A Geometric Account of Activation Steering

Anonymous Authors[1]

Abstract

Linear activation steering has gained popularity as a simple and empirically effective way to control language model behavior. More recently, spherical steering paradigms have been proposed to address limitations of additive interventions, often motivated by the assumption that hidden-state norm does not carry concept-relevant information. In this work, we revisit this assumption through a theoretical analysis of existing steering methods and a controlled empirical study designed to disentangle the roles of angular and radial components. We show that steering methods differ mainly in how they couple two geometric effects: changing a token's angular alignment with a concept direction and changing its hidden-state norm. Across seven language models, we find that concepts are represented primarily in angular structure, supporting the motivation for spherical methods, but that norm remains important for the stability and downstream effects of steering. Our results explain why methods with similar concept-level effects can behave differently, and suggest that activation steering should be understood as a geometric control problem over token representations rather than as the choice of a single additive strength.

1. Introduction

Linear activation steering has become a popular approach for controlling the behavior of language models at inference time. In its simplest form, a steering direction is estimated from contrastive examples and then added to hidden states with some scalar strength,

$$x \mapsto x + \alpha s,$$

where x is a token representation, s is a steering direction, and α is a global coefficient. This family of methods is attractive because it is lightweight, training-free, and often empirically effective. However, despite its simplicity, the mechanism behind activation steering remains poorly understood. In particular, the scalar coefficient α does not directly describe what happens geometrically to the representation being steered.

The main difficulty is that additive steering entangles two different effects. First, it changes the angular alignment between the hidden state and the steering direction, which can be interpreted as changing the degree to which the target concept is expressed. Second, it changes the hidden-state norm, which can affect how the modified representation is processed by the remaining layers of the model. These two effects are coupled by the additive update: increasing α simultaneously changes both the angle and the norm, and the same value of α can have very different consequences for different tokens depending on their initial norm and alignment. Thus, the usual steering coefficient is not an interpretable control parameter.

Recent spherical and angular steering paradigms attempt to address this issue by replacing additive updates with norm-preserving transformations. These approaches are motivated by the idea that concepts are represented primarily in direction rather than in norm. If true, then steering should be understood mostly as an angular operation: one should rotate a hidden state toward or away from a concept direction while preserving its magnitude. This view provides a compelling alternative to linear steering, but it also relies on a strong assumption: that norm is not important for concept representation or for successful intervention. In this work, we revisit this assumption.

We argue that activation steering should be analyzed as a two-component geometric problem. For each token representation, a steering method implicitly decides two things: how much to change the token's angular alignment with the concept direction, and how much to change its norm. The angular component determines the strength of concept expression, while the radial component determines how the scale of the representation is modified. This decomposition allows us to separate two questions that are usually confounded: whether a method successfully controls the target concept, and whether it preserves the representation scale needed for stable model behavior.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Using this perspective, we provide a theoretical analysis of existing steering methods. We show that common methods differ in whether they match a desired per-token concept score and whether they preserve hidden-state norm. Standard additive steering controls neither quantity directly. Normalized or angular variants may preserve norm, but they do not necessarily ensure that each token reaches the same concept score. Matched additive variants can control the concept score, but generally do so by changing the norm. Spherical steering occupies the corner of this design space that both preserves norm and matches a target angular concept score. This taxonomy clarifies that many empirical differences between steering methods arise from how they couple or decouple angular and radial changes.

This framework also lets us examine the central assumption behind spherical steering. Across seven language models, we test whether concept information is encoded in direction or in norm by comparing probes on raw hidden states, normalized hidden states, and norm-only features. We find that concepts are represented primarily in angular structure, supporting the motivation for spherical methods. However, norm is not irrelevant: even when it does not carry the target concept directly, changing the norm can substantially affect downstream stability and degradation. Thus, norm should not be treated as concept-free noise; it is a separate geometric component that matters for the effect of intervention.

Our goal is therefore not simply to introduce another steering method, but to understand what steering methods are doing and why some work better than others. The analysis suggests that the core problem of activation steering is not choosing a single global additive coefficient. Rather, it is deciding, for each token, how far its representation should move in angle toward the concept direction and how its norm should be handled during that move. This view replaces opaque additive strengths with interpretable geometric quantities and provides a controlled framework for comparing existing steering methods.

Our contributions are as follows:

- We analyze existing activation steering methods through a two-component geometric decomposition into angular concept change and radial norm change.
- We show that common steering methods can be organized by two interpretable properties: per-token concept-score matching and norm preservation.
- We empirically test the assumption that concepts live in angle rather than norm across seven language models, finding that concept information is primarily angular but that norm remains important for intervention stability.
- We use this framework to explain why steering methods with similar apparent strengths can behave differently, depending on how they couple angular movement with norm change.

2. Related Work

Activation steering and representation engineering. Activation steering modifies internal activations at inference time rather than updating model weights. Methods such as ITI, ActAdd, CAA, and Representation Engineering use directions in activation space to elicit behaviors including truthfulness, sentiment, topic, refusal, and other high-level attributes (Li et al., 2023; Turner et al., 2023; Panickssery et al., 2023; Zou et al., 2023). CAA is especially relevant to our work: it estimates a steering vector as a contrastive mean difference between positive and negative examples and applies it as an additive update during generation (Panickssery et al., 2023; Rimsky et al., 2024). These methods are simple and empirically effective, but their usual control parameter is an additive strength whose geometric meaning is unclear.

Linear concept representations. The success of activation steering is connected to the linear representation hypothesis: many model behaviors can be associated with directions in hidden-state space, and projections onto these directions can act as concept scores (Park et al., 2024). This motivates contrastive direction extraction and probing-based analyses. However, a useful direction does not by itself specify how to intervene. Adding along the direction changes both the normalized alignment with the concept and the hidden-state norm. Our work focuses on separating these two effects, asking when concepts are angular and when norm still matters for the consequences of intervention.

Angular, spherical, and adaptive steering. Recent work has proposed replacing additive shifts with angular or spherical transformations. Angular Steering rotates activations in a behavior-related subspace (Vu & Nguyen, 2025), while Spherical Steering performs a norm-preserving geodesic rotation toward a target direction (You et al., 2026). These works provide practical evidence that norm-aware or norm-preserving interventions can be more stable than vanilla addition. However, the existing analysis is mostly method-specific: it does not give a unified theory of how additive, normalized, matched, angular, and spherical methods relate, nor does it formally separate concept-score matching from norm preservation. Adaptive and selective steering methods likewise show that a single global coefficient is often insufficient (Dang & Ngo, 2026), but leave open which geometric quantity should be adapted. We address this gap by analyzing existing methods in a common angular–radial framework.

3. Methodology

We analyze activation steering methods as geometric transformations of token representations. Let $x \in \mathbb{R}^d$ denote a hidden state at a chosen layer and token position, and let $s \in \mathbb{R}^d$ denote a steering direction. Our goal is to understand how different steering methods modify two quantities: the angular alignment of x with the concept direction s, and the norm $\|x\|_2$. The former controls the degree to which the target concept is expressed, while the latter controls the radial scale of the representation.

3.1. Contrastive Activation Addition

Contrastive Activation Addition (CAA) constructs a steering direction from paired examples that differ in a target concept (Rimsky et al., 2024). Given positive examples $\mathcal{D}^+$ and negative examples $\mathcal{D}^-$, let $h_\ell(z)$ denote the hidden representation of example z at layer ℓ. A contrastive steering vector is computed as

$$\Delta_\ell = \frac{1}{|\mathcal{D}^+|} \sum_{z \in \mathcal{D}^+} h_\ell(z) - \frac{1}{|\mathcal{D}^-|} \sum_{z \in \mathcal{D}^-} h_\ell(z). \tag{1}$$

We use the normalized direction

$$s = \frac{\Delta_\ell}{\|\Delta_\ell\|_2}. \tag{2}$$

At inference time, CAA applies an additive intervention

$$\widetilde{x} = x + \alpha s, \tag{3}$$

where $\alpha \in \mathbb{R}$ is a scalar steering strength.

Although α is the standard control parameter, it does not separately determine the angular and radial effects of the intervention. To see this, write

$$r := \|x\|_2, \qquad u := \frac{x}{\|x\|_2}, \qquad c := \langle u, s \rangle. \tag{4}$$

Here c is the normalized concept score of the token representation. After the additive update, the new norm is

$$\|\widetilde{x}\|_2 = \sqrt{r^2 + 2\alpha r c + \alpha^2}, \tag{5}$$

and the new normalized concept score is

$$\widetilde{c} = \left\langle \frac{\widetilde{x}}{\|\widetilde{x}\|_2}, s \right\rangle = \frac{rc + \alpha}{\sqrt{r^2 + 2\alpha r c + \alpha^2}}. \tag{6}$$

Thus, a fixed additive coefficient α simultaneously changes both the concept score and the norm. Moreover, the achieved score $\widetilde{c}$ depends on the token-specific initial norm r and initial concept score c. Consequently, the same value of α does not correspond to the same conceptual intervention across tokens.

Equations (5) and (6) show why additive steering is geometrically entangled. CAA is not only an angular movement toward the concept direction; it also induces a token-dependent radial change. In particular, for positive steering strengths,

$$\alpha \to \infty \quad \Longrightarrow \quad \widetilde{c} \to 1 \quad \text{and} \quad \|\widetilde{x}\|_2 \to \infty. \tag{7}$$

Thus, driving the representation closer to perfect alignment with the concept direction by increasing α necessarily sends the hidden-state norm outside the typical activation scale. This makes α difficult to interpret: increasing α may improve concept control, but it may also move hidden states to norm regimes that differ substantially from those produced by the original model. Our analysis therefore treats CAA as the baseline method in which angular and radial effects are coupled by construction.

3.2. Spherical Steering

Spherical steering replaces additive displacement with a norm-preserving angular transformation (You et al., 2026). The motivation is that, if concept information is primarily encoded in direction, then steering should directly control the angle between the hidden state and the concept direction while avoiding unnecessary changes in norm.

Let $x \neq 0$, and define

$$r := \|x\|_2, \qquad u := \frac{x}{\|x\|_2}, \qquad c := \langle u, s \rangle. \tag{8}$$

We decompose u into its component along s and its orthogonal component:

$$u = cs + \sqrt{1 - c^2}\, v, \qquad v := \frac{u - cs}{\|u - cs\|_2}, \tag{9}$$

where $v \in s^\perp$ whenever $|c| \neq 1$. The vector v is not treated as an independent steering parameter; it is simply the remaining direction needed to preserve the part of the representation not explained by the current concept direction.

Given a target normalized concept score $\gamma \in [-1, 1]$, spherical steering defines the intervention

$$S_\gamma(x) = r\left(\gamma s + \sqrt{1 - \gamma^2}\, v\right). \tag{10}$$

This update has two immediate properties:

$$\|S_\gamma(x)\|_2 = \|x\|_2, \tag{11}$$

$$\left\langle \frac{S_\gamma(x)}{\|S_\gamma(x)\|_2}, s \right\rangle = \gamma. \tag{12}$$

Thus, spherical steering exactly separates the two quantities that are entangled in CAA: the angular concept score is set to γ, while the norm is preserved.

This operation can also be characterized as a minimal angular edit. Let

$$d_{\mathbb{S}}(u, w) := \arccos(\langle u, w \rangle) \tag{13}$$

denote geodesic distance on the unit sphere. Then $S_\gamma(x)$ is the solution of

$$\begin{aligned} \min_{y \in \mathbb{R}^d} \quad & d_{\mathbb{S}}\left(\frac{x}{\|x\|_2}, \frac{y}{\|y\|_2}\right) \\ \text{s.t.} \quad & \|y\|_2 = \|x\|_2, \\ & \left\langle \frac{y}{\|y\|_2}, s \right\rangle = \gamma. \end{aligned} \tag{14}$$

Therefore, among all same-norm representations that reach the desired concept score, spherical steering chooses the closest one in angular distance.

The geodesic formulation is useful because it makes explicit what spherical steering controls. Let

$$\theta := \arccos(c) \tag{15}$$

be the initial angle between x and the concept direction. Moving a fraction $t \in [0, 1]$ of the way from u toward s along the geodesic corresponds to the target score

$$\gamma_t = \cos((1 - t)\theta). \tag{16}$$

Substituting γ_t into (10) yields the usual geodesic rotation toward the concept direction. In contrast to the additive coefficient α, the parameter γ, or equivalently the target angle $\arccos(\gamma)$, directly specifies the intended angular concept level.

Importantly, norm preservation should not be interpreted as the claim that norm is irrelevant. Rather, spherical steering provides a controlled reference intervention: it changes the angular concept score while holding the norm fixed. Comparing spherical steering with additive methods therefore lets us isolate the effect of radial norm changes. This is central to our analysis: concepts may primarily live in angular structure, but norm can still affect the stability and downstream consequences of an intervention.

3.3. CAA and Spherical Modifications

The previous two subsections describe the two endpoints of our analysis: additive steering, which couples angular and radial changes through a single coefficient α, and spherical steering, which fixes a target concept score γ while preserving norm. We now introduce the family of variants used in our controlled comparisons. These variants are not intended as independent new algorithms; rather, they isolate the two geometric choices that a steering method makes: whether it matches a desired per-token concept score, and whether it preserves the hidden-state norm. Table 1 summarizes this two-axis design.

Table 1. Two-axis taxonomy of the steering methods used in our controlled analysis. Here "per-token γ" means that the method explicitly matches a desired normalized concept score for each token.

Method	Norm-preserving	Per-token γ
CAA	$-$	$-$
CAA-r	$+$	$-$
CAA-m	$-$	$+$
CAA-mr	$+$	$+$
S	$+$	$+$
AS	$+$	$-$
SN	$-$	$+$

Let

$$q := \sqrt{1 - c^2}, \qquad \theta := \arccos(c), \tag{17}$$

where $c = \langle x/\|x\|_2, s \rangle$ is the initial normalized concept score. For a target score $\gamma \in (-1, 1)$, the additive coefficient that makes CAA reach exactly $\widetilde{c} = \gamma$ is

$$\alpha_\gamma = r\left(\frac{\gamma\sqrt{1 - c^2}}{\sqrt{1 - \gamma^2}} - c\right). \tag{18}$$

We call the resulting intervention *matched CAA*:

$$\text{CAA-m}_\gamma(x) = x + \alpha_\gamma s. \tag{19}$$

By construction, CAA-m matches the same target concept score as spherical steering. However, it does not preserve norm. Its norm is

$$\|\text{CAA-m}_\gamma(x)\|_2 = r\frac{\sqrt{1 - c^2}}{\sqrt{1 - \gamma^2}}, \tag{20}$$

so the norm diverges as $\gamma \to 1$ unless the token is already perfectly aligned with s. Thus, CAA-m isolates the effect of radial distortion under fixed angular control.

A natural modification is to renormalize the matched additive update back to the original radius:

$$\text{CAA-mr}_\gamma(x) = r\frac{x + \alpha_\gamma s}{\|x + \alpha_\gamma s\|_2}. \tag{21}$$

This method both matches the target score and preserves norm. It is algebraically identical to spherical steering,

$$\text{CAA-mr}_\gamma(x) = S_\gamma(x), \tag{22}$$

so spherical steering can be understood either as a direct geodesic construction or as matched additive steering followed by radial renormalization.

We also consider two norm-preserving variants that do not match a fixed target score. The first is renormalized CAA,

$$\text{CAA-r}_\alpha(x) = r\frac{x + \alpha s}{\|x + \alpha s\|_2}, \tag{23}$$

which preserves norm after a fixed additive update but still produces token-dependent achieved scores. The second is angular steering with a fixed rotation amount δ:

$$\mathrm{AS}_\delta(x) = r\left(\cos(\theta-\delta)s + \sin(\theta-\delta)v\right), \tag{24}$$

with clipping when $\theta - \delta$ leaves the valid angular range. AS preserves norm, but a fixed angular step δ does not generally correspond to a fixed target score γ across tokens.

Finally, to test whether norm changes are useful beyond angular concept control, we use a radial extension of spherical steering:

$$\mathrm{SN}_{\gamma,\beta}(x) = \beta r\left(\gamma s + \sqrt{1-\gamma^2}\, v\right), \tag{25}$$

where β is an explicit radial scale. SN reaches the same target concept score as S_γ, but changes the norm from r to βr, with $\beta = 1$ recovering standard spherical steering. This variant separates the question of concept control from the question of whether changing norm improves or damages model behavior.

4. Experiments

Our experiments are designed to disentangle the two geometric components isolated in Section 3: angular concept control and radial norm change. We use the same steering direction construction for all methods and compare them in controlled settings where only one geometric property changes at a time.

4.1. Experimental Setup

We evaluate all experiments on seven language models spanning model families, sizes, and instruction-tuning regimes: Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, Gemma-2-9B-IT, Llama-3.1-8B, Llama-3.2-1B-Instruct, Qwen2.5-3B-Instruct, and Llama-3.1-70B-Instruct. For all models we steer the residual stream at the midpoint layer. Steering directions are estimated by contrastive mean difference on held-out positive and negative pairs, using the last prompt-token activation at the steering layer.

For every steering method, the intervention is applied by a forward hook that replaces each token hidden state during generation. Unless otherwise specified, all methods use the same concept direction, layer, datasets, and evaluation protocol. This ensures that differences between methods come from their geometric update rule rather than from differences in direction extraction or experimental conditions.

4.2. Hidden-State Norm Variation

The hidden-state norm variation experiment characterizes the radial component that steering methods either preserve or modify. We measure hidden-state norm variation across layers, corpora, and token populations using the coefficient of variation of token norms, defined as $\mathrm{CV}(\|x\|_2) = \frac{\mathrm{std}(\|x\|_2)}{\mathrm{mean}(\|x\|_2)}$. We compute this quantity for three token populations: the last prompt token, all prompt tokens, and generated tokens. The generated-token population is the most operationally relevant, since steering hooks modify representations during decoding.

This experiment does not assume that all hidden states lie on a single global sphere. Instead, it quantifies the norm landscape in which steering operates. Large norm variation makes radial changes a potentially important side effect, while norm-preserving methods preserve each token's own radius rather than assuming a common radius across tokens.

4.3. Directional Concept Encoding

The directional-encoding experiment tests whether concept information is primarily angular or radial. For each model and concept dataset, we train probes on three representations of the same hidden states: raw activations x, normalized activations $x/\|x\|_2$, and norm-only features $\|x\|_2$. We compare both linear and nonlinear probes to check whether norm carries concept information either linearly or through simple nonlinear transformations.

If normalized probes match raw probes while norm-only probes remain near chance, then the concept is primarily encoded in direction. This experiment is the empirical motivation for using angular concept score as the semantic component of steering. At the same time, it does not imply that norm is irrelevant: norm may carry little concept information while still affecting downstream stability when it is changed by an intervention.

4.4. Disentangling Concept Matching and Norm Preservation

The steering-method comparison is the main controlled comparison in the paper. It is organized around the two axes in Table 1: whether a method matches a desired per-token concept score and whether it preserves the hidden-state norm. In the main text, we focus on two completed controlled comparisons that isolate why existing steering methods behave differently. The basic concept-score precision check is reported in Section A.2, and the all-method Pareto synthesis is described in Section A.4.

4.4.1. Matched Concept Scores with and without Norm Preservation

The matched-score norm comparison evaluates Spherical Steering against matched CAA. This is the cleanest experiment for isolating the role of norm because both methods reach the same per-token target concept score, defined as

$\left\langle \frac{\widetilde{x}}{\|\widetilde{x}\|_2}, s \right\rangle = \gamma$, but they handle the radial component differently. S preserves the original token norm, while CAA-m changes the norm as a consequence of reaching the target score by additive displacement. For CAA-m, the induced norm ratio is

$$\frac{\|\text{CAA-m}_\gamma(x)\|_2}{\|x\|_2} = \frac{\sqrt{1-c^2}}{\sqrt{1-\gamma^2}}, \tag{26}$$

where $c = \langle x/\|x\|_2, s\rangle$ is the initial concept score. Thus, as the target score approaches perfect alignment, CAA-m can send the hidden-state norm far outside the model's usual activation scale, while S keeps the same radius.

This comparison holds angular concept control fixed and varies only radial behavior. We therefore measure both the induced norm ratio and downstream effects such as task performance, perplexity, or other degradation metrics as a function of γ. The goal is to determine when differences between S and CAA-m are explained by radial distortion rather than by a difference in achieved concept score.

4.4.2. Fixed-Strength Steering under Norm Preservation

The fixed-strength comparison evaluates CAA, CAA-r, and AS. These methods are useful because they do not perform per-token concept-score matching, but they differ in how they treat the norm. Vanilla CAA applies a fixed additive displacement and therefore controls neither the achieved concept score nor the norm. CAA-r uses the same fixed additive update but renormalizes the result back to the original token norm. AS applies a fixed angular rotation toward the concept direction and also preserves norm, but the resulting concept score still depends on the token's initial angle.

This comparison asks whether norm preservation alone is sufficient for good steering when the method does not explicitly match a target concept score. For each method, we measure the achieved concept score distribution, the norm behavior, and downstream metrics across a sweep of steering strengths. Comparing CAA with CAA-r isolates the effect of renormalizing an additive update, while comparing CAA-r with AS tests whether fixed additive and fixed angular movements behave differently when both are constrained to preserve norm.

4.5. Radial Scaling Beyond Angular Control

The radial-scaling experiment studies whether adding an explicit radial degree of freedom improves steering once angular concept control is fixed. We compare S with SN, which reaches the same target concept score but rescales the norm by β:

$$\text{SN}_{\gamma,\beta}(x) = \beta r \left(\gamma s + \sqrt{1-\gamma^2}\, v\right). \tag{27}$$

Table 2. Coefficient of variation of hidden-state norms (%) at the steering layer. Generation tokens are the population modified by steering hooks during decoding.

Model	Last tok	All prompt	Gen toks
Llama-3.1-8B-Inst	6.5	264.9	9.4
Qwen2.5-7B-Inst	7.6	654.2	18.9
Gemma-2-9B-IT	53.6	95.6	17.1
Llama-3.1-8B base	6.2	251.1	200.8
Llama-3.2-1B-Inst	6.4	539.3	9.1
Qwen2.5-3B-Inst	8.6	293.4	17.9
Llama-3.1-70B-Inst	9.7	260.6	17.4

This experiment directly tests the claim that norm is unnecessary for steering. If optimizing β improves control or reduces degradation, then norm cannot be ignored even when concepts are primarily angular. If $\beta = 1$ is consistently sufficient, then norm preservation is a good default once the angular target has been chosen.

5. Experimental Results

We organize results around the two axes of Table 1: angular concept control and radial norm change. The norm-variation and directional-encoding experiments characterize the representation space; the steering-method comparisons isolate method differences under controlled settings; the radial-scaling analysis tests whether an explicit radial degree of freedom helps once angular control is fixed. To save space, all figures are placed in Section A.1 and referenced from the corresponding paragraphs.

5.1. Hidden-State Norms Vary Moderately During Generation

Table 2 shows that hidden-state norms are not globally constant. Pooling all prompt tokens gives very large CV because early attention-sink positions have unusually large norms (Xiao et al., 2024); the per-position plots in Figure 24 show that this effect is localized near the start of the prompt. The population that matters most for steering is generated tokens, where instruct-tuned models have moderate CV (9–19%; Figure 1). Thus, norm preservation should not be interpreted as assuming a single global sphere. Instead, norm-preserving methods keep each token on its own radius, while additive methods may move tokens into atypical radial regimes.

5.2. Concepts Are Represented in Angular Structure

Table 3 compares probes on raw activations, unit-normalized activations, and norm-only features. Unit normalization loses less than 1 pp accuracy on average, while norm-only probes are close to chance for sentiment, toxicity, truthful-

Table 3. Probe accuracy (%) at the steering layer, averaged across five concept datasets. Full per-dataset results are in Section A.9.

Model	raw	unit	norm-only	raw–unit
Llama-8B-Inst	89.3	89.1	60.1	−0.2
Qwen-7B-Inst	85.8	85.7	59.9	−0.1
Gemma-9B-IT	72.3	71.8	49.6	−0.5
Llama-8B base	89.2	89.3	57.7	+0.1
Llama-1B-Inst	84.9	85.0	55.9	+0.1
Qwen-3B-Inst	81.8	82.0	61.5	+0.2
Llama-70B-Inst	90.6	90.7	57.7	+0.1

ness, and IMDB. MLP probes give the same conclusion, indicating that the missing concept signal is not recovered by a simple nonlinear function of norm.

Refusal is the main exception: norm-only probes reach 73–93% accuracy in several models (Figure 2), suggesting that refusal also modulates activation scale. Even there, normalized probes match raw probes. We therefore use the normalized concept score $c = \langle x/\|x\|_2, s\rangle$ as the semantic steering quantity, while treating norm as a separate stability variable.

5.3. Disentangling Concept Matching and Norm Preservation

5.3.1. Matched-Score Comparison: S vs. CAA-m

The matched-score comparison isolates norm by comparing S and CAA-m at the same per-token target γ. The two methods reach the same angular concept score, but CAA-m changes the radius according to

$$\frac{\|\text{CAA-m}_\gamma(x)\|_2}{\|x\|_2} = \frac{\sqrt{1-c^2}}{\sqrt{1-\gamma^2}},$$

where c is the initial concept score. As predicted, Figure 3 shows mild norm inflation at $\gamma = 0.3$ and much larger inflation at $\gamma = 0.7$ for Llama and Qwen, while Gemma is less affected due to its architecture-specific norm behavior.

The downstream comparison shows why this radial difference matters. At moderate targets ($\gamma \leq 0.3$), S and CAA-m are similar. At aggressive targets, the gap becomes large and unstable across tasks (Figure 4). Since the achieved angular score is matched, these differences are attributable to radial behavior rather than stronger concept expression. Quality metrics sharpen the point: high-γ S can sharply increase perplexity, while CAA-m often remains closer to baseline (Figure 5; MMLU in Figure 33). This does not contradict the directional-encoding result. Concepts are primarily angular, but downstream layers are sensitive to activation scale, so norm can affect whether a large angular edit remains usable.

Table 4. Fixed-strength methods at matched mean concept score. Values are averaged over 31 model–dataset cells. CAA and CAA-r use the same calibrated α, so they achieve the same normalized concept scores but differ in norm. AS is calibrated separately via $\Delta\theta$.

	$\|y\|/\|x\|$			PPL ratio			KL		
$\bar{\gamma}$	CAA	CAA-r	AS	CAA	CAA-r	AS	CAA	CAA-r	AS
0.1	1.00	1.00	1.00	1.02	1.02	1.06	0.03	0.03	0.06
0.3	1.04	1.00	1.00	1.12	1.12	1.36	0.19	0.19	0.41
0.5	1.13	1.00	1.00	1.42	1.43	3.45	0.55	0.57	1.16
0.7	1.34	1.00	1.00	4.32	5.47	34.7	1.23	1.23	2.38

5.3.2. Fixed-Strength Steering Under Norm Preservation

The fixed-strength comparison evaluates fixed-strength methods—CAA, CAA-r, and AS—which do not match a per-token concept score. For each method, we choose its strength parameter (α for CAA/CAA-r and $\Delta\theta$ for AS) by binary search so that the method reaches a target mean concept score $\bar{\gamma} \in \{0.1, 0.3, 0.5, 0.7\}$ on held-out hidden states. We then measure norm ratio, perplexity, KL divergence, and downstream accuracy. Results are averaged over 31 model–dataset cells: six models on all five datasets plus Llama-70B on TruthfulQA.

CAA vs. CAA-r. CAA and CAA-r have identical normalized directions after the additive update: CAA-r only rescales $x + \alpha s$ back to the original radius. Therefore, because the concept score depends only on direction, the same calibrated α gives the same achieved concept scores for both methods. Table 4 shows that their PPL and KL are nearly identical up to $\bar{\gamma} = 0.5$, even though vanilla CAA already increases norms by 13% at this target. At $\bar{\gamma} = 0.7$, CAA inflates norms by 34% on average, but its PPL ratio is slightly lower than CAA-r's (4.32× vs. 5.47×). Thus, in this fixed-strength comparison, renormalization is not uniformly beneficial: preserving norm removes radial distortion, but it does not by itself guarantee better downstream behavior.

CAA-r vs. AS. CAA-r and AS both preserve norm exactly, but they differ substantially in downstream degradation. At the same matched mean concept score, AS has much higher PPL and KL: at $\bar{\gamma} = 0.5$, AS reaches 3.45× baseline PPL compared with 1.43× for CAA-r; at $\bar{\gamma} = 0.7$, the gap grows to 34.7× versus 5.47×. AS is worse in all 31 model–dataset cells at both high targets. This shows that norm preservation alone is not sufficient. The reason is that CAA-r and AS induce different token-wise angular profiles. CAA-r normalizes a fixed additive perturbation, so its angular displacement depends on each token's initial norm and alignment with s. AS instead applies a fixed geodesic rotation $\Delta\theta$ to every token, regardless of its initial position.

Table 5. Peak TruthfulQA MC1 in the S/SN grid. "Gain" is SN's best β minus $\beta = 1$ ($\equiv$ S), using 2 folds.

Model	No int.	S peak	SN peak	Gain	CAA peak
Llama-8B-Inst	36.3	44.3	44.8	+0.5	39.0
Qwen-7B-Inst	39.3	43.4	43.8	+0.3	38.8
Gemma-9B-IT	46.1	50.8	50.9	+0.1	46.9
Llama-8B base	29.9	41.6	41.8	+0.2	35.6

Therefore, two methods can match the same mean concept score and preserve the same norm while still producing very different token-level interventions.

Downstream and Pareto. Figure 15 plots the aggregated steering–degradation tradeoff. At moderate targets ($\bar{\gamma} \leq 0.3$), all three methods cluster near the baseline with small PPL increase and modest task improvement. At $\bar{\gamma} = 0.5$, AS achieves slightly higher downstream gain but at $3.4\times$ the PPL cost of CAA/CAA-r. At $\bar{\gamma} = 0.7$, AS collapses: PPL explodes to $35\times$ while the downstream gain vanishes. Per-dataset breakdowns and per-cell scatter plots are in Section A.3, along with the full downstream table (Table 6).

Takeaway. Among fixed-strength methods, renormalization is essentially free (no quality cost, no quality gain) and α is model-dependent. The more important observation is that all three methods spread the per-token concept score over a broad range ($\sigma \sim 0.05$–0.06; see Figure 7), in contrast to score-matched methods where $\sigma < 10^{-3}$. This motivates the all-method Pareto synthesis in Section A.4, which is designed to test whether per-token control, norm behavior, or their combination best explains the steering–quality tradeoff.

5.4. Explicit Radial Tuning Adds Little on Top of Angular Control

The radial-scaling analysis asks whether norm should be optimized as a second explicit control once the angular target is fixed. We compare S (β=1, strict norm preservation) with SN at $\beta \in \{0.9, 1.1, 1.2\}$, all reaching the same per-token γ. On a focused TruthfulQA grid (2 folds), SN improves peak MC1 over S by at most 0.5 pp (Table 5), and CAA peaks lower than S on three of four models, confirming the value of direct angular control.

Multi-dataset evidence ($\beta \in \{0.9, 1.0, 1.1, 1.2\}$, 6 models $\times$ 5 datasets). Across 29 model–dataset cells (Figure 6), *task accuracy is nearly invariant to* β: the mean Δmc1 spread across all four β values is ≤ 0.04 pp at every γ. No single β dominates—β=0.9 wins the most cells at low γ (16/29 at γ=0.1), while β=1.2 wins slightly at high γ (12/29 at γ=0.7).

Perplexity tells a different story. Strict norm preservation (β=1.0) gives the *highest* PPL ratio at $\gamma \geq 0.3$, winning 0/29 cells at $\gamma \geq 0.5$. All SN variants outperform it: at γ=0.7, mean PPL ratios are $61.0\times$ for β=1.0 vs. $28.1\times$ for β=1.2 (lowest in 27/29 cells). Even modest inflation (β=1.1) halves the PPL penalty. The pattern suggests that large angular displacement ($\gamma \geq 0.5$) systematically underscales the hidden state when norm is preserved exactly, and a small radial correction partially compensates.

The asymmetry between task accuracy and PPL is instructive: γ determines conceptual content, while β affects model stability in a way that is largely invisible to task metrics but visible to distributional measures. The PPL improvement from tuning β is nonetheless modest compared to the PPL cost of increasing γ itself (the ratio exceeds $28\times$ even at the best β for γ=0.7). A global radial multiplier is therefore not inert, but it remains a second-order knob relative to γ.

Summary of findings. The results support the two-component view. The directional-encoding results show that concept information is primarily angular, making γ the natural semantic control. The norm-variation and matched-score comparisons show why norm still matters: steering operates in a nontrivial norm landscape, and radial changes can explain downstream differences even when angular score is matched. The radial-scaling results show that a global radial multiplier β has little effect on task accuracy but can reduce PPL at high γ, consistent with norm being a stability lever rather than a semantic one. The main design problem is therefore to choose appropriate token-wise angular movement while handling norm carefully enough to avoid destabilizing the model.

6. Conclusion

We presented a geometric analysis of activation steering as an intervention with two coupled components: angular movement toward a concept direction and radial change in hidden-state norm. This perspective explains why a single additive coefficient is difficult to interpret: it changes both quantities in a token-dependent way. Across seven models, our experiments show that concepts are primarily represented in angular structure, supporting the motivation for spherical methods. At the same time, norm remains important because radial changes can strongly affect downstream stability even when angular concept score is held fixed. Thus, the central problem is not simply to choose a larger or smaller additive strength, but to decide how token representations should move in angle and how their norm should be handled during that move. We hope this framework provides a cleaner basis for comparing steering methods and for designing future token-wise steering policies.

References

Dang, Q.-A. and Ngo, C. Selective steering: Norm-preserving control through discriminative layer selection, 2026. URL https://arxiv.org/abs/2601.19375.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model, 2023. URL https://arxiv.org/abs/2306.03341.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering Llama 2 via contrastive activation addition, 2023. URL https://arxiv.org/abs/2312.06681.

Park, K., Choe, Y. J., and Veitch, V. 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. URL https://proceedings.mlr.press/v235/park24c.html.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. 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, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Turner, A. M., Thiergart, L., Udell, D., Leech, G., Vazquez, J. J., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization, 2023. URL https://arxiv.org/abs/2308.10248.

Vu, H. M. and Nguyen, T. M. Angular steering: Behavior control via rotation in activation space, 2025. URL https://arxiv.org/abs/2510.26243.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

You, Z., Deng, C., and Chen, H. Spherical steering: Geometry-aware activation rotation for language models, 2026. URL https://arxiv.org/abs/2602.08169.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to AI transparency, 2023. URL https://arxiv.org/abs/2310.01405.

A. Additional Experiments

The main paper focuses on the experiments that directly disentangle angular concept control and radial norm change. We place the following experiments in the appendix because they support the geometric framework without changing the two primary steering variables, or because they synthesize the main comparisons rather than adding a new controlled contrast. In particular, concept-score precision validates the taxonomy, the all-method Pareto comparison gives the completed synthesis, initial-score distributions explain why fixed-strength methods are heterogeneous, S/CAA-mr equivalence verifies the algebra, and the spherical-arc experiment checks that the standard construction is a reasonable reference point rather than an additional steering hyperparameter.

A.1. Main-result Figures

For space, the main paper reports the numerical tables and discusses each result in text, while the large diagnostic figures are collected here. The labels are the same as those cited from the main text.

A.2. Concept-score Precision

This supporting experiment verifies which methods actually reach a desired per-token target score. For every steered hidden state $\widetilde{x}$, we measure the achieved concept score

$$\widetilde{c} = \left\langle \frac{\widetilde{x}}{\|\widetilde{x}\|_2}, s \right\rangle. \tag{28}$$

We then compute the standard deviation of $\widetilde{c}$ across tokens for a fixed target γ. Score-matched methods should collapse this distribution around γ, whereas fixed-strength methods should produce a broader distribution because the achieved score depends on each token's initial norm and initial alignment.

This experiment serves as a diagnostic check for the taxonomy in Table 1. It verifies the distinction between methods that explicitly target a per-token concept score and methods that use a fixed global strength. We place it in the appendix because the main text focuses on downstream comparisons: once the precision difference is established, the central question becomes how angular matching and radial behavior affect model quality.

Figure 7 confirms the expected separation. S and CAA-mr achieve $\sigma \sim 10^{-4}$ (the binary-search convergence floor), while CAA-r and AS produce concept-score distributions that are $\sim 200\times$ wider. CAA-m and SN are score-matched by construction and are omitted from this compact diagnostic plot for readability. The separation holds uniformly across all five datasets and all seven models including the 1B and 70B scale probes.

The combined KDE plot (Figure 8) shows the overall pattern, while per-dataset KDE plots confirm consistency: TruthfulQA (Figure 10), SST-2 (Figure 11), CivilComments (Figure 12), IMDB (Figure 13), and Refusal (Figure 14). In every case, S collapses to a near-delta function at γ while CAA-r and AS smear the concept score across a broad range.

The dose–response curves in Figure 9 illustrate why fixed-strength methods produce heterogeneous scores: CAA-r's additive α has norm-dependent effect (Gemma saturates before reaching $\gamma = 0.7$ due to its $\sim 100\times$ larger norms), while AS's angular $\Delta\theta$ is norm-invariant and reaches all targets uniformly.

A.3. Fixed-Strength Downstream Metrics and Pareto Frontiers

This section reports downstream task performance for the fixed-strength methods evaluated in the fixed-strength comparison (Section 5.3.2). Each method is calibrated to matched mean concept score $\bar{\gamma}$ via binary search, then evaluated on task-specific metrics (MC1 for TruthfulQA, positive rate for SST-2/IMDB, non-toxicity for CivilComments, refusal rate for Refusal). Cells where calibration failed to reach the target (concept score deviating > 0.05) are excluded from aggregation.

Table 6 reveals that no single fixed-strength method dominates across datasets. On TruthfulQA at $\bar{\gamma} = 0.3$ all three methods produce similar gains (+1–4 pp), while at $\bar{\gamma} = 0.7$ AS degrades sharply (-21 pp for Llama-70B). On sentiment tasks (SST-2, IMDB), AS occasionally achieves large gains (Llama-1B SST-2: 0.905 vs. baseline 0.380) but is unstable across models (Qwen-3B collapses to 0.050 at $\bar{\gamma} = 0.7$). CAA and CAA-r are nearly interchangeable, confirming that renormalization does not affect downstream accuracy.

The Pareto frontiers (Figure 17) show the steering–degradation tradeoff across datasets. CAA and CAA-r trace nearly identical curves, confirming that the quality cost of steering is insensitive to renormalization. AS consistently lies above

the other two methods on the PPL axis, especially at $\bar{\gamma} \geq 0.5$. The per-cell scatter (Figure 18) confirms this is not driven by outlier datasets: AS points cluster at higher PPL across the board. The aggregated Pareto in the main text (Figure 15) summarizes this pattern.

A.4. All-Method Pareto Comparison

This section contains the completed all-method Pareto synthesis. We place it in the appendix because it aggregates the controlled comparisons from the matched-score and fixed-strength comparisons rather than introducing a new geometric variable.

Protocol. The all-method Pareto comparison evaluates steering methods by Pareto frontier instead of by a single arbitrary raw parameter (α, γ, δ, or β). For each method and steering strength, we measure (i) downstream task-metric change, where larger values indicate stronger steering, and (ii) perplexity ratio, where smaller values indicate less degradation. Fixed-strength methods are calibrated to matched mean concept scores as in the fixed-strength comparison, while score-matched methods use the target concept score directly. The completed grid shown here includes CAA, CAA-r, AS, S, and CAA-mr; CAA-m and SN are analyzed separately in the matched-score norm comparison and the radial-scaling analysis because they serve as explicit radial controls rather than additional fixed-strength Pareto baselines.

Aggregate Pareto structure. Figure 19 shows three main patterns. First, all methods are close to the baseline at $\bar{\gamma} = 0.1$: the task gain is small and PPL remains near $1\times$. Second, CAA and CAA-r nearly overlap through most of the sweep. This reinforces the fixed-strength comparison's conclusion that simply projecting a fixed additive update back to the original norm does not materially change the aggregate steering–quality tradeoff; the limiting factor for this family is the lack of per-token concept-score control. Third, the methods with explicit angular targets move farther to the right at moderate settings, but the cost of aggressive angular movement is highly nonlinear. S reaches the largest aggregate gain at intermediate targets, whereas high-γ S moves sharply upward on the PPL axis. CAA-mr traces a more stable high-target curve in this completed sweep, and AS is generally dominated at aggressive targets: it incurs substantially higher PPL for comparable or worse task change.

The resulting interpretation is that per-token concept-score control is the strongest predictor of useful steering at moderate strengths, while norm handling determines how quickly that control becomes unstable as the target is pushed toward high alignment. Norm preservation alone is not sufficient—AS preserves norm but can still degrade heavily—and fixed additive strength is not sufficient either, because CAA/CAA-r remain left of the matched methods when stronger steering is needed.

Per-dataset variation. The per-dataset frontiers in Figure 20 explain why the aggregate curve should be read as a Pareto analysis rather than as a single universal ranking. SST-2 and IMDB provide the clearest positive steering gains, but the best point is not always the strongest target: high-γ settings can increase PPL by an order of magnitude or more after the task metric has already saturated. TruthfulQA and Refusal show the opposite failure mode most clearly, where aggressive S or AS can move upward on degradation while moving little or even left on the task metric. CivilComments has a narrower task-metric range, so large PPL increases are rarely justified by the small changes in non-toxicity.

Across datasets, CAA and CAA-r remain close, confirming that the fixed-additive family is not primarily limited by radial rescaling. AS confirms the complementary point: preserving norm does not guarantee a good Pareto point if the token-wise angular profile is poorly matched. The most reliable region is therefore the moderate matched-angular regime, where concept control improves before the nonlinear PPL penalty of large angular edits dominates.

A.5. Initial Concept-score Distributions

This experiment reports the distribution of initial concept scores $c = \langle x/\|x\|_2, s\rangle$ across datasets, models, and token populations (Figure 21). These distributions help explain why fixed-strength methods produce heterogeneous achieved scores: tokens start from different locations relative to the concept direction. Most distributions are approximately centered near zero with moderate spread. We use these results as diagnostic support for the fixed-strength comparisons, but keep them in the appendix because they do not directly compare steering methods.

A.6. Equivalence of S and CAA-mr

This experiment verifies the algebraic identity $\text{CAA-mr}_\gamma(x) = S_\gamma(x)$ from Equation (22). We apply both methods to the same hidden states in `float64` and compare geodesic distances, concept-score deviations (Table 7), and downstream metric differences (Table 8).

Across all 280 data points (7 models $\times$ 5 datasets $\times$ 4 γ $\times$ 2 folds), the geodesic distance is exactly 4.47×10^{-4} with zero variance: the residual is floating-point rounding between S's direct construction ($\gamma s + \sqrt{1-\gamma^2}\, v$) and CAA-mr's closed-form α path ($x + \alpha s$ then renormalize). Residual alignment $\langle v_S, v_{\text{CAA-mr}} \rangle_{\min} = 1.000000$ everywhere. Concept-score deviation peaks at 3.3×10^{-4} (Qwen-7B/CivilComments at $\gamma = 0.5$).

Downstream, TruthfulQA (closed-form MC1) shows $|\Delta\text{MC1}| \leq 0.005$ across all models. Generation-based tasks show slightly larger differences (up to 0.031) due to sampling stochasticity. The global mean $|\Delta\text{metric}|$ across all 280 data points is 0.011, with the 95th percentile at 0.045. The equivalence holds both geometrically and functionally across all seven models (1B–70B), five datasets, and four γ values.

A.7. Optimality of the Spherical Arc

This experiment checks an auxiliary property of the spherical reference construction. It should not be read as introducing a third primary steering parameter. The main paper treats steering as choosing angular and radial behavior; here we only ask whether the particular same-norm, same-score point chosen by S is a stable default. We fix a target concept score γ and perturb the output away from the spherical solution by rotating the orthogonal remainder v by angle δ toward a direction $q \perp s,\ q \perp v$:

$$y(\delta) = \|x\| \big(\gamma\, s + \sqrt{1-\gamma^2}\, (\cos\delta\ v + \sin\delta\ q)\big).$$

All points on this sweep have the same norm and concept score; only the otherwise-uncontrolled orthogonal remainder changes. We sweep $\delta \in \{-0.2, -0.1, 0, +0.1, +0.2\}$ with $\gamma \in \{0.1, 0.3, 0.5, 0.7\}$ across 7 models and 5 datasets, using 8 perturbation directions per cell (2 random, 2 PCA, 4 cross-dataset). Table 9 reports aggregate degradation, and Table 10 breaks the results down by perturbation-direction type.

PPL achieves its minimum at $\delta = 0$ in 37 of 40 cells; the three exceptions have negligible gaps ($< 0.7\%$ relative PPL difference). KL is minimized at $\delta = 0$ in 31 of 40 cells, with all nine exceptions within 3% of the true minimum. The valley is approximately symmetric and deepens monotonically with $|\delta|$: at $\delta = \pm 0.2$, PPL rises by $\sim$22% and KL roughly doubles.

PCA directions produce the steepest valleys ($\times 1.43$ at $\delta = -0.2$), consistent with the top PCs of the residual subspace encoding the most task-relevant variance. Random directions are mildest ($\times 1.15$–1.21); cross-dataset directions are intermediate, suggesting that concept axes from related tasks partially overlap with the residual PCs. The result holds across scales: from Llama-1B ($\times 1.26$ at $\delta = \pm 0.2$) to Llama-70B ($\times 1.43$), the valley at $\delta = 0$ is consistent. This confirms that the spherical arc is not only geodesically minimal but also empirically optimal for preserving model behavior.

A.8. Detailed Norm Variation

This section provides the full per-model and per-position norm analysis summarized in the main text (Section 5.1). Figure 22 shows per-model CV curves across layers and corpora, while Table 11 gives the per-corpus breakdown at the steering layer. Figure 23 compares last-token vs. all-token CV, confirming that the all-token inflation is due to attention sinks. Per-position analysis is shown for prompt tokens (Figure 24) and generation tokens (Figure 25). Cumulative CV over prompt (Figure 26) and generation (Figure 27) tokens converges quickly. Mean norm vs. position is shown for prompt tokens (Figure 28) and generation tokens (Figure 29). Figure 30 decomposes CV at sublayer boundaries.

A.9. Detailed Directional Concept Encoding

This section provides the full per-dataset probe results (Table 12) and the layer-wise accuracy trajectory (Figure 31) for the directional encoding analysis in Section 5.2.

A.10. Detailed Matched-Score S vs. CAA-m Quality and Scatter Analysis

The main text reports the key matched-score findings with the norm-ratio and PPL images. Here we provide the remaining diagnostic plots: the norm-ratio scatter (Figure 32), MMLU under steering (Figure 33), and the super-linear growth of ΔPPL

(Figure 34) and ΔMMLU (Figure 35) with γ.

B. Discussion

The results separate two claims that are often merged in discussions of spherical steering. First, concept information is primarily angular: normalized hidden states retain almost all probe accuracy, while norm-only probes are weak for most concepts. Second, norm still matters operationally: once an intervention is applied, downstream layers process the full vector, including its radius. This is why S and CAA-m can express the same target concept score but differ substantially in perplexity or task accuracy at large γ.

This distinction also clarifies how to interpret norm preservation. Preserving norm is not justified because all tokens live on one global sphere; the norm-variation results show that norms vary across tokens, models, and positions. Rather, norm preservation is a controlled reference operation: it changes the angular concept score while keeping each token on its original radius. Deviations from this reference, such as CAA-m or SN, can then be analyzed as explicit radial effects.

Finally, the current evidence suggests that a single global radial multiplier is too coarse to be a useful steering knob. the matched-score comparison shows that radial changes can matter, but the radial-scaling analysis shows that optimizing one global β on top of S adds little in the completed grid. This motivates a more refined view: the useful object is not a universal norm offset, but a token-wise policy for when radial compensation is beneficial.

C. Future work

The most direct next step is to learn token-wise steering policies. In the notation of this paper, such a policy would choose an angular target γ_t and possibly a radial scale β_t for each token, rather than using a global additive coefficient. The policy could depend on the token's initial concept score, norm, position, uncertainty, or local generation context.

A second direction is to study multi-concept steering. Real interventions often involve several behavior directions that need not be orthogonal. Extending the angular–radial decomposition to a subspace of concept directions would make it possible to separate concept tradeoffs from norm effects.

A third direction is to improve concept-direction estimation. Our analysis assumes a fixed direction s obtained by contrastive mean difference. If s is noisy or mixes several concepts, then even a geometrically clean intervention may target the wrong semantic axis. Better direction estimation could reduce the need for aggressive angular movement and therefore reduce the stability problems observed at large γ.

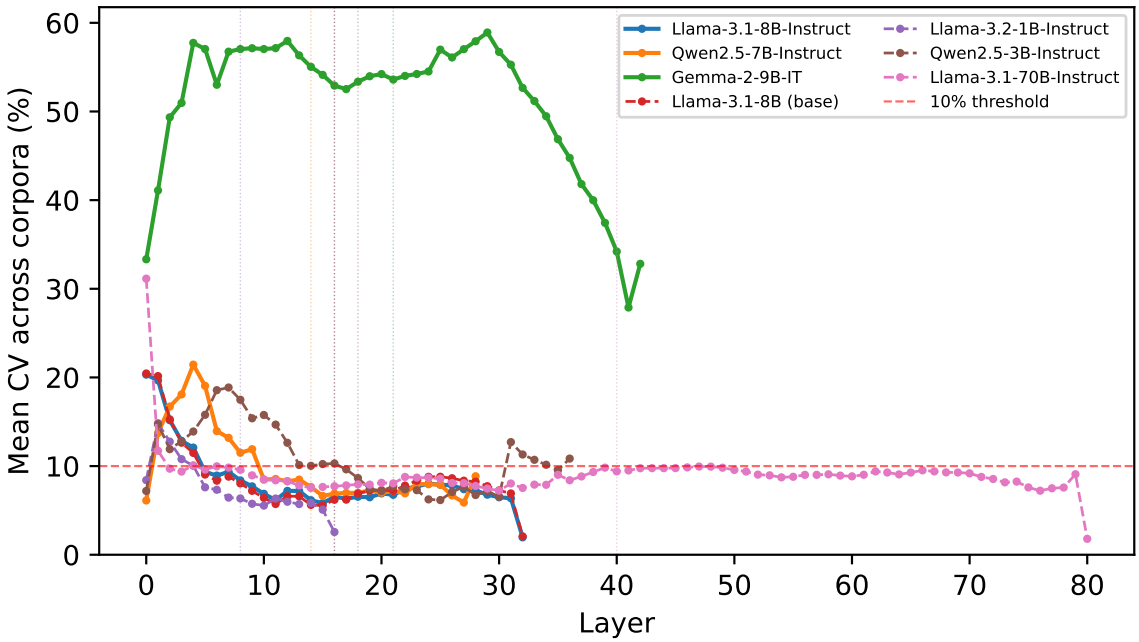

Figure 1. Mean CV of hidden-state norms across 10 corpora vs. layer depth for all seven models. Llama and Qwen stay below 10% at most layers; Gemma-2's post-norm architecture produces elevated CV throughout. Red dashed line = 10% threshold.

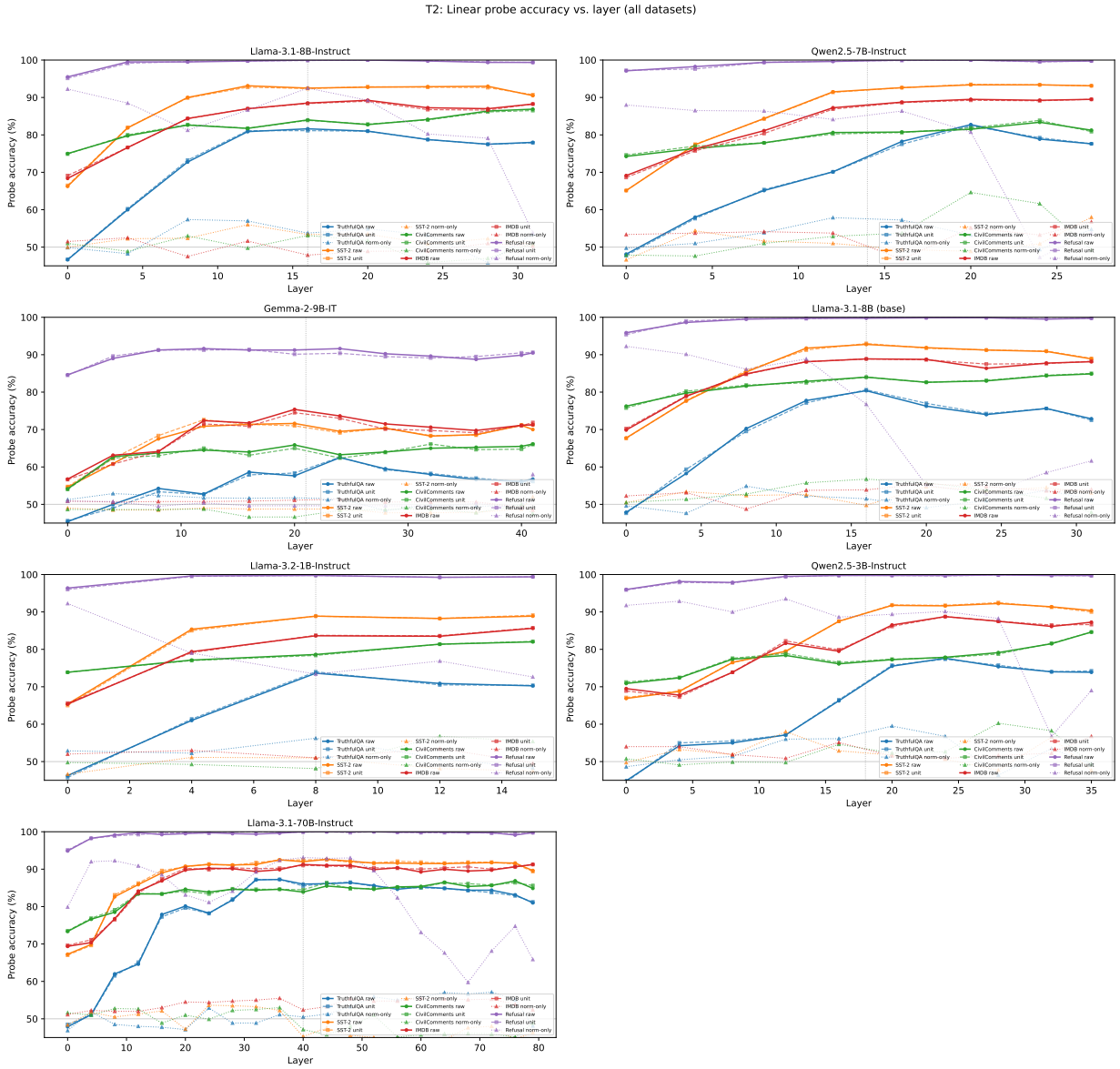

Figure 2. Linear probe accuracy vs. layer for all five concept datasets. Each dataset shows raw h (solid), unit $h/\|h\|$ (dashed), and norm-only $\|h\|$ (dotted). Raw and unit curves overlap across datasets; norm-only curves sit near chance except for refusal.

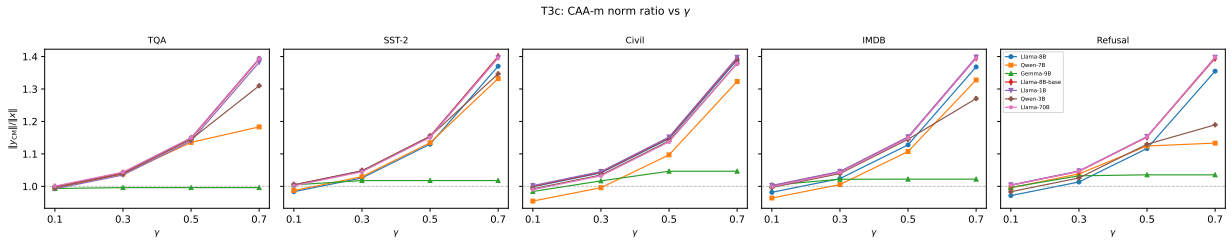

Figure 3. CAA-m norm ratio vs. target γ, seven models $\times$ five datasets. Matched additive steering reaches the same concept score as S, but does so by changing the norm; the inflation grows quickly as γ increases.

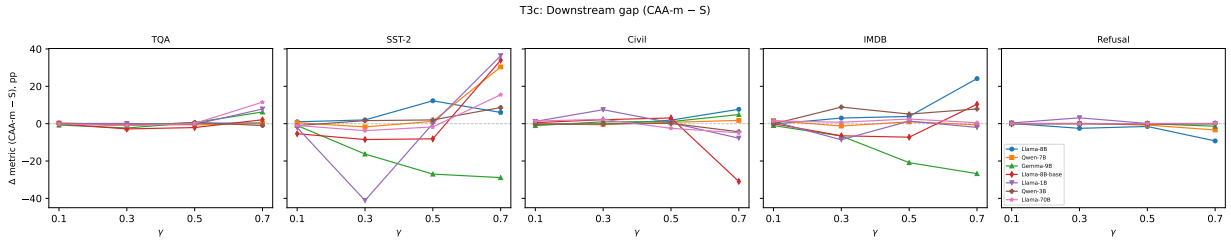

Figure 4. Downstream metric gap (CAA-m − S) vs. γ, seven models $\times$ five datasets. At moderate γ the gap is small; at aggressive targets radial distortion produces larger and less predictable shifts.

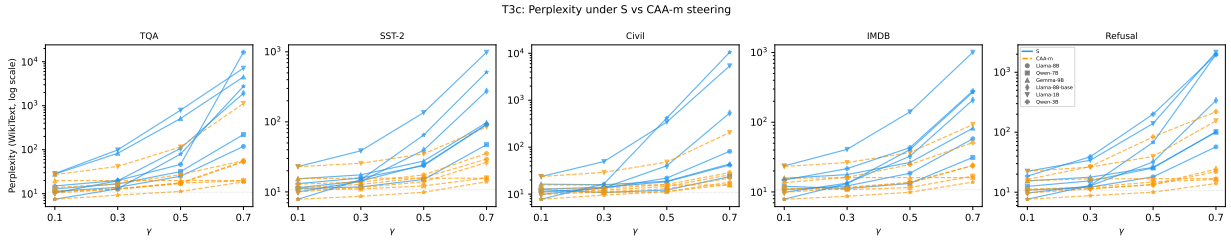

Figure 5. Perplexity under S (solid) vs. CAA-m (dashed) across γ values, seven models $\times$ five steering datasets. Log scale. The comparison holds angular target fixed and changes only radial behavior.

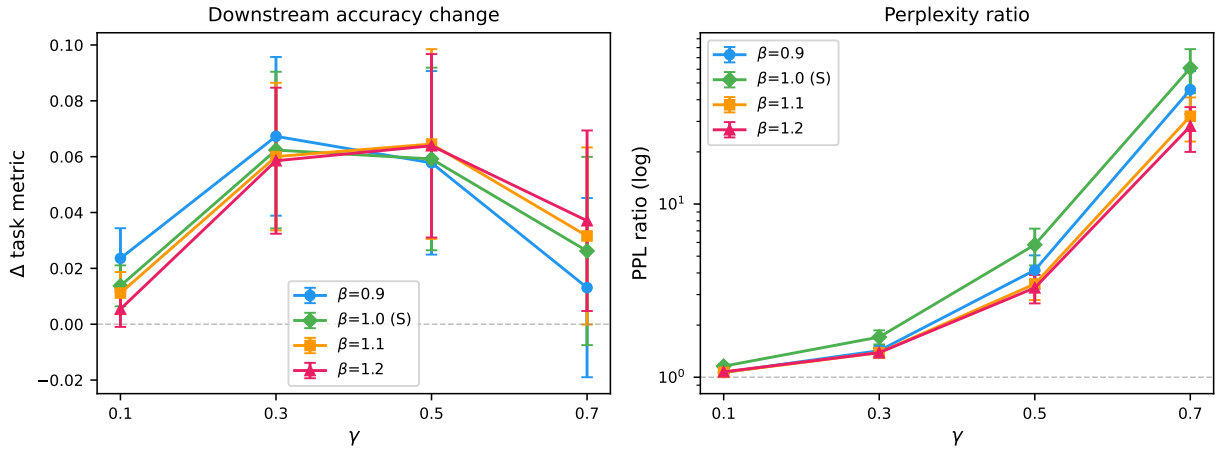

Figure 6. Δ task metric (left) and PPL ratio (right) vs. γ for $\beta \in \{0.9, 1.0, 1.1, 1.2\}$, averaged over 29 model–dataset cells. β=1.0 is S (norm-preserving). Task accuracy is nearly invariant to β; PPL shows a systematic preference for $\beta > 1$ at high γ.

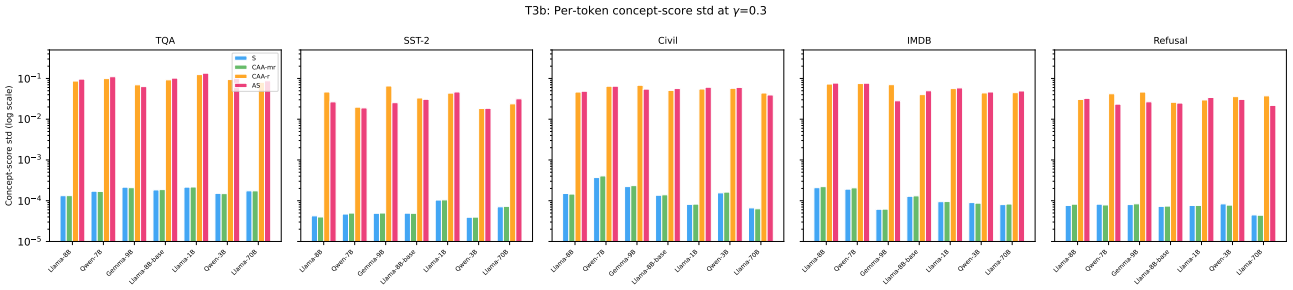

Figure 7. Per-token concept-score standard deviation (log scale) at $\gamma = 0.3$, 7 models $\times$ 5 datasets. Score-matched methods (S, CAA-mr) achieve $\sigma \sim 10^{-4}$ (binary-search floor); fixed-strength methods (CAA-r, AS) spread $\sim 200\times$ wider. The pattern is consistent across all datasets and model scales (1B–70B).

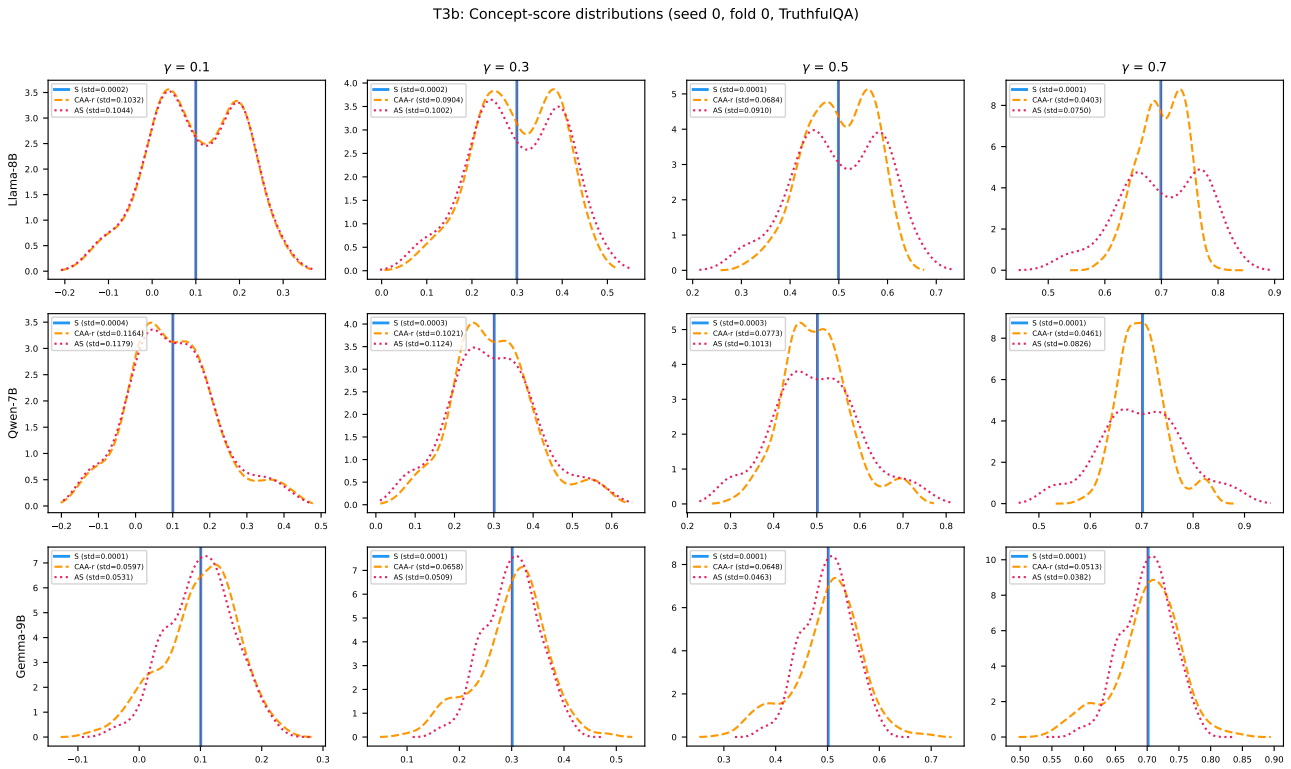

Figure 8. Combined concept-score distributions at matched mean (seed 0, fold 0, TruthfulQA). S collapses to a vertical line at γ; CAA-r and AS spread over a wide range.

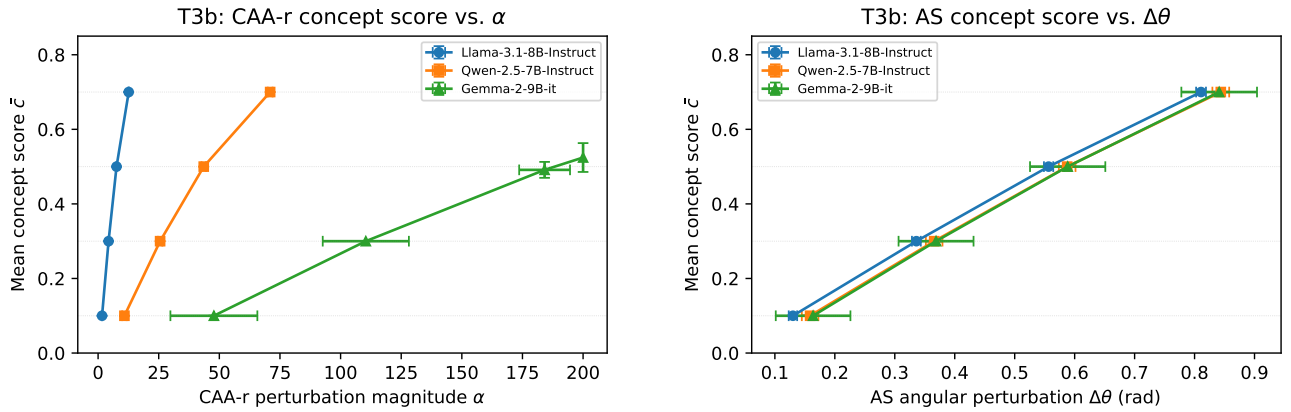

Figure 9. Dose–response curves for fixed-strength methods. Left: CAA-r mean concept score vs. α — Gemma saturates and falls short of $\gamma = 0.7$ due to its $\sim 100\times$ larger norms. Right: AS mean concept score vs. $\Delta\theta$ — all models reach the target because angular perturbation is norm-invariant.

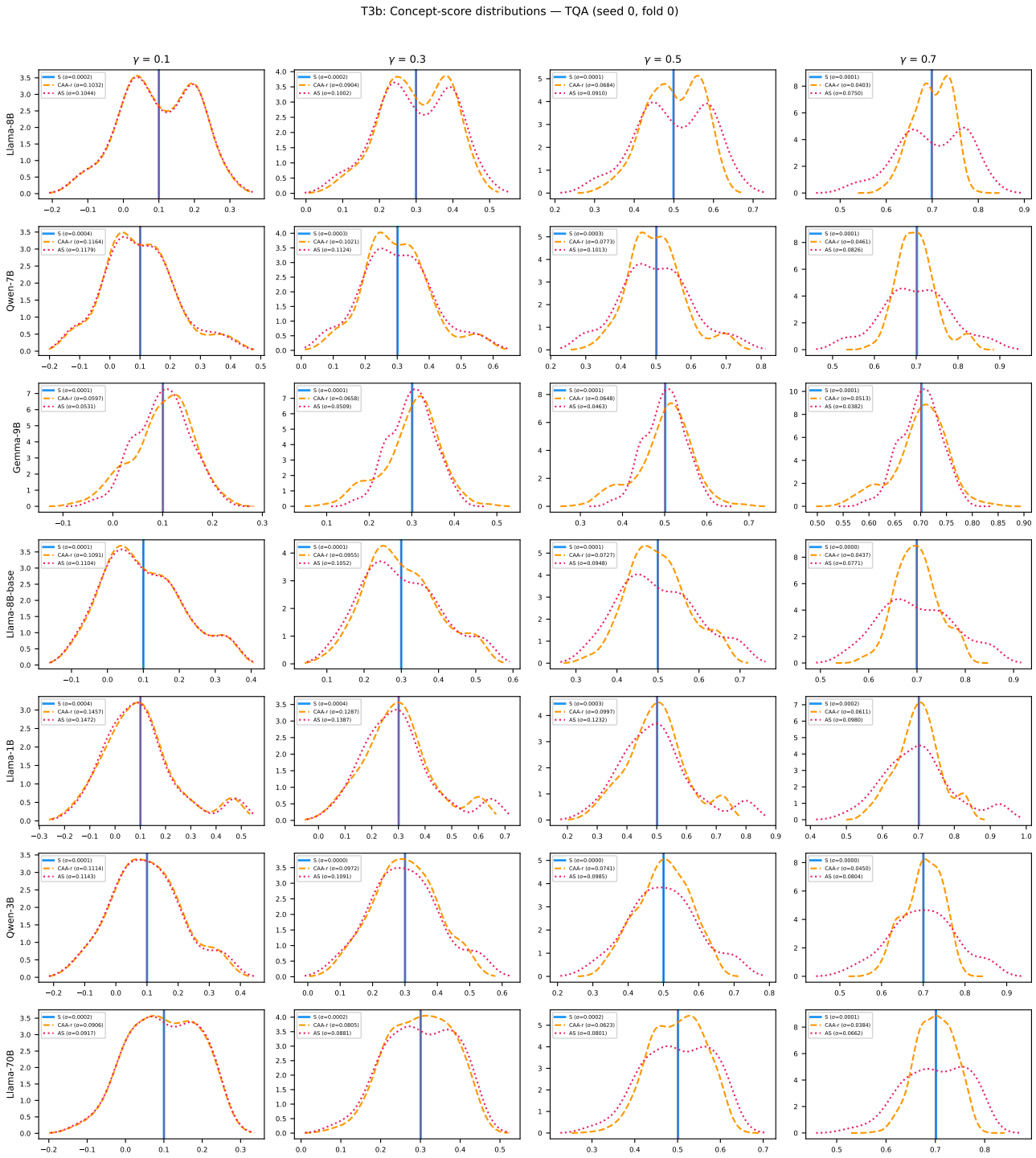

Figure 10. Concept-score distributions on TruthfulQA (seed 0, fold 0, all 7 models). S collapses to a vertical line at γ; CAA-r and AS spread over a wide range.

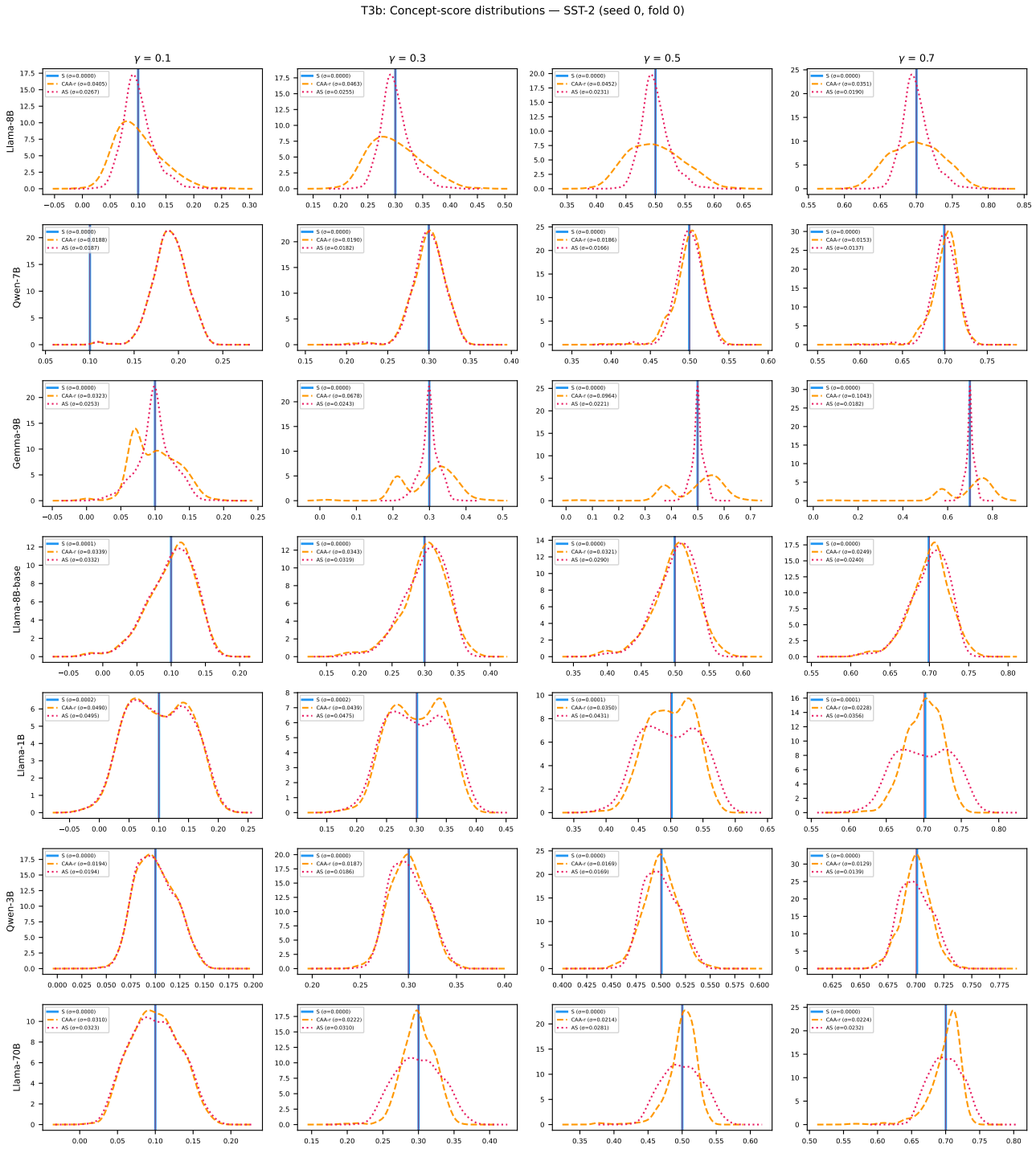

Figure 11. Concept-score distributions on SST-2 (seed 0, fold 0, all 7 models).

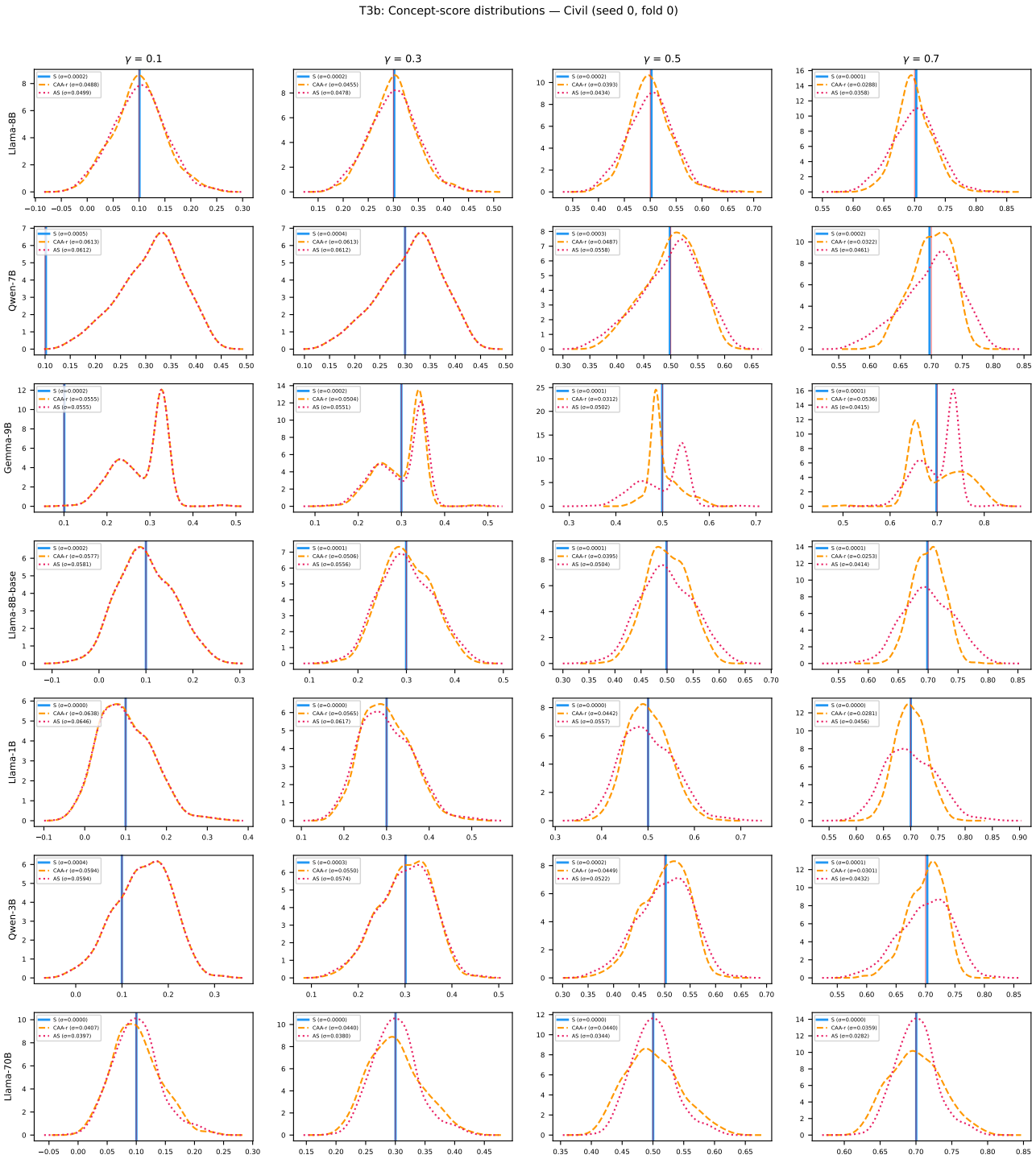

Figure 12. Concept-score distributions on CivilComments (seed 0, fold 0, all 7 models).

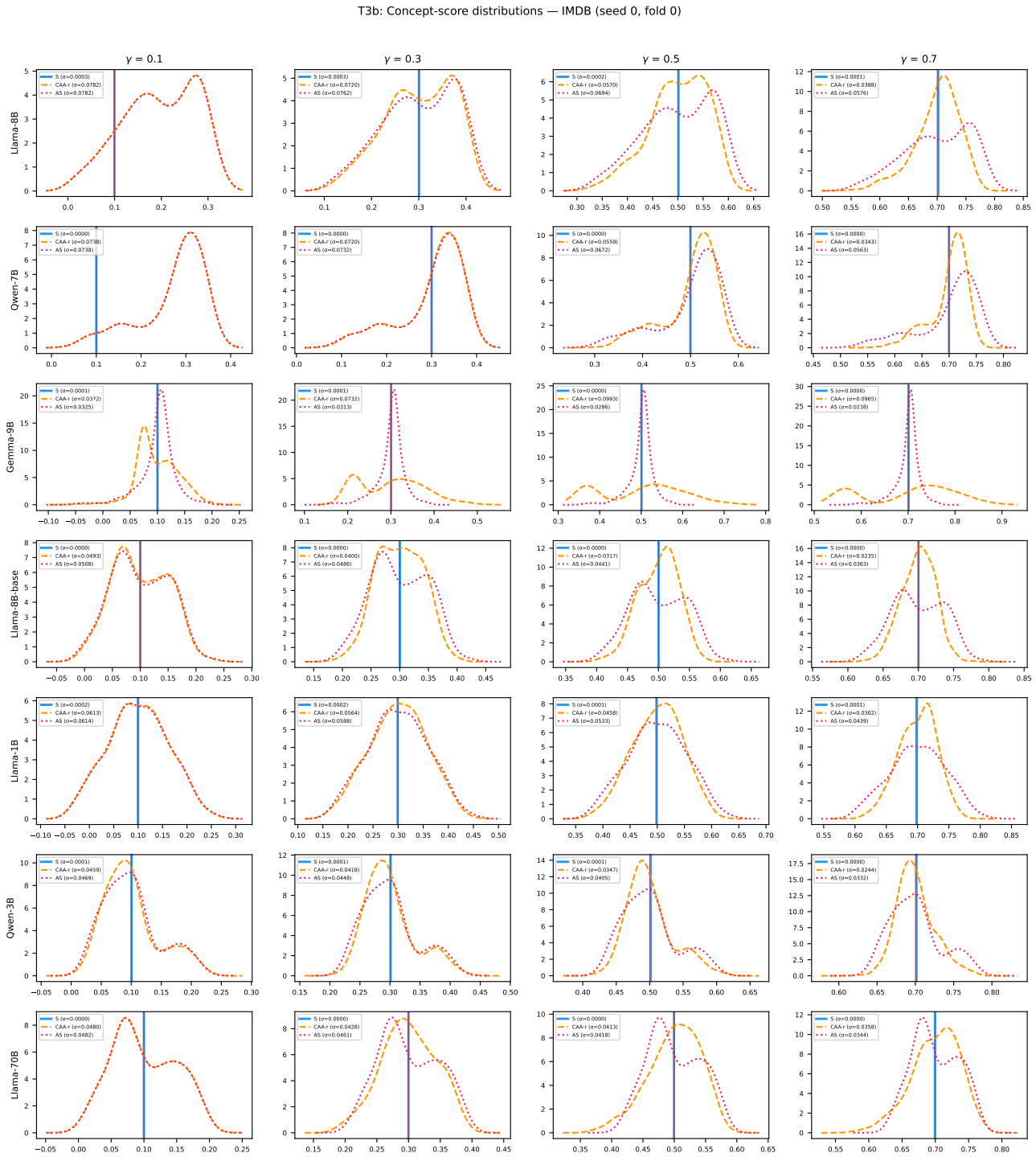

Figure 13. Concept-score distributions on IMDB (seed 0, fold 0, all 7 models).

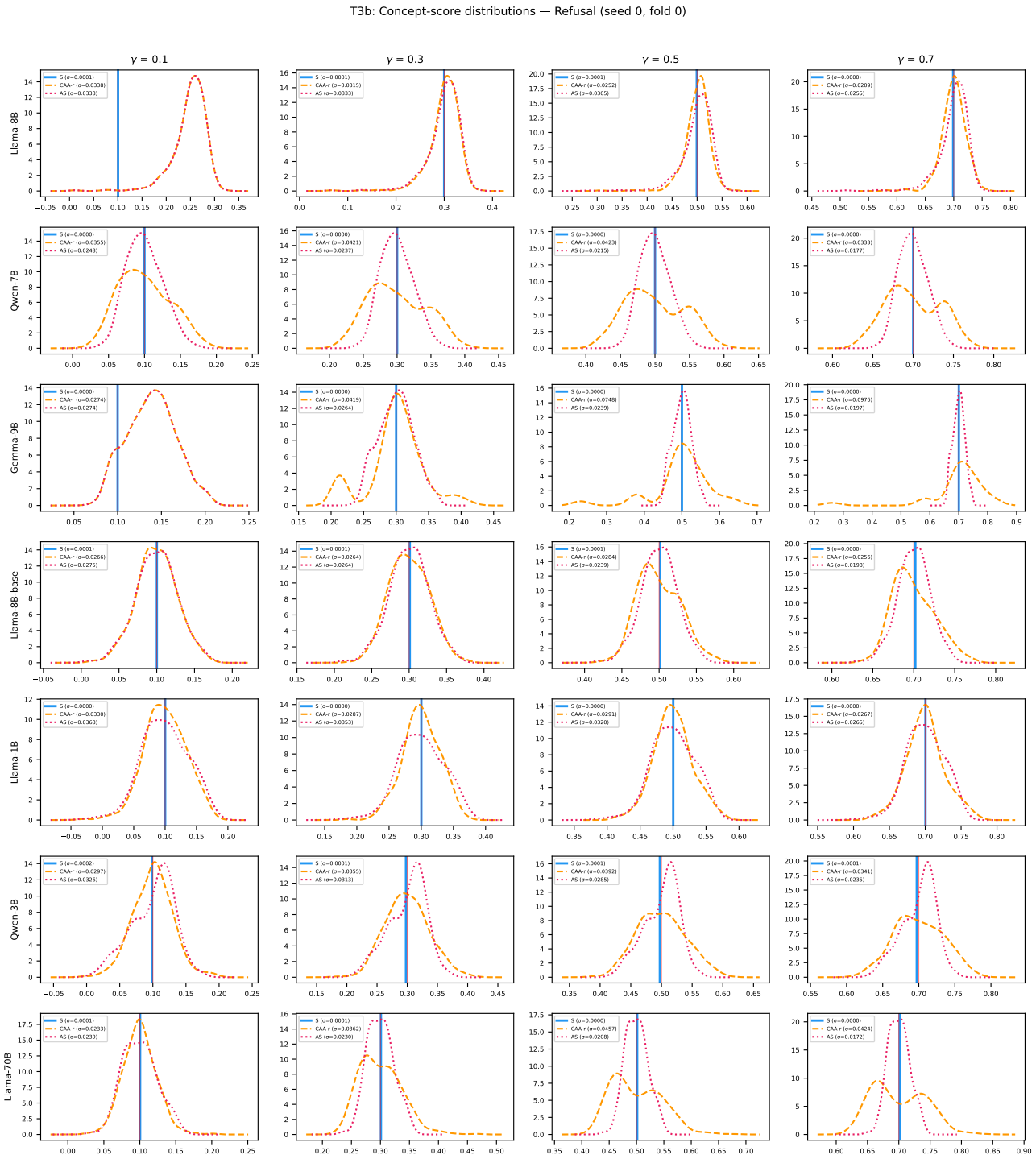

Figure 14. Concept-score distributions on Refusal (seed 0, fold 0, all 7 models).

Table 6. Fixed-strength downstream metrics at $\bar{\gamma} \in \{0.3,\ 0.7\}$ across 7 models and 5 datasets. Metric is dataset-specific (MC1, positive rate, non-toxicity, refusal rate).

			$\bar{\gamma} = 0.3$			$\bar{\gamma} = 0.7$		
Dataset	Model	Base	CAA	CAA-r	AS	CAA	CAA-r	AS
TQA	Llama-8B	.373	.373	.375	.365	.338	.338	.297
	Qwen-7B	.387	.380	.380	.365	.363	.375	.311
	Gemma-9B	.478	.480	.483	.488	.480	.483	.360
	Llama-8B-base	.319	.341	.338	.392	.365	.360	.343
	Llama-1B	.314	.358	.358	.360	.358	.324	.277
	Qwen-3B	.304	.311	.306	.299	.326	.292	.184
	Llama-70B	.458	.471	.473	.488	.461	.463	.245
SST-2	Llama-8B	.562	.545	.542	.480	.370	.355	.305
	Qwen-7B	.470	.463	.482	.460	.555	.542	.345
	Gemma-9B	.632	.685	.718	.723	.812	.752	.960
	Llama-8B-base	.525	.645	.623	.693	.703	.690	.420
	Llama-1B	.380	.465	.502	.905	.812	.855	.502
	Qwen-3B	.500	.482	.450	.477	.128	.100	.050
Civil	Llama-8B	.896	.922	.904	.889	.935	.913	.799
	Qwen-7B	.914	.920	.918	.924	.854	.851	.844
	Gemma-9B	.808	.793	.791	.801	.786	.802	.786
	Llama-8B-base	.862	.834	.841	.827	.636	.567	.901
	Llama-1B	.950	.934	.931	.865	.890	.931	.967
	Qwen-3B	.917	.888	.890	.900	.913	.914	.946
IMDB	Llama-8B	.550	.583	.605	.593	.757	.807	.680
	Qwen-7B	.522	.507	.507	.502	.588	.537	.535
	Gemma-9B	.590	.670	.715	.710	.720	.755	.917
	Llama-8B-base	.525	.650	.665	.672	.820	.823	.752
	Llama-1B	.455	.733	.740	.868	.927	.930	.965
	Qwen-3B	.500	.615	.595	.613	.723	.328	.507
Refusal	Llama-8B	.030	.040	.037	.037	.045	.050	.092
	Qwen-7B	.007	.003	.003	.003	.010	.007	.025
	Gemma-9B	.003	.005	.005	.005	.005	.005	.013
	Llama-8B-base	.000	.000	.000	.000	.000	.000	.000
	Llama-1B	.113	.048	.043	.007	.000	.000	.000
	Qwen-3B	.010	.000	.000	.003	.000	.000	.000

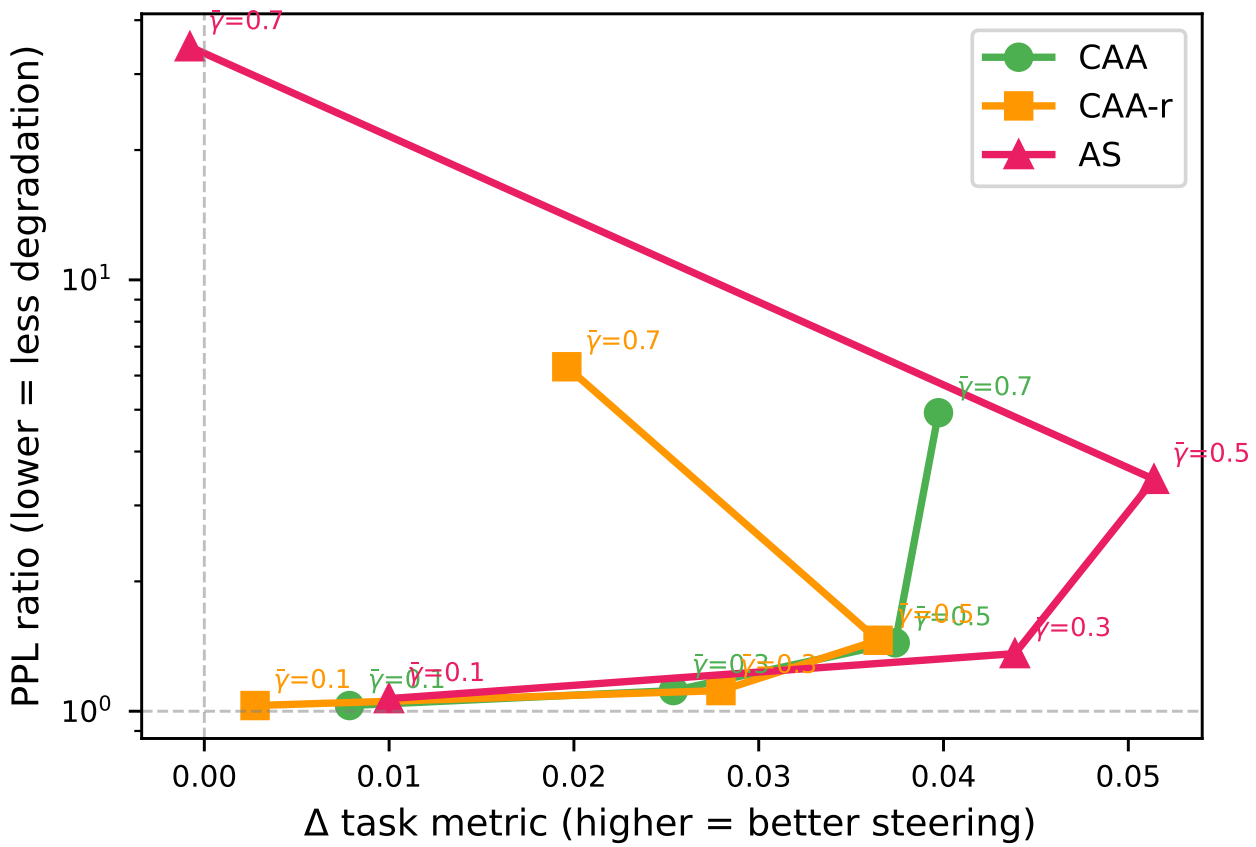

Figure 15. Aggregated Pareto frontier for fixed-strength methods (31 model×dataset cells). Each curve traces one method from $\bar{\gamma} = 0.1$ to 0.7. CAA and CAA-r overlap; AS achieves marginally higher task gain at $\bar{\gamma} = 0.5$ but at much higher PPL cost, and collapses at $\bar{\gamma} = 0.7$.

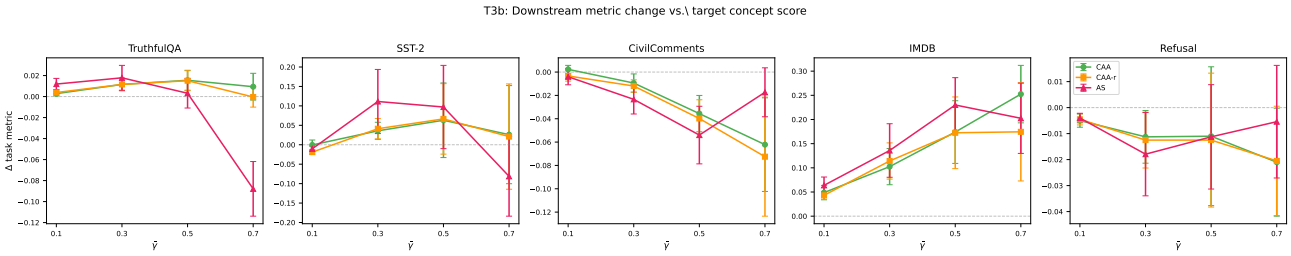

Figure 16. Mean downstream metric change (Δ task) vs. target concept score, averaged across models per dataset. Error bars show standard error. All three methods produce similar gains at moderate targets; AS diverges at high $\bar{\gamma}$ due to its larger per-token disruption.

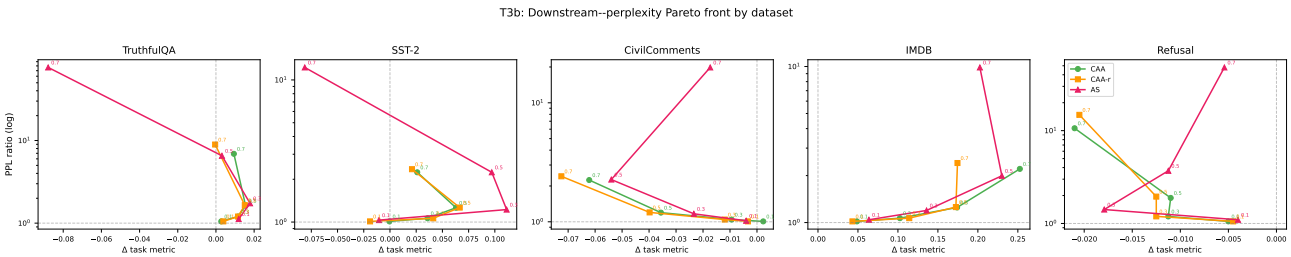

Figure 17. Per-dataset Pareto frontiers for fixed-strength methods. Each curve traces one method from $\bar{\gamma} = 0.1$ (lower-left, near baseline) to $\bar{\gamma} = 0.7$ (upper region, high PPL). CAA and CAA-r overlap closely; AS pays substantially higher perplexity for similar or lower task improvement. Labels indicate $\bar{\gamma}$ at each point.

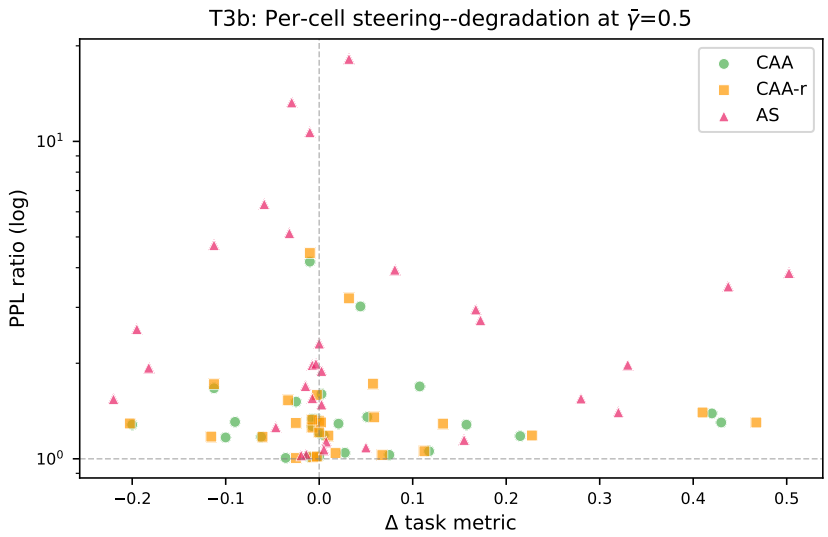

Figure 18. Per-cell scatter at $\bar{\gamma} = 0.5$. Each point is one model×dataset combination. AS (triangles) clusters at higher PPL ratios than CAA/CAA-r at comparable downstream changes.

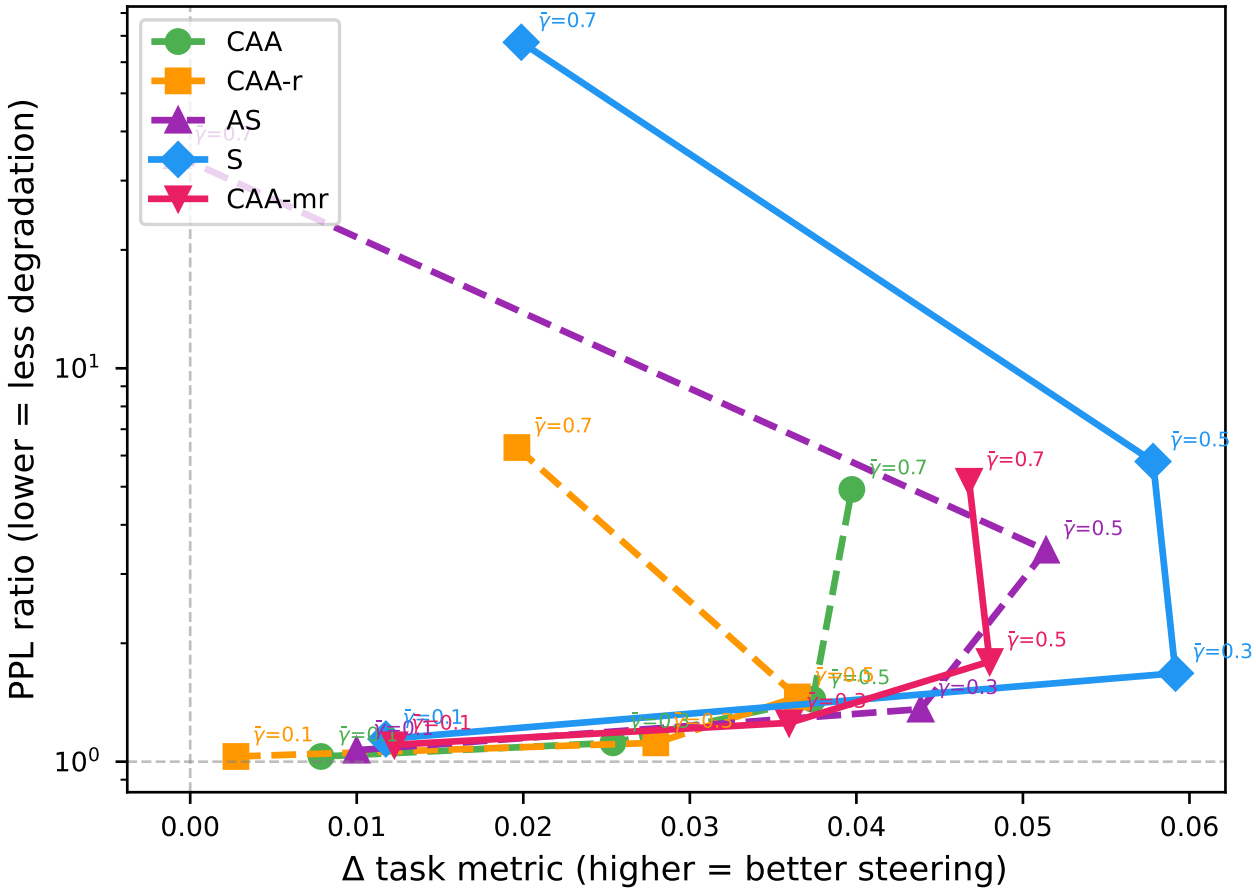

Figure 19. Aggregate Pareto frontier across the completed model–dataset grid. The x-axis is downstream task-metric change, so moving right indicates better steering; the y-axis is perplexity ratio on a log scale, so moving down indicates less degradation. Labels mark the target mean or per-token concept score.

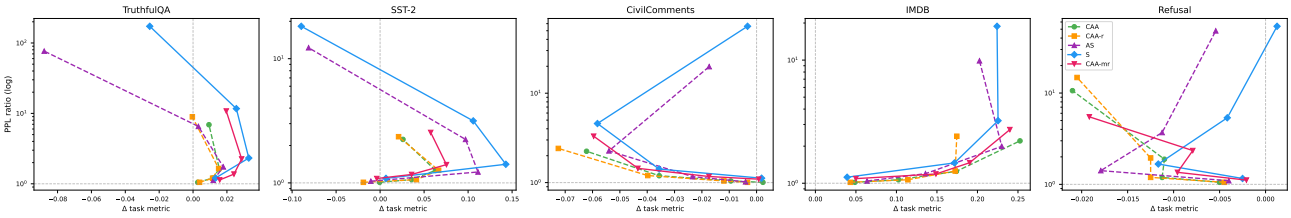

Figure 20. Per-dataset Pareto frontiers for the all-method comparison. The aggregate pattern is not driven by a single task: sentiment datasets allow the largest rightward movement, while TruthfulQA, CivilComments, and Refusal expose the cost of over-steering more clearly. High-target S or AS often moves upward much more than rightward, indicating degradation without commensurate task improvement.

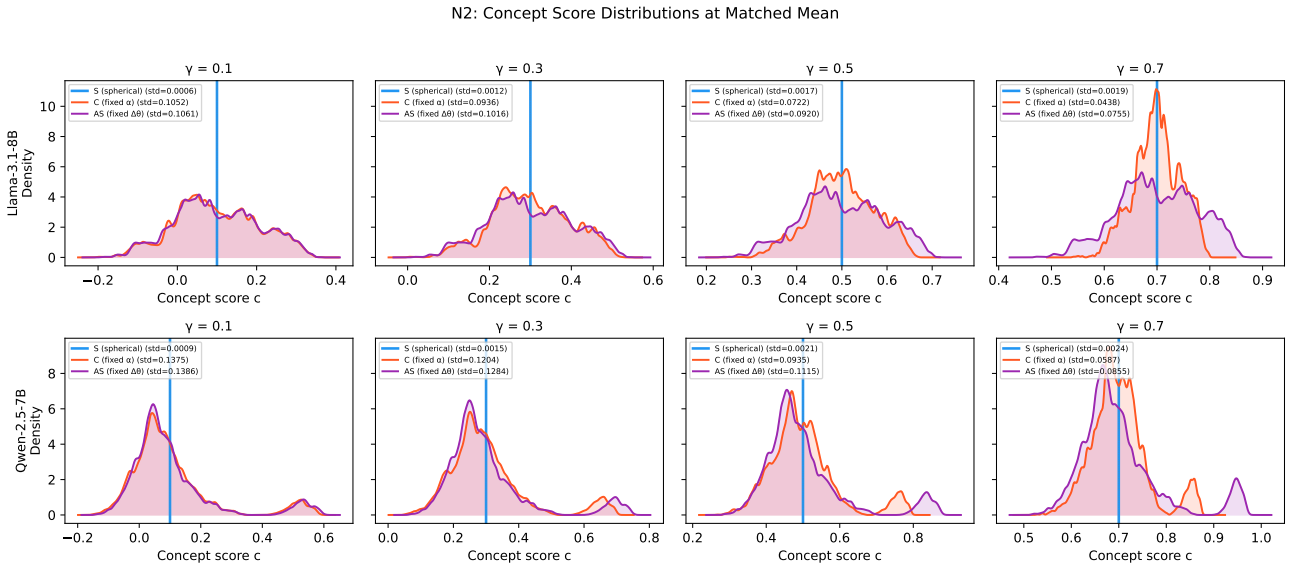

Figure 21. Initial concept-score distributions $c = \langle x/\|x\|_2, s\rangle$ across datasets and models. Most distributions are approximately centered near zero with moderate spread, explaining why a fixed additive α produces heterogeneous achieved concept scores after steering.

Table 7. Geodesic distance and concept-score deviation between S and CAA-mr, 7 models $\times$ 5 datasets (means over 2 folds, $\gamma = 0.3$). $d_{\mathbb{S}}$ is constant at 4.47×10^{-4} everywhere; only $|\Delta c|_{\max}$ varies.

	$\lvert\Delta c\rvert_{\max}$ at $\gamma = 0.3$				
Model	TruthfulQA	SST-2	CivilComm.	IMDB	Refusal
Llama-8B-Inst	1.8×10^{-5}	4.2×10^{-6}	3.0×10^{-5}	7.1×10^{-5}	3.3×10^{-5}
Qwen-7B-Inst	3.1×10^{-5}	8.6×10^{-6}	2.4×10^{-4}	6.3×10^{-5}	1.9×10^{-5}
Gemma-9B-IT	3.3×10^{-5}	1.4×10^{-6}	8.3×10^{-5}	4.0×10^{-6}	1.2×10^{-5}
Llama-8B base	8.3×10^{-6}	3.0×10^{-6}	1.9×10^{-5}	1.5×10^{-5}	1.0×10^{-5}
Llama-1B-Inst	4.7×10^{-5}	5.9×10^{-6}	2.2×10^{-5}	8.8×10^{-6}	4.5×10^{-6}
Qwen-3B-Inst	2.2×10^{-5}	1.6×10^{-6}	7.8×10^{-5}	1.4×10^{-5}	4.3×10^{-5}
Llama-70B-Inst	1.5×10^{-5}	4.8×10^{-6}	1.8×10^{-5}	3.0×10^{-6}	7.3×10^{-6}

Table 8. Downstream metric difference Δmetric = CAA-mr − S at $\gamma = 0.3$, 7 models × 5 datasets. Per-dataset metric: MC1 (TruthfulQA), positive rate (SST-2, IMDB), non-toxicity (CivilComments), refusal rate (Refusal).

Model	TruthfulQA	SST-2	CivilComm.	IMDB	Refusal
Llama-8B-Inst	−0.001	+0.011	−0.008	+0.009	−0.003
Qwen-7B-Inst	−0.001	−0.026	+0.008	−0.021	0.000
Gemma-9B-IT	−0.003	−0.023	0.000	−0.009	0.000
Llama-8B base	−0.003	+0.031	+0.013	−0.023	0.000
Llama-1B-Inst	0.000	−0.004	+0.006	−0.004	0.000
Qwen-3B-Inst	−0.003	+0.009	+0.006	−0.013	0.000
Llama-70B-Inst	+0.005	−0.031	−0.006	+0.004	0.000

Table 9. Aggregate degradation when perturbing off the spherical arc. Values are ratios (PPL, KL) or absolute changes (MMLU, downstream) relative to $\delta = 0$, averaged across 40 (model, dataset, γ) cells and up to 8 perturbation directions per cell.

	δ=−0.2	δ=−0.1	δ=+0.1	δ=+0.2
PPL ratio	×1.22	×1.05	×1.06	×1.23
KL ratio	×1.97	×1.23	×1.28	×2.19
ΔMMLU	−0.014	−0.003	−0.003	−0.012
ΔDown	−0.005	+0.001	−0.004	−0.013

Table 10. PPL ratio by perturbation-direction type, averaged across all cells. PCA directions produce the steepest valleys, consistent with the residual subspace encoding the most task-relevant variance.

Direction type	δ=−0.2	δ=−0.1	δ=+0.1	δ=+0.2
Random (n=2)	×1.15	×1.03	×1.06	×1.21
PCA (n=2)	×1.43	×1.09	×1.07	×1.32
Cross-dataset (n=4)	×1.20	×1.05	×1.06	×1.23

Table 11. Per-corpus CV (%) of hidden-state norms at the 50%-depth steering layer (last prompt token, mean across 1000 samples per corpus).

Corpus	Llama-8B	Qwen-7B	Gemma-9B	Llama-8B base	Llama-1B	Qwen-3B	Llama-70B
OpenWebText	8.61	10.29	59.44	7.92	7.31	11.17	14.43
Alpaca	6.51	8.12	59.16	6.43	6.43	8.14	11.55
arXiv	5.65	7.12	78.61	5.57	5.50	8.96	8.82
WritingPrompts	7.94	8.01	57.54	7.04	6.66	10.23	14.35
TruthfulQA	6.47	8.31	51.12	6.86	6.85	8.41	6.39
NaturalQuestions	6.85	7.85	53.64	6.58	8.61	9.70	5.59
CivilComments	5.57	7.45	73.03	5.65	5.79	7.65	8.17
CNN/DailyMail	8.01	7.42	24.97	6.96	6.78	8.89	13.50
PubMedQA	4.08	4.27	26.25	3.40	4.29	6.18	4.84
CodeSearchNet	5.76	7.64	52.28	5.63	5.28	7.17	6.34

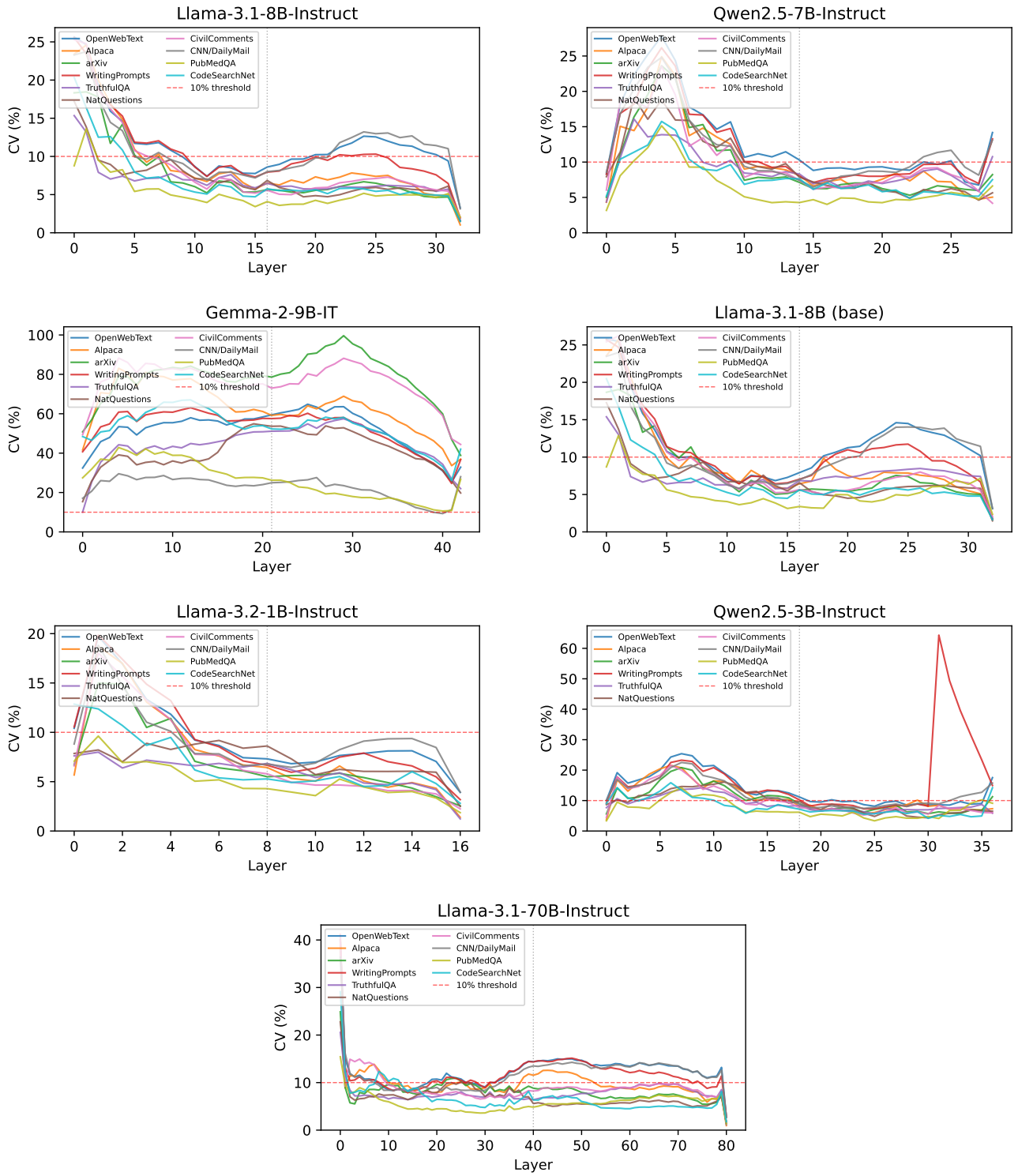

Figure 22. CV of hidden-state norms vs. layer for all 7 models, 10 corpora each. Red dashed line = 10% threshold; grey dotted line = 50%-depth steering layer. Llama and Qwen models stay below 10% at most layers; Gemma-2's post-norm architecture produces elevated CV throughout.

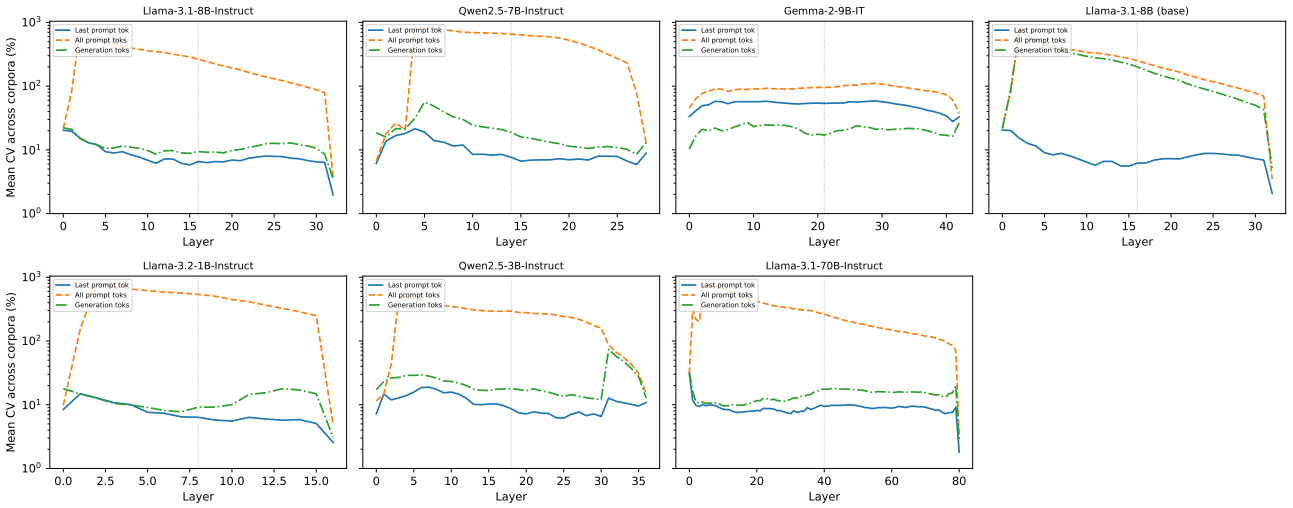

Figure 23. Mean CV across corpora for last prompt token (solid) vs. all prompt tokens (dashed). Position-dependent norm variation (attention sinks) inflates the all-token CV but does not affect angular structure.

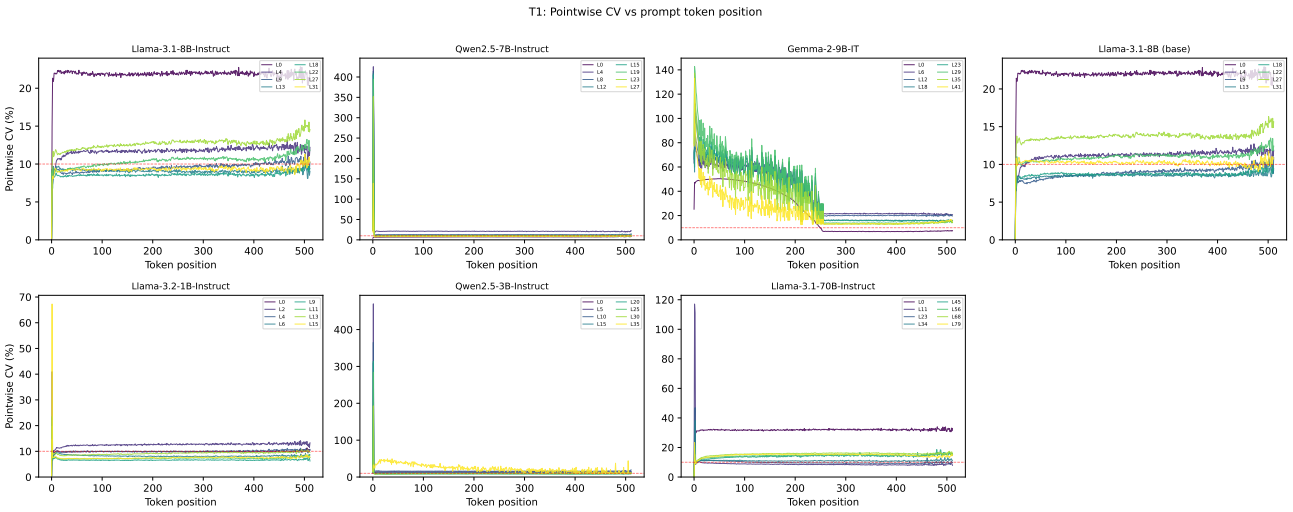

Figure 24. Pointwise CV at each prompt-token position. The spike at position 0 reflects the attention-sink token; CV settles to a stable plateau by position ≈10.

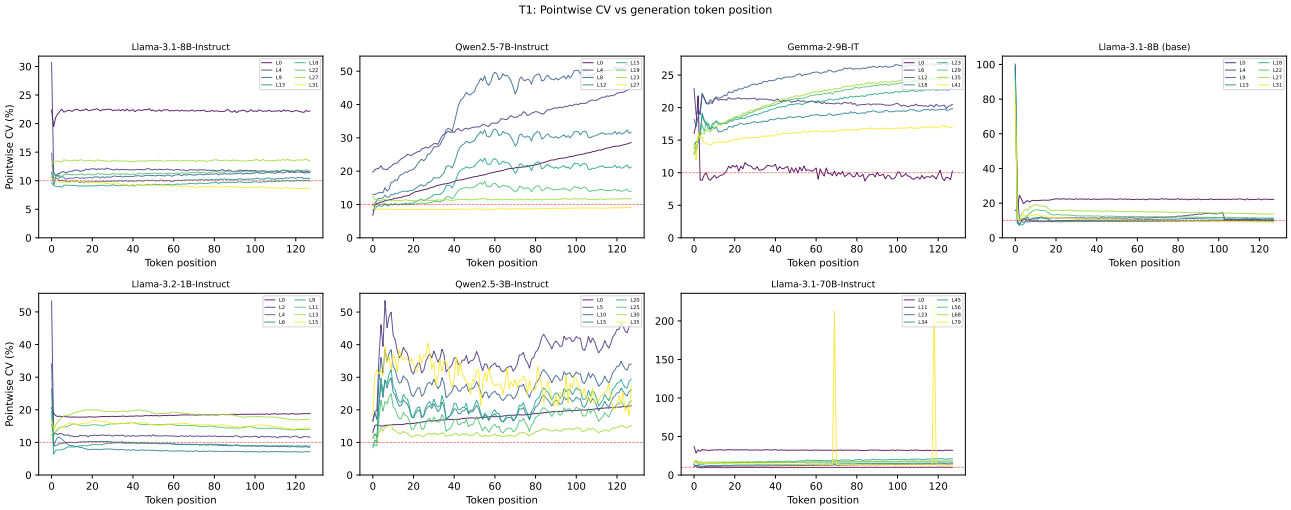

Figure 25. Pointwise CV at each generation-token position (128 tokens). For instruct-tuned models CV stays flat at 10–20%; Llama-8B base shows elevated CV at later layers.

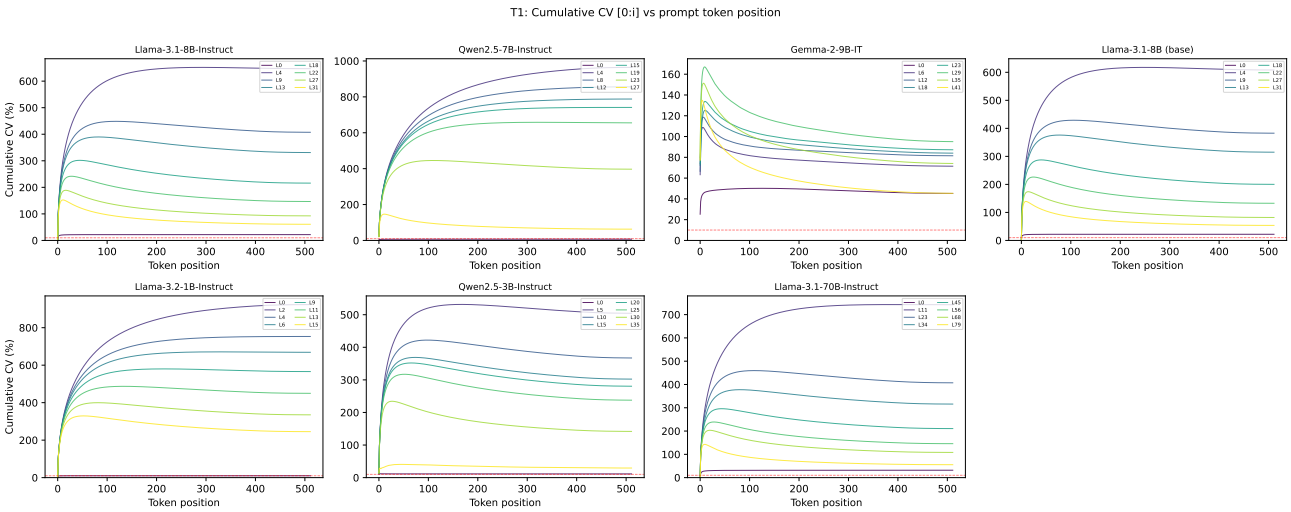

Figure 26. Cumulative CV over prompt tokens $[0{:}i]$. The initial spike from the attention-sink token is rapidly diluted as more positions are included.

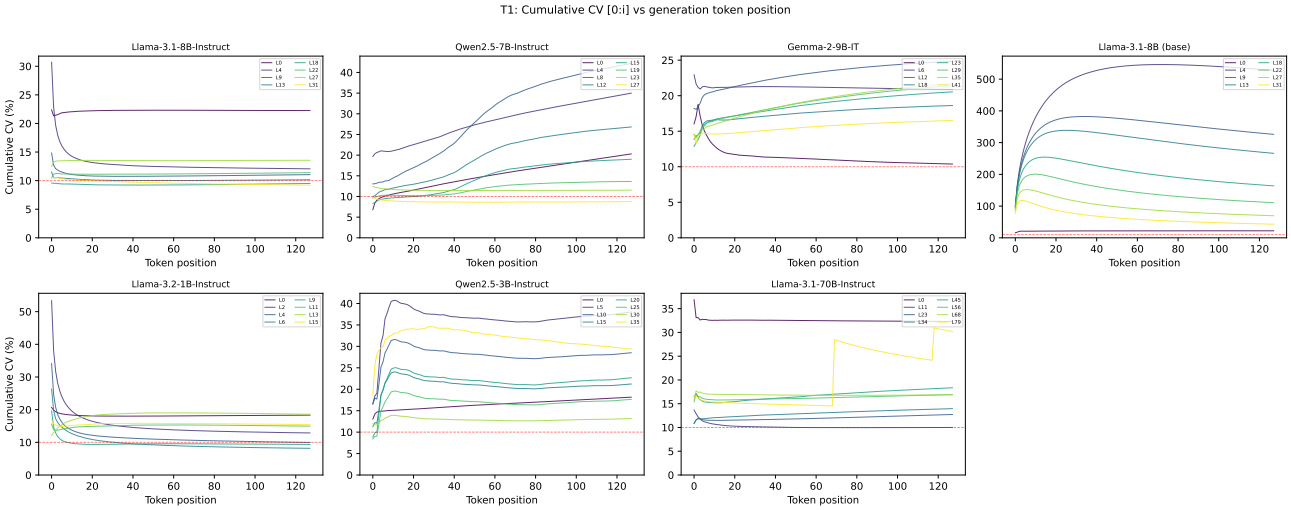

Figure 27. Cumulative CV over generation tokens $[0{:}i]$. Converges quickly, confirming the all-token CV is not dominated by outlier positions.

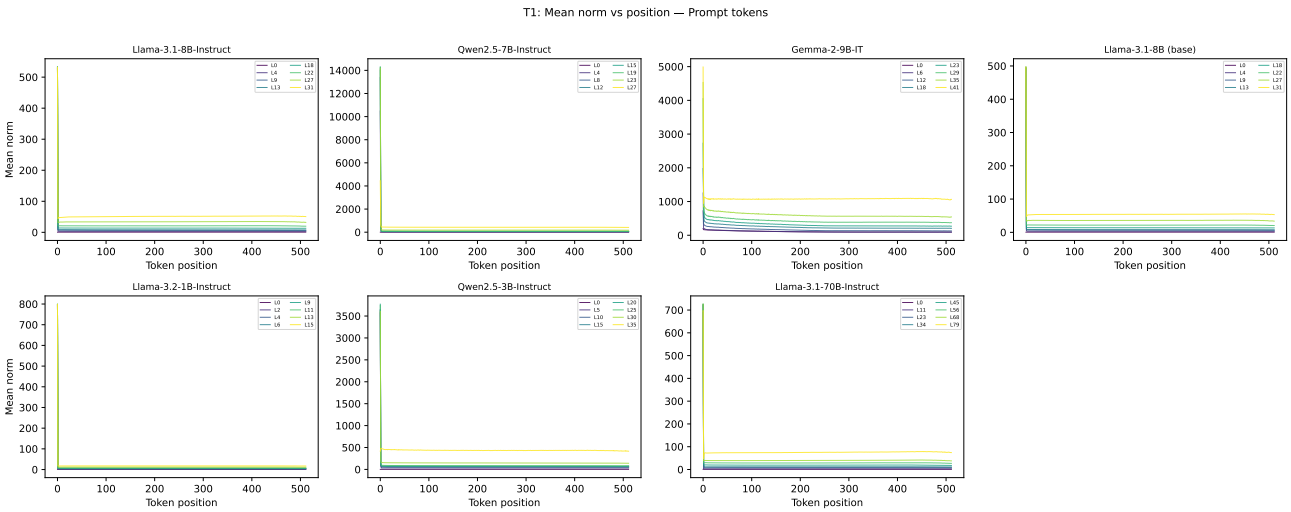

Figure 28. Mean hidden-state norm vs. prompt-token position. Norms grow monotonically with layer depth. The attention-sink token (position 0) has disproportionately large norms in Llama/Qwen architectures.

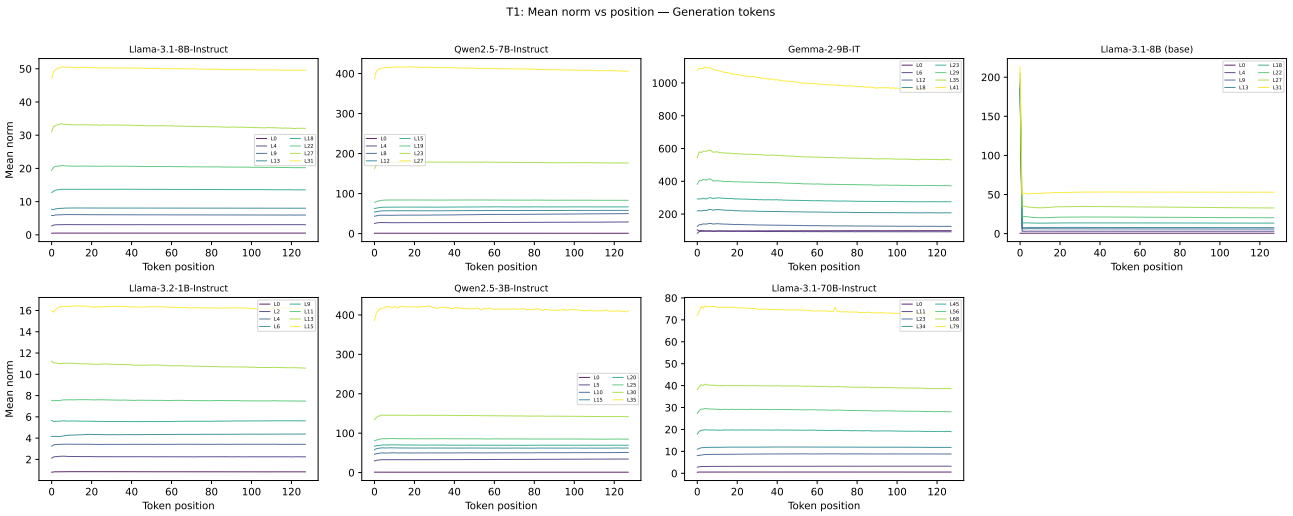

Figure 29. Mean hidden-state norm vs. generation-token position (128 tokens). Norms are nearly constant across positions at each layer, confirming that the steering hook encounters a stable norm landscape during decoding.

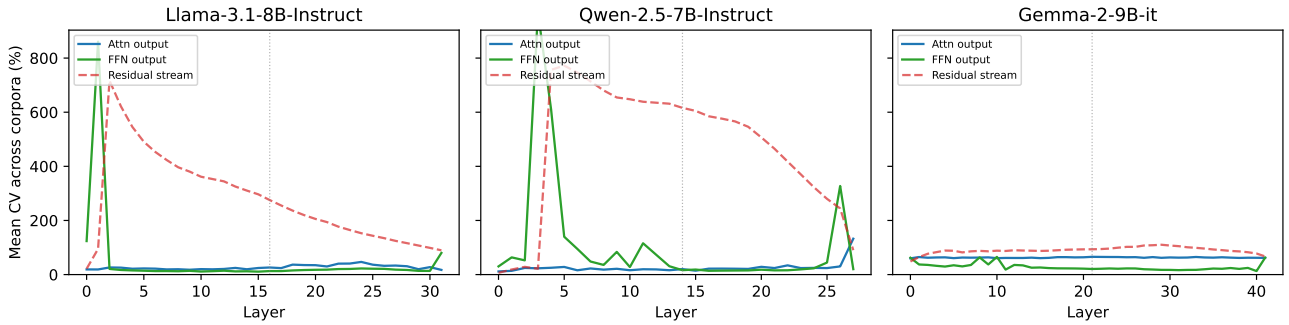

Figure 30. Sublayer CV decomposition showing norm variation at attention and MLP sublayer boundaries. This helps identify where within the transformer block norm changes are most concentrated.

Table 12. Full probe accuracies (%) at the 50%-depth steering layer. 5-fold stratified CV; 6 probes per cell (linear/MLP × raw/normalized/norm-only), 5 concept datasets, 7 models.

Model	Dataset	Lin(raw)	Lin(norm)	Lin($\|h\|$)	MLP(raw)	MLP(norm)	MLP($\|h\|$)
Llama-8B-Inst	TruthfulQA	81.6	81.1	53.8	83.3	83.1	52.1
Llama-8B-Inst	SST-2	92.5	92.4	53.4	92.0	91.7	53.1
Llama-8B-Inst	CivilComments	84.0	83.9	53.1	81.5	81.6	52.0
Llama-8B-Inst	IMDB	88.5	88.4	47.9	86.7	86.5	49.4
Llama-8B-Inst	Refusal	100.0	99.9	92.5	100.0	100.0	91.9
Qwen-7B-Inst	TruthfulQA	70.1	70.1	57.9	68.4	68.9	58.9
Qwen-7B-Inst	SST-2	91.5	91.4	51.0	91.2	91.2	49.9
Qwen-7B-Inst	CivilComments	80.6	80.2	52.9	80.1	80.0	50.3
Qwen-7B-Inst	IMDB	87.3	86.9	53.7	83.8	83.4	51.0
Qwen-7B-Inst	Refusal	99.6	99.8	84.1	99.9	99.9	84.9
Gemma-9B-IT	TruthfulQA	57.6	58.4	51.7	64.3	64.5	47.6
Gemma-9B-IT	SST-2	71.6	71.0	48.8	71.1	71.5	53.2
Gemma-9B-IT	CivilComments	65.9	65.0	46.6	63.4	63.3	49.6
Gemma-9B-IT	IMDB	75.4	74.5	51.1	71.1	70.5	49.9
Gemma-9B-IT	Refusal	91.2	90.1	49.6	91.1	91.4	58.9
Llama-8B base	TruthfulQA	80.4	80.6	51.5	80.1	79.9	50.8
Llama-8B base	SST-2	92.8	93.0	49.7	91.2	91.0	51.6
Llama-8B base	CivilComments	84.0	84.0	56.8	82.8	83.0	54.2
Llama-8B base	IMDB	88.9	88.9	53.9	86.4	86.4	49.9
Llama-8B base	Refusal	99.8	100.0	76.8	99.9	99.9	77.5
Llama-1B-Inst	TruthfulQA	73.6	74.0	56.2	74.6	75.3	56.0
Llama-1B-Inst	SST-2	88.9	88.9	51.0	90.0	90.1	50.3
Llama-1B-Inst	CivilComments	78.6	78.4	48.1	78.4	78.2	49.1
Llama-1B-Inst	IMDB	83.6	83.8	51.0	83.4	83.0	48.4
Llama-1B-Inst	Refusal	99.8	99.8	73.4	99.6	99.6	74.5
Qwen-3B-Inst	TruthfulQA	66.2	66.5	56.1	65.5	65.0	56.0
Qwen-3B-Inst	SST-2	87.5	87.5	52.9	89.7	89.4	50.5
Qwen-3B-Inst	CivilComments	76.1	76.5	54.6	77.7	78.6	52.4
Qwen-3B-Inst	IMDB	79.5	79.9	55.1	79.9	79.0	55.4
Qwen-3B-Inst	Refusal	99.8	99.8	88.6	99.9	99.7	90.0
Llama-70B-Inst	TruthfulQA	86.0	85.5	50.5	86.7	85.4	50.1
Llama-70B-Inst	SST-2	92.0	92.5	45.2	92.1	91.9	55.0
Llama-70B-Inst	CivilComments	83.9	84.5	47.1	82.4	82.0	48.4
Llama-70B-Inst	IMDB	91.2	91.0	52.4	89.1	88.1	52.5
Llama-70B-Inst	Refusal	100.0	99.9	93.1	99.7	99.9	93.1

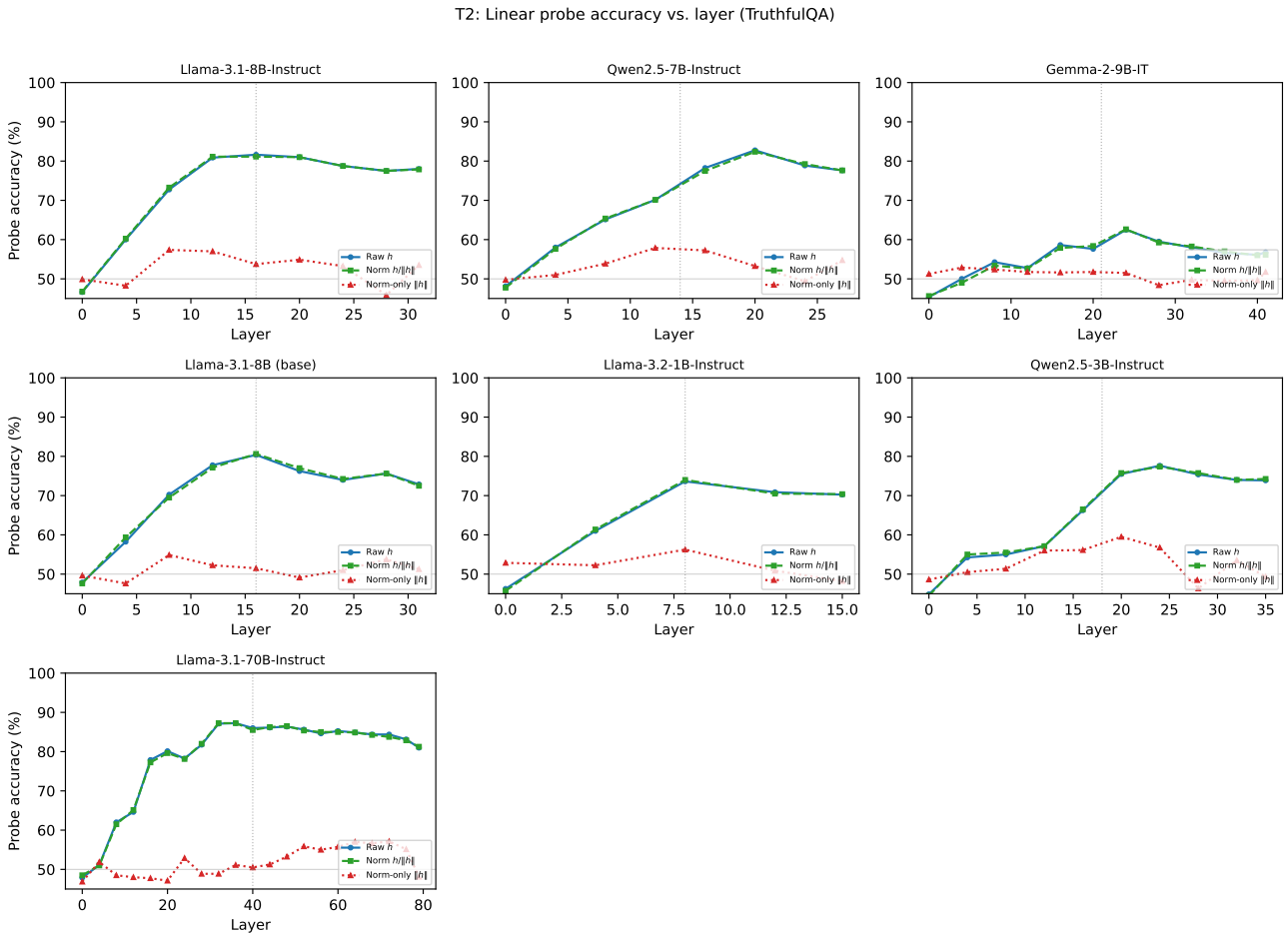

Figure 31. Probe accuracy vs. layer across the full model depth. Concept discriminability emerges gradually, peaks in the middle layers, and remains high through the final layers. The raw–normalized gap stays near zero at every layer.

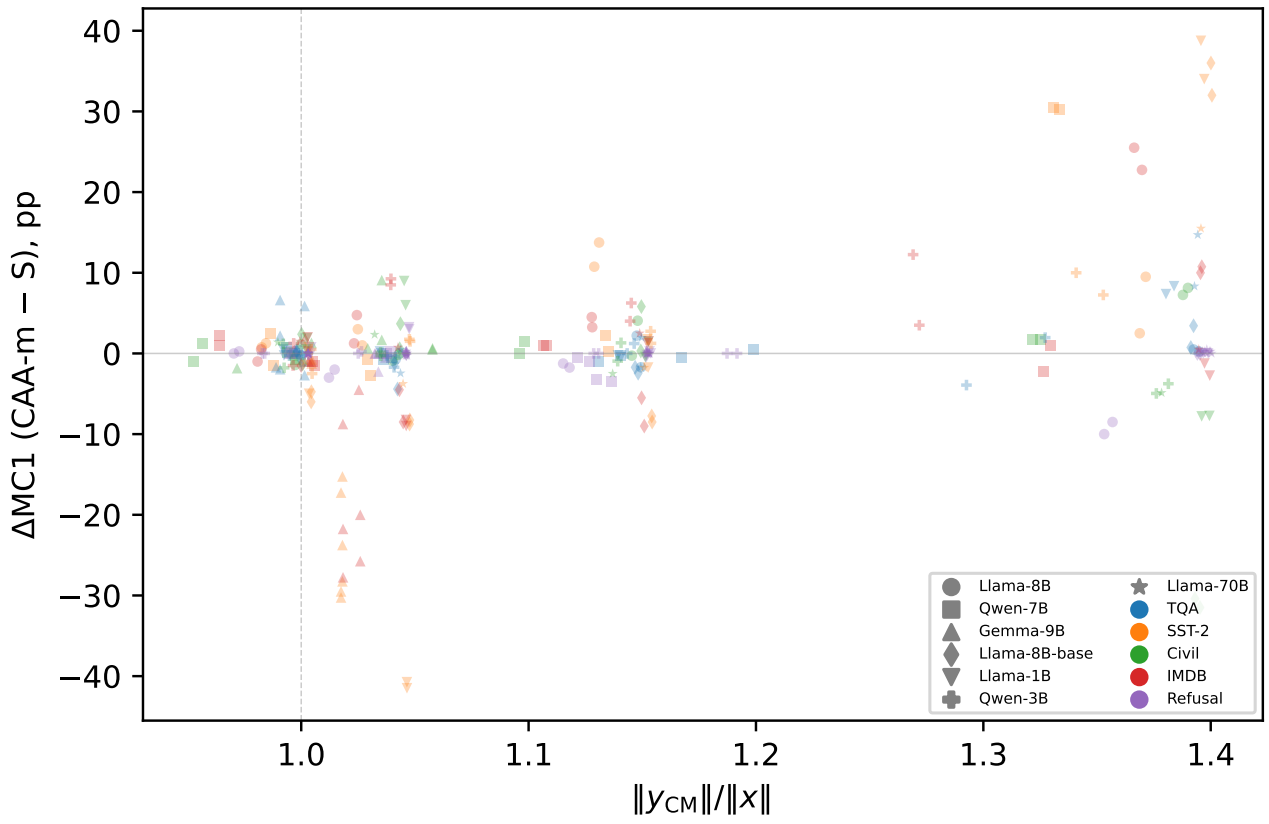

Figure 32. Per-fold norm ratio vs. Δmetric across all models, datasets, and γ values. Points near $(1.0, 0)$ indicate no distortion; the fan-out at higher norm ratios shows that skipping renormalization hurts unpredictably.

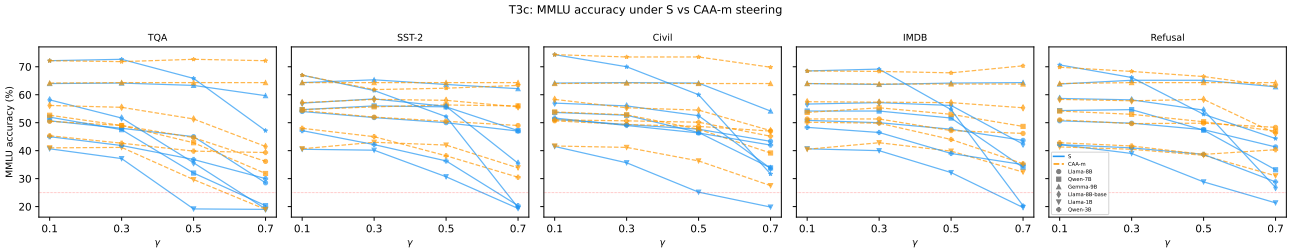

Figure 33. MMLU accuracy under S (solid) vs. CAA-m (dashed), 7 models × 5 steering datasets. CAA-m retains 5–35 pp more knowledge at $\gamma = 0.7$. Red line = chance (25%).

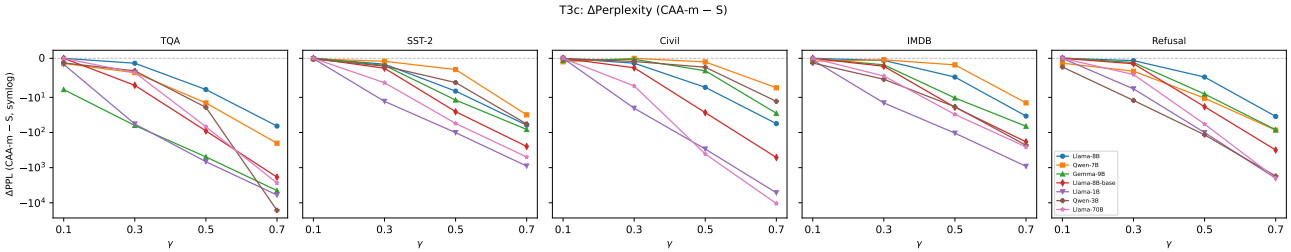

Figure 34. ΔPPL (CAA-m − S) vs. γ, 7 models × 5 datasets. Negative values mean CAA-m has lower perplexity. The gap grows super-linearly with γ.

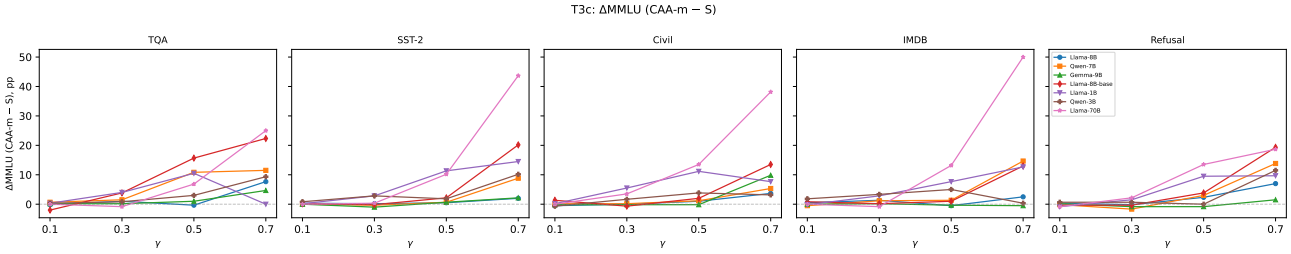

Figure 35. ΔMMLU (CAA-m − S) in pp vs. γ. Positive values mean CAA-m retains more knowledge. At $\gamma = 0.7$, CAA-m preserves up to $+35$ pp more MMLU accuracy (Llama-70B: 67.7% vs. 32.6%).