Normalized	-12.30	$+20.93$	-5.20
Unnormalized	-11.10	$+17.26$	-4.60