PROJECTION STEERING: SPECIFYING OUTPUT DISTRIBUTIONS OVER STRUCTURED CONCEPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper formalizes what counts as good steering for *structured concepts*, whose values form a cycle, an order, or a tree—months of the year, numerals, a taxonomy. We identify three core desiderata that the steering literature has addressed as separate components: steering should achieve the intended target probabilities, follow a path that respects the concept's geometric structure, and disrupt the off-target behavior as little as possible. These desiderata are derived directly from the model's output distribution and the structure of the concept. We then propose *projection steering* as a convex optimization program whose objective minimizes the disruption to the model's softmax output distribution, subject to constraints that specify the concept's conditional distribution (e.g. 60% probability of April when restricted to months of the year) and the total probability mass assigned to the concept (e.g. probability allocated to month tokens). This formulation enables attaining the target distribution exactly whenever it is reachable. We then develop a general construction that uses the geometry of the target concept (e.g. a circle for months of the year) to turn a semantic path into a family of target output distributions. Crucially, choosing this semantic path is separate from determining how to steer the model to realize it. We verify that this three-step framework has clear reachability conditions, attains all proposed desiderata, and improves on existing steering baselines on both a synthetic task and a real language model.

1 INTRODUCTION

In activation steering, one typically adds a scaled direction to the model's internal representation in the hope of inducing a desired output behavior (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023). However, steering directions may transfer unreliably across contexts, depend sensitively on their magnitude, and distort unrelated parts of the output distribution (Tan et al., 2024; Pres et al., 2024; Wu et al., 2025). A growing line of work improves activation steering by taking the geometry of model representations into account. These steering methods follow curves fitted to representation manifolds or geodesics under learned metrics (Sarfati et al., 2026; Wurgaft et al., 2026; Oozeer et al., 2026). Such geometry-aware paths can produce more coherent output behavior, but there remains a fundamental question: *how should the model behave along the path?*

Structured concepts enable one to concretely answer this question. A multi-labeled concept is *structured* when its labels carry a relational structure—a cycle, an order, or a tree, as in months of the year, numerals, or a taxonomy. For such concepts, the desired behavior is naturally expressed as a distribution over related labels. Consider steering a model's belief state from January toward April. We may want the final output to place most of the month probability on April, intermediate states to move through February and March in order, and unrelated outputs to remain as unchanged as possible. Concretely, we formalize three desiderata for what constitutes good steering:

- **Target accuracy.** Does the model reach the intended *behavior*—the intended distribution over the concept's labels, with the intended share of the total output probability? In the month example, the final distribution should be peaked on April, may place some probability on nearby months such as March and May, and should assign the intended total probability mass to month tokens.

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.

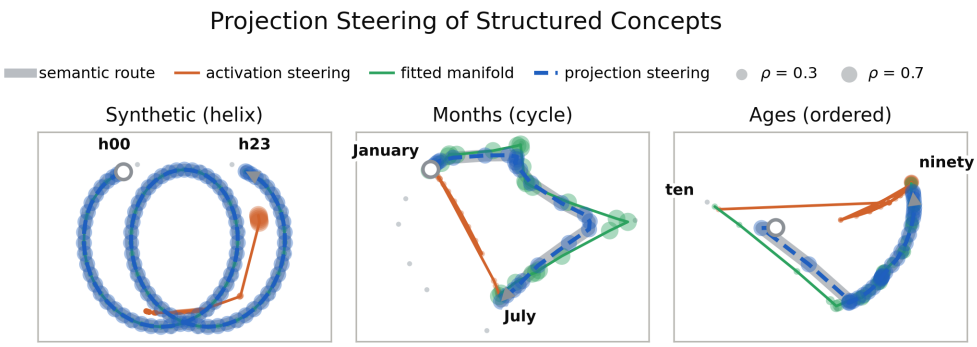


Figure 1: **Steering structured concepts along semantic routes.** Each panel shows steering from the open circle to the endpoint along a specified semantic route (gray). *Projection steering* (blue) closely follows this route while maintaining the required probability on the concept, whereas activation steering (orange) cuts across the structure and steering along a fitted manifold (green) can follow a different path. To visualize the trajectories, each label z_j is assigned a representer point d_j (§6), and a distribution q over the concept labels is plotted at $\sum_j q(z_j)d_j$. Marker area shows ρ, the total probability assigned to the concept labels.

- **Off-target preservation.** Is the model's off-target behavior disturbed as little as possible? In the month example, the distribution over other concepts, such as animals, verbs, punctuation, should be changed only as much as necessary to achieve the intended month behavior.
- **Route fidelity.** Do intermediate interventions follow the concept's structure in the intended order? From January to April, the intermediate states should pass through February and March rather than jumping straight from January to April.

Each desideratum has appeared as a component of prior steering work: activation steering already measures target success, information-geometric approaches minimize off-target changes in the output distribution, and geometry-aware methods concern whether intermediate behavior follows the concept's structure (Turner et al., 2023; Rimsky et al., 2024; Park et al., 2026; Wurgaft et al., 2026). No existing method accounts for all three simultaneously, so we bring these components together into a single formulation for steering structured concepts. The key idea is to *specify the desired behavior directly as constraints, rather than measuring it after an intervention*.

Projection steering. To do so, we propose *projection steering* as a three-step framework. Consider steering a model from January to April. First, we specify the desired behavior along the *semantic route* by assigning each position on the January–February–March–April path a target distribution over month labels, with nearby months receiving more probability than distant ones. Second, we turn each target into *behavioral constraints* that specify both the desired distribution among month labels (conditional on the month concept) and a minimum amount of total output probability assigned to months (mass); for example, 60% of the probability among month tokens on April, while months as a whole receive at least 70% of the model's output probability. Third, *forward-KL projection* finds a hidden state that satisfies these constraints while changing the model's original output distribution as little as possible. Figure 1 illustrates the resulting behavior across several structured concepts.

Our contributions are as follows:

- **Unified desiderata for steering structured concepts** (§3.2). We identify three requirements for steering structured concepts: target accuracy, route fidelity, and off-target preservation. They follow from decomposing the model's output distribution into how much total probability is assigned to the concept, how that probability is distributed among its labels, and how the remaining probability is distributed over everything else. This provides a common language for specifying and evaluating steering of structured concepts.
- **Projection steering with exact behavioral control** (§§4–5). We construct a three-step steering framework that converts a concept's *semantic route* to target distributions over its labels, then imposes *behavioral constraints* on both that distribution and the total probability

assigned to the concept, and finally uses *forward-KL projection* to find the hidden states that satisfy the constraints while changing the original output distribution as little as possible. This is a convex problem, so projection steering exactly realizes any feasible behavioral target.

- **A general construction for semantic routes and readout geometry** (§6). We show how positions along a structured concept can be mapped to distributions over its labels. The method is generally applicable to any concept structure (cycles, orders, trees, and more). Crucially, specifying these semantic routes is independent of the rest of the steering framework, so any geometry can be used with the same steering procedure. In particular, the model's readout itself supplies a candidate geometry.

2 PRELIMINARIES

Let $\mathcal{X}$ be the set of prompts and $h : \mathcal{X} \to \mathbb{R}^d$ be the activation map of a model at the readout before taking softmax, and write $\lambda = h(x) \in \mathbb{R}^d$ for the representation, or *embedding*, of an input x. Let $\mathcal{Y}$ be the output vocabulary, $V = |\mathcal{Y}|$, and $\Gamma \in \mathbb{R}^{V \times d}$ the unembedding matrix with rows $\gamma_y^\top$. The logit of y and the softmax output distribution are

$$\ell_y(\lambda) = \gamma_y^\top \lambda, \qquad p_\lambda(y) = \frac{\exp\big(\gamma_y^\top \lambda\big)}{\sum_{y' \in \mathcal{Y}} \exp\big(\gamma_{y'}^\top \lambda\big)} = \exp\big(\ell_y(\lambda) - A(\lambda)\big), \tag{1}$$

where $A(\lambda) := \log \sum_{y' \in \mathcal{Y}} \exp(\gamma_{y'}^\top \lambda)$ is the *log-normalizer* that makes p_λ sum to one. For the final layer Γ is the unembedding; for intermediate layers it may be a learned multiclass probe. We will not use A and its derivatives until §5, where they supply the mean unembedding $\nabla A(\lambda) = \mathbb{E}_{p_\lambda}[\gamma]$, the covariance $\nabla^2 A(\lambda) = \mathrm{Cov}_{p_\lambda}[\gamma]$, and the KL divergence between two softmax distributions.

3 WHAT SHOULD COUNT AS GOOD STEERING OF A STRUCTURED CONCEPT?

3.1 STRUCTURED CONCEPT DOMAINS AND THE THREE-WAY DECOMPOSITION

A *structured concept domain* is a set $Z = \{z_1, \ldots, z_m\} \subseteq \mathcal{Y}$ of m output labels carrying a relational structure: a cycle (weekdays, months, compass headings), an order (number words, decades, magnitudes), or a graph (a kinship tree, a geography hierarchy, a state machine). Language models are known to represent cyclic and ordered features as low-dimensional manifolds (Engels et al., 2025; Gurnee & Tegmark, 2024) and hierarchies as structured subspaces (Park et al., 2025; Sakata et al., 2026).

The rest of the vocabulary, $\mathcal{O} := \mathcal{Y} \setminus Z$, we call the *outsiders*, which are off-target concepts that one does not intend to change. For example, if Z is the set of month labels, then $\mathcal{O}$ contains all other tokens in the vocabulary. Every output distribution splits along this partition into a *domain mass*, an *in-domain conditional*, and an *outsider conditional*:

$$\rho_\lambda := p_\lambda(Z), \qquad q_\lambda := p_\lambda(\cdot \mid Z) \in \Delta^m, \qquad r_\lambda := p_\lambda(\cdot \mid \mathcal{O}) \in \Delta^{V-m}, \tag{2}$$

and p_λ is recovered from the three:

$$p_\lambda(y) = \begin{cases} \rho_\lambda \, q_\lambda(y), & y \in Z, \\ (1 - \rho_\lambda) \, r_\lambda(y), & y \in \mathcal{O}. \end{cases} \tag{3}$$

For strictly positive p_λ, it can be shown that the correspondence $p_\lambda \leftrightarrow (\rho_\lambda, q_\lambda, r_\lambda)$ is a bijection. We abbreviate the reference states by $\rho_0 := \rho_{\lambda_0}$, q_{λ_0}, and r_{λ_0}.

The domain mass ρ_λ says how much probability the entire concept domain holds, q_λ says how that mass is distributed among the m labels, and r_λ says how the model distributes mass over the rest of the vocabulary. In the month example, ρ_λ is the total probability assigned to months, q_λ determines how that probability is divided among January through December, and r_λ describes the probabilities conditioned on non-month tokens. From the desiderata, ρ_λ and q_λ are what an intervention specifies, and r_λ is what it should preserve.

Lemma 1 (Partition identity). *Let p and p' be strictly positive distributions on $\mathcal{Y}$, with decompositions (ρ, q, r) and (ρ', q', r') over the same domain Z. Then*

$$D_{\mathrm{KL}}(p \,\|\, p') = D_{\mathrm{KL}}(\mathrm{Bern}(\rho) \,\|\, \mathrm{Bern}(\rho')) + \rho\, D_{\mathrm{KL}}(q \,\|\, q') + (1-\rho)\, D_{\mathrm{KL}}(r \,\|\, r')\,,$$

where $\mathrm{Bern}(\rho)$ is the distribution of the indicator $\mathbf{1}[y \in Z]$ under p.

Lemma 1 decomposes the KL divergence between two output distributions into three parts: a *mass term*, measuring the change in total probability assigned to the domain Z; a *shape term*, measuring how the conditional distribution within the domain changes; and an *off-target term*, measuring how the conditional distribution over the rest of the vocabulary changes. The identity follows from the chain rule for KL divergence applied to the indicator of being in Z; all proofs are in Appendix A.

3.2 Three Desiderata

Lemma 1 decomposes the KL divergence into a mass term, a shape term, and an off-target term, so the change from an unsteered distribution to a steered one is exactly determined by (ρ, q, r). Inspired by such results, we acquire the following desiderata for steering with the three variables.

Target accuracy. We want the model to achieve the specified on-target probability, which we can describe in terms of ρ and q: the model should attribute at least $\rho^\star$ of the output mass to the domain and distribute that mass as $q^\star$. Given a desired conditional $q^\star$ and a mass lower bound $\rho^\star$, the realized state λ is scored by the target error:

$$\mathrm{TE}(\lambda) := \Big(D_{\mathrm{KL}}(q^\star \,\|\, q_\lambda)\,,\ \ (\rho^\star - \rho_\lambda)_+ \Big), \qquad (x)_+ := \max(x, 0),$$

reported together with the full-vocabulary argmax indicator $\mathbf{1}[\arg\max_y p_\lambda(y) = z_t]$ for the intended label z_t. The mass component is not a divergence because $\rho^\star$ is a lower bound.

Off-target preservation. Given $q^\star$ and $\rho^\star$, the realized state should change the off-target conditional as little as possible, as measured by the off-target error:

$$\mathrm{OTE}(\lambda) := D_{\mathrm{KL}}(r_{\lambda_0} \,\|\, r_\lambda)\,,$$

which is the third term of Lemma 1 up to the factor $(1 - \rho_0)$.

Route fidelity. For structured concepts, steering should respect a route through labels. We represent the concept domain Z in a space $\overline{Z}$ carrying the declared concept structure. A *route* is a path through $\overline{Z}$ with positions ζ. Let $\zeta \mapsto q^\star_\zeta$ assign a target conditional to each route position. Given positions $\zeta_1, \ldots, \zeta_n$ and realized states $\lambda_1, \ldots, \lambda_n$, we measure agreement with the target route by the route error:

$$\mathrm{RE} := \tfrac{1}{n} \sum\nolimits_{t=1}^{n} W_1^c\big(q_{\lambda_t},\, q^\star_{\zeta_t}\big),$$

where c is the declared ground cost on Z, such as cyclic distance for a cycle or shortest-path distance for a graph. The 1-Wasserstein distance is used over KL because it is the metric in which misplacing probability one semantic step from the target location is penalized less than misplacing it four steps away. This way, RE measures the distance between the realized and target *paths*, rather than between individual probability distributions. Notably, no existing method meets all three: see §7 for details.

4 Behavioral Constraints and Reachability

The behavior we want to control is expressed in the model's output distribution, but the intervention itself is applied in activation space. As a result, methods based on directions or fitted curves can only reveal their effect on the model's behavior after the intervention is made. We instead specify the desired output behavior directly as constraints to an optimization problem. Any activation satisfying those constraints then realizes the desired behavior by construction. This section defines the resulting feasible set and characterizes when it is nonempty.

Definition 2 (Behavioral constraints). *Behavioral constraints are a pair $(q^\star, \rho^\star)$ with $q^\star \in \operatorname{int} \Delta^m$ a strictly positive in-domain conditional over Z and $\rho^\star \in (0, 1)$ a domain mass lower bound. Their feasible set in the support Q is*

$$\mathcal{C}(q^\star, \rho^\star) := \big\{\lambda \in \lambda_0 + \mathrm{range}(Q) : q_\lambda = q^\star,\ \rho_\lambda \geq \rho^\star\big\}. \tag{4}$$

Here, $Q \in \mathbb{R}^{d\times s}$ specifies the allowed edit directions, so $\lambda_0 + \mathrm{range}(Q)$ is the set of reachable states; $Q = I_d$ allows unrestricted edits. Whether a target is achievable therefore depends on the readout restricted to $\mathrm{range}(Q)$, which we will characterize in Proposition 7. This is useful when steering is restricted to a k-dimensional subspace (PCA or manifold subspace), whose basis vectors are collected in Q. The rest of this section connects the behavioral constraints to the model's representations and readout, and shows that they form a convex set of activations.

First, note that centering the in-domain logits does not change the conditional they induce. Let $\Gamma_Z \in \mathbb{R}^{m\times d}$ be the readout matrix restricted to Z and $H_m := I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$ the centering matrix. The centered restricted readout and the centered logits of a state are

$$\widetilde{\Gamma} := H_m\Gamma_Z \in \mathbb{R}^{m\times d}, \qquad \phi(\lambda) := \widetilde{\Gamma}\lambda, \qquad r_\Gamma := \mathrm{rank}(\widetilde{\Gamma}) \le \min(m-1,\, d). \tag{5}$$

Lemma 3 (Centered logits parametrize in-domain conditional)**.** *For any λ, $q_\lambda = \mathrm{softmax}(\Gamma_Z\lambda) = \mathrm{softmax}(\phi(\lambda))$.*

This way, $\phi(\lambda)$ is the unique centered representative of the in-domain logits, and $\mathrm{range}(\widetilde{\Gamma})$ determines exactly which conditionals can be read from a state. This leads to our next statement.

Proposition 4 (The shape constraint is affine)**.** *Let $\phi^\star := H_m \log q^\star$. Then, for any λ, $q_\lambda = q^\star \iff \widetilde{\Gamma}\lambda = \phi^\star$. Moreover, there exists an edit x such that $q_{\lambda_0+Qx} = q^\star \iff \phi^\star - \phi(\lambda_0) \in \mathrm{range}(\widetilde{\Gamma}Q)$.*

Hence, one obtains an affine set, which is convex. Also, a target conditional is met exactly when a target logit is met, and which targets are available is determined by the readout structure. We now turn to the mass constraint.

Lemma 5 (Log-mass is concave on the shape constraints)**.** *Fix any $z_i \in Z$, and let $\widetilde{\Gamma}\lambda = \phi^\star$. Then $\log\rho_\lambda = \log p_\lambda(z_i) - \log q^\star(z_i)$, where $\log\rho_\lambda$ is concave.*

When λ satisfies the shape constraints, $p_\lambda(z_i) = \rho_\lambda\, q^\star(z_i)$, so $\log\rho_\lambda$ differs by a constant from $\log p_\lambda(z_i) = \gamma_{z_i}^\top\lambda - A(\lambda)$, which is concave because A is convex. For details see Appendix A.2.

Proposition 6 (The feasible set is convex)**.** *Fix any $z_i \in Z$. Then*

$$\mathcal{C}(q^\star,\rho^\star) \;=\; \big(\lambda_0 + \mathrm{range}(Q)\big) \;\cap\; \{\lambda : \widetilde{\Gamma}\lambda = \phi^\star\} \;\cap\; \{\lambda : \log p_\lambda(z_i) \ge \log\rho^\star + \log q^\star(z_i)\},$$

which is convex.

An affine set intersected with a convex sublevel set is convex. We conclude that the behavior constraints $(q^\star,\rho^\star)$ form a convex feasible set. Moreover, we deduce when the solution is reachable:

Proposition 7 (Reachability)**.** *Let $\rho_{\max} := \sup\{\rho_\lambda : \lambda \in \lambda_0 + \mathrm{range}(Q),\ q_\lambda = q^\star\}$, with $\rho_{\max} := 0$ if no such λ exists. Then $\mathcal{C}(q^\star,\rho^\star) \neq \varnothing$ if and only if*

$$\phi^\star - \phi(\lambda_0) \in \mathrm{range}(\widetilde{\Gamma}Q) \qquad \textit{and} \qquad \rho^\star \le \rho_{\max},$$

with the latter strict unless the supremum is attained.

Hence, when the inequality holds strictly, one obtains a nonempty convex set, and the projection (6) admits a minimizer. Therefore, under the reachability conditions one is guaranteed to attain target accuracy. However, note that while the realized distribution $p_{\lambda^\star}$ is unique, $\lambda^\star$ need not be unique. Appendix B details such arguments and treats the boundary case $\rho^\star = \rho_{\max}$, where the supremum need not be attained.

A more nuanced question is how much $\rho^\star$ one needs, for a given $q^\star$, to make an intended label z_t the model's most probable output over the whole vocabulary. We show the following.

Proposition 8 (Required mass for reaching target)**.** *Let z_t be the unique mode of $q^\star$ and $\tau := q^\star(z_t)$. If $\rho^\star > (1+\tau)^{-1}$, then $z_t = \arg\max_{y\in\mathcal{Y}} p_\lambda(y)$ for every $\lambda \in \mathcal{C}(q^\star,\rho^\star)$.*

In the month example, suppose $q^\star$ assigns probability $\tau = 0.6$ to April. Then $\rho^\star > 0.625$ guarantees that April is the most probable token overall. For example, at $\rho^\star = 0.7$, every feasible state assigns April at least 0.42 probability, while any outsider token has probability at most 0.30.

Overall, the *target accuracy* desideratum defines a convex set of feasible activations; when the reachability conditions hold, it is nonempty and every activation in it meets the target exactly. What remains is to choose an activation in that set that best preserves off-target behavior (§5) and to specify the family of targets induced by a semantic route (§6).

5 PROJECTION STEERING

The behavioral constraints confine the target behavior, but say nothing about what the model should do off-target. The idea is to choose, among the hidden states that achieve the demanded behavior, the one that disrupts the rest of the distribution least. We now formally define projection steering:

Definition 9 (Projection steering). *Given a reference state λ_0 and behavioral constraints $(q^\star, \rho^\star)$, projection steering is defined by the forward-KL projection*

$$\lambda^\star \in \operatorname*{arg\,min}_{\lambda \in \mathcal{C}(q^\star, \rho^\star)} D_{\mathrm{KL}}(p_{\lambda_0} \,\|\, p_\lambda)\,. \tag{6}$$

This is an I-projection in the sense of Csiszár (1975), generalizing dual steering's binary KL projection, where the feasible set is the hyperplane $\{\lambda : \beta_W^\top \lambda = c\}$ defined by a binary probe β_W (Park et al., 2026). Write

$$\mu(\lambda) := \nabla A(\lambda) = \mathbb{E}_{p_\lambda}[\gamma], \qquad G(\lambda) := \nabla^2 A(\lambda) = \mathrm{Cov}_{p_\lambda}[\gamma] \succeq 0 \tag{7}$$

for the dual coordinate and the Fisher metric. For a fixed reference, the objective is the Bregman divergence of the log-normalizer,

$$D_{\mathrm{KL}}(p_{\lambda_0} \,\|\, p_\lambda) = A(\lambda) - A(\lambda_0) - \mu(\lambda_0)^\top(\lambda - \lambda_0), \qquad \nabla^2_\lambda D_{\mathrm{KL}}(p_{\lambda_0} \,\|\, p_\lambda) = G(\lambda), \tag{8}$$

so it is convex in λ, with the Fisher metric as its Hessian. Together with Proposition 6 this makes (6) a convex program, and by strict convexity along distribution-changing directions the realized output distribution $p_{\lambda^\star}$ is unique even when $\lambda^\star$ is not. We then show that, among states realizing the same specified behavior, projection steering minimizes off-target distortion.

Theorem 10 (Off-target optimality). *Let $\lambda^\star$ solve (6), and let $\hat{\rho} := \rho_{\lambda^\star}$. Then $\lambda^\star$ minimizes $D_{\mathrm{KL}}(r_{\lambda_0} \,\|\, r_\lambda)$ over all $\lambda \in \lambda_0 + \mathrm{range}(Q)$ satisfying $q_\lambda = q^\star$ and $\rho_\lambda = \hat{\rho}$.*

Proof. By Lemma 1, on $q_\lambda = q^\star$,

$$D_{\mathrm{KL}}(p_{\lambda_0} \,\|\, p_\lambda) = D_{\mathrm{KL}}(\mathrm{Bern}(\rho_0) \,\|\, \mathrm{Bern}(\rho_\lambda)) + \rho_0 D_{\mathrm{KL}}(q_{\lambda_0} \,\|\, q^\star) + (1-\rho_0) D_{\mathrm{KL}}(r_{\lambda_0} \,\|\, r_\lambda)\,.$$

On $\rho_\lambda = \hat{\rho}$, the first two terms are constant. Since $\lambda^\star$ minimizes the full KL over the feasible set, it therefore minimizes $D_{\mathrm{KL}}(r_{\lambda_0} \,\|\, r_\lambda)$. □

Notably, the forward direction of the KL divergence is deliberate. The proof above turns on the off-target term carrying the constant weight $1 - \rho_0$. If the arguments are switched, that weight is $1 - \rho_\lambda$, which is controlled by the optimizer: it can simply reduce $D_{\mathrm{KL}}(r_{\lambda_0} \,\|\, r_\lambda)$ by increasing ρ_λ. This breaks the off-target preservation guarantee. Details about the solver are given in Appendix B.

6 SEMANTIC ROUTES AND GEOMETRY

What the KL projection on behavioral constraints achieves is exactly reaching the desired target while preserving off-target behavior. What remains is *route fidelity*: whether steering respects the concept's structure along the way. We answer this by giving a general construction for specifying a route, stated independently of how it is realized.

Let $\overline{Z}$ denote the space carrying the concept's structure, with labels Z embedded in it. A route is a path $s : \mathcal{T} \to \overline{Z}$, where $\zeta \in \mathcal{T}$ indexes positions along the path. One intends to find a map for each position to a target conditional distribution, $S : \zeta \mapsto q^\star_\zeta$. Each ζ therefore defines behavioral constraints $(q^\star_\zeta, \rho^\star)$, which projection steering realizes independently from the same λ_0. We introduce a technique to specify how a position ζ determines the target conditional $q^\star_\zeta$.

Definition 11 (Kernel-distance bump). *Let (M, d_S) be a metric space, with points $x_1, \ldots, x_m$ corresponding to the labels $z_1, \ldots, z_m$, and let $s : \mathcal{T} \to M$ be a route through these points. At position ζ, define*

$$q^\star_{\zeta,\beta}(z_j) \propto \exp\!\left(-\frac{\beta}{2} d_S(x_j, s(\zeta))^2\right), \qquad \phi^\star_{\zeta,\beta} = H_m\left[-\frac{\beta}{2} d_S(\cdot, s(\zeta))^2\right]. \tag{9}$$

The construction gives more probability to labels that are closer to the current semantic position. For example, on the month cycle, if adjacent months have distance one, a position near February assigns more probability to January and March than to July or August. The metric d_S determines which labels are close, while β controls how concentrated the target distribution is around the current position. Importantly, the same construction applies to cycles, orders, and graphs by changing only the underlying metric d_S and route passing through the labels. This construction therefore keeps the choice of semantic route separate from the projection used to realize it.

Proposition 12 (Ranking and calibration)**.** *Let* $\zeta = t_i$.

(i) for any $\beta > 0$ *and* $x_j \neq x_i$, $\log q^\star_{t_i,\beta}(z_i) - \log q^\star_{t_i,\beta}(z_j) = \frac{\beta}{2} d_S(x_j, x_i)^2 > 0$;

(ii) $q^\star_{t_i,\beta}(z_i)$ *is continuous and strictly increasing in* β*, from* $1/m$ *to* 1.

Part (i) guarantees that z_i is the unique target. Part (ii) means that any desired confidence $\tau \in (1/m, 1)$ determines a unique calibration hyperparameter β, which can be found via a root search. For example, specifying 60% confidence on April fixes the corresponding β. Between labeled positions, the bump varies smoothly and concentrates probability on nearby labels, so intermediate states such as "halfway between January and February" are well defined; see Appendix A.3.1.

The bump technique requires choosing a metric space (M, d_S) that represents the concept. We identify three available candidates of semantic geometries:

- **Declared.** $M = \overline{Z}$, the concept's known domain, with d_S given directly by its structure: for example, cyclic distance $d_S(z_i, z_j) = \min(|i-j|, m - |i-j|)$ on a month cycle, or shortest-path distance on a taxonomy.
- **Empirical.** $M = \mathbb{R}^k$, where an embedding $x_i \in \mathbb{R}^k$ is estimated from data, for example from activation centroids or fitted probe representations, and each x_i corresponds to a label z_i. We then use the Euclidean distance, $d_S(z_i, z_j) = \|x_i - x_j\|_2$.
- **Readout.** $M = \mathbb{R}^{r_\Gamma}$, where r_Γ is the rank of the centered readout matrix $\widetilde{\Gamma}$. We use the geometry induced by the model's readout rows by mapping the rows to their *representer coordinates* d_i, explained in the next paragraph, and use $d_S(z_i, z_j) = \|d_i - d_j\|_2$.

The first two choices obtain semantic geometry either from known concept structure or from representations estimated from data. We additionally consider whether the readout itself provides a useful geometry over the labels. This view is motivated by feature-field theory (Yocum et al., 2025), in which the readout rows can be viewed as an embedding of the concept domain and their induced kernel characterizes the representable fields (logits). Concretely, writing $\widetilde{\Gamma} = U\Sigma E^\top$, we use the rows of $D := U\Sigma$ as representer coordinates and define $d_S(z_i, z_j) = \|d_i - d_j\|_2$. We may additionally truncate these coordinates to their first k dimensions before fitting a semantic route. The readout thus plays two roles simultaneously in our framework: its range determines which target conditionals are reachable (Proposition 4), while its kernel also induces a geometry over the concept labels.

Note that there is no guarantee that Euclidean distance is the right metric on either the activation space or the readouts in representer coordinates. We note the elegance and potential of this construction and leave future work to supply the right metric for this default. Next, we design experiments to validate our steering program, compare against other baselines, and report empirical results in §7.

7 Experiments

This section reports a set of experiments on Gemma-3-4B (Gemma Team, 2025). We compare projection steering with linear-displacement methods, including activation steering (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023) and dual steering (Park et al., 2026), as well as geometry-aware methods, including fitted-manifold steering (Wurgaft et al., 2026) and geodesic steering under a learned metric (Oozeer et al., 2026). Regarding metrics for comparison, we report *target error*, *route error*, and *off-target error* (§3.2), as well as intermediate route positions.

Our readout matrix Γ comes from three sources: unembedding rows, multiclass probe rows, and synthetic rows arranged on a helix. For projection steering, we construct target conditionals from the *readout geometry* (§6) for the unembedding and synthetic tasks and *the declared geometry* for the

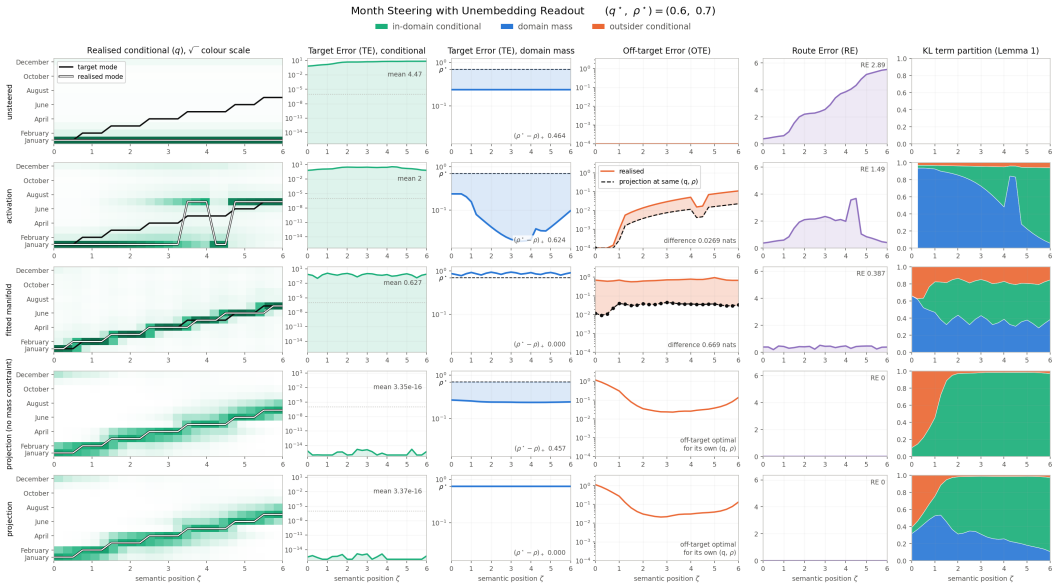


Figure 2: **Month steering with the unembedding readout**. Projection steering reaches the specified targets to solver precision, follows the semantic route, and preserves off-target behavior more effectively than the baselines. Rows compare the steering methods. Columns show: **(a)** the heatmap of conditional distributions over month labels along the semantic route; **(b)** *target error* (TE) for the in-domain conditional; **(c)** *target error* (TE) for the domain mass; **(d)** *off-target error* (OTE) measured by outsider KL; **(e)** *route error* (RE) measured by Wasserstein-1 distance to the target route; and **(f)** the proportion of domain-mass, in-domain conditional, and outsider-conditional terms decomposed from the KL term in Lemma 1.

taxonomy probe. We report additional experiments, including a synthetic task with known ground truth, used to validate the algebra and solver, in Appendix C.

7.1 Unembeddings: Months of the Year

In this task, we consider months of the year, which is known to exhibit cyclic features (Engels et al., 2025; Gurnee & Tegmark, 2024). We use the unembedding readout in projection steering, which will use the model's own vocabulary and determine the model's own output distribution. Detailed data generation, prompt templates, and probe construction are provided in Appendix C.3.

As shown in Figure 2, projection steering attains *target accuracy*, reaching the target in-domain conditional $q^\star$ up to solver precision and satisfying the lower bound on domain mass $\rho^\star$; the other steering methods either miss the target conditional or allocate insufficient probability mass to the concept. The heatmap in **(a)** further shows that projection steering transports probability smoothly along the specified semantic route, yielding almost zero route error in **(e)**. Linear methods fail to maintain route fidelity. Fitted-manifold steering (geometry-aware) better preserves the route than activation steering, but causes greater off-target distortion than projection steering, and its transitions between labels are less smooth than those of projection steering, as shown in the heatmaps. Projection steering also achieves lower off-target error than the baselines at the same realized behavior (q, ρ), preserving the outsider conditional more effectively.

7.2 Probes: Animal Taxonomy

Language model representations are known to encode hierarchical structure (Park et al., 2025; Sakata et al., 2026). We next test projection steering on a hierarchical concept at an intermediate layer. We fit an affine multiclass probe at layer 16 to predict animal family from Wikipedia species articles. A taxonomic rank is a level in the biological hierarchy: kingdoms contain phyla, which contain classes,

orders, and then families. Our concept domain consists of animal families, while non-animal probe classes form the outsiders. The detailed setup and results are given in Appendix C.8.

Figure 10 shows a route between two distant families. The route first moves through progressively broader groups and then descends through the other side of the taxonomy. Projection steering follows this hierarchy, moving probability through the corresponding groups while matching the target conditional along the route. Activation steering and fitted-manifold steering deviate from the specified path, and at matched realized (q, ρ) they cause greater off-target change. All quantities are defined on the probe output; this experiment tests control of an intermediate-layer readout, not downstream token probabilities or generated text.

8 Related Work and Discussion

Linear Representations and Linear Interventions. The linear representation hypothesis (Park et al., 2024; Marks & Tegmark, 2023) motivates difference-in-means and contrastive activation steering (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023). These methods assume concepts as binary-labeled directions, and intervene them by adding scaled directions to activations. Closest to our work is Park et al. (2026), which formulates dual steering through the information geometry of the softmax output and uses forward-KL projection to control a binary concept while minimizing off-target change. Our work extends their framework from binary concepts to structured m-label concepts, where the desired behavior is a distribution over related labels. Instead of specifying a scaled direction, we view steering as a convex program with constraints on the concept's conditional distribution and total mass while minimizing changes to off-target behavior, without requiring the concept-factorization assumption.

Manifold Representations and Geometry-Aware Interventions. Another line of work studies concepts as structured geometries like manifolds, hierarchies, or lattices in representation space (Gurnee & Tegmark, 2024; Engels et al., 2025; Park et al., 2025; Sakata et al., 2026; Xiong & Staab, 2025). This has motivated geometry-aware steering methods that follow fitted representation manifolds from data-generation processes (Sarfati et al., 2026), corresponding manifold structures in activations and output behavior (Wurgaft et al., 2026), or geodesics under a learned metric (Oozeer et al., 2026). These methods use representation geometry to improve steering, whereas our steering framework directly specifies the desired output trajectory and finds an activation that realizes it with minimal distortion to the remaining behavior. Importantly, we separate the geometry used to define a semantic route from the optimization used to realize it. Feature-field theory (Yocum et al., 2025) views readout rows as embeddings of a concept domain, with their Gram kernel characterizing the linearly representable feature fields over that domain. We use the readout as a candidate semantic geometry for projection steering, with the same readout structure also determining which target distributions are reachable.

Limitations and Future Work. There is not yet a theory for choosing the best semantic path, partly because the correct metric for the encoded semantic geometry remains unclear. More generally, outside structured concepts, the desired behavioral specification may itself be ambiguous, and a coarser directional steering may sometimes be more appropriate. Moreover, for multiclass probes, one must also explicitly specify which vocabulary outputs should remain unchanged. Furthermore, our results apply directly to single-token outputs at the final layer through the unembedding and to probe readouts at intermediate layers, where controlling the probe distributions does not guarantee the corresponding downstream model behavior. Finally, projection steering can in principle extend to unstructured or binary concepts, and combining multiple readouts into joint constraints may provide a path toward more scalable control. We leave these extensions for future work.

AI Use Statement

In this work, we used generative AI tools to assist with drafting and editing the manuscript, developing proofs (Appendix A), designing algorithms and experiments, executing experiments, and analyzing results. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the final text, methods, claims, results, and other artifacts presented in this work.

REFERENCES

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025.

Gemma Team. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv preprint arXiv:2310.06824, 2023.

Narmeen Oozeer, Shivam Raval, Philip Quirke, Manikandan Ravikiran, Jeff Phillips, Shriyash Upadhyay, and Amirali Abdullah. Riemannian-manifold steering: Geometry-aware generative autoencoders for label-free steering. arXiv preprint arXiv:2605.24942, 2026.

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*, pp. 39643–39666. PMLR, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. arXiv:2602.15293.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in LLMs. arXiv preprint arXiv:2410.17245, 2024.

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

Masaki Sakata, Benjamin Heinzerling, Takumi Ito, Sho Yokoi, and Kentaro Inui. Linear representations of hierarchical concepts in language models. arXiv preprint arXiv:2604.07886, 2026.

Raphaël Sarfati, Eric Bigelow, Daniel Wurgaft, Siddharth Boppana, Jack Merullo, Atticus Geiger, Owen Lewis, Tom McGrath, and Ekdeep Singh Lubana. The shape of beliefs: Geometry, dynamics, and interventions along representation manifolds of language models' posteriors. arXiv preprint arXiv:2602.02315, 2026.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, 2024.

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

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. arXiv preprint arXiv:2501.17148, 2025.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric J. Bigelow, Raphael Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah D. Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior. arXiv preprint arXiv:2605.05115, 2026.

Bo Xiong and Steffen Staab. From tokens to lattices: Emergent lattice structures in language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Julian Yocum, Cameron Allen, Bruno Olshausen, and Stuart Russell. Neural manifold geometry encodes feature fields. In *NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023.

A PROOFS

Proofs are grouped by the section in which each result is stated; the supporting results for semantic routes are stated in Appendix A.3 together with their proofs. The proof of Theorem 10 is given in §5.

Notation. For distributions a, b on $\mathcal{Y}$ write cross entropy $\mathrm{CE}(a,b) := -\sum_y a(y)\log b(y)$ and entropy $H(a) := \mathrm{CE}(a,a)$. Let $\tilde{q}^\star \in \Delta^V$ extend $q^\star$ by zeros on $\mathcal{O}$, write $\mathrm{CE}(q^\star, p_\lambda) := \mathrm{CE}(\tilde{q}^\star, p_\lambda)$, and let $\bar{\gamma}_{q^\star} := \Gamma^\top \tilde{q}^\star = \sum_i q^\star(z_i)\,\gamma_{z_i}$. For every λ,

$$\mathrm{CE}(q^\star, p_\lambda) = A(\lambda) - \bar{\gamma}_{q^\star}^\top \lambda, \tag{10}$$

which is convex. The *shape set* in the support is

$$\mathcal{C}_0 := \{\lambda \in \lambda_0 + \mathrm{range}(Q) : \widetilde{\Gamma}\lambda = \phi^\star\}, \qquad \mathcal{C}(q^\star, \rho^\star) = \{\lambda \in \mathcal{C}_0 : \rho_\lambda \geq \rho^\star\}.$$

On $\mathcal{C}_0$, Lemma 5 is equivalent to the cross-entropy form (see its proof)

$$\log \rho_\lambda = H(q^\star) - \mathrm{CE}(q^\star, p_\lambda) = \bar{\gamma}_{q^\star}^\top \lambda - A(\lambda) + H(q^\star), \tag{11}$$

which is the form used by the solver. Write $F(\lambda) := D_{\mathrm{KL}}(p_{\lambda_0} \,\|\, p_\lambda)$ for the projection objective.

A.1 PROOFS FOR SECTION 3

Proof of Lemma 1. Let $\chi := \mathbf{1}[y \in Z]$. Since χ is a function of y, $p(y) = p(\chi)\,p(y \mid \chi)$, and likewise for p', so the chain rule for KL divergence gives

$$D_{\mathrm{KL}}(p \,\|\, p') = D_{\mathrm{KL}}(p(\chi) \,\|\, p'(\chi)) + \mathbb{E}_{\chi\sim p}\Big[D_{\mathrm{KL}}(p(\cdot \mid \chi) \,\|\, p'(\cdot \mid \chi))\Big].$$

The laws of χ under p and p' are $\mathrm{Bern}(\rho)$ and $\mathrm{Bern}(\rho')$, which gives the first term. By (2) the conditionals are q, q' given $\chi = 1$ and r, r' given $\chi = 0$, events of probability ρ and $1-\rho$ under p, so the expectation equals $\rho\, D_{\mathrm{KL}}(q \,\|\, q') + (1-\rho)\, D_{\mathrm{KL}}(r \,\|\, r')$. Strict positivity makes every conditional well defined and every term finite. □

A.2 PROOFS FOR SECTION 4

Proof of Lemma 3. By (2), $q_\lambda(z) = p_\lambda(z)/p_\lambda(Z)$ for $z \in Z$, and the normalizer $e^{A(\lambda)}$ of (1) cancels, leaving $e^{\gamma_z^\top \lambda} \big/ \sum_{z' \in Z} e^{\gamma_{z'}^\top \lambda}$; hence $q_\lambda = \mathrm{softmax}(\Gamma_Z \lambda)$. Moreover $\Gamma_Z \lambda = H_m \Gamma_Z \lambda + \frac{1}{m}\mathbf{1}\mathbf{1}^\top \Gamma_Z \lambda = \phi(\lambda) + c\mathbf{1}$ with $c := \frac{1}{m}\mathbf{1}^\top \Gamma_Z \lambda$, and softmax is invariant to adding $c\mathbf{1}$. □

Proof of Proposition 4. Softmax is injective modulo constants: $\mathrm{softmax}(u) = \mathrm{softmax}(u')$ iff every coordinate of $u - u'$ equals the difference of the two log-normalizers, i.e. iff $u - u' \in \mathrm{span}(\mathbf{1}) = \ker H_m$, i.e. iff $H_m u = H_m u'$. Since $q^\star$ is strictly positive, $q^\star = \mathrm{softmax}(\log q^\star)$, so by Lemma 3, $q_\lambda = q^\star$ iff $H_m \Gamma_Z \lambda = H_m \log q^\star$, i.e. $\widetilde{\Gamma}\lambda = \phi^\star$, a linear equation with affine solution set. For $\lambda = \lambda_0 + Qx$ it reads $\widetilde{\Gamma} Q x = \phi^\star - \phi(\lambda_0)$, which has a solution x iff $\phi^\star - \phi(\lambda_0) \in \mathrm{range}(\widetilde{\Gamma} Q)$. □

Proof of Lemma 5. By Proposition 4, $\widetilde{\Gamma}\lambda = \phi^\star$ gives $q_\lambda = q^\star$, so $p_\lambda(z_i) = \rho_\lambda\, q^\star(z_i)$ by (3), and taking logarithms gives the identity. By (1), $\log p_\lambda(z_i) = \gamma_{z_i}^\top \lambda - A(\lambda)$, which is concave because A, a log-sum-exp of linear functions, is convex (its Hessian is $G(\lambda) \succeq 0$ by (7)); restricting to the affine shape set and subtracting the constant $\log q^\star(z_i)$ preserve concavity. Averaging the identity over i with weights $q^\star(z_i)$ gives $\log \rho_\lambda = H(q^\star) - \mathrm{CE}(q^\star, p_\lambda)$, which is (11) by (10) and does not depend on the choice of z_i. □

Proof of Proposition 6. By Proposition 4, the condition $q_\lambda = q^\star$ in (4) is $\widetilde{\Gamma}\lambda = \phi^\star$. On this set, Lemma 5 and monotonicity of the logarithm give $\rho_\lambda \geq \rho^\star$ iff $\log p_\lambda(z_i) \geq \log \rho^\star + \log q^\star(z_i)$, which proves the identity. The first two sets are affine, and the third is a sublevel set of the convex function $A(\lambda) - \gamma_{z_i}^\top \lambda$, hence convex; so is their intersection. □

Proof of Proposition 7. By Proposition 4, $\mathcal{C}_0 = \{\lambda \in \lambda_0 + \mathrm{range}(Q) : q_\lambda = q^\star\}$, which is nonempty iff $\phi^\star - \phi(\lambda_0) \in \mathrm{range}(\widetilde{\Gamma} Q)$. If $\mathcal{C}_0 = \varnothing$, then $\mathcal{C} = \varnothing$ and the first condition fails. Otherwise $\rho_{\max} = \sup_{\mathcal{C}_0} \rho_\lambda$, and $\mathcal{C} = \{\lambda \in \mathcal{C}_0 : \rho_\lambda \geq \rho^\star\}$ is nonempty iff some $\lambda \in \mathcal{C}_0$ has $\rho_\lambda \geq \rho^\star$, which by the definition of the supremum holds iff $\rho^\star < \rho_{\max}$, or $\rho^\star = \rho_{\max}$ and the supremum is attained. □

Proof of Proposition 8. The target label z_t is the unique mode of $q^\star$, so $q^\star(z) < \tau$ for $z \in Z \setminus \{z_t\}$. Every $\lambda \in \mathcal{C}$ has $q_\lambda = q^\star$ and $\rho_\lambda \geq \rho^\star$, so by (3), $p_\lambda(z) = \rho_\lambda q^\star(z) < \rho_\lambda \tau = p_\lambda(z_t)$ and $p_\lambda(z_t) \geq \rho^\star \tau$, while every $o \in \mathcal{O}$ has $p_\lambda(o) \leq p_\lambda(\mathcal{O}) = 1 - \rho_\lambda \leq 1 - \rho^\star$. Hence z_t is the unique maximizer whenever $\rho^\star \tau > 1 - \rho^\star$, i.e. $\rho^\star > (1 + \tau)^{-1}$. □

A.3 Section 6: Proofs and Supporting Results

Proof of Proposition 12. At $\zeta = t_i$, $s(t_i) = x_i$; write $a_j := d_S(x_j, x_i)^2$, so $a_i = 0$. (i) The normalizing constant of (9) cancels in the log-ratio, leaving $\frac{\beta}{2} a_j$, which is positive since $\beta > 0$ and d_S is a metric with $x_j \neq x_i$. (ii) Assume $x_j \neq x_i$ for all $j \neq i$, as the limit 1 requires. Then

$$q^\star_{t_i,\beta}(z_i) = \Big(1 + \sum\nolimits_{j \neq i} e^{-\beta a_j / 2}\Big)^{-1}$$

with every $a_j > 0$, so each summand is continuous and strictly decreasing in β, and $q^\star_{t_i,\beta}(z_i)$ increases continuously and strictly from $1/m$ at $\beta = 0$ towards 1 as $\beta \to \infty$. By the intermediate value theorem and strict monotonicity, each $\tau \in (1/m, 1)$ is attained at a unique root $\beta(\tau)$. □

A.3.1 Behavior Between Knots

The ranking in Proposition 12(i) holds at every position of the route.

Proposition 13 (Ranking along the route)**.** *For every $\zeta \in \mathcal{T}$, $\beta > 0$, and $z_j, z_k \in Z$,*

$$\log q^\star_{\zeta,\beta}(z_j) - \log q^\star_{\zeta,\beta}(z_k) = \tfrac{\beta}{2}\Big(d_S\big(x_k, s(\zeta)\big)^2 - d_S\big(x_j, s(\zeta)\big)^2\Big).$$

Hence $q^\star_{\zeta,\beta}$ ranks the labels by their distance to $s(\zeta)$, and its modes are the labels nearest $s(\zeta)$.

Proof. The normalizing constant of (9) cancels in the log-ratio, which is positive iff $d_S(x_j, s(\zeta)) < d_S(x_k, s(\zeta))$ since $\beta > 0$. □

Between adjacent knots t_i and t_{i+1} the mode is therefore z_i or z_{i+1} unless the route passes nearer a third label, which cannot happen when the route follows a declared cycle or order but must be checked for a graph or a fitted route. Calibration (Proposition 12(ii)) holds only at the knot where β is solved; elsewhere the modal confidence is set by the geometry.

A.4 Proofs for Appendix B

Proof of Proposition 14. For every $m \in \Delta^V$, (1) gives $\mathrm{CE}(m, p_\lambda) = -\sum_y m(y)\big(\gamma_y^\top \lambda - A(\lambda)\big) = A(\lambda) - (\Gamma^\top m)^\top \lambda$, of which (10) is the case $m = \tilde{q}^\star$. By the definition of KL divergence, $F(\lambda) = \mathrm{CE}(p_{\lambda_0}, p_\lambda) - H(p_{\lambda_0})$; applying the formula with $m = p_{\lambda_0}$, $\tilde{q}^\star$, and m_ν, and using $p_{\lambda_0} + \nu\tilde{q}^\star = (1+\nu)m_\nu$,

$$\begin{aligned} F(\lambda) + \nu\ \mathrm{CE}(q^\star, p_\lambda) &= (1+\nu)A(\lambda) - \big(\Gamma^\top (p_{\lambda_0} + \nu\tilde{q}^\star)\big)^\top \lambda - H(p_{\lambda_0}) \\ &= (1+\nu)\ \mathrm{CE}(m_\nu, p_\lambda) - H(p_{\lambda_0}). \end{aligned}$$

The second form follows from $\mathrm{CE}(m_\nu, p_\lambda) = D_{\mathrm{KL}}(m_\nu \,\|\, p_\lambda) + H(m_\nu)$, with $c(\nu) = (1+\nu)H(m_\nu) - H(p_{\lambda_0})$. □

Proof of Lemma 15. Here $\mathcal{C}_0 \neq \varnothing$, as the statement presupposes. Write $g(\lambda) := \mathrm{CE}(q^\star, p_\lambda)$, so $\rho_\lambda = \exp(H(q^\star) - g(\lambda))$ on $\mathcal{C}_0$ by (11), and by Proposition 14 $\lambda(\nu)$ minimizes $F + \nu g$ over $\mathcal{C}_0$. Both F and g depend on λ only through p_λ, so they are invariant along *distribution-preserving* directions v, those with $\Gamma v \in \mathrm{span}(\mathbf{1})$.

Well defined. $F + \nu g$ is convex by (8) and (10), and $F + \nu g \geq F$ since $g \geq 0$. Modulo distribution-preserving directions, F is coercive on $\mathcal{C}_0$: with $\delta := \min_y p_{\lambda_0}(y) > 0$, $F(\lambda) \geq -\delta \log \min_y p_\lambda(y) - H(p_{\lambda_0})$, and $\min_y p_\lambda(y) \to 0$ as λ leaves every bounded set of the quotient, since the spread of the logits $\Gamma\lambda$ then diverges. The sublevel sets of $F + \nu g$ are thus compact in the quotient, and a minimizer exists by the extreme value theorem. The Hessian of $F + \nu g$ is $(1+\nu)G(\lambda)$ by (8) and (10), and $v^\top G(\lambda) v = \mathrm{Var}_{p_\lambda}(\gamma^\top v) > 0$ unless $\Gamma v \in \mathrm{span}(\mathbf{1})$, so $F + \nu g$ is strictly convex on every segment joining states with different output distributions. The minimizing distribution, hence $\rho(\nu)$, is therefore unique. As $m_0 = p_{\lambda_0}$, $\rho(0)$ is the mass of the shape-only projection.

Monotone. For $\nu_1 < \nu_2$ with minimizers λ_1, λ_2, optimality gives $F(\lambda_1) + \nu_1 g(\lambda_1) \leq F(\lambda_2) + \nu_1 g(\lambda_2)$ and $F(\lambda_2) + \nu_2 g(\lambda_2) \leq F(\lambda_1) + \nu_2 g(\lambda_1)$; adding, $(\nu_2 - \nu_1)(g(\lambda_2) - g(\lambda_1)) \leq 0$, so $g(\lambda(\nu))$ is nonincreasing and $\rho(\nu)$ nondecreasing.

Continuous. Fix $\bar{\nu} > 0$ and $\hat{\lambda} \in \mathcal{C}_0$. For $\nu \in [0, \bar{\nu}]$, optimality and $g \geq 0$ place every minimizer in $K := \{\lambda \in \mathcal{C}_0 : F(\lambda) \leq F(\hat{\lambda}) + \bar{\nu} g(\hat{\lambda})\}$, which is compact modulo distribution-preserving directions by coercivity. By Berge's maximum theorem on K, the minimizer, unique in the quotient, is continuous in ν, and hence so is $\rho(\nu)$.

Limit. For $\nu > 0$ and any $\lambda \in \mathcal{C}_0$, optimality and $F \geq 0$ give $g(\lambda(\nu)) \leq g(\lambda) + F(\lambda)/\nu$. The limit of $g(\lambda(\nu))$ exists by monotonicity; letting $\nu \to \infty$ and then taking the infimum over λ shows that it is at most $\inf_{\mathcal{C}_0} g$, and it is at least $\inf_{\mathcal{C}_0} g$ since $\lambda(\nu) \in \mathcal{C}_0$. By Proposition 4, $\mathcal{C}_0$ is the set over which $\rho_{\max}$ is defined in Proposition 7, so $\inf_{\mathcal{C}_0} g = H(q^\star) - \log \rho_{\max}$ by (11), and $\rho(\nu) \to \rho_{\max}$. □

Proof of Proposition 16. Let g be as in the proof of Lemma 15 and $g^\star := H(q^\star) - \log \rho^\star$, so that $\mathcal{C} = \{\lambda \in \mathcal{C}_0 : g(\lambda) \leq g^\star\}$ by (11); the first hypothesis gives $\mathcal{C}_0 \neq \varnothing$ by Proposition 4.

Existence of $\nu^\star$. If $\rho(0) < \rho^\star$, then ρ is continuous with limit $\rho_{\max} > \rho^\star$ (Lemma 15), so the intermediate value theorem gives $\nu^\star > 0$ with $\rho(\nu^\star) = \rho^\star$.

Optimality. By Proposition 14, $\lambda^\star = \lambda(\nu^\star)$ minimizes $F + \nu^\star g$ over $\mathcal{C}_0$. If $\nu^\star = 0$, then $\rho(0) \geq \rho^\star$, so $\lambda^\star$ minimizes F over $\mathcal{C}_0 \supseteq \mathcal{C}$ and is feasible. Otherwise $\rho_{\lambda^\star} = \rho^\star$, so $\lambda^\star \in \mathcal{C}$, and every $\lambda \in \mathcal{C}$ has $g(\lambda) \leq g^\star = g(\lambda^\star)$, whence $F(\lambda^\star) \leq F(\lambda) + \nu^\star(g(\lambda) - g(\lambda^\star)) \leq F(\lambda)$. In both cases $\lambda^\star$ solves (6) and $\nu^\star(\rho_{\lambda^\star} - \rho^\star) = 0$.

Uniqueness of the distribution. $\mathcal{C}$ is convex (Proposition 6), and F is strictly convex on every segment joining states with different output distributions (proof of Lemma 15 with $\nu = 0$), so the midpoint of two solutions with different distributions would be feasible with strictly smaller objective.

Uniqueness of $\nu^\star$. Suppose $0 < \nu_1 < \nu_2$ with $\rho(\nu_1) = \rho(\nu_2) = \rho^\star > \rho(0)$. Then $g(\lambda(\nu_1)) = g(\lambda(\nu_2))$, the two optimality inequalities in the proof of Lemma 15 force $F(\lambda(\nu_1)) = F(\lambda(\nu_2))$, and so $\lambda_1 := \lambda(\nu_1)$ minimizes $F + \nu_k g$ over $\mathcal{C}_0$ for $k = 1, 2$. With P the orthogonal projector onto $\{v \in \mathrm{range}(Q) : \widetilde{\Gamma} v = 0\}$, the direction space of $\mathcal{C}_0$, first-order optimality gives $P\nabla F(\lambda_1) + \nu_k P \nabla g(\lambda_1) = 0$ for $k = 1, 2$, hence $P\nabla g(\lambda_1) = P\nabla F(\lambda_1) = 0$. By convexity, λ_1 then minimizes F over $\mathcal{C}_0$, so $\rho(\nu_1) = \rho(0)$ by the uniqueness in Lemma 15, a contradiction. □

B SOLVER

We solve (6) through the multiplier ν of its mass constraint: at fixed ν the inner problem is a KL projection onto the shape set $\mathcal{C}_0$, and the achieved mass is monotone in ν, so ν is found by bisection. Notation is as in Appendix A; proofs are in Appendix A.4.

B.1 MIXTURE FORM AND MONOTONE MASS

Proposition 14 (Mixture form of the Lagrangian). *For $\nu \geq 0$ let $m_\nu := (p_{\lambda_0} + \nu\tilde{q}^\star)/(1+\nu) \in \Delta^V$. Then for every λ,*

$$F(\lambda) + \nu\ \mathrm{CE}(q^\star, p_\lambda) = (1+\nu)\ \mathrm{CE}(m_\nu, p_\lambda) - H(p_{\lambda_0}) = (1+\nu)\, D_{\mathrm{KL}}(m_\nu \,\|\, p_\lambda) + c(\nu),$$

with $c(\nu)$ independent of λ.

The multiplier thus mixes the target conditional into the reference, and the inner problem projects the mixture m_ν onto $\{p_\lambda : \lambda \in \mathcal{C}_0\}$. Its objective is nonnegative and its Hessian is $G(\lambda)$ for every ν, so Newton's method on it is Fisher-metric descent. The achieved mass is monotone in ν:

Lemma 15 (Monotone mass). *Let $\lambda(\nu)$ minimize $D_{\mathrm{KL}}(m_\nu \,\|\, p_\lambda)$ over $\mathcal{C}_0 \neq \varnothing$. Then $\rho(\nu) := \rho_{\lambda(\nu)}$ is well defined, continuous, and nondecreasing, with $\rho(0)$ the mass of the shape-only projection and $\rho(\nu) \to \rho_{\max}$ as $\nu \to \infty$.*

Proposition 16 (Projection by bisection on the multiplier). *Suppose $\phi^\star - \phi(\lambda_0) \in \mathrm{range}(\widetilde{\Gamma}Q)$ and $\rho^\star < \rho_{\max}$. (i) If $\rho(0) < \rho^\star$, there is a unique $\nu^\star > 0$ with $\rho(\nu^\star) = \rho^\star$; otherwise set $\nu^\star := 0$. (ii) $\lambda^\star := \lambda(\nu^\star)$ solves (6), and every solution has output distribution $p_{\lambda^\star}$. (iii) $\nu^\star$ is a multiplier for the mass constraint: $\lambda^\star$ minimizes $F + \nu^\star\,\mathrm{CE}(q^\star, p_\cdot)$ over $\mathcal{C}_0$, and $\nu^\star(\rho_{\lambda^\star} - \rho^\star) = 0$.*

Bisection on ν therefore converges to $\nu^\star$. At the boundary $\rho^\star = \rho_{\max}$, feasibility also requires the supremum defining $\rho_{\max}$ to be attained (Proposition 7), and $\rho(\nu)$ may reach $\rho^\star$ only as $\nu \to \infty$; the solver reports BOUNDARY rather than bisecting on a diverging multiplier.

B.2 ALGORITHM

Algorithm 1 Projection steering (Definition 9)

Require: reference λ_0, support Q, constraints $(q^\star, \rho^\star)$, tolerances $\epsilon_{\mathrm{shape}}, \epsilon_\rho$

$\phi^\star \leftarrow H_m \log q^\star$; solve $\widetilde{\Gamma}Qx_p = \phi^\star - \widetilde{\Gamma}\lambda_0$ by least squares
if residual $> \epsilon_{\mathrm{shape}}$ **return** INFEASIBLE(SHAPE)
$N \leftarrow$ basis of $\ker(\widetilde{\Gamma}Q)$; parameterize $\mathcal{C}_0$ by $\lambda(y) = \lambda_0 + Q(x_p + Ny)$
$\rho_{\max} \leftarrow \exp\left(H(q^\star) + \sup_y[\bar{\gamma}_{q^\star}^\top \lambda(y) - A(\lambda(y))]\right)$ (reachability)
if $\rho_{\max} < \rho^\star$ **return** INFEASIBLE(MASS) with $\rho_{\max}$
if $\rho_{\max} \leq \rho^\star + \epsilon_\rho$ **return** BOUNDARY with $\rho_{\max}$
$y_0 \leftarrow \arg\min_y D_{\mathrm{KL}}\left(p_{\lambda_0} \,\|\, p_{\lambda(y)}\right)$ (shape-only projection)
if $\rho_{\lambda(y_0)} \geq \rho^\star$ **return** $\lambda(y_0)$, $\nu^\star = 0$
bracket $\nu \in [\nu_{\mathrm{lo}}, \nu_{\mathrm{hi}}]$ with $\rho(\nu_{\mathrm{lo}}) < \rho^\star < \rho(\nu_{\mathrm{hi}})$, doubling ν_{hi} as needed
repeat
  $\nu \leftarrow (\nu_{\mathrm{lo}} + \nu_{\mathrm{hi}})/2$; $y \leftarrow \arg\min_y D_{\mathrm{KL}}\left(m_\nu \,\|\, p_{\lambda(y)}\right)$, warm-started from the previous y
  if $\rho_{\lambda(y)} < \rho^\star$ **then** $\nu_{\mathrm{lo}} \leftarrow \nu$ **else** $\nu_{\mathrm{hi}} \leftarrow \nu$
until $|\rho_{\lambda(y)} - \rho^\star| \leq \epsilon_\rho$
return $\lambda(y)$, $\nu^\star = \nu$

Notes. (i) The reachability problem is a concave maximization, solved by Newton's method or L-BFGS. Every inner problem minimizes $\mathrm{CE}(m_\nu, p_{\lambda(y)})$ over y by damped Newton, with $m_0 = p_{\lambda_0}$ for the shape-only projection; its gradient is $N^\top Q^\top(\mu(\lambda) - \Gamma^\top m_\nu)$ and its Hessian $N^\top Q^\top G(\lambda) Q N$ for every ν. With $V \approx 2.6 \times 10^5$, G is formed from the top-K tokens ($K \approx 2 \times 10^4$) with negligible error (Park et al., 2026), and the Newton system is solved by conjugate gradients.

(ii) The reachability step certifies feasibility, and the shape-only step detects an inactive floor ($\nu^\star = 0$); both are needed.

(iii) Every returned λ comes with a certificate: $\|\widetilde{\Gamma}\lambda - \phi^\star\|$, ρ_λ, $\nu^\star$, $\rho_{\max}$, the margin $p_\lambda(z_t) - \max_o p_\lambda(o)$, and the three terms of Lemma 1 at λ.

C Additional Experiments and Details

C.1 Metric Details

Metrics. Route error W_1^c is computed by an exact transport LP using the declared concept structure. We report the two components of target error separately at the final route position and average route error over the full route. A position counts as a *success* when its in-domain conditional KL is at most 10^{-6}, its mass shortfall is at most 2×10^{-7}, and its numerical certificate succeeds.

Off-target comparisons. Off-target error alone is only descriptive when methods realize different behaviors (q, ρ). We therefore also run projection steering from the same reference state using a baseline's realized behavior (q_b, ρ_b) as the behavioral constraints. We compare off-target errors only when the projection matches both quantities (to 10^{-7} and 2×10^{-7}, respectively) and the mass floor is active, as in Theorem 10.

Figures. Appendix figures follow the layout of Figure 2 and show the first held-out prompt. In the off-target panel, the dashed line shows the matched projection described above. Values below 10^{-16} for the in-domain conditional KL and 10^{-4} for off-target error are treated as numerical zeros.

C.2 Semantic Route Construction

Geometry. The unembedding and synthetic tasks use the readout geometry of §6: label z_i is placed at its representer coordinate d_i, truncated to the dimension in Table 1. The probe taxonomy uses the declared geometry, with d_S the path length in the taxonomic tree: its routes pass through orders, classes, and phyla, which are vertices of the declared tree but have no readout row. Route error is scored as in Appendix C.1.

Routes. For cycles we fit a smooth closed curve through the label points, so the step from December to January is like every other step; for orders the same curve is left open. On the declared tree, $s(\zeta)$ moves along edges at unit speed, so an internal vertex is itself a route position; there the bump assigns equal mass to all families at equal path length, which is what makes a broader group a meaningful intermediate state. One step between neighboring vertices is one unit of ζ, and every route is evaluated at quarter-steps, endpoints included.

Calibration. The bump of Definition 11 needs a width β, which the geometry does not supply. At each labeled position t_i we solve $q^\star_{t_i,\beta_i}(z_i) = \tau$, which has a unique root by Proposition 12(ii), and interpolate $\log \beta_i$ linearly in between. Confidence is therefore τ at every labeled position and set by the geometry elsewhere (Appendix A.3.1); τ is fixed in advance. To keep the tail of the target representable in float64, we cap $\beta(\zeta) \max_j d_S(x_j, s(\zeta))^2/2$ at 700.

C.3 Data, Prompts, and Baselines

Readouts and labels. Unembedding experiments use the last-position input to the output head of Gemma-3-4B, after final normalization, so Γ contains all 262,208 vocabulary rows. Readout evaluations and projections use float64. Each label is a single vocabulary token that decodes exactly to the label. Case variants, abbreviations, and other variants remain outsiders, and multi-token labels are excluded.

Prompts. Each label has eight templates, split by wording into four training, two development, and two test templates (Table 1), so evaluation tests transfer across wordings. Training templates include lists, enumerations, previous- or next-item statements; some contain the label. Test templates use held-out wordings such as "The day before [next day] is", "The month after [previous month] is", and "[previous number word] plus ten more is".

Readout/domain	$\lvert Z\rvert$	Geometry dim.	Support dim.	Train	Dev.	Test
Unembedding: weekdays	7	2	27	28	14	14
Unembedding: months	12	2	32	48	24	24
Unembedding: ages	9	2	29	36	18	18
Probe: taxonomy	46	—	60	3,050	741	817
Synthetic helix, per seed	24	3	128	192	72	72

Table 1: Data and supports. Geometry dimension is the number of representer coordinates used by the readout geometry; the probe uses the declared taxonomy instead. Support dimension is $\mathrm{rank}(Q)$. For unembedding experiments, Q contains the centered label rows, the mean label row, twelve training principal components, and eight seeded random directions. For the probe, Q is the complete centered span of its 61 class rows; additional directions lie in the readout null space and therefore cannot change the probe distribution. Probe counts are Wikipedia articles, including the 15 outsider classes.

Support and behavioral constraints. For unembeddings, Q orthonormalizes the centered label rows, the mean label row, twelve principal components of the training activations, and eight random directions. For the probe, Q is the centered span of all 61 probe class rows. All methods share the same support. Because $\mathrm{range}(Q)$ contains the row space of the centered in-domain readout, which has full rank $m-1$ in every task, every target conditional is reachable (Proposition 4); only the mass constraint can be infeasible. Unless stated otherwise, $(\tau, \rho^\star) = (0.6, 0.7)$, and every route position is projected independently from the reference state λ_0.

Baselines. *Activation steering* adds the difference between destination and source training centroids, while *linear-probe steering* whitens this difference by the within-class covariance at equal norm. For both, the scale at each position is chosen on development prompts from $\{2^{-8}, \dots, 2^5\}$. *Dual steering* takes Fisher steps along a ridge-regularized binary probe direction toward the modal target probability $\rho^\star q_\zeta^\star(z_t)$, for at most 3,000 steps. *Fitted-manifold steering* sets the support coordinates to the training-centroid spline (the harmonic path on trees), while *geodesic steering* follows a geodesic to the destination centroid under a metric learned from the Hellinger geometry of unsteered output distributions. We also report the unsteered model and projection without a mass constraint (generalized dual steering). Centroids, probes, supports, targets, and scales are fixed before test evaluation.

C.4 Synthetic: Twenty-Four-Label Helix

Setup. The readout has $d = 128$, $V = 4{,}000$, and $m = 24$ labels. With $\theta_i = 4\pi i/23$ and an orthonormal basis $e_1, \dots, e_{128}$,

$$\gamma_{z_i} = 3\cos\theta_i\, e_1 + 3\sin\theta_i\, e_2 + \frac{3\theta_i}{2\pi} e_3 + 3e_4 + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}\left(0, 0.02^2 \sum_{k=5}^{128} e_k e_k^\top\right),$$

and outsider rows are $\mathcal{N}(0, 0.3^2/d)$ with zero fourth coordinate. The first three coordinates form a two-turn helix, while e_4 changes domain mass without changing the in-domain or outsider conditional. The declared chain $z_0 \to \cdots \to z_{23}$ follows the helix. Each source label is routed to the farther endpoint, with $Q = I_{128}$. Training activations realize a noisy squared-distance logit field with domain mass 0.03, and test reference states are $\mathcal{N}(0, I_{128})$ draws.

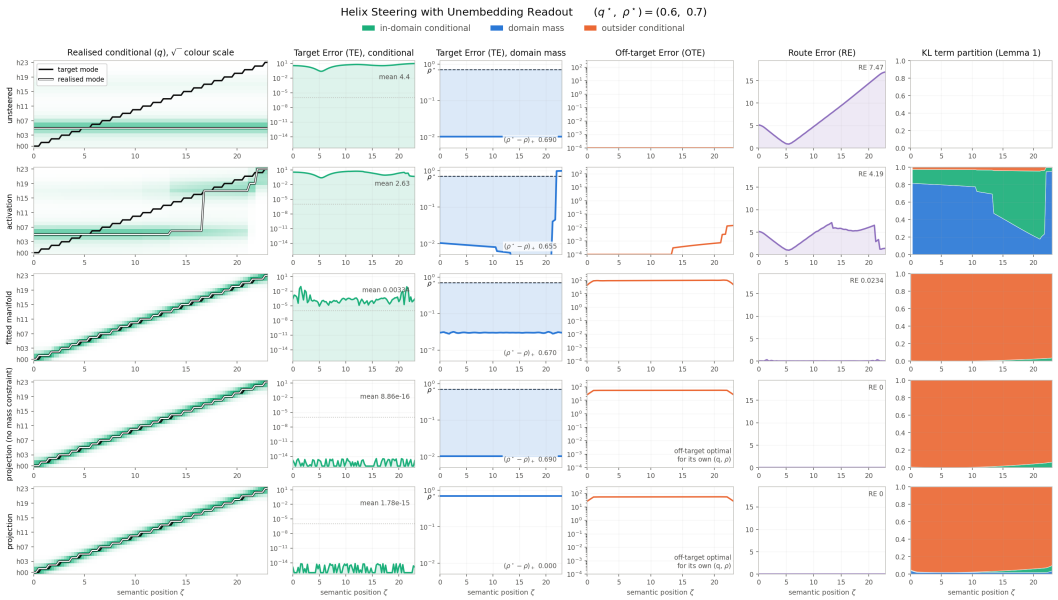


Figure 3: **Synthetic helix steering** along the semantic route $z_0 \to \cdots \to z_{23}$ at $(\tau, \rho^\star) = (0.6, 0.7)$. Both projections follow the target conditional; only projection steering also satisfies the mass constraint.

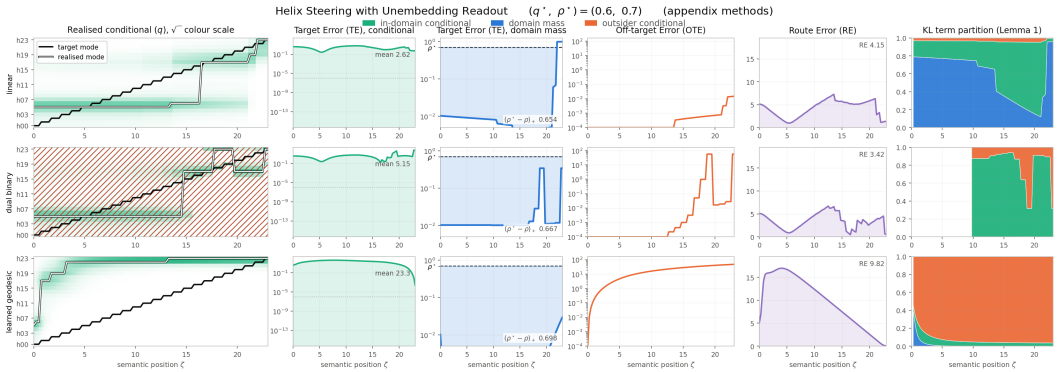


Figure 4: **Synthetic helix steering: remaining methods.** Linear-probe steering, dual steering, and geodesic steering on the route and reference state of Figure 3.

C.5 UNEMBEDDINGS: MONTHS OF THE YEAR

The labels are January through December, arranged as a cycle. Section 7 reports the main comparison in Figure 2; Figure 5 shows the remaining methods.

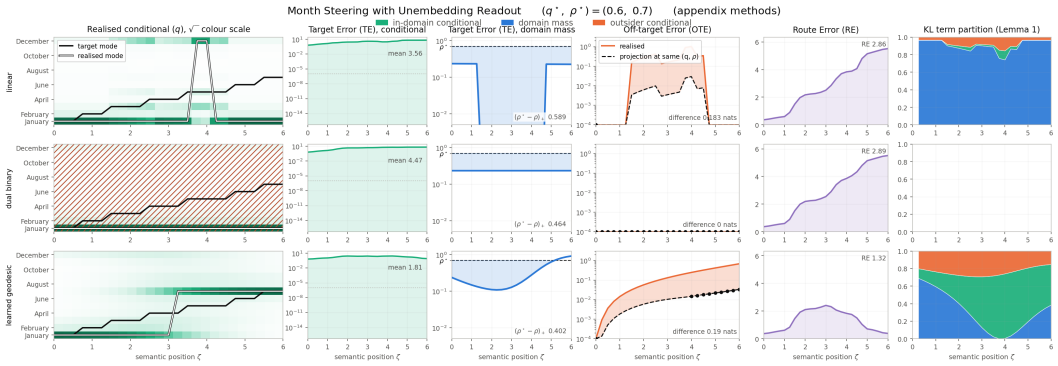


Figure 5: **Month steering with the unembedding readout: remaining methods.** Linear-probe steering, dual steering, and geodesic steering on the route and prompt of Figure 2.

C.6 UNEMBEDDINGS: DAYS OF THE WEEK

The labels are the seven weekdays arranged as a cycle, and each source label advances three days. As shown in Figure 6, projection steering satisfies the specified behavioral constraints across held-out prompts. Figure 7 shows the remaining methods.

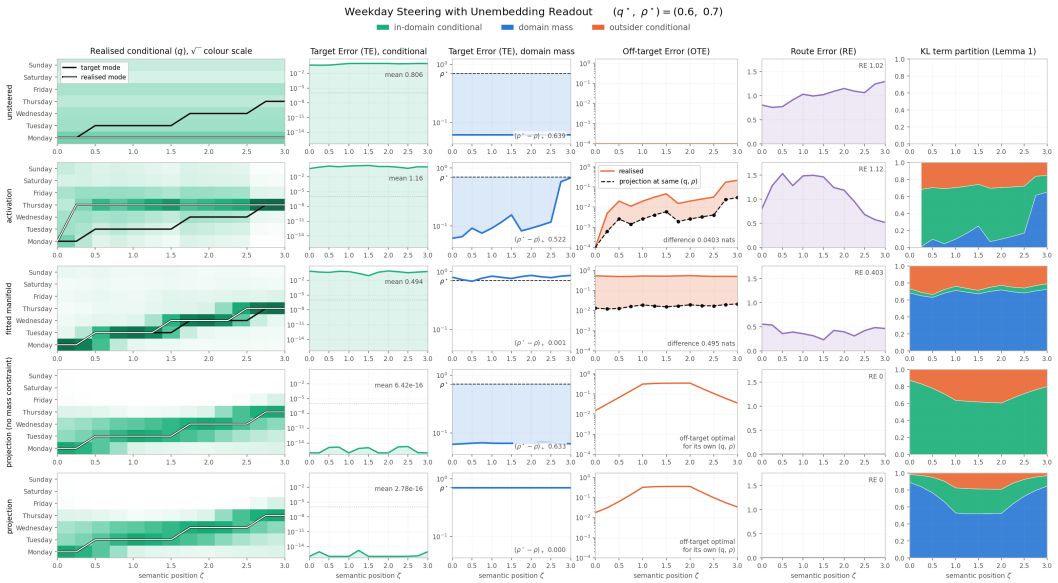


Figure 6: **Weekday steering with the unembedding readout** along the semantic route from Monday to Thursday, for the prompt "The day before Tuesday is".

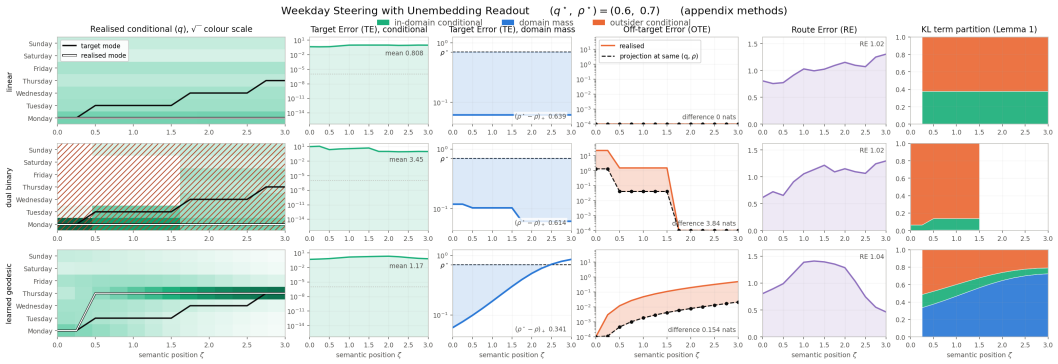


Figure 7: **Weekday steering with the unembedding readout: remaining methods.** Linear-probe steering, dual steering, and geodesic steering on the route and prompt of Figure 6.

C.7 Unembeddings: Ages

The labels are the decade words *ten*, . . . , *ninety*, arranged as a unit-spaced order and elicited using counting and spelling prompts rather than age inference. Each source label is routed to the farther endpoint. Figure 8 shows projection steering along this order, and Figure 9 shows the remaining methods.

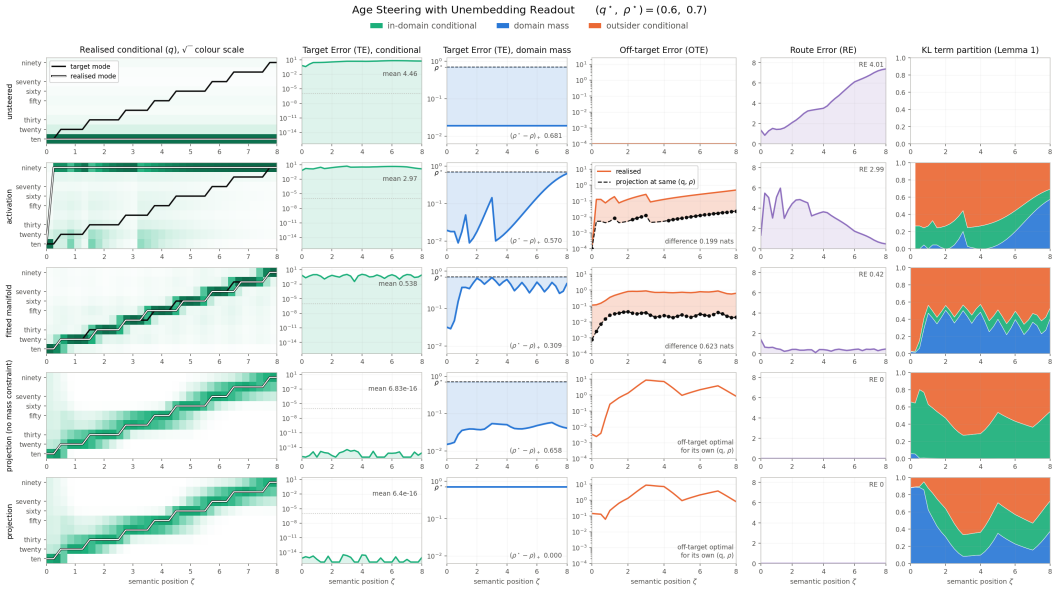


Figure 8: **Age steering with the unembedding readout** along the semantic route from *ten* to *ninety*. Projection steering realizes the ordered target distributions while satisfying the behavioral constraints.

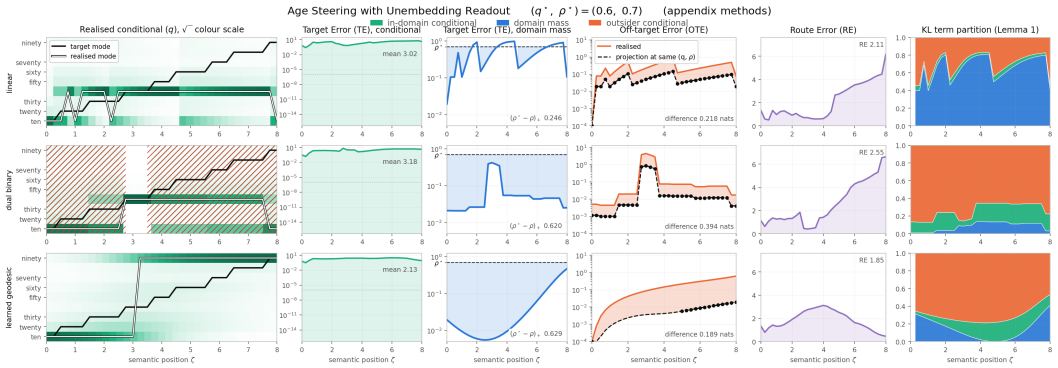


Figure 9: **Age steering with the unembedding readout: remaining methods.** Linear-probe steering, dual steering, and geodesic steering on the route and prompt of Figure 8.

C.8 Probes: Animal Taxonomy

Taxonomy. The concept domain contains 46 animal families arranged in a declared taxonomic tree. Families are leaves, while broader groups such as orders, classes, phyla, and *Animalia* are internal vertices. These internal groups are not probe labels, but they define meaningful intermediate positions along a route between families.

Data. Each animal example is an English Wikipedia article for one species, labeled by its family. To prevent the probe from reading the answer directly, we mask names that identify the species or its taxonomic groups, including scientific and common names, the genus, article title, and ancestor names. Thus an input may look like *"The [MASK] is a large [MASK] native to Africa and India…"*, and the probe must infer the family from the remaining description. The dataset contains 3,562 animal articles across the 46 families and 1,046 articles from 15 non-animal outsider classes. Splits are made by species.

Probe and routes. We fit a 61-class multiclass probe at layer 16, with the 46 animal families as the concept domain and the remaining 15 classes as outsiders. The probe reaches 69.9% accuracy on 817 held-out articles. Each source family is connected to a distant family by the unique path through the declared taxonomy, and intermediate positions may therefore lie at broader taxonomic groups rather than at family labels. Targets use the declared geometry, with d_S the path length in the taxonomy.

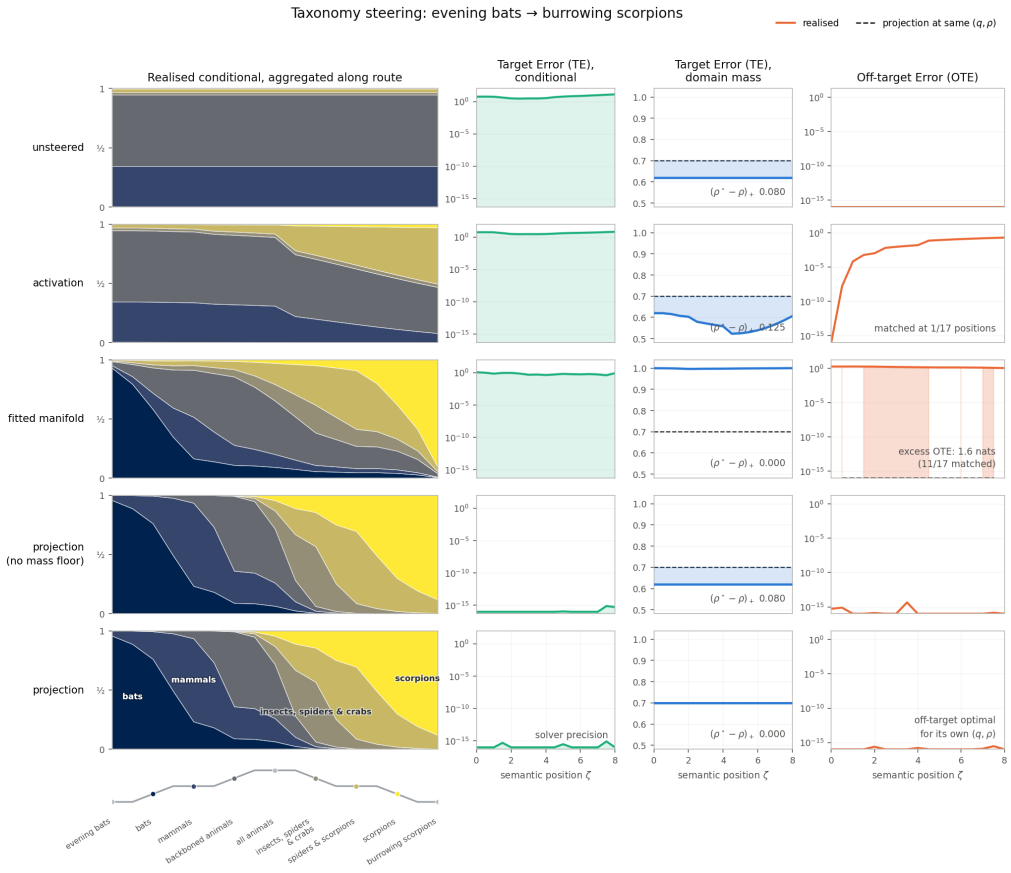


Figure 10: **Taxonomy steering with a multiclass probe readout.** The displayed route runs from evening bats to burrowing scorpions through broader groups in the animal taxonomy. Columns: **(a)** Realized conditional over disjoint groups along the route; colors follow the path from source to destination, and the curve below shows its taxonomic rank. **(b)** Target error for the conditional distribution. **(c)** Realized domain mass ρ against the required minimum $\rho^\star$. **(d)** Off-target error. The dashed curve shows projection steering matched to each method's realized (q, ρ), and shading shows the excess off-target error where this comparison applies.