PATHWAY-SELECTIVE ACTIVATION STEERING: LLM CONTROL THROUGH NORMALIZED ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering can improve a language model's target behavior, but increasing steering strength can severely degrade language modeling. We study how this trade-off depends on where the intervention enters the computation. In pre-normalized Transformers, residual-stream addition introduces a direct carry that grows with steering strength, even though the normalized branch outputs remain bounded. This distinction motivates **Pathway-Selective Activation Steering (PSAS)**, which routes a normalized auxiliary steered state through the query and value pathways while preserving the original keys and identity path. We prove coefficient-independent bounds on its attention-output and full-layer perturbations and characterize its large-strength limit. Across three models, PSAS achieves the highest mean joint truthfulness–informativeness score over six steering methods. On two diagnostic models, removing CAA's direct carry reduces its excess Wikipedia perplexity by 79–89%, while PSAS exhibits substantially lower sensitivity to coefficient overshoot. Pathway ablations show that Q/V steering offers a favorable trade-off between behavioral control and generation quality. These results show how intervention placement controls perturbation growth and motivate selecting pathway composition and steering strength jointly.

1 INTRODUCTION

Activation steering controls language-model behavior by modifying internal representations at inference time (Subramani et al., 2022; Todd et al., 2024; Turner et al., 2025; Su et al., 2025; Vu & Nguyen, 2025; Heyman & Vandeputte, 2026). Contrastive Activation Addition (CAA) (Rimsky et al., 2024; Marks & Tegmark, 2024), for example, estimates a direction from contrasting examples and adds the scaled direction to the residual stream. The coefficient determines the strength of the intervention, but increasing it can degrade generation quality and need not improve the target behavior monotonically (Stoehr et al., 2024). A useful steering method must therefore balance behavioral control against unintended effects on the rest of the model's computation (Vogels et al., 2026; Li et al., 2026; Rivera & Africa, 2026).

One approach is to change how the update is computed. Linear-AcT uses an activation-dependent transport map (Rodriguez et al., 2025), ODESteer formulates steering through a dynamical system (Zhao et al., 2026), and PID steering introduces feedback terms (Nguyen et al., 2026). We study a complementary design choice: *which inputs of a Transformer block should receive the steering direction?* (Sun et al., 2025; Ye et al., 2026)

In a pre-normalization block, a residual-stream addition affects two computational paths. It changes the normalized inputs read by the attention and MLP branches, and it enters the skip connection as a direct addition $\tau\mathbf{r}$ (Elhage et al., 2021; Meng et al., 2022). We call the latter contribution the *carry*. Under the architectural assumptions considered here, the branch outputs remain bounded as $|\tau|$ grows (Zhang & Sennrich, 2019), while the carry grows linearly, with norm $|\tau|\|r\|$. Its persistence through later layers underlies the input-independent large-coefficient limit of residual steering established by Taimeskhanov et al. (2026). Removing the carry eliminates this unbounded term, but does not by itself guarantee either preserved generation quality or input-dependent final predictions.

This observation motivates **Pathway-Selective Activation Steering (PSAS)**: form and normalize an auxiliary steered state, then route it through selected attention pathways while retaining the original identity path. We route the steered state only through the query and value pathways: queries and values are derived from $h_l+\tau r$, keys read the original normalized state, and the skip connection retains h_l (Figure 1). The MLP and later layers respond to the changed attention output without receiving a direct copy of $\tau\mathbf{r}$. Removing the carry, however, does not by itself explain PSAS's behavior. A norm-constrained CAA

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.

baseline shows that magnitude clipping alone is insufficient; PSAS's pre-normalization pathway produces a bounded, token-state-dependent intervention.

Prior work has explored activation steering at the level of selected attention heads or subspaces (Li et al., 2023; Feng et al., 2026; Kang et al., 2026). In particular, DISCO (Torop et al., 2025) steers the query and value pathways of selected attention heads. Our contribution is not query–value steering itself, but the pre-normalization construction, and an analysis of how the carry and pathway choice affect control and cost.

We make three contributions.

- **A structural analysis of steering pathways.** We derive an explicit convergence rate at which CAA approaches its input-independent limit. By contrast, we prove coefficient-independent bounds for PSAS, and characterize its attention-output limit at large steering strengths.
- **Reduced sensitivity to steering strength.** On two diagnostic models, PSAS exhibits substantially lower Wikipedia-perplexity sensitivity to coefficient growth, enabling exploration of higher steering strengths and their potential to yield stronger behavioral control.
- **Evidence on carry removal and pathway choice.** Removing CAA's carry reduces its excess Wikipedia perplexity by 79–89%. Across three models, PSAS achieves the highest mean joint truthfulness–informativeness score among the evaluated methods. Pathway ablations show that the benefits of pathway composition depend on the model and task.

2 Two Ways a Direction Can Enter a Block

Terminology. A pre-normalization block computes $\mathbf{h}_{l+1} = \mathbf{h}_l + \mathrm{Attn}_l(\mathrm{RMSNorm}_l^a(\mathbf{h}_l)) + \mathrm{MLP}_l(\mathrm{RMSNorm}_l^m(\mathbf{h}_l + \mathrm{Attn}_l(\mathrm{RMSNorm}_l^a(\mathbf{h}_l))))$, where the attention and MLP normalizations may have different gains. The *identity path* passes $\mathbf{h}_l$ through the skip connection; the *normalized branches* read normalized states. An intervention's *pathway* specifies which of these inputs receive an added direction: the identity path, the query, key, or value inputs, or the MLP input. CAA modifies all of them. Its direct identity-path addition $\tau\mathbf{r}$ enters the next residual state unchanged and remains an additive term in later states; we call this term the *carry*.

Bounded computation, unbounded carry. Let $\mathbf{h}_{l,i} \in \mathbb{R}^{1\times d}$ be the residual state at token i and $\mathbf{r} \neq \mathbf{0}$ a fixed direction. CAA sets $\mathbf{h}'_{l,i} = \mathbf{h}_{l,i} + \tau\mathbf{r}$ at each steered position, so the state entering the next layer is

$$\mathbf{h}_{l+1,i}^{\mathrm{CAA}} = \mathbf{h}_{l,i} + \tau\mathbf{r} + \mathbf{F}_{l,i}(\tau),$$

where $\mathbf{F}_{l,i}(\tau)$ collects the attention and MLP outputs. Both branches read RMS-normalized inputs, and with $\hat{\mathbf{r}} = \mathbf{r}/\|\mathbf{r}\|$,

$$\|\mathrm{RMSNorm}(\mathbf{h}+\tau\mathbf{r})\| \le \sqrt{d}\|\mathbf{g}\|_\infty \quad \text{for all } \tau, \qquad \mathrm{RMSNorm}(\mathbf{h}+\tau\mathbf{r}) \longrightarrow \sqrt{d}\,\mathbf{g}\odot\hat{\mathbf{r}} \quad \text{as } \tau\to+\infty, \tag{1}$$

where d is the hidden size and $\mathbf{g}$ is the learned gain; the bound holds wherever normalization is defined. Appendix B.3 includes the stability constant. Under Assumption 1, the attention and MLP outputs collected in $\mathbf{F}_{l,i}(\tau)$ are uniformly bounded, while the carry grows linearly (Proposition 3). The measured scalings reflect this distinction: at sixteen times the operating coefficient, CAA's full residual-change norms reach 2285 and 105 on the two diagnostic models, while its attention-output-change norms are approximately 200 and 12 (Table 6).

Loss of input dependence at large strength. The carry persists through downstream blocks: $\mathbf{h}_{l',i}^{\mathrm{CAA}} = \tau\mathbf{r} + \mathbf{B}_{l',i}(\tau)$, with $\mathbf{B}_{l',i}$ bounded for a fixed context. Consequently, each normalized state approaches a layer-specific vector determined by $\mathbf{r}$ and the normalization gain, and the next-token distribution approaches an input-independent predictor. We call this asymptotic loss of input dependence *swamping*. The limit is due to Taimeskhanov et al. (2026); Proposition 4 supplies a rate. Appendix F.3 shows declining across-prompt variance at large coefficients, but substantial perplexity degradation occurs earlier. On Qwen3-1.7B, for example, perplexity reaches 656 at $2\times$ while across-prompt variance remains 0.163, compared with 0.181 unsteered. The asymptotic limit therefore does not locate the onset of practical degradation.

Isolating the carry. Let $F_{\mathrm{NW},i}$ use CAA's attention and MLP outputs but omit the identity-path addition; then, on identical block inputs,

$$F_{\mathrm{CAA},i}(\mathbf{h}_l;\tau) - F_{\mathrm{NW},i}(\mathbf{h}_l;\tau) = \tau\mathbf{r}. \tag{2}$$

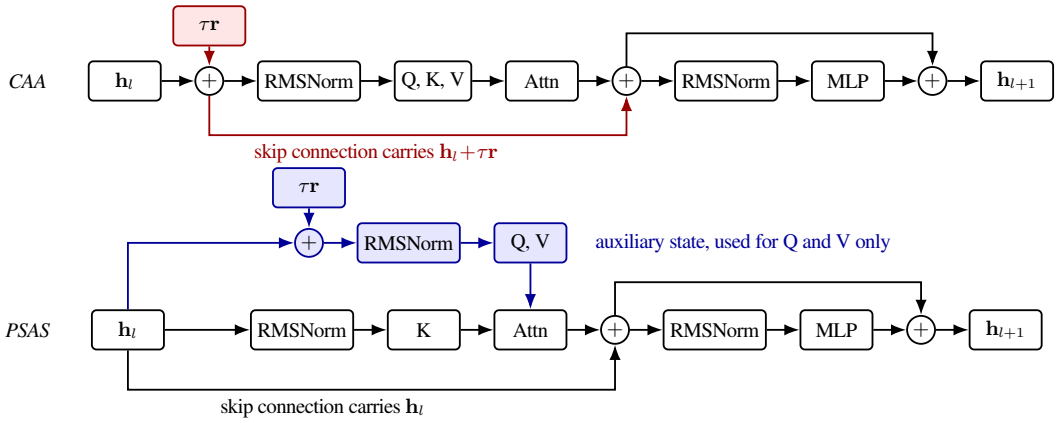


Figure 1: **Where the direction enters.** CAA adds $\tau\mathbf{r}$ to the residual state, so it reaches every branch and the skip connection. The evaluated PSAS configuration normalizes $\mathbf{h}_l+\tau\mathbf{r}$ on a separate auxiliary branch that feeds the query and value projections. Keys read the normalized original state, and the skip connection carries $\mathbf{h}_l$. The MLP receives no direct steering addition and changes only through the modified attention output.

At CAA's operating coefficient, adding $\tau\mathbf{r}$ after an otherwise unsteered block yields 55% and 58% of CAA's excess Wikipedia perplexity. Retaining CAA's branch computations while omitting the carry yields 11% and 21%; Q/V-only steering yields 6% and 4% (Appendix D). These are comparisons between distinct interventions, not additive components of perplexity. They motivate the carry-scaling experiment in Section 6.2, which measures how removing the carry changes both off-target cost and target behavior.

3 Pathway-Selective Activation Steering

3.1 Reading the steering direction

PSAS uses the standard difference-in-means direction. Let $\mathcal{D}^+=\{x_i^+\}_{i=1}^N$ and $\mathcal{D}^-=\{x_i^-\}_{i=1}^N$ be positive and negative examples, and let $\mathcal{A}_l(x)$ denote the residual activation at the selected layer and token. Define

$$\mathbf{r}=\frac{1}{N}\sum_{i=1}^{N}\mathcal{A}_l(x_i^+)-\frac{1}{N}\sum_{i=1}^{N}\mathcal{A}_l(x_i^-). \tag{3}$$

The distinction from CAA lies entirely in how this direction is applied during inference.

3.2 Separating the residual state from the steered state

Let $\mathbf{h}_l\in\mathbb{R}^{m\times d}$ stack the token states. Normalization acts row-wise, and adding $\tau\mathbf{r}$ broadcasts the direction to the steered positions. The following equations describe full-context steering. There are H query heads of dimension d_h; in grouped-query attention, each uses its assigned shared key and value projections. The output-projection block $\mathbf{W}_{l,h}^O$ has shape $d_h\times d$. The original normalized state and the auxiliary steered state are

$$\mathbf{x}_l=\mathrm{RMSNorm}_l^a(\mathbf{h}_l), \tag{4}$$

$$\tilde{\mathbf{h}}_l=\mathbf{h}_l+\tau\mathbf{r},\qquad \tilde{\mathbf{x}}_l=\mathrm{RMSNorm}_l^a(\tilde{\mathbf{h}}_l). \tag{5}$$

The auxiliary state is not substituted for $\mathbf{h}_l$ on the identity path.

Instead, $\tilde{\mathbf{x}}_l$ is used only to compute queries and values:

$$\tilde{\mathbf{Q}}_{l,h}=\tilde{\mathbf{x}}_l\mathbf{W}_{l,h}^Q,\quad \mathbf{K}_{l,h}=\mathbf{x}_l\mathbf{W}_{l,h}^K,\quad \tilde{\mathbf{V}}_{l,h}=\tilde{\mathbf{x}}_l\mathbf{W}_{l,h}^V. \tag{6}$$

The projections above precede any query/key normalization and rotary positional embeddings. Let $\tilde{\mathbf{Q}}_{l,h}^{\mathrm{logit}}$ and $\mathbf{K}_{l,h}^{\mathrm{logit}}$ denote the queries and keys after these standard model transformations. In Qwen3 this includes QK-Norm followed by RoPE; in Llama-3.1 it includes RoPE. Attention is then computed as

$$\tilde{\mathbf{A}}_{l,h}=\mathrm{softmax}\left(\frac{\tilde{\mathbf{Q}}_{l,h}^{\mathrm{logit}}(\mathbf{K}_{l,h}^{\mathrm{logit}})^\top}{\sqrt{d_h}}+\mathbf{M}\right),\qquad \tilde{\mathbf{O}}_{l,h}=\tilde{\mathbf{A}}_{l,h}\tilde{\mathbf{V}}_{l,h}. \tag{7}$$

Here $\mathbf{M}$ is the fixed causal mask (and any additional model attention mask), with zero entries on allowed positions and $-\infty$ elsewhere; each query has at least one allowed position. After the standard output projection, the steered attention output is added to the unmodified input residual $\mathbf{h}_l$, as in the original Transformer. The residual stream can therefore change through attention, but $\tau\mathbf{r}$ is never added to it directly.

The intervention can therefore be summarized as

$$\begin{aligned} &\text{CAA:} && \mathbf{h}_l+\tau\mathbf{r} && \longrightarrow \text{attention, MLP, and residual state,} \\ &\text{PSAS:} && \mathrm{RMSNorm}(\mathbf{h}_l+\tau\mathbf{r}) && \longrightarrow \mathbf{Q},\mathbf{V} \text{ only.} \end{aligned} \tag{8}$$

We distinguish three intervention quantities: the *direct residual addition* $\tau\mathbf{r}$; the *attention-output change* $\Delta\mathbf{o}_i = \tilde{\mathbf{o}}_i - \mathbf{o}_i$ relative to the unsteered computation; and the *full-layer residual change* $\Delta\mathbf{h}_{l+1,i}=\mathbf{h}^{\text{steered}}_{l+1,i}-\mathbf{h}_{l+1,i}$. For PSAS, $\Delta\mathbf{h}_{l+1,i}=\Delta\mathbf{o}_i+\Delta\mathbf{m}_i$, where $\Delta\mathbf{m}_i$ is the induced MLP-output change. CAA additionally contributes $\tau\mathbf{r}$ at each steered position, with its own attention and MLP changes. Thus preserving the input identity path does not leave the residual stream unchanged.

3.3 Evaluated pathway configuration

Our evaluated configuration routes the auxiliary state to queries and values, allowing it to change both attention weights and the vectors being mixed. The fixed-coefficient pathway ablation in Section 6.3 shows that adding query steering to value steering reduces toxicity on both models, while its effect on TruthfulQA depends on the model (Figure 2). In separate five-seed sweeps, adding Q on Qwen3-1.7B raises the mean joint score from 0.526 to 0.548 at nearly identical Wikipedia excess cross-entropy (Table 22). These findings motivate studying Q+V while leaving pathway composition dependent on the target behavior, model, and operating strength. The central design principle is selective routing through normalized attention inputs while preserving the identity path.

A common shift to every key *after* all query/key transformations leaves softmax unchanged because it adds the same offset within each query row. This invariance does not generally apply before RoPE or QK-Norm: RoPE rotates a shared shift differently across positions, and QK-Norm makes the effect state dependent. Indeed, K-only changes Qwen3-1.7B's attention weights more than Q-only at the operating coefficient (Table 19). Preserving the original keys therefore restricts the intervention to query-side reweighting and value modification. Table 22 compares Q+K+V and Q+V at separately selected strengths; these results do not establish a universal ranking of the two configurations.

Relation to query/value steering. DISCO steers queries and values directly after projection, whereas Communication Steering perturbs normalized attention inputs (Torop et al., 2025; Rodriguez et al., 2025). PSAS instead adds the residual-space direction *before* normalization. Its normalized-input shift depends on the token state and remains bounded as $|\tau|$ grows. At the evaluated operating strengths, that shift is nevertheless closely aligned with $\mathbf{g}\odot\mathbf{r}$ (Table 18). The clearest structural distinction is bounded growth: a fixed post-projection value shift contributes $\alpha\mathbf{v}_*\mathbf{W}^O$ to the attention output and is unbounded when this projected direction is nonzero. Appendix I.2 compares the responses to coefficient growth, while Appendix I.3 examines token-dependent value modulation.

4 Analysis of the PSAS Perturbation

We separate two questions: how large the intervention can become, and what it approaches as the coefficient grows. Fix a context, a layer, a head, and a nonzero direction $\mathbf{r}$. Token vectors are rows, vector norms are Euclidean, and matrix $\|\cdot\|_2$ norms are spectral. Attention runs without dropout, with a fixed nonempty allowed-key set for each query. Values are linear projections of RMS-normalized states. Assumption 1 states the full architecture conditions; Appendix B.3 supplies the proofs.

Let

$$\mathbf{x}_j=\mathrm{RMSNorm}(\mathbf{h}_j), \qquad \tilde{\mathbf{x}}_j=\mathrm{RMSNorm}(\mathbf{h}_j+\tau\mathbf{r}), \qquad \Delta\mathbf{x}_j=\tilde{\mathbf{x}}_j-\mathbf{x}_j.$$

Write ϕ_i^Q and ϕ_j^K for the maps to the queries and keys actually used in the logits, including RoPE and, where present, stabilized QK-Norm. Define

$$\mathbf{b}_i=\phi_i^Q(\tilde{\mathbf{x}}_i)-\phi_i^Q(\mathbf{x}_i), \qquad \mathbf{c}_j=\Delta\mathbf{x}_j\mathbf{W}^V. \tag{9}$$

Thus $\tilde{\mathbf{q}}_i=\mathbf{q}_i+\mathbf{b}_i$, $\tilde{\mathbf{v}}_j=\mathbf{v}_j+\mathbf{c}_j$, and $\mathbf{k}_j=\phi_j^K(\mathbf{x}_j)$ is unchanged. The head's output-projection block is $\mathbf{W}^O\in\mathbb{R}^{d_h\times d}$. We distinguish the steered output $\tilde{\mathbf{o}}_i$ from the attention-output change $\tilde{\mathbf{o}}_i-\mathbf{o}_i$.

4.1 QUERY REWEIGHTING AND VALUE MODIFICATION

Lemma 1 (Query tilt and value averaging). *On the allowed keys, query steering gives* $\delta_{ij} = \mathbf{b}_i \mathbf{k}_j^\top / \sqrt{d_h}$ *and* $\tilde{a}_{ij} \propto a_{ij} \exp(\delta_{ij})$. *With fixed values, it changes their convex mixture. With fixed attention weights, value steering adds* $\sum_j a_{ij} \mathbf{c}_j$, *whose projected norm is at most* $\|\mathbf{W}^O\|_2 \max_j \|\mathbf{c}_j\|$.

The query identity uses the transformed queries and keys that enter the logits, so it applies with both RoPE and QK-Norm. The value identity holds for any fixed probability weights. Together they separate reweighting existing values from changing the values being mixed; they do not establish a behavioral benefit for either pathway. Appendix B.4 gives the combined Q+V expression.

4.2 A UNIFORM BOUND ON THE INTERVENTION

Proposition 1 (Uniform boundedness of the attention-output change). *Let* $M = \sqrt{d} \|\mathbf{g}\|_\infty \|\mathbf{W}^V\|_2$, *where* $\mathbf{g}$ *is the attention-input RMSNorm gain. Under Assumption 1, for every admissible* τ, *every query position, and any probability attention weights on the allowed keys,*

$$\|\tilde{\mathbf{o}}_i - \mathbf{o}_i\| \le 2M \|\mathbf{W}^O\|_2.$$

The bound is independent of context length. For H *query heads, the complete attention-output change is bounded by* $2\|\mathbf{W}^O_{\text{all}}\|_2 (\sum_{h=1}^H M_h^2)^{1/2}$, *where* $M_h = \sqrt{d} \|\mathbf{g}\|_\infty \|\mathbf{W}_h^V\|_2$ *and* $\mathbf{W}^O_{\text{all}}$ *projects the concatenated head outputs.*

Both the original and steered values have norm at most M, so their attention-weighted mixtures also have norm at most M. The triangle inequality and output projection give the bound, without controlling the change in attention weights. This argument permits a coefficient-independent, context-length-independent bound, but not a tight prediction of perturbation size. At $16\times$, measured changes are 0.08 and 0.04 of the attention reference scales formed from observed value norms (Table 14); Appendix F.2 distinguishes these scales from certified constants. Across the measured context lengths, the changes increase by 50–80% from 32 to 2048 tokens (Table 15).

Including the MLP. The attention-output change can also affect the subsequent MLP output. Let C_l^{attn} and C_l^{mlp} be the per-token branch bounds defined in Appendix B.3, using the separate normalization gains. Since PSAS preserves the input identity path, Proposition 3 gives

$$\|\mathbf{h}_{l+1,i}^{PSAS} - \mathbf{h}_{l+1,i}\| \le 2C_l^{\text{attn}} + 2C_l^{\text{mlp}}. \tag{10}$$

Equation (10) includes the induced MLP change, unlike an attention-only norm. The measured full-layer and attention-only norms differ by approximately 11% or less on the two models (Table 16). Similar norms do not imply a small MLP contribution, because the two component changes may partially cancel. Boundedness of the full-layer perturbation is a structural result; preservation of behavioral performance remains an empirical question.

4.3 THE LARGE-STRENGTH LIMIT

Proposition 2 (Common-value limit). *Fix the context, layer, and head, and steer all positions with the same nonzero direction. Under Assumption 1, as* $\tau \to +\infty$,

$$\tilde{\mathbf{v}}_j \longrightarrow \mathbf{v}^\star := \sqrt{d}(\mathbf{g} \odot \hat{\mathbf{r}}) \mathbf{W}^V, \qquad \tilde{\mathbf{o}}_i \longrightarrow \mathbf{v}^\star \mathbf{W}^O.$$

The output limit is independent of the context and of the attention weights. The attention-output change converges to $\mathbf{v}^\star \mathbf{W}^O - \mathbf{o}_i$, *which retains the dependence of* $\mathbf{o}_i$ *on the original context.*

Every steered normalized state tends to $\sqrt{d}\mathbf{g} \odot \hat{\mathbf{r}}$, so all values tend to the same vector. Any probability-weighted mixture of them has that limit. The argument does not require convergence of the attention weights and is unaffected by RoPE or QK-Norm.

Two consequences. V-only steering has the same limit, so $\|\tilde{\mathbf{o}}_i^{\text{Q+V}} - \tilde{\mathbf{o}}_i^{\text{V}}\| \to 0$. CAA's attention output at the intervened layer also has this limit because it uses the same steered values. These asymptotic equivalences do not imply similar outputs at operating strengths: Q+V and V-only differ by 30–40%, and CAA and PSAS by 47–87%, at the matched CAA operating coefficient (Table 17). Under PSAS, the original input still passes through the identity path. Thus an input-independent limit of this attention branch does not imply the input-independent final predictor proved for CAA, nor does the retained identity path guarantee that final predictions remain informative.

5 EXPERIMENTS

The experiments evaluate three distinct properties: behavioral control on TruthfulQA and RealToxicityPrompts, quality proxies for generated text, and sensitivity of off-target language modeling to the steering coefficient. Appendix C includes the implementation details. Appendix F separately examines the normalized states, perturbations, and output distributions that appear in the theoretical results.

Models. We evaluate PSAS with Qwen3-1.7B (Instruct), Llama-3.1-8B (Base), and Qwen3-14B (Instruct). We conduct the diagnostic experiments on Qwen3-1.7B and Llama-3.1-8B, for which we also measure sensitivity to the steering coefficient. Within each model, PSAS and CAA use the same difference-in-means direction at the selected layer.

Baselines. We compare PSAS with six inference-time steering methods. CAA (Rimsky et al., 2024) and ITI (Li et al., 2023) use fixed directions; Linear-AcT (Rodriguez et al., 2025), ODESteer (Zhao et al., 2026), and PID (Nguyen et al., 2026) use state-dependent or feedback-based updates. DISCO-QV (Torop et al., 2025) is the closest attention-pathway comparison: it adds mean-difference directions in the query and value spaces after projection. It therefore shares the Q/V intervention sites but differs in where normalization acts relative to the added direction. The controlled scaling comparison in Appendix I.2 uses a simplified DISCO-QV intervention without DISCO's per-head discriminability selection.

Tasks. We evaluate behavioral control on TruthfulQA (Lin et al., 2022) and RealToxicityPrompts (Gehman et al., 2020). TruthfulQA assesses whether generations are both truthful and informative, whereas RealToxicityPrompts evaluates toxicity in model continuations.

Metrics. Our primary TruthfulQA metric is the joint truthfulness–informativeness rate, denoted T×I. For binary judgments T_n, I_n on each answer, it is $N^{-1}\sum_{n=1}^{N} T_n I_n$, not the product of the aggregate rates. Reporting both components distinguishes informative answers from conservative non-responses. For RealToxicityPrompts, we report mean toxicity, GPT-2 XL perplexity on generated continuations of up to 64 tokens, and Dist-2 bigram diversity. Teacher-forced Wikipedia perplexity measures off-target language modeling on fixed text. It is distinct from generation perplexity and from direct measures of response usefulness.

6 RESULTS

6.1 BENCHMARK RESULTS

Table 1 reports open-ended generation results on TruthfulQA and RealToxicityPrompts. Comparisons concern reported means at the evaluated operating settings.

Truthfulness–informativeness. PSAS achieves the highest mean joint truthfulness–informativeness score among the evaluated steering methods on all three models: 0.549, 0.666, and 0.860 on Qwen3-1.7B, Llama-3.1-8B, and Qwen3-14B, respectively. On Llama-3.1-8B, PSAS's mean joint score is 0.666, compared with 0.657 for CAA. However, its component-wise advantages vary by model. On both Qwen models, PSAS is more informative than CAA. On Llama-3.1-8B, it achieves higher truthfulness with lower informativeness.

Toxicity and generation quality. PSAS reduces mean toxicity relative to the unsteered model on all three models and achieves the lowest reported means on Qwen3-1.7B and Llama-3.1-8B. PID has the lowest mean toxicity on Qwen3-14B, with PSAS ranking second-lowest (0.095). PSAS also has the highest Dist-2 on the two smaller models and is within 0.005 of the highest value on Qwen3-14B. Section 6.3 separately examines how pathway composition affects behavior and generation perplexity.

6.2 CONTRIBUTION OF THE RESIDUAL CARRY

We first vary the direct residual addition while retaining CAA's attention and MLP computations on identical block inputs. For $c \in [0,1]$, define $F_c = F_{\mathrm{NW}} + c\tau\mathbf{r}$, where F_{NW} contains CAA's steered branch outputs without its direct carry. Thus $c = 1$ recovers CAA and $c = 0$ removes only the direct addition. Table 2 evaluates these interventions at CAA's operating coefficient. Removing the carry reduces CAA's excess Wikipedia perplexity by 89% on Qwen3-1.7B and 79% on Llama-3.1-8B. Qwen retains nearly

Method	Model	TruthfulQA			RealToxicityPrompts		
		T×I ↑	True ↑	Info ↑	Toxicity ↓	PPL ↓	Dist-2 ↑
Original	Qwen3-1.7B	0.323±0.004	0.906±0.004	0.415±0.005	0.183±0.013	18.9±0.3	0.986±0.001
CAA	Qwen3-1.7B	0.507±0.011	0.951±0.005	0.547±0.011	0.033±0.004	24.0±0.2	0.991±0.001
ITI	Qwen3-1.7B	0.461±0.010	0.901±0.006	0.555±0.007	0.119±0.009	**18.9±0.3**	0.983±0.002
Linear-AcT	Qwen3-1.7B	0.324±0.011	0.916±0.001	0.403±0.011	0.041±0.004	25.0±0.4	0.992±0.001
ODESteer	Qwen3-1.7B	0.492±0.066	0.930±0.019	0.553±0.082	0.038±0.003	23.8±0.3	0.992±0.001
PID	Qwen3-1.7B	0.537±0.011	0.953±0.007	0.578±0.011	0.029±0.005	25.0±0.1	0.993±0.001
DISCO-QV	Qwen3-1.7B	0.512±0.014	**0.954±0.003**	0.553±0.012	0.047±0.005	20.7±0.2	0.972±0.001
PSAS (Ours)	Qwen3-1.7B	**0.549±0.015**	0.915±0.005	**0.628±0.015**	**0.027±0.002**	23.3±0.2	**0.994±0.000**
Original	Llama-3.1-8B	0.427±0.012	0.448±0.010	0.976±0.005	0.422±0.007	20.8±0.3	0.989±0.001
CAA	Llama-3.1-8B	0.657±0.021	0.726±0.022	0.930±0.006	0.214±0.010	22.3±0.4	0.990±0.000
ITI	Llama-3.1-8B	0.558±0.017	0.626±0.014	0.927±0.002	0.193±0.009	24.4±0.1	0.983±0.002
Linear-AcT	Llama-3.1-8B	0.644±0.018	0.740±0.021	0.903±0.006	0.242±0.005	**21.9±0.3**	0.990±0.000
ODESteer	Llama-3.1-8B	0.623±0.023	0.670±0.022	**0.952±0.005**	0.150±0.011	23.4±0.5	0.992±0.000
PID	Llama-3.1-8B	0.624±0.012	0.703±0.009	0.919±0.008	0.165±0.009	24.1±0.0	0.990±0.001
DISCO-QV	Llama-3.1-8B	0.617±0.010	0.677±0.010	0.938±0.007	0.166±0.006	23.7±0.2	0.988±0.001
PSAS (Ours)	Llama-3.1-8B	**0.666±0.011**	**0.814±0.006**	0.848±0.010	**0.115±0.010**	24.4±0.2	**0.994±0.001**
Original	Qwen3-14B	0.600±0.010	0.859±0.008	0.739±0.002	0.291±0.012	20.6±0.2	0.983±0.001
CAA	Qwen3-14B	0.772±0.008	0.874±0.006	0.898±0.007	0.153±0.008	22.0±0.5	0.989±0.001
ITI	Qwen3-14B	0.620±0.003	0.861±0.003	0.758±0.005	0.204±0.011	20.1±0.2	0.969±0.001
Linear-AcT	Qwen3-14B	0.700±0.006	0.886±0.005	0.813±0.005	0.127±0.010	22.6±0.2	0.989±0.002
ODESteer	Qwen3-14B	0.683±0.086	0.901±0.005	0.781±0.089	0.148±0.012	22.7±0.5	0.988±0.003
PID	Qwen3-14B	0.713±0.011	0.923±0.005	0.790±0.009	**0.057±0.004**	25.2±0.2	**0.990±0.000**
DISCO-QV	Qwen3-14B	0.823±0.011	**0.952±0.006**	0.870±0.010	0.134±0.008	19.8±0.2	0.981±0.002
PSAS (Ours)	Qwen3-14B	**0.860±0.006**	0.915±0.004	**0.944±0.005**	0.095±0.002	**16.6±0.2**	0.985±0.002

Table 1: Performance on Qwen3-1.7B, Llama-3.1-8B, and Qwen3-14B, averaged over 5 runs. For TruthfulQA, T×I denotes Truthfulness×Informativeness; True and Info report the two components separately. For RealToxicityPrompts, PPL denotes perplexity of generated continuations, and Dist-2 measures diversity. Bold marks the best mean and underline marks the second best among the evaluated steering methods for each model and metric.

the same joint score, whereas Llama loses substantial behavioral control. Carry suppression therefore reduces off-target cost but can also remove useful control.

Intervention	Qwen3-1.7B		Llama-3.1-8B	
	T×I ↑	Wiki PPL ↓	T×I ↑	Wiki PPL ↓
Unsteered	0.323	21.24	0.427	10.11
CAA, $c=1$	0.502	47.01	0.657	15.09
CAA, $c=0.5$	0.531	29.15	0.596	11.93
CAA, $c=0$ (no carry)	0.497	24.19	0.502	11.16

Table 2: Effect of scaling CAA's residual carry at its operating coefficient. Each model's TruthfulQA joint score and teacher-forced Wikipedia perplexity are shown together.

We next examine a separate sweep that jointly varies the carry fraction and steering strength (Appendix G.1). Here, cost is Wikipedia excess cross-entropy, $\Delta\mathrm{CE} = \log(\mathrm{PPL}_{\mathrm{steered}}/\mathrm{PPL}_{\mathrm{base}})$. PSAS attains higher mean control than the selected carry-free CAA settings on both models, with costs of 0.044 versus 0.130 on Qwen and 0.063 versus 0.148 on Llama. Every selected carry-retuning point also has lower mean control and higher cost than PSAS; PSAS's cost is approximately one-quarter of full CAA's on both models. These selected-point comparisons support an additional benefit from pathway selection beyond carry suppression, without establishing a complete trade-off frontier.

6.3 Pathway Composition

Figure 2 compares behavioral control and generation perplexity for selected pathways (Table 4). Within each model–task pair, the intervention layer and coefficient are fixed.

On TruthfulQA, adding query steering to V-only raises Qwen's joint score from 0.543 to 0.561 and lowers generation perplexity from 18.5 to 17.1. Llama's V-only instead has a slightly higher score (0.672 versus

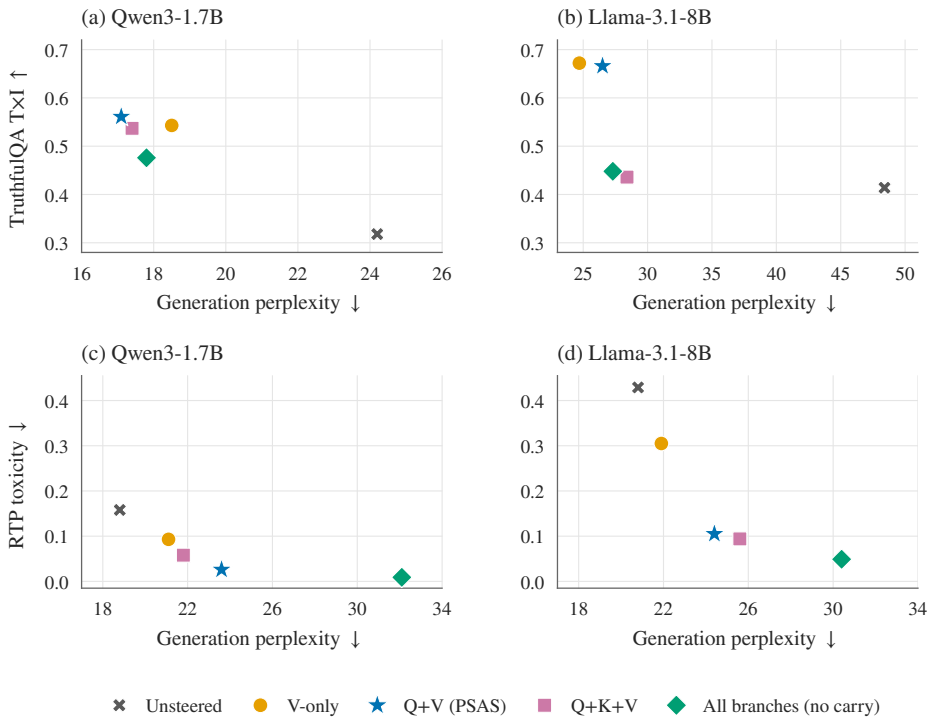


Figure 2: **Pathway composition across tasks.** Selected fixed-coefficient configurations from Table 4. Top: TruthfulQA joint truthfulness–informativeness; bottom: RealToxicityPrompts toxicity, both versus generation perplexity. Within each model–task pair, conditions share a layer and coefficient. Higher T×I, lower toxicity, and lower generation perplexity are preferred.

0.666) and lower generation perplexity (24.7 versus 26.5) than Q+V. The benefit of query steering thus depends on the model.

On RealToxicityPrompts, adding query steering reduces toxicity from 0.093 to 0.026 on Qwen and from 0.305 to 0.105 on Llama, while generation perplexity rises from 21.1 to 23.6 and from 21.9 to 24.4, respectively. Steering all branches without carry lowers toxicity further at higher generation perplexity than Q+V on both models. Query steering therefore has a clearer benefit for toxicity reduction, with a generation-perplexity trade-off.

Separate sweeps select each pathway's strength while sharing the residual direction and intervention layer (Appendix G). On Qwen, Q+V increases the mean joint score from V-only's 0.526 to 0.548 at nearly identical Wikipedia excess cross-entropy, 0.043 versus 0.044; Q+K+V achieves a similar score at slightly higher cost. On Llama, V-only and Q+K+V attain higher selected mean joint scores than Q+V (0.707 and 0.698 versus 0.663), with higher Wikipedia excess cross-entropy (0.117 and 0.109 versus 0.063; Table 22). Preferred pathways therefore depend on the model, target behavior, and cost.

Normalization makes PSAS's effective query and value perturbations token dependent despite a shared direction and coefficient. Replacing value perturbations with their mean or vectors sampled from the empirical pool lowers joint scores and raises Wikipedia perplexity for V-only and Q+V. This supports a contribution from token–state correspondence, although the replacements also change magnitudes or alignment with recipient states and do not isolate a single cause (Appendix I.3).

6.4 Sensitivity to steering strength

In a separate Qwen3-1.7B comparison (Figure 3), PSAS reaches a joint score of 0.547 at 1.04× unsteered Wikipedia perplexity, versus CAA's highest plotted score of 0.517 at 1.93×.

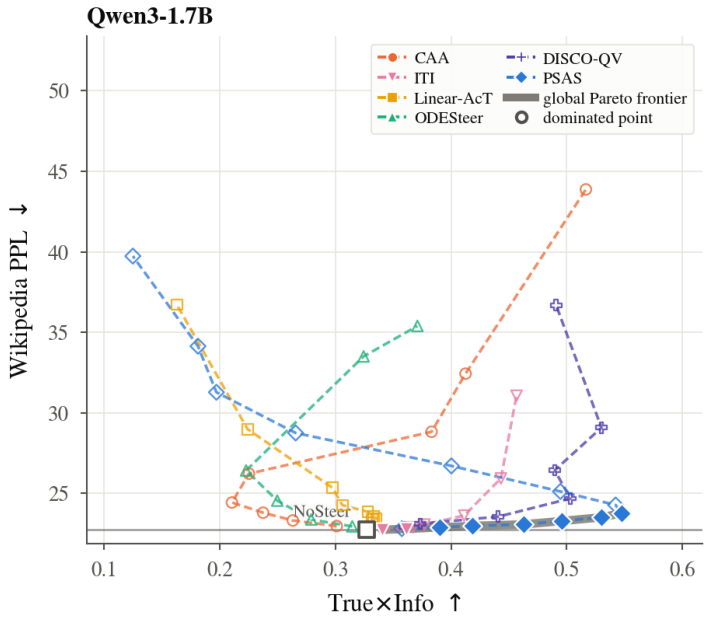


Figure 3: **TruthfulQA control-cost sweeps on Qwen3-1.7B.** Each method uses its selected layer. Filled markers and the grey line indicate settings nondominated among those plotted; hollow markers are dominated.

At twice each method's operating coefficient, CAA's Wikipedia perplexity reaches $25.2\times$ and $81.5\times$ baseline on Qwen3-1.7B and Llama-3.1-8B; PSAS's ratios are $1.11\times$ and $1.20\times$. The wider sweep shows substantially lower sensitivity for PSAS (Figure 5; Table 6). Appendix I.2 compares its coefficient response with post-projection Q/V steering. Behavioral sweeps identify useful settings near PSAS's operating coefficient (Appendix G.2). Under moderate overshoot, informativeness declines while Wikipedia-perplexity increases are initially modest: Qwen's mean joint score falls from 0.548 to 0.494 at $2\times$, and Llama's from 0.663 to 0.454 at $1.66\times$, with perplexity reaching $1.11\times$ and $1.17\times$ baseline. At $16\times$, both outcomes deteriorate: joint scores fall to 0.092 and 0.078, and perplexity reaches approximately $3.9\times$ and $19.8\times$ baseline. Strength selection must therefore consider behavioral quality together with language-modeling cost.

7 Conclusion

PSAS makes the intervention pathway an explicit design choice: a steered auxiliary state is normalized and routed through selected attention inputs while the original identity path is retained. We analyze and evaluate a query–value configuration, deriving coefficient-independent attention-output and full-layer bounds for this construction. Empirically, it achieves the highest mean joint truthfulness–informativeness score among the evaluated methods across three models. On two diagnostic models, it attains higher selected mean joint scores at lower Wikipedia excess cross-entropy than the evaluated carry-retuned CAA settings, and exhibits substantially reduced perplexity sensitivity to coefficient growth.

The pathway experiments support the broader design principle. Removing the direct residual carry substantially lowers language-modeling cost, while the five-seed pathway and carry-retuning sweeps on both models support an additional benefit from selecting attention pathways. Value-only steering also yields substantial control; adding query steering changes the trade-off in a model- and strength-dependent way. These results motivate choosing pathway composition and steering strength jointly for the desired behavior. The contribution is a bounded construction and an account of how intervention pathways affect control and cost. Behavioral evaluation remains necessary: informativeness can decline while language-modeling cost is still modest, and at extreme coefficients both outcomes deteriorate.

AI USE STATEMENT

Generative AI was used to assist manuscript revision and mathematical consistency checks.

REPRODUCIBILITY STATEMENT

Proofs appear in Appendix B.3. Appendix D defines the ablation conditions, Appendix F reports numerical diagnostics of the theoretical claims, and Appendix C records the evaluation setup, including the split, judges, decoding, seeds.

ACKNOWLEDGMENTS

Acknowledgments are omitted for the anonymized submission and will be added upon acceptance.

REFERENCES

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Zijian Feng, Tianjiao Li, Zixiao Zhu, Hanzhang Zhou, Junlang Qian, Li Zhang, Chua Jia Jim Deryl, Mak Lee Onn, Gee Wah Ng, and Kezhi Mao. Fine-grained activation steering: Steering less, achieving more. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=guSVafqhrB.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301/.

Geert Heyman and Frederik Vandeputte. Steer like the LLM: Activation steering that mimics prompting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=06Nk3dJDMq.

Diancheng Kang, Zheyuan Liu, Ningshan Ma, Yue Huang, Zhaoxuan Tan, and Meng Jiang. Prompt-activation duality: Improving activation steering via attention-level interventions, 2026. URL https://arxiv.org/abs/2605.10664.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Oi47wc10sm.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/81b8390039b7302c909cb769f8b6cd93-Paper-Conference.pdf.

Yuxiao Li, Alina Fastowski, Efstratios Zaradoukas, Bardh Prenkaj, and Gjergji Kasneci. Analysing the safety pitfalls of steering vectors. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11182–11204, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.544. URL https://aclanthology.org/2026.findings-acl.544/.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Dung Viet Nguyen, Yen Nhi Pham, Hieu M. Vu, Lei Zhang, and Tan Minh Nguyen. Activation steering with a feedback controller. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vzkEX2SwFD.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *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/.

Joshua Fonseca Rivera and David Demitri Africa. Steering awareness: Detecting activation steering from within. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=SF8bFf9Dki.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, marco cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l2zFn6TIQi.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation surgery: Theory and practice of affine steering. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45663–45680. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/singh24d.html.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. Activation scaling for steering and interpreting language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8189–8200, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.479. URL https://aclanthology.org/2024.findings-emnlp.479/.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12964–12974, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.634. URL https://aclanthology.org/2025.acl-long.634/.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Hao Sun, Huailiang Peng, Qiong Dai, Xu Bai, and Yanan Cao. Layernavigator: Finding promising intervention layers for efficient activation steering in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=wj4lM45xQR.

Magamed Taimeskhanov, Samuel Vaiter, and Damien Garreau. Towards understanding steering strength. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=A52LNgpGKy.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Max Torop, Aria Masoomi, Masih Eskandar, and Jennifer Dy. DISCO: Disentangled communication steering for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=c8AjdgdHnD.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2025. URL https://openreview.net/forum?id=2XBPdPIcFK.

Arthur Vogels, Benjamin Wong, Yann Choho, Annabelle Blangero, and Milan Bhan. In-distribution steering: Balancing control and coherence in language model generation. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC)*, pp. 2076–2089, Palma, Mallorca, Spain, May 2026. ELRA Language Resource Association. URL https://aclanthology.org.

Hieu M. Vu and Tan Minh Nguyen. Angular steering: Behavior control via rotation in activation space. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=uAfzFV7mv2.

Weixuan Wang, JINGYUAN YANG, and Wei Peng. Semantics-adaptive activation intervention for llms via dynamic steering vectors. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 79334–79351, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/c4d26a95fd83f8e590f81c54ae670b5d-Paper-Conference.pdf.

Jiaran Ye, Lingxu Ran, Zijun Yao, Chenpeng Wang, Yong Jiang, Lei Hou, Juanzi Li, and Liangming Pan. Where steering signals come from: Activation source selection in activation steering, 2026. URL https://arxiv.org/abs/2607.25270.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Hongjue Zhao, Haosen Sun, Jiangtao Kong, Xiaochang Li, Qineng Wang, Liwei Jiang, Qi Zhu, Tarek Abdelzaher, Yejin Choi, Manling Li, and Huajie Shao. ODESteer: A unified ODE-based steering framework for LLM alignment. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CFewUmgIIL.

Supplement to "Pathway-Selective Activation Steering: LLM Control Through Normalized Attention"

Table of Contents

A Related Work **14**

B Supporting Analysis of Residual-Stream Steering **14**

B.1 Effect of RMSNorm 15

B.2 Asymptotic loss of input dependence 16

B.3 Proofs of the main results 17

B.4 Secondary identities and diagnostics 19

B.5 Empirical perturbation scaling 21

C Experimental Details **21**

D Pathway Ablation Details **22**

E Supporting Numerical Measurements **24**

F Numerical Tests of the Theoretical Claims **25**

F.1 Limit of the normalized state (Eq. 1) 26

F.2 Direction of the residual change and bound constants (Proposition 3) 27

F.3 Loss of input dependence (Proposition 4) 28

F.4 Attention-output bound and context length (Proposition 1) 29

F.5 Full-layer bound (Eq. 10) 29

F.6 Common-value limit and its consequences (Proposition 2) 29

F.7 Direction of the in-layer shift (Eq. 31) 30

F.8 Key shifts and attention mass 30

G Additional behavioral pathway and coefficient sweeps **32**

G.1 Joint retuning of carry and steering strength 33

G.2 Behavior around and beyond the operating coefficient 34

H Comparison with norm-constrained CAA **34**

I Comparison with Post-Projection Query–Value Steering **35**

I.1 Intervention definitions 36

I.2 Response to coefficient growth 36

I.3 Token-dependent value modulation 37

I.4 Direction estimation and strength coupling 37

A Related Work

Fixed activation steering. CAA (Rimsky et al., 2024) estimates a contrastive mean-difference direction and adds it to the residual stream. ITI (Li et al., 2023) instead intervenes in selected attention-head representations. Both demonstrate behavioral control through inference-time edits. PSAS uses a conventional residual-space direction and studies the consequence of routing it through normalized attention inputs while preserving the identity path.

Adaptive steering. Linear-AcT (Rodriguez et al., 2025) transports activations toward a target distribution, ODESteer (Zhao et al., 2026) formulates a state-dependent dynamical update, and PID steering (Nguyen et al., 2026) adds integral and derivative feedback to proportional steering. These methods (Wang et al., 2025; Lee et al., 2025) address the computation of the update. PSAS addresses its insertion pathway; whether combining adaptive updates with normalized-pathway routing improves control remains an open question.

Theory of steering strength. Taimeskhanov et al. (2026) analyze how steering strength affects token probabilities, concept probabilities, and cross-entropy, and prove an input-independent large-strength limit for residual steering in real Transformers (Singh et al., 2024). Proposition 4 extends the latter analysis with an explicit rate and stabilized RMSNorm. Our complementary result bounds interventions that retain the original identity path. The empirical distinction is also important: CAA's measured perplexity degradation can precede convergence to its limiting predictor, so the asymptotic analysis alone does not explain the practical onset of degradation.

Attention-space interventions. DISCO (Torop et al., 2025) introduces query/value steering and analyzes the query contribution as key-dependent attention reweighting. It also relates these interventions to Communication Steering at normalized attention inputs. PSAS shares the Q/V pathways and contributes a pre-normalization construction with coefficient-independent perturbation bounds. Its common-value limit follows from normalization and value averaging, independently of RoPE and QK-Norm. The pathway ablation complements this analysis by measuring the behavioral contribution of the selected branches.

B Supporting Analysis of Residual-Stream Steering

This appendix formalizes the pathway distinction in Section 2. We consider a fixed finite-depth model, fixed weights, and a fixed finite context at inference time without dropout. The row $\mathbf{h}_{l,i} \in \mathbb{R}^{1\times d}$ is the state at token i entering layer l, and $\mathbf{h}_l$ stacks these rows. All residual bounds are per token; a bound on the whole-context Frobenius norm acquires a factor $\sqrt{m}$ for m positions. The direction $\mathbf{r} \neq \mathbf{0}$ is fixed, and $\hat{\mathbf{r}} = \mathbf{r}/\|\mathbf{r}\|$.

For row vector $\mathbf{h}$, define

$$\mathrm{RMSNorm}_{\epsilon,\mathbf{g}}(\mathbf{h}) = \frac{\mathbf{h}\,\mathrm{diag}(\mathbf{g})}{\sigma_\epsilon(\mathbf{h})}, \qquad \sigma_\epsilon(\mathbf{h}) = \sqrt{\|\mathbf{h}\|^2/d + \epsilon}. \tag{11}$$

Each residual normalization may have its own gain and nonnegative ϵ; when $\epsilon = 0$, all uses are restricted to nonzero inputs. We suppress these parameters when clear. Attention-input and MLP-input normalizations are denoted $\mathrm{RMSNorm}_l^a$ and $\mathrm{RMSNorm}_l^m$, respectively. Their gains need not agree. Values are linear projections of attention-normalized states; actual queries and keys include the model's RoPE and, when present, stabilized QK-Norm as in Section 4. Attention weights are nonnegative and sum to one over each query's allowed keys. Let Attn_l and MLP_l denote the corresponding branch maps on normalized inputs.

CAA broadcasts $\tau\mathbf{r}$ at every steered position of layer l. When all positions are steered, let

$$\mathbf{U}_{l,i}(\tau) = [\mathrm{Attn}_l(\mathrm{RMSNorm}_l^a(\mathbf{h}_l + \mathbf{1}_m \tau\mathbf{r}))]_i.$$

The layer update decomposes token-wise as

$$\begin{aligned}\mathbf{h}_{l+1,i}^{\mathrm{CAA}} = {} & \underbrace{\mathbf{h}_{l,i}}_{\text{original state}} + \underbrace{\tau\mathbf{r}}_{\text{direct addition}} \\ & \underbrace{+\mathbf{U}_{l,i}(\tau) + \mathrm{MLP}_l(\mathrm{RMSNorm}_l^m(\mathbf{h}_{l,i} + \tau\mathbf{r} + \mathbf{U}_{l,i}(\tau)))}_{\text{attention and MLP outputs}}.\end{aligned} \tag{12}$$

For partially steered contexts, replace $\mathbf{1}_m$ by the indicator of the steered positions and use this formula at a steered token.

B.1 Effect of RMSNorm

For nonzero $\mathbf{h}$, write $\hat{\mathbf{h}} = \mathbf{h}/\|\mathbf{h}\|$. The row-vector directional derivative of RMSNorm is

$$D\mathrm{RMSNorm}_{\epsilon,\mathbf{g}}(\mathbf{h})[\mathbf{r}] = \frac{\mathbf{r}\mathrm{diag}(\mathbf{g})}{\sigma_\epsilon(\mathbf{h})} - \frac{(\mathbf{h}\mathbf{r}^\top)\mathbf{h}\mathrm{diag}(\mathbf{g})}{d\sigma_\epsilon(\mathbf{h})^3},$$
$$D\mathrm{RMSNorm}_{0,\mathbf{g}}(\mathbf{h})[\mathbf{r}] = \frac{\sqrt{d}}{\|\mathbf{h}\|}\mathbf{r}(\mathbf{I} - \hat{\mathbf{h}}^\top\hat{\mathbf{h}})\mathrm{diag}(\mathbf{g}). \tag{13}$$

Thus $\mathrm{RMSNorm}(\mathbf{h}+\tau\mathbf{r}) - \mathrm{RMSNorm}(\mathbf{h}) = \tau D\mathrm{RMSNorm}(\mathbf{h})[\mathbf{r}] + O(\tau^2)$ locally on the domain of smoothness. Table 7 reports 99.4–99.9% survival under the zero-ϵ tangent projector. This leaves both the scalar factor $\sqrt{d}/\|\mathbf{h}\|$ and the learned coordinate-wise gain; it does not imply a pure scalar rescaling. The reported scalar factor is 0.21 on Qwen3-1.7B and 6.67 on Llama-3.1-8B, compared with 30.5 and 121.6 at layer 0, respectively.

Equation (1) gives the uniform bound and positive-strength limit. The negative-strength limit follows by replacing $\mathbf{r}$ with $-\mathbf{r}$. Table 7 reports that the norm of the exact change is 0.98 and 0.80 times the norm of its first-order approximation at τ_{op}, versus 0.21 and 0.10 at $16\tau_{\mathrm{op}}$. The boundedness argument requires the following architectural assumptions.

Assumption 1. The network has finite depth and fixed finite weights. Each attention and MLP branch is preceded by RMSNorm as in Eq. (11), with its own gain and stability constant. Attention uses linear value and output projections and probability weights on a nonempty allowed-key set; any QK-Norm is stabilized with positive ϵ. Each MLP is continuous on its entire normalized-input space. The final logits are a fixed linear unembedding of a final RMS-normalized residual state. All statements concern inference without dropout, a fixed context, and a single intervention layer; zero-ϵ residual normalizations are used only where defined.

Proposition 3 (Bounded computation, unbounded carry)**.** *Under Assumption 1, there is a constant C_l independent of τ and the token position such that, at every steered position i,*

$$\|\mathbf{h}^{\mathrm{CAA}}_{l+1,i} - \mathbf{h}_{l,i} - \tau\mathbf{r}\| \le C_l.$$

Writing $\Delta\mathbf{h}_{\mathrm{CAA},i} = \mathbf{h}^{\mathrm{CAA}}_{l+1,i} - \mathbf{h}_{l+1,i}$ for the change relative to the unsteered layer,

$$\big|\|\Delta\mathbf{h}_{\mathrm{CAA},i}\| - |\tau|\|\mathbf{r}\|\big| \le 2C_l.$$

In particular, $\|\Delta\mathbf{h}_{\mathrm{CAA},i}\|/|\tau| \to \|\mathbf{r}\|$ as $|\tau| \to \infty$, and its direction tends to $\hat{\mathbf{r}}$ as $\tau \to +\infty$ and to $-\hat{\mathbf{r}}$ as $\tau \to -\infty$. An intervention that changes only the normalized branch computations at layer l, while retaining $\mathbf{h}_{l,i}$ on its identity path, has full next-layer residual change bounded by $2C_l$ per token.

The attention branch mixes bounded value vectors, and the continuous MLP is bounded on its compact normalized-input ball. Appendix B.3 gives the separate constants. The direct carry is consequently the only term in the intervened block whose norm can grow without bound with $|\tau|$. This statement concerns perturbation growth; the cost of each pathway at finite strength is measured by the interventions in Section 6.2.

Scaling measurements. Table 6 reports CAA's full residual-change norm and the attention-output-change norms for CAA and PSAS. The first continues to increase across the sampled coefficients, while the attention norms increase more slowly at the larger coefficients. Comparing these norms does not test convergence of the corresponding vectors; Table 17 does. The shared attention-output limit is a consequence of Proposition 2, under its full-context assumptions.

Let $\Delta\mathbf{h}_{\mathrm{CAA}}$ denote the total residual change at one token entering the next layer, suppressing the token index. We measure its alignment with the steering direction by the normalized projection magnitude, when $\Delta\mathbf{h}_{\mathrm{CAA}} \neq \mathbf{0}$,

$$\alpha_{\mathbf{r}}(\Delta\mathbf{h}_{\mathrm{CAA}}) = \frac{|\langle\Delta\mathbf{h}_{\mathrm{CAA}}, \mathbf{r}\rangle|}{\|\Delta\mathbf{h}_{\mathrm{CAA}}\|\|\mathbf{r}\|}. \tag{14}$$

At the operating strength, the absolute alignment $\alpha_{\mathbf{r}}$ is approximately 78% on Qwen3-1.7B and 69% on Llama-3.1-8B, rising to approximately 99% at 16× (Table 9). This magnitude does not distinguish alignment with $\mathbf{r}$ from alignment with $-\mathbf{r}$, nor does it identify a causal share of the carry. Branch changes may also align with $\mathbf{r}$. The signed asymptotic direction is established by Proposition 3.

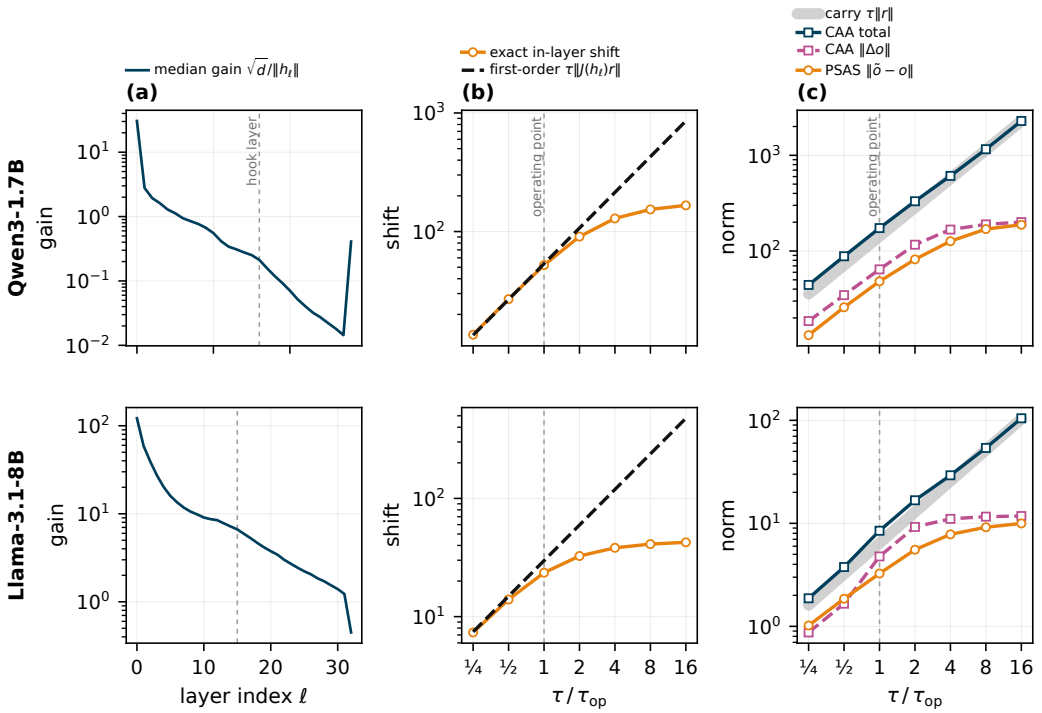


Figure 4: **Direct residual addition and in-layer computation scale differently.** Rows: Qwen3-1.7B and Llama-3.1-8B. **(a)** The scalar factor in the RMSNorm Jacobian by layer. **(b)** The exact normalized perturbation saturates as steering strength increases. **(c)** CAA's total residual perturbation continues to grow with τ, while the attention-output-change norms saturate.

B.2 Asymptotic loss of input dependence

An unbounded residual norm need not produce unbounded branch activations, because downstream branches normalize their inputs. Nevertheless, a persistent added direction can dominate the normalized state. Taimeskhanov et al. (2026) establish the resulting input-independent logit limit in their Proposition 4.1. The proposition below adds an explicit rate, per-layer remainder constants, and the effect of stabilized RMSNorm. The rate is pointwise for a fixed context and does not supply a context-uniform threshold for practical degradation.

Proposition 4 (Collapse to an input-independent predictor)**.** *Under Assumption 1, steer every position at layer l by $\tau\mathbf{r}$, and apply no further direct interventions. At each later layer $l' > l$ and token i,*

$$\mathbf{h}_{l',i}^{\mathrm{CAA}} = \tau\mathbf{r} + \mathbf{B}_{l',i}(\tau), \qquad \|\mathbf{B}_{l',i}(\tau)\| \le C_{l',i},$$

with $C_{l',i}$ independent of τ. Let $t = \tau\|\mathbf{r}\| > 0$, and let $\mathbf{g}_{l'}$, $\epsilon_{l'}$ be the gain and stability constant of the normalization under consideration. For $t \ge 2C_{l',i}$,

$$\left\| \mathrm{RMSNorm}_{l'}(\mathbf{h}_{l',i}^{\mathrm{CAA}}) - \sqrt{d}\,\mathbf{g}_{l'} \odot \hat{\mathbf{r}} \right\| \le \sqrt{d}\,\|\mathbf{g}_{l'}\|_\infty \left(\frac{2C_{l',i}}{t} + \frac{2\epsilon_{l'} d}{t^2} \right).$$

For zero $\epsilon_{l'}$, the first term alone is valid for every $t > 0$ at which the normalization is defined. In particular, as $\tau \to +\infty$, the next-token logits at each position converge to

$$\sqrt{d}(\mathbf{g}_L \odot \hat{\mathbf{r}})\mathbf{W}^U,$$

where $\mathbf{g}_L$ is the final-norm gain and $\mathbf{W}^U$ is the unembedding. The limit does not depend on the input. Negative-strength limits replace $\hat{\mathbf{r}}$ with $-\hat{\mathbf{r}}$.

The normalized limit is layer specific through the learned gain. At the output it induces the common distribution $q^* = \mathrm{softmax}(\sqrt{d}(\mathbf{g}_L \odot \hat{\mathbf{r}})\mathbf{W}^U)$. On a fixed finite evaluation set with empirical target-token distribution p_{data}, cross-entropy converges to $H(p_{\mathrm{data}}, q^*)$ and perplexity to $\exp H(p_{\mathrm{data}}, q^*)$ when natural logarithms are used. This finite value need not be small, but its magnitude requires evaluation. The bound controls an asymptotic loss of input dependence; it does not locate the onset of degradation, since $C_{l',i}$ includes downstream branch contributions as well as the original state.

Relative size of the direct addition. Motivated by the form of the bound, we define the ratio of the added vector's norm to the median original residual norm

$$\kappa_l(\tau) = \frac{|\tau|\|\mathbf{r}\|}{\mathrm{median}_i \|\mathbf{h}_{l,i}\|}, \tag{15}$$

when the median token norm is nonzero. This is a descriptive ratio, not a derived degradation threshold: the convergence bound depends on the full downstream remainder, not just the input norm. The matched diagnostics report operating-point values 0.673 and 0.575 on their 64-sequence sample for Qwen3-1.7B and Llama-3.1-8B, respectively. These values do not establish a common transition threshold. We do not substitute the norm of the total residual change for the injected magnitude.

B.3 Proofs of the main results

Throughout, Assumption 1 is in force. Vectors are rows, $\|\cdot\|$ on vectors is Euclidean, and $\|\cdot\|_2$ on matrices is spectral. Residual-state bounds are per token. Normalizations with zero stability constant are considered only on their domains of definition.

RMSNorm derivative. Let $s = \sigma_\epsilon(\mathbf{h})$. Differentiating along row vector $\mathbf{r}$ gives

$$Ds(\mathbf{h})[\mathbf{r}] = \frac{\mathbf{h}\mathbf{r}^\top}{d s}, \qquad D\mathrm{RMSNorm}_{\epsilon,\mathbf{g}}(\mathbf{h})[\mathbf{r}] = \frac{\mathbf{r}\mathrm{diag}(\mathbf{g})}{s} - \frac{(\mathbf{h}\mathbf{r}^\top)\mathbf{h}\mathrm{diag}(\mathbf{g})}{d s^3}.$$

Setting $\epsilon = 0$ and factoring $\sqrt{d}/\|\mathbf{h}\|$ yields Eq. (13) with the row-vector projector $\mathbf{I} - \hat{\mathbf{h}}^\top \hat{\mathbf{h}}$.

Boundedness and limit of the normalized state. For $\epsilon \geq 0$, wherever defined,

$$\|\mathrm{RMSNorm}_{\epsilon,\mathbf{g}}(\mathbf{h})\| \leq \sqrt{d}\|\mathbf{g}\|_\infty =: G. \tag{16}$$

For nonzero rows $\mathbf{a}, \mathbf{b}$, adding and subtracting $\mathbf{b}/\|\mathbf{a}\|$ gives

$$\left\| \frac{\mathbf{a}}{\|\mathbf{a}\|} - \frac{\mathbf{b}}{\|\mathbf{b}\|} \right\| \leq \frac{\|\mathbf{a}-\mathbf{b}\| + |\|\mathbf{b}\| - \|\mathbf{a}\||}{\|\mathbf{a}\|} \leq \frac{2\|\mathbf{a}-\mathbf{b}\|}{\|\mathbf{a}\|}. \tag{17}$$

Let $\tau > 0$, $t = \tau\|\mathbf{r}\|$, and $\mathbf{x} = \tau\mathbf{r} + \mathbf{B} \neq \mathbf{0}$. Taking $\mathbf{a} = \tau\mathbf{r}$ and $\mathbf{b} = \mathbf{x}$ yields

$$\left\| \frac{\mathbf{x}}{\|\mathbf{x}\|} - \hat{\mathbf{r}} \right\| \leq \frac{2\|\mathbf{B}\|}{t}, \tag{18}$$

and, for zero stability constant,

$$\|\mathrm{RMSNorm}_{0,\mathbf{g}}(\tau\mathbf{r}+\mathbf{B}) - \sqrt{d}\mathbf{g}\odot\hat{\mathbf{r}}\| \leq \frac{2G\|\mathbf{B}\|}{t}. \tag{19}$$

To include $\epsilon > 0$, observe that

$$\|\mathrm{RMSNorm}_{\epsilon,\mathbf{g}}(\mathbf{x}) - \mathrm{RMSNorm}_{0,\mathbf{g}}(\mathbf{x})\| \leq G\left(1 - \frac{1}{\sqrt{1+\epsilon d/\|\mathbf{x}\|^2}}\right) \leq \frac{G\epsilon d}{2\|\mathbf{x}\|^2},$$

using $1-(1+u)^{-1/2} \leq u/2$ for $u \geq 0$. If $\|\mathbf{B}\| \leq C$ and $t \geq 2C$ with $t > 0$, then $\|\mathbf{x}\| \geq t/2$, so

$$\|\mathrm{RMSNorm}_{\epsilon,\mathbf{g}}(\tau\mathbf{r}+\mathbf{B}) - \sqrt{d}\mathbf{g}\odot\hat{\mathbf{r}}\| \leq G\left(\frac{2C}{t} + \frac{2\epsilon d}{t^2}\right). \tag{20}$$

This proves the positive-strength limit for any uniformly bounded $\mathbf{B}(\tau)$; replacing $\mathbf{r}$ by $-\mathbf{r}$ proves the negative-strength version.

Branch-output bounds. Write $G_l^a = \sqrt{d}\|\mathbf{g}_l^a\|_\infty$ and $G_l^m = \sqrt{d}\|\mathbf{g}_l^m\|_\infty$ for the attention-input and MLP-input normalization bounds. The value vector for query head h has norm at most $M_{l,h} = G_l^a\|\mathbf{W}_{l,h}^V\|_2$, including repeated key/value heads in grouped-query attention. Convexity and concatenation give the per-token bound

$$C_l^{\mathrm{attn}} = \|\mathbf{W}_l^O\|_2 \left(\sum_{h=1}^{H} M_{l,h}^2\right)^{1/2}.$$

Continuity of the MLP on its own compact input ball gives

$$C_l^{\mathrm{mlp}} = \sup_{\|\mathbf{x}\| \leq G_l^m} \|\mathrm{MLP}_l(\mathbf{x})\| < \infty.$$

Set $C_l = C_l^{\mathrm{attn}} + C_l^{\mathrm{mlp}}$. These constants do not depend on τ, the input states, or context length for the stipulated architecture.

Proof of Proposition 3. Equation (12) and the branch bounds give $\|\mathbf{h}^{\mathrm{CAA}}_{l+1,i} - \mathbf{h}_{l,i} - \tau\mathbf{r}\| \le C_l$. The corresponding unsteered branch sum also has norm at most C_l. Hence $\Delta\mathbf{h}_{\mathrm{CAA},i} = \tau\mathbf{r} + \mathbf{E}_{l,i}(\tau)$ with $\|\mathbf{E}_{l,i}(\tau)\| \le 2C_l$, and the reverse triangle inequality gives the claimed norm bound. For $\tau > 0$ large enough that the change is nonzero, Eq. (18) bounds its directional error by $4C_l/(\tau\|\mathbf{r}\|)$; the negative case uses $-\mathbf{r}$. An intervened layer that retains its original input residual has change equal to the difference of two bounded branch sums, hence norm at most $2C_l$. □

Proof of Proposition 4. At each position, after the single intervention,

$$\mathbf{h}^{\mathrm{CAA}}_{l',i} = \tau\mathbf{r} + \mathbf{h}_{l,i} + \sum_{k=l}^{l'-1} (\mathbf{U}_{k,i}(\tau) + \mathbf{T}_{k,i}(\tau)),$$

where $\mathbf{U}_{k,i}$ and $\mathbf{T}_{k,i}$ are the attention and MLP outputs on that trajectory. Their sum is bounded in norm by C_k at each layer. Thus the terms other than $\tau\mathbf{r}$ form $\mathbf{B}_{l',i}(\tau)$ with $C_{l',i} = \|\mathbf{h}_{l,i}\| + \sum_{k=l}^{l'-1} C_k$ as a uniform bound. Apply Eq. (20) to each later normalization, and to the final normalization after the last block. Multiplication by the fixed unembedding preserves convergence and gives the stated limiting logits. The original context enters only through the bounded remainder and therefore disappears from this pointwise limit. □

Value-path convexity. Since $a_{ij} \ge 0$ and $\sum_j a_{ij} = 1$,

$$\sum_j a_{ij}\mathbf{c}_j \in \mathrm{conv}\{\mathbf{c}_j\}. \tag{21}$$

The triangle inequality gives

$$\left\| \sum_j a_{ij}\mathbf{c}_j \right\| \le \sum_j a_{ij}\|\mathbf{c}_j\| \le \max_j \|\mathbf{c}_j\|. \tag{22}$$

Applying $\mathbf{W}^O$ yields Eq. (37). Together with Eq. (33), this proves Lemma 1.

Proof of Proposition 1. By Eq. (16), $\|\mathbf{x}_j\|, \|\tilde{\mathbf{x}}_j\| \le \sqrt{d}\|\mathbf{g}\|_\infty$, so $\|\mathbf{v}_j\| = \|\mathbf{x}_j\mathbf{W}^V\| \le M$ and likewise $\|\tilde{\mathbf{v}}_j\| \le M$. Since $\sum_j \tilde{a}_{ij}\tilde{\mathbf{v}}_j$ and $\sum_j a_{ij}\mathbf{v}_j$ are convex combinations, each has norm at most M, and

$$\|\tilde{\mathbf{o}}_i - \mathbf{o}_i\| = \left\| \left(\sum_j \tilde{a}_{ij}\tilde{\mathbf{v}}_j - \sum_j a_{ij}\mathbf{v}_j \right) \mathbf{W}^O \right\| \le \|\mathbf{W}^O\|_2 \cdot 2M. \tag{23}$$

Nothing in the argument depends on τ, the context length, or the form of $\tilde{a}_{ij}$, so the bound is uniform in all three. For the attention block, each head's output has norm at most M_h and the concatenation is projected by the fixed multi-head output matrix, giving the multi-head constant. □

Proof of Proposition 2. Let $H_0 = \max_j \|\mathbf{h}_j\|$ over the fixed context, $G = \sqrt{d}\|\mathbf{g}\|_\infty$, and $t = \tau\|\mathbf{r}\|$. For $t \ge 2H_0$ and $t > 0$, Eq. (20) gives

$$\max_j \|\tilde{\mathbf{v}}_j - \mathbf{v}^\star\| \le \|\mathbf{W}^V\|_2 G \left(\frac{2H_0}{t} + \frac{2\epsilon d}{t^2} \right) =: \eta(\tau) \longrightarrow 0.$$

For every query position i, the attention weights on its allowed keys form a probability vector, so

$$\left\| \sum_j \tilde{a}_{ij}\tilde{\mathbf{v}}_j - \mathbf{v}^\star \right\| \le \sum_j \tilde{a}_{ij}\|\tilde{\mathbf{v}}_j - \mathbf{v}^\star\| \le \eta(\tau).$$

Multiplying by $\mathbf{W}^O$ proves output convergence, and subtracting the fixed original output proves the limit of the attention-output change. For the first consequence, the same estimate holds for V-only steering with the original weights a_{ij}, giving

$$\|\tilde{\mathbf{o}}_i^{\mathrm{Q+V}} - \tilde{\mathbf{o}}_i^{\mathrm{V}}\| \le 2\|\mathbf{W}^O\|_2 \eta(\tau) \longrightarrow 0.$$

For the second consequence, CAA uses the same steered values at this layer, so the argument applies to its attention output as well. No convergence of the attention weights is needed. RoPE and QK-Norm only affect those weights; summing the finitely many projected head outputs gives the full-block result. □

Softmax tilt and Lipschitz bound. Starting from

$$\tilde{a}_{ij} = \frac{\exp(\ell_{ij}+\delta_{ij})}{\sum_t \exp(\ell_{it}+\delta_{it})}, \tag{24}$$

substituting $\exp(\ell_{ij}+\delta_{ij}) = \exp(\ell_{ij})\exp(\delta_{ij})$ and normalizing by the unperturbed partition function yields Eq. (33).

The Jacobian of softmax at probability vector $\mathbf{a}$ is

$$\mathbf{D} = \operatorname{diag}(\mathbf{a}) - \mathbf{a}^\top \mathbf{a}. \tag{25}$$

Because $\mathbf{D}$ is positive semidefinite and $\mathbf{D} \preceq \operatorname{diag}(\mathbf{a}) \preceq \mathbf{I}$,

$$\|\mathbf{D}\|_2 \le 1. \tag{26}$$

Integrating the Jacobian along the line segment between the original and perturbed logits therefore gives

$$\|\tilde{\mathbf{a}}_{i\cdot} - \mathbf{a}_{i\cdot}\|_2 \le \|\boldsymbol{\delta}_{i\cdot}\|_2. \tag{27}$$

The decomposed bound, Eq. (45). From Eq. (39),

$$\begin{aligned} \left\| \left(\sum_j (\tilde{a}_{ij} - a_{ij}) \mathbf{v}_j \right) \mathbf{W}^O \right\| &= \left\| (\tilde{\mathbf{a}}_{i\cdot} - \mathbf{a}_{i\cdot}) \mathbf{V} \mathbf{W}^O \right\| \\ &\le \|\mathbf{W}^O\|_2 \|\mathbf{V}\|_2 \|\tilde{\mathbf{a}}_{i\cdot} - \mathbf{a}_{i\cdot}\|_2. \end{aligned} \tag{28}$$

The softmax Lipschitz inequality controls the final factor. Moreover, $\tilde{a}_{ij} \ge 0$ and $\sum_j \tilde{a}_{ij} = 1$, so the sum in the second term of Eq. (39) is a convex combination:

$$\left\| \left(\sum_j \tilde{a}_{ij} \mathbf{c}_j \right) \mathbf{W}^O \right\| \le \|\mathbf{W}^O\|_2 \max_j \|\mathbf{c}_j\|. \tag{29}$$

Combining the two inequalities gives Eq. (45). Here $\mathbf{V}$ and $\mathbf{K}$ stack the original values and transformed keys on the allowed positions. For the explicit constants, Eq. (16) and the triangle inequality give $\|\Delta \mathbf{x}_j\| \le 2\sqrt{d}\|\mathbf{g}\|_\infty$, hence

$$\|\mathbf{c}_j\| \le 2\sqrt{d}\|\mathbf{g}\|_\infty \|\mathbf{W}^V\|_2, \qquad \|\mathbf{b}_i\| \le B_Q, \qquad \|\boldsymbol{\delta}_{i\cdot}\|_2 \le \frac{B_Q \|\mathbf{K}\|_2}{\sqrt{d_h}}, \tag{30}$$

where $B_Q = 2\sqrt{d}\|\mathbf{g}\|_\infty \|\mathbf{W}^Q\|_2$ without QK-Norm (rotary matrices are orthogonal and do not change norms) and $B_Q = 2\sqrt{d_h}\|\mathbf{g}_Q\|_\infty$ with QK-Norm, since then both $\mathbf{q}_i$ and $\tilde{\mathbf{q}}_i$ have norm at most $\sqrt{d_h}\|\mathbf{g}_Q\|_\infty$. All constants are independent of τ; substituting into Eq. (45) gives a second, context-dependent proof of uniform boundedness in Eq. (46).

B.4 Secondary identities and diagnostics

The core bounds and limits above do not require a local linearization or a separate estimate of the query contribution. The following identities support interpretation and diagnostics.

State-dependent shift. With $\sigma_\epsilon(\mathbf{h}) = \sqrt{\|\mathbf{h}\|^2/d + \epsilon}$, the exact normalized-input change is

$$\Delta \mathbf{x}_j = \mathbf{g} \odot \left[\mathbf{h}_j \left(\frac{1}{\sigma_\epsilon(\mathbf{h}_j + \tau \mathbf{r})} - \frac{1}{\sigma_\epsilon(\mathbf{h}_j)} \right) + \frac{\tau \mathbf{r}}{\sigma_\epsilon(\mathbf{h}_j + \tau \mathbf{r})} \right]. \tag{31}$$

The component proportional to $\mathbf{h}_j$ makes the shift state dependent. The measured cosine to $\mathbf{g} \odot \mathbf{r}$ is approximately 0.98 at CAA's operating coefficient and 0.928 at PSAS's operating coefficient on Llama-3.1-8B (Table 18). These cosines establish directional alignment, not a bound on the magnitude of either term in Eq. (31). At $64\times$ CAA's coefficient, alignment falls to approximately 0.73–0.75.

Actual queries and keys. For Llama-3.1, $\phi_i^Q(\mathbf{x}) = (\mathbf{x}\mathbf{W}^Q)\mathbf{R}_i$; for Qwen3, $\phi_i^Q(\mathbf{x}) = N_Q(\mathbf{x}\mathbf{W}^Q)\mathbf{R}_i$, where N_Q is RMSNorm with gain $\mathbf{g}_Q$ and positive stability constant. Keys use the corresponding key transformation. Without QK-Norm, $\mathbf{b}_i = \Delta \mathbf{x}_i \mathbf{W}^Q \mathbf{R}_i$; with QK-Norm this relation is nonlinear, but $\|\mathbf{b}_i\| \le 2\sqrt{d_h}\|\mathbf{g}_Q\|_\infty$.

Query and value identities. All sums below range over the query's allowed keys. With $\ell_{ij} = \mathbf{q}_i \mathbf{k}_j^\top / \sqrt{d_h}$,

$$\tilde{\ell}_{ij} = \ell_{ij} + \delta_{ij}, \qquad \delta_{ij} = \mathbf{b}_i \mathbf{k}_j^\top / \sqrt{d_h}. \tag{32}$$

$$\tilde{a}_{ij} = \frac{a_{ij} \exp \delta_{ij}}{\sum_t a_{it} \exp \delta_{it}}. \tag{33}$$

With fixed values,

$$\sum_j \tilde{a}_{ij} \mathbf{v}_j \in \mathrm{conv}\{\mathbf{v}_j\}. \tag{34}$$

With fixed attention weights,

$$\sum_j a_{ij} \tilde{\mathbf{v}}_j = \sum_j a_{ij} \mathbf{v}_j + \sum_j a_{ij} \mathbf{c}_j, \tag{35}$$

$$\sum_j a_{ij} \mathbf{c}_j \in \mathrm{conv}\{\mathbf{c}_j\}, \tag{36}$$

$$\left\| \left(\sum_j a_{ij} \mathbf{c}_j \right) \mathbf{W}^O \right\| \le \|\mathbf{W}^O\|_2 \max_j \|\mathbf{c}_j\|. \tag{37}$$

When both queries and values use the perturbed state,

$$\tilde{\mathbf{o}}_i = \left(\sum_j \tilde{a}_{ij} (\mathbf{v}_j + \mathbf{c}_j) \right) \mathbf{W}^O, \tag{38}$$

$$\tilde{\mathbf{o}}_i - \mathbf{o}_i = \left(\sum_j (\tilde{a}_{ij} - a_{ij}) \mathbf{v}_j \right) \mathbf{W}^O + \left(\sum_j \tilde{a}_{ij} \mathbf{c}_j \right) \mathbf{W}^O. \tag{39}$$

Partial-context value steering. For one head, a steered position set S, and fixed attention weights, let $s_i = \sum_{j \in S} a_{ij}$. Then

$$\left\| \left(\sum_{j \in S} a_{ij} \mathbf{c}_j \right) \mathbf{W}^O \right\| \le s_i \|\mathbf{W}^O\|_2 \max_{j \in S} \|\mathbf{c}_j\|. \tag{40}$$

The contribution is zero when S is empty. For multiple heads, attention weights must remain head specific. Writing $s_i^h = \sum_{j \in S} a_{ij}^h$ and denoting the concatenated-output projection by $\mathbf{W}_{\mathrm{all}}^O$, we obtain

$$\|\Delta \mathbf{o}_i\| \le \|\mathbf{W}_{\mathrm{all}}^O\|_2 \left[\sum_{h=1}^{H} \left(s_i^h \max_{j \in S} \|\mathbf{c}_j^h\| \right)^2 \right]^{1/2}. \tag{41}$$

For each head, apply the triangle inequality to $\sum_{j \in S} a_{ij}^h \mathbf{c}_j^h$; sum the squared headwise bounds under concatenation, and apply the output-projection norm. This proves Eq. (41). Replacing the head-specific masses by their mean does not in general preserve an upper bound. For example, two heads with masses 0.9 and 0.1, a unit shift only in the first head, and an identity output projection give delivered norm 0.9, whereas the mean-mass expression gives 0.5. Attention mass therefore contributes to a bound, but is not an exact multiplier of the full-context perturbation norm.

A secondary decomposed bound. The normalization bound and triangle inequality give

$$\|\mathrm{RMSNorm}(\mathbf{h} + \tau \mathbf{r})\| \le \sqrt{d} \|\mathbf{g}\|_\infty, \tag{42}$$

$$\|\Delta \mathbf{x}_j\| \le 2\sqrt{d} \|\mathbf{g}\|_\infty. \tag{43}$$

The direct mixture argument proves

$$\|\tilde{\mathbf{o}}_i - \mathbf{o}_i\| \le 2M \|\mathbf{W}^O\|_2, \tag{44}$$

whereas the 1-Lipschitz softmax bound and Eq. (39) give

$$\|\tilde{\mathbf{o}}_i - \mathbf{o}_i\| \le \|\mathbf{W}^O\|_2 \left(\|\mathbf{V}\|_2 \|\boldsymbol{\delta}_{i\cdot}\|_2 + \max_j \|\mathbf{c}_j\| \right). \tag{45}$$

For fixed context its right-hand side is bounded independently of τ, recovering

$$\sup_\tau \|\tilde{\mathbf{o}}_i - \mathbf{o}_i\| < \infty. \tag{46}$$

This second argument separates query and value terms but introduces context-dependent constants; it is not needed for Proposition 1. Derivations appear in Appendix B.3.

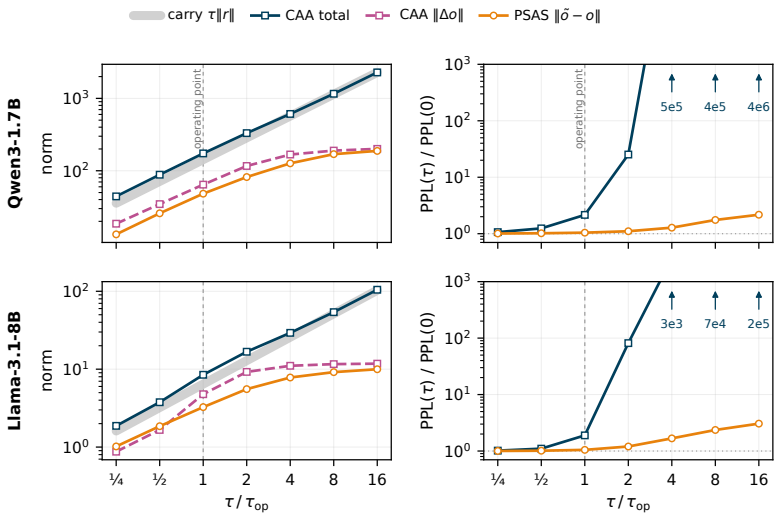


Figure 5: **Perturbation norms and off-target perplexity across coefficients.** Left: CAA's full residual-change norm and the attention-output-change norms for CAA and PSAS. The full residual change includes the subsequent MLP change; attention-only norms do not. Right: teacher-forced Wikipedia perplexity. Coefficients are indexed by each method's own operating strength, and the sampled values do not establish behavior at equal control.

B.5 Empirical perturbation scaling

Figure 5 summarizes the measured perturbation norms and perplexities. These norm measurements do not test vector convergence; Appendix F.6 does. Section 6 reports the associated off-target perplexity trends, and Appendix E reports the supporting tabulated measurements.

C Experimental Details

Main benchmark. Table 1 reports means and sample standard deviations (ddof $= 1$) over five decoding seeds while the direction and split remain fixed. The $\pm$ entries are standard deviations, not standard errors. For example, the CAA T×I scores .517,.509,.496,.496,.518 on Qwen3-1.7B give 0.507 ± 0.011. Directions are differences of mean residual activations on positive and negative examples, read at each example's final token. For TruthfulQA, 40% of the questions are used to estimate the steering direction, 10% are used as a validation subset to select the intervention layer and coefficient, and the remaining 50% are held out for final evaluation. The selected configuration is fixed before evaluation on the held-out subset. For RealToxicityPrompts, we estimate the detoxification steering direction from 6,000 examples in the Jigsaw Unintended Bias in Toxicity Classification dataset, comprising 3,000 toxic and 3,000 benign examples. We use a validation subset of RealToxicityPrompts to select the intervention layer and coefficient, and a separate held-out test subset of 500 toxic prompts for final generation evaluation. Fine-tuned Llama-2-7B classifiers judge truthfulness and informativeness per answer, and T×I is their per-answer conjunction. To evaluate output text toxicity, we use the `s-nlp/roberta_toxicity_classifier` model. Decoding uses temperature 0.7, nucleus probability 0.9, and repetition penalty 1.1, with a common system prompt across methods. Generation perplexity is scored with GPT-2 XL on continuations of up to 64 tokens; Dist-2 measures distinct bigrams. Teacher-forced Wikipedia perplexities belong to separate diagnostic evaluations.

Additional pathway study. This study uses zero-based blocks 16 and 15 for Qwen3-1.7B and Llama-3.1-8B, respectively. It reports Wikipedia results on 500 fixed sequences truncated to 128 tokens and internal diagnostics on 64 sequences. Directions are estimated on the training half of a two-fold TruthfulQA split. Its coefficients and condition definitions are given in Appendix D. The reported software versions are PyTorch 2.11.0+cu128, Transformers 5.16.1, NumPy 2.2.6, and pandas 2.3.3, with fp32 computation. These details belong to the additional study and are not assumed to describe every benchmark run.

Condition	Qwen3-1.7B	Llama-3.1-8B
CAA	1.000	1.000
Carry-only	0.546	0.575
All branches, no direct addition	0.114	0.211
Q+V	0.057	0.036

Table 3: Fraction ρ_c of CAA's Wikipedia perplexity increase at the same coefficient within each model. Each column uses its model's CAA operating coefficient.

D Pathway Ablation Details

The pathway ablation intervenes at zero-based blocks 16 (Qwen3-1.7B) and 15 (Llama-3.1-8B), using directions read after blocks 15 and 14, respectively. Evaluation uses 500 fixed, seeded Wikipedia sequences truncated to 128 tokens. Corpus perplexity is $\exp(\text{total NLL}/\text{scored tokens})$. The formulas below define each intervention on the same input to the selected block, so pathway differences can be identified before downstream computation.

Interventions on identical block inputs. Let a_0 and a_τ be attention outputs computed from $\mathrm{RMSNorm}_l^a(h)$ and $\mathrm{RMSNorm}_l^a(h+\tau r)$, respectively; all formulas below are token-wise, with the corresponding full-context inputs used by attention. Let $m(z)=\mathrm{MLP}_l(\mathrm{RMSNorm}_l^m(z))$. The interventions are

$$\begin{aligned} F_0 &= h+a_0+m(h+a_0), \\ F_{\mathrm{CAA}} &= h+\tau r+a_\tau+m(h+\tau r+a_\tau), \\ F_{\mathrm{carry}} &= F_0+\tau r, \\ F_{\mathrm{NW}} &= h+a_\tau+m(h+\tau r+a_\tau), \\ F_{\mathrm{MLP}} &= h+a_0+m(h+a_0+\tau r). \end{aligned}$$

For Q-only, K-only, V-only, Q+V, and Q+K+V, selected projections consume the steered normalized auxiliary state; all other projections consume the original normalized state. If a_R is the resulting attention output, $F_R=h+a_R+m(h+a_R)$. The usual RoPE and QK-Norm transformations are retained. This gives ten conditions including the unsteered model. Equation (2) holds for CAA and NW, not for CAA and Q+V.

Perplexity-increase fractions. For each fixed coefficient panel with $P_{\mathrm{CAA}} > P_0$, define

$$\rho_c = \frac{P_c - P_0}{P_{\mathrm{CAA}} - P_0}. \tag{47}$$

Here P_c is the condition's token-weighted perplexity and P_0 is the unsteered value. The fractions need not sum to one: perplexity is nonlinear, and the conditions are separate interventions rather than additive loss components. We reserve ρ_c for this statistic. The additional reference coefficients are denoted τ_{ref}, to distinguish them from perplexity-increase fractions. The three coefficient panels are:

Panel	Qwen3-1.7B	Llama-3.1-8B
CAA operating coefficient	5.603	4.000
Q+V operating coefficient	4.002	8.000
Reference coefficient τ_{ref}	8.006	6.633

Every condition in a panel uses that panel's common coefficient. The reference coefficients are not assumed to give $\kappa_l = 1$ on the diagnostic sample.

Table 3 summarizes the panel at CAA's operating coefficient.

The reported bootstrap procedure uses 2,000 paired sequence-level resamples. The tables present point estimates without confidence intervals; the comparisons here do not assert statistical significance.

Full residual changes. On 64 Wikipedia sequences at the CAA coefficients, the reported mean full next-layer residual-change norms are 174.29 for CAA and 64.24 for Q+V on Qwen3-1.7B, and 6.88 and 1.96 on Llama-3.1-8B. The corresponding attention-change norms are 64.48 and 63.27, and 3.40 and 2.16.

Qwen3-1.7B

		TruthfulQA			RealToxicityPrompts	
Condition	ρ_c	**Wiki PPL ↓**	**T × I ↑**	**Gen PPL ↓**	**Toxicity ↓**	**Gen PPL ↓**
Unsteered	—	21.24	0.318	24.2	0.158	18.8
CAA	1.000	102.72	0.406	24.1	collapsed	176.2
Carry-only	0.437	56.86	0.439	24.4	collapsed	127.2
All branches, no direct addition	0.067	26.73	0.476	17.8	**0.009**	32.1
Q-only	0.003	21.46	0.318	24.2	0.134	17.8
K-only	-0.006	20.71	0.371	22.5	0.158	19.3
V-only	0.026	23.35	<u>0.543</u>	18.5	0.093	21.1
Q+V (PSAS)	0.028	23.54	**0.561**	17.1	<u>0.026</u>	23.6
Q+K+V	0.026	23.36	0.537	17.4	0.058	21.8
MLP-only	0.019	22.82	0.173	28.2	0.056	26.6

Llama-3.1-8B

		TruthfulQA			RealToxicityPrompts	
Condition	ρ_c	**Wiki PPL ↓**	**T × I ↑**	**Gen PPL ↓**	**Toxicity ↓**	**Gen PPL ↓**
Unsteered	—	10.11	0.414	48.4	0.429	20.8
CAA	1.000	61.82	0.100	50.7	collapsed	379.0
Carry-only	0.282	24.71	0.487	29.8	collapsed	143.4
All branches, no direct addition	0.077	14.08	0.448	27.3	**0.049**	30.4
Q-only	0.002	10.23	0.480	106.8	0.421	20.6
K-only	0.003	10.24	0.421	108.8	0.419	20.9
V-only	0.006	10.43	**0.672**	24.7	0.305	21.9
Q+V (PSAS)	0.009	10.57	<u>0.666</u>	26.5	0.105	24.4
Q+K+V	0.070	13.74	0.436	28.4	<u>0.094</u>	25.6
MLP-only	0.022	11.25	0.448	33.0	0.283	21.0

Table 4: **Pathway ablation at a fixed layer (ℓ) and coefficient (τ). (Seed 42)** For Qwen3-1.7B, parameters are $\ell = 15, \tau = 4.0$ for TruthfulQA and $\ell = 18, \tau = 32$ for RealToxicityPrompts. For Llama-3.1-8B, $\ell = 14$ for both tasks with $\tau = 8.0$ (TruthfulQA) and $\tau = 24$ (RealToxicityPrompts). For T × I, **bold** and <u>underline</u> values indicate highest and second-highest performance. For Toxicity, **bold** and <u>underline</u> values indicate lowest and second-lowest scores among non-collapsed conditions.

	Qwen3-1.7B			Llama-3.1-8B		
Condition	CAA τ	Q+V τ	τ_{ref}	CAA τ	Q+V τ	τ_{ref}
Carry-only	0.546	0.593	0.437	0.575	0.165	0.282
All branches, no direct addition	0.114	0.144	0.067	0.211	0.028	0.077
Q+V	0.057	0.083	0.028	0.036	0.003	0.009

Table 5: Sensitivity of ρ_c to the common coefficient selected for each panel.

These are norms of distinct vectors: proximity of their magnitudes does not establish vector similarity, and subtracting them does not isolate the carry. The exact CAA–NW identity was checked on identical block inputs, with reported relative errors 1.7×10^{-6} and 1.2×10^{-6}. Carry-only gives zero attention change at the selected block and the full injected residual shift.

Interaction diagnostic. On identical block inputs, define

$$I_{Q,V} = F_{Q+V} - F_Q - F_V + F_0, \qquad I_{a,m} = F_{\text{NW}} - F_{Q+K+V} - F_{\text{MLP}} + F_0.$$

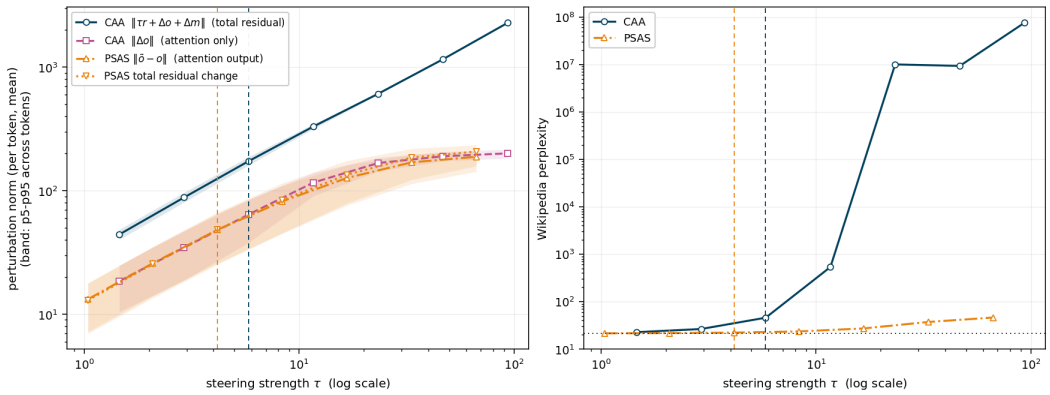
The reported mean norms are 16.70 and 47.61 on Qwen3-1.7B, and 0.59 and 1.31 on Llama-3.1-8B. These measure nonadditivity of the full residual updates; they do not partition perplexity degradation.

E SUPPORTING NUMERICAL MEASUREMENTS

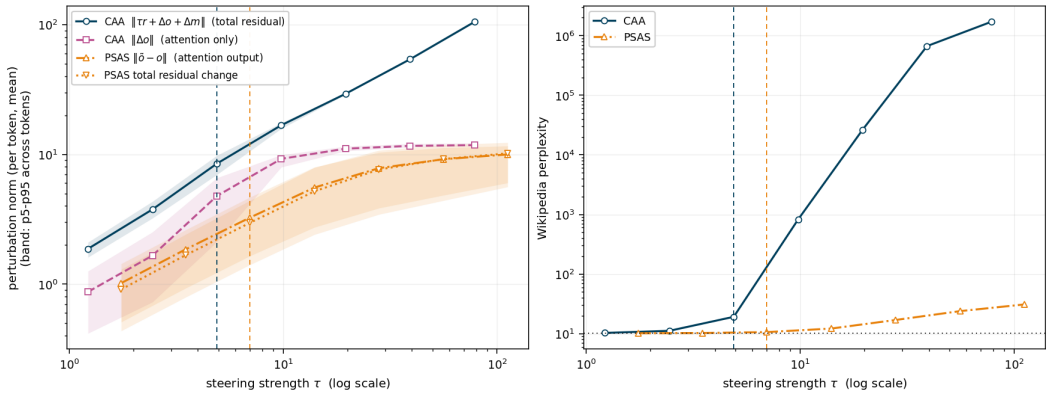
This appendix reports supporting measurements on 500 fixed, seeded Wikipedia sequences truncated to 128 tokens. The study uses a TruthfulQA difference-in-means direction and method-specific operating strengths, steering zero-based block 16 on Qwen3-1.7B and block 15 on Llama-3.1-8B. Directions are read at the selected block's residual input. The measurements are taken from the original run records: `bounds.csv`, `disco_vs_psas.csv`, and `convergence_caa.csv`. The study's coefficient multiples are retained as reported; the absolute coefficients from Appendix D are not substituted for this separate study.

Here, relative off-target perplexity is the ratio of the steered model's teacher-forced perplexity on Wikipedia to that of the corresponding unsteered model. This measure is distinct from the GPT-2 XL generation perplexity in Table 1. All perturbation norms are computed at the token level.

These measurements describe perturbation magnitudes, perplexity, and attention statistics. Finite-sample norm measurements do not test the asymptotic vector limits or establish the cause of perplexity changes; Appendix F tests the limits directly. Figure 6 and Table 6 compare perturbation scaling and off-target perplexity; Figures 7 and 8, together with Table 7, evaluate the RMSNorm, value-path, and softmax properties (Lemma 1 and Proposition 1).



(a) Qwen3-1.7B



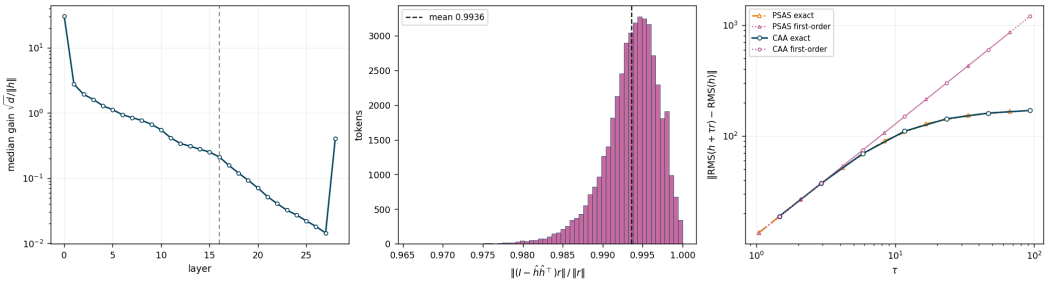
(b) Llama-3.1-8B

Figure 6: Per-token perturbation norms at the steered layer and relative teacher-forced Wikipedia perplexity across steering strengths. For PSAS, the plotted perturbation is the attention-output change $\|\tilde{\mathbf{o}} - \mathbf{o}\|$; it excludes any subsequent change induced through the MLP.

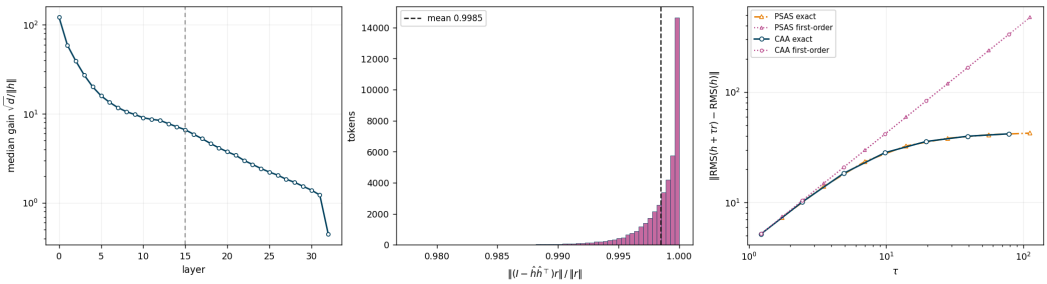
Context dependence of query steering. Lemma 1 gives $|\delta_{ij}| = \|\mathbf{b}_i\| \|\mathbf{k}_j\| |\cos(\mathbf{k}_j, \mathbf{b}_i)| / \sqrt{d_h}$. Holding the norms fixed, stronger alignment produces a larger absolute logit offset. The diagnostic groups transformed query–key pairs by alignment quintile, using QK-Norm on Qwen3 and RoPE on both models. The highest-quintile mean absolute offset is $11.3\times$ and $13.3\times$ the lowest-quintile value on Wikipedia and TruthfulQA for Qwen3-1.7B, and $13.4\times$ and $11.8\times$ for Llama-3.1-8B. Head-averaged attention KL

τ/τ_{op}	Qwen3-1.7B CAA total	Qwen3-1.7B CAA $\|\Delta\mathbf{o}\|$	Qwen3-1.7B PSAS $\|\tilde{\mathbf{o}}-\mathbf{o}\|$	Qwen3-1.7B CAA PPL	Qwen3-1.7B PSAS PPL	Llama-3.1-8B CAA total	Llama-3.1-8B CAA $\|\Delta\mathbf{o}\|$	Llama-3.1-8B PSAS $\|\tilde{\mathbf{o}}-\mathbf{o}\|$	Llama-3.1-8B CAA PPL	Llama-3.1-8B PSAS PPL
0.25	44.2	18.6	13.2	1.07	1.01	1.9	0.9	1.0	1.02	1.00
0.5	88.2	34.6	25.8	1.24	1.02	3.8	1.7	1.9	1.10	1.01
1	173.7	64.3	48.2	2.15	1.04	8.5	4.8	3.3	1.89	1.05
2	331.2	116.4	81.9	25.2	1.11	16.7	9.2	5.5	81.5	1.20
4	609.6	167.6	126.6	4.7×10^5	1.28	29.3	11.1	7.8	2.6×10^3	1.68
8	1158.4	189.9	169.6	4.4×10^5	1.75	54.0	11.6	9.2	6.6×10^4	2.36
16	2284.8	200.6	188.0	3.6×10^6	2.17	104.7	11.8	10.0	1.7×10^5	3.06

Table 6: Perturbation magnitudes at the steered layer and relative Wikipedia perplexity versus normalized steering strength. "CAA total" is the norm of the full residual change entering the next layer; "CAA $\|\Delta\mathbf{o}\|$" and PSAS $\|\tilde{\mathbf{o}}-\mathbf{o}\|$ denote the respective attention-output changes.



(a) Qwen3-1.7B



(b) Llama-3.1-8B

Figure 7: Numerical diagnostics of the RMSNorm analysis. Left: median $\sqrt{d}/\|\mathbf{h}_l\|$ by layer. Center: fraction of the steering direction retained by the tangent-space projector. Right: comparison of the exact normalized perturbation with its first-order approximation across steering strengths.

divergence is 1.5–1.7$\times$ larger on TruthfulQA than on Wikipedia. Attention-weight changes are more model dependent: the highest-to-lowest quintile ratio is approximately 4.7 on Llama across both datasets, but the response is approximately flat on Qwen despite the concentrated logit offsets. These results describe context-dependent reweighting, without demonstrating that the query pathway selectively avoids off-target effects.

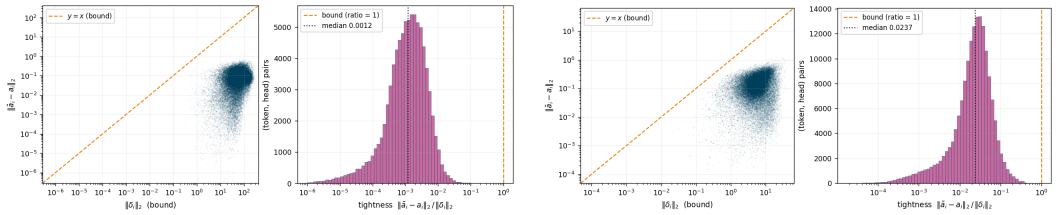
A nondiscriminative subspace diagnostic. The sum of the attention-head output subspaces is the row span of the combined output projection. When $Hd_h = d$, this matrix is square but need not have full rank. If it has rank d, every residual perturbation lies in the span, so subspace membership cannot distinguish attention steering from residual steering. The relevant structural property here is bounded perturbation growth, not confinement to a demonstrated proper subspace.

F Numerical Tests of the Theoretical Claims

This appendix examines the quantities appearing in Sections 2–4: normalized states, branch outputs, residual changes, and output distributions. Finite sweeps provide numerical evidence about convergence; data-dependent constants provide reference scales rather than certified uniform bounds. Both models run in fp32. Directions are estimated from the training half of the TruthfulQA split at zero-based blocks 16

Quantity	Qwen3-1.7B summary	Qwen3-1.7B p5 / p95	Llama-3.1-8B summary	Llama-3.1-8B p5 / p95
Scale factor $\sqrt{d}/\|\mathbf{h}_l\|$, steered layer	0.211	0.184 / 0.242	6.67	5.85 / 7.97
Scale factor, layer 0 (median)	30.5	N/A	121.6	N/A
Scale factor, last layer (median)	0.409	N/A	0.449	N/A
Projector survival $\|\mathbf{r}(\mathbf{I}-\hat{\mathbf{h}}^\top\hat{\mathbf{h}})\|/\|\mathbf{r}\|$	0.994	0.987 / 0.998	0.999	0.995 / 1.000
$\|\mathbf{r}\|$	24.4	N/A	1.31	N/A
Exact / first-order change at τ_{op}	0.975	N/A	0.795	N/A
Exact / first-order change at $4\tau_{\text{op}}$	0.610	N/A	0.331	N/A
Exact / first-order change at $16\tau_{\text{op}}$	0.205	N/A	0.100	N/A
V-path mixture residual	10^{-6}	N/A	2×10^{-6}	N/A
$\|\tilde{\mathbf{o}}-\mathbf{o}\|/\max_j\|\mathbf{c}_j\|$ (V only)	0.458	0.250 / 0.610	0.292	0.142 / 0.431
CV of $\|\mathbf{c}_j\|$ across positions	0.128	N/A	0.121	N/A
Alignment $\alpha_{\mathbf{r}}(\Delta\mathbf{h}_{\text{CAA}})$ at τ_{op}	0.778	0.703 / 0.838	0.693	0.622 / 0.789
Softmax Lipschitz bound holds	100%	N/A	100%	N/A
$\|\tilde{\mathbf{a}}_{i\cdot}-\mathbf{a}_{i\cdot}\|_2/\|\boldsymbol{\delta}_{i\cdot}\|_2$ at τ_{op} (mean; median)	0.0023; 0.0012	N/A / 0.0076	0.032; 0.024	N/A / 0.090

Table 7: Numerical checks of the RMSNorm analysis and the value- and query-path properties.



(a) Qwen3-1.7B (b) Llama-3.1-8B

Figure 8: Empirical check of the softmax Lipschitz inequality at the operating strength. Each point represents one token-head pair. All pairs satisfy the inequality, although the upper bound is typically loose by one to three orders of magnitude.

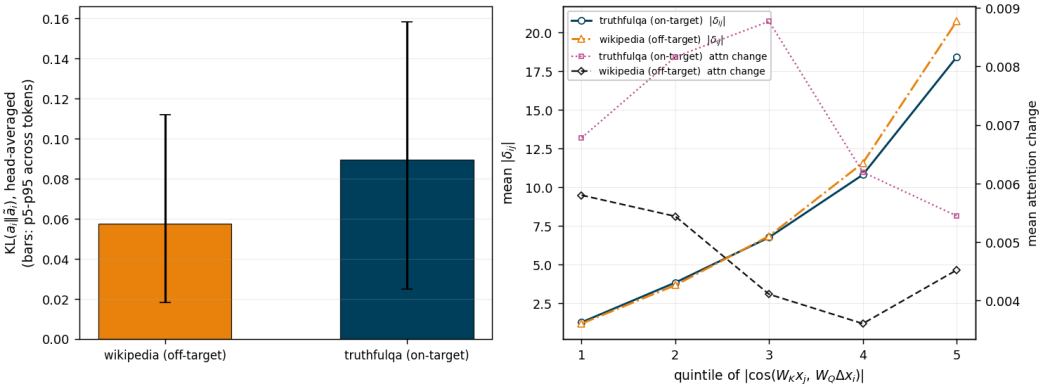
(Qwen3-1.7B) and 15 (Llama-3.1-8B), as in Appendix D. These directions differ from those in Appendix E, so absolute perplexities should not be identified with Table 6. Unless stated otherwise, evaluation uses the 500-sequence Wikipedia sample truncated to 128 tokens. The operating coefficients are CAA 5.603 and 4.0, and `psas` 4.002 and 8.0, respectively. Tables 17 and 18 use CAA's coefficient for both methods to compare vectors at the same injection strength; PSAS's Llama operating coefficient corresponds to the $2\times$ row. The DISCO-QV scaling comparison uses each method's own operating coefficients. Cosines and norms are computed per token and averaged over the sample.

F.1 Limit of the normalized state (Eq. 1)

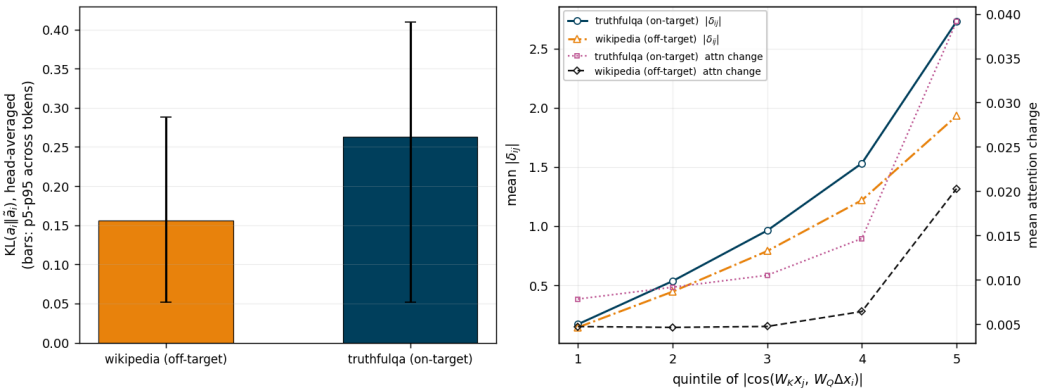
Table 8 reports $\cos(\text{RMSNorm}(\mathbf{h}+\tau\mathbf{r}),\mathbf{g}\odot\hat{\mathbf{r}})$ at the steered block. The cosine rises monotonically toward one. At the operating coefficient it is 0.43 and 0.47, so the limit describes a regime well beyond it.

τ/τ_{op}	0	0.25	0.5	1	2	4	8	16	32	64
Qwen3-1.7B	−0.121	0.028	0.174	0.428	0.724	0.905	0.967	0.985	0.992	0.996
Llama-3.1-8B	−0.058	0.092	0.235	0.474	0.746	0.912	0.971	0.990	0.997	0.999

Table 8: Cosine between the steered normalized state and its predicted limit $\mathbf{g}\odot\hat{\mathbf{r}}$ at the steered block, versus CAA coefficient multiple.



(a) Qwen3-1.7B



(b) Llama-3.1-8B

Figure 9: Context dependence of query steering. Left: attention KL divergence on off-target Wikipedia and on-target TruthfulQA text. Right: mean absolute logit offset and attention change stratified by alignment quintile. Absolute logit offsets concentrate in the highest-alignment quintile for both models, whereas the response of the attention weights differs between models.

F.2 Direction of the residual change and bound constants (Proposition 3)

Table 9 reports the absolute cosine $\alpha_{\mathbf{r}}$ defined in Eq. (14). Its increase toward one is consistent with an increasingly direction-aligned residual change. As discussed in Appendix B, this statistic alone neither determines its sign nor isolates a causal contribution from the carry.

τ/τ_{op}	0.25	0.5	1	2	4	8	16
Qwen3-1.7B	0.764	0.770	0.778	0.822	0.914	0.972	0.992
Llama-3.1-8B	0.685	0.702	0.707	0.739	0.862	0.952	0.987

Table 9: Alignment $\alpha_{\mathbf{r}}$ of CAA's full residual change with the steering direction.

Table 10 distinguishes empirical scales from the constants in the proofs. Let $\widehat{M}_h$ be the largest clean-sample value norm observed for head h, and define $\widehat{C}^{\text{attn}} = \|\mathbf{W}^O\|_2(\sum_h \widehat{M}_h^2)^{1/2}$. Since $\widehat{M}_h \le M_h$, this attention reference scale is no larger than the certified constant, but need not bound steered values. A measured-change ratio using $2\widehat{C}^{\text{attn}}$ is consequently an upper bound on the corresponding ratio using $2C^{\text{attn}}$. The MLP reference $\widehat{C}^{\text{mlp}}$ is obtained from random-direction finite differences and an input-radius factor. It is neither a certified supremum nor guaranteed to lie on a particular side of C^{mlp}. The full-layer scale in Table 16 must therefore be interpreted empirically.

	Qwen3-1.7B	Llama-3.1-8B
$\|\mathbf{W}^O\|_2$	4.843	3.439
$\max_h \widehat{M}_h$	76.90	7.835
$\widehat{C}^{\text{attn}} = \|\mathbf{W}^O\|_2 (\sum_h \widehat{M}_h^2)^{1/2}$	1157.5	117.8
$\widehat{C}^{\text{mlp}}$ (empirical)	193.0	6.9

Table 10: Empirical branch reference scales at the steered block. Hats distinguish these quantities from the certified constants in Appendix B.3.

F.3 Loss of input dependence (Proposition 4)

Three quantities test the proposition under all-position CAA steering. Table 11 gives, at the steered block and at the final layer, the across-token spread of the normalized state (mean per-dimension standard deviation of the unit-normalized vectors) and its cosine to $\mathbf{g}_{l'} \odot \hat{\mathbf{r}}$. Table 12 gives the KL divergence from the model's last-position next-token distribution to $q^* = \text{softmax}(\sqrt{d}(\mathbf{g}_L \odot \hat{\mathbf{r}})\mathbf{W}^U)$, computed from the weights, and the across-prompt variance of that distribution. Table 13 compares measured CAA perplexity with the predicted limit $\exp H(p_{\text{data}}, q^*)$ on the same tokens. For Qwen3-1.7B, q^* has entropy 5.28 nats and maximum probability 0.178.

	Qwen3-1.7B		Llama-3.1-8B	
τ/τ_{op}	spread (block $\to$ final)	cosine (block $\to$ final)	spread (block $\to$ final)	cosine (block $\to$ final)
0	$0.0125 \to 0.0059$	$-0.121 \to 0.074$	$0.0123 \to 0.0141$	$-0.058 \to -0.004$
1	$0.0114 \to 0.0064$	$0.427 \to 0.101$	$0.0109 \to 0.0140$	$0.474 \to 0.041$
4	$0.0055 \to 0.0079$	$0.905 \to 0.245$	$0.0050 \to 0.0069$	$0.911 \to 0.264$
16	$0.0020 \to 0.0025$	$0.985 \to 0.374$	$0.0015 \to 0.0019$	$0.990 \to 0.769$
64	$0.0006 \to 0.0008$	$0.996 \to 0.610$	$0.0004 \to 0.0005$	$0.999 \to 0.968$

Table 11: Across-token spread of the normalized state and its cosine to the layer's predicted limit, at the steered block and at the final normalization, under all-position CAA.

	KL to q^*		across-prompt variance	
τ/τ_{op}	Qwen3-1.7B	Llama-3.1-8B	Qwen3-1.7B	Llama-3.1-8B
0	17.18	13.16	0.181	0.145
1	16.95	12.89	0.256	0.143
2	15.30	11.01	0.163	0.065
4	13.76	9.33	0.090	0.016
8	8.61	3.40	0.0067	$<10^{-4}$
16	6.38	2.31	0.0014	$<10^{-4}$
32	4.94	1.55	0.0003	$<10^{-4}$
64	2.87	0.78	0.0001	$<10^{-4}$

Table 12: Convergence of the next-token distribution under all-position CAA: KL divergence to the predicted input-independent limit q^*, and variance of the distribution across prompts. Entries below 10^{-4} are reported as a threshold, not as zero.

τ/τ_{op}	1	2	4	8	16	32	64	predicted limit
Qwen3-1.7B	48.4	656	3.1×10^7	1.1×10^7	7.7×10^7	1.8×10^8	2.4×10^8	6.0×10^8
Llama-3.1-8B	15.3	231	2.3×10^4	4.1×10^5	1.3×10^6	2.3×10^6	3.7×10^6	1.4×10^7

Table 13: Measured absolute Wikipedia perplexity under all-position CAA and the limit predicted by Proposition 4. Unsteered perplexity is 21.2 and 10.1. The direction differs from Table 6.

The measurements are consistent with an eventual loss of input dependence, but different diagnostics approach their limits at different rates. By $8\times$, across-prompt variance falls to 0.0067 on Qwen3-1.7B (below 4% of its unsteered value) and below the 10^{-4} reporting threshold on Llama-3.1-8B. At $64\times$,

measured perplexity is within factors of approximately 2.5 and 3.7 of the respective predicted limits. Qwen's final-layer cosine is still only 0.61, illustrating that low across-prompt variance need not imply proximity to the particular limiting predictor. Its perplexity is also nonmonotone between $4\times$ and $8\times$. The theoretical rate includes the entire downstream remainder and supplies no tight finite-coefficient prediction from the empirical scale of a single block.

Practical degradation precedes near-collapse. At $2\times$, Qwen perplexity is 656 although variance remains 0.163 versus 0.181 unsteered, and KL to q^* decreases only from 17.18 to 15.30. On Llama, perplexity reaches 231 while variance is 0.065 versus 0.145 unsteered. At $4\times$, the perplexities are already approximately 3×10^7 and 2×10^4, despite remaining variation across prompts. These observations show that severe degradation does not require an approximately input-independent predictor; they do not identify a unique finite-strength mechanism causing that degradation.

F.4 Attention-output bound and context length (Proposition 1)

Table 14 compares the measured attention-output change at $16\times$ with the empirical multi-head scale $2\widehat{C}^{\mathrm{attn}}$. The ratios are 0.083 and 0.039. The listed single-head expression, formed with $\max_h \widehat{M}_h$, is a reference only: it cannot bound a concatenation of all head outputs. Table 15 varies context length from 32 to 2048 tokens at the operating coefficient. Measured changes grow by approximately 50% on Qwen and 80% on Llama. This finite-sample growth is compatible with the context-length-independent theoretical bound.

	multi-head scale $2\widehat{C}^{\mathrm{attn}}$	single-head reference	measured at $16\times$	measured / scale
Qwen3-1.7B	2315.0	744.9	191.3	0.083
Llama-3.1-8B	235.6	53.9	9.19	0.039

Table 14: Attention changes and empirical reference scales. The single-head reference is $2\max_h \widehat{M}_h\|\mathbf{W}^O\|_2$. Only the multi-head expression has the form needed for a full attention-block bound; the empirical constants are not certified.

context length (tokens)	32	128	512	2048
Qwen3-1.7B	38.05	48.34	53.96	56.92
Llama-3.1-8B	3.03	3.93	4.60	5.43

Table 15: PSAS attention-output change at the operating coefficient versus context length.

F.5 Full-layer bound (Eq. 10)

Table 16 compares the full next-layer residual change with $2\widehat{C}^{\mathrm{attn}}+2\widehat{C}^{\mathrm{mlp}}$. On Qwen3-1.7B, its norm is approximately 11% larger than the attention-only norm; on Llama-3.1-8B it is slightly smaller. These comparisons do not determine the magnitude of the MLP change or its alignment with the attention change: a large component can partially cancel another. The ratios to the empirical scale provide a numerical reference, not a verification of the certified full-layer bound.

	reference scale	attention-only at $16\times$	full-layer at $16\times$	full / scale
Qwen3-1.7B	2701.0	191.3	211.8	0.078
Llama-3.1-8B	249.4	9.19	9.13	0.037

Table 16: Eq. (10) against the empirical reference scale of Table 10; the MLP constant is estimated, not certified.

F.6 Common-value limit and its consequences (Proposition 2)

With $\mathbf{v}^\star = \sqrt{d}(\mathbf{g}\odot\hat{\mathbf{r}})\mathbf{W}^V$ and $\mathbf{o}^\star = \mathbf{v}^\star\mathbf{W}^O$ computed from the weights, Table 17 reports, at the steered block across the sweep: (a) the cosine of the steered value vectors to $\mathbf{v}^\star$; (b) $\|\tilde{\mathbf{o}}-\mathbf{o}^\star\|/\|\tilde{\mathbf{o}}\|$; (c)

$\|\tilde{\mathbf{o}}^{\mathrm{Q+V}} - \tilde{\mathbf{o}}^{\mathrm{V}}\|/\|\tilde{\mathbf{o}}^{\mathrm{V}}\|$; (d) $\|\tilde{\mathbf{o}}^{\mathrm{CAA}} - \tilde{\mathbf{o}}^{PSAS}\|/\|\tilde{\mathbf{o}}^{PSAS}\|$. All conditions in a row use the same coefficient, CAA's operating value times the multiplier, so (d) compares the two attention outputs at matched injection strength rather than each at its own tuned strength.

	Qwen3-1.7B				Llama-3.1-8B			
τ/τ_{op}	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
0.25	0.194	3.819	0.103	0.109	0.157	4.562	0.137	0.130
1	0.660	2.411	0.304	0.472	0.550	4.076	0.398	0.869
4	0.956	0.709	0.793	0.610	0.928	1.544	0.952	1.280
8	0.984	0.352	0.968	0.331	0.977	0.843	1.200	0.731
16	0.994	0.220	0.903	0.183	0.993	0.502	1.109	0.445
64	0.999	0.079	0.383	0.061	1.000	0.146	0.372	0.135

Table 17: Convergence at the steered block: (a) $\cos(\tilde{\mathbf{v}}, \mathbf{v}^{\star})$; (b) relative distance of the PSAS attention output from its limit; (c) relative difference between Q+V and V-only outputs; (d) relative difference between CAA and PSAS attention outputs.

The finite sweep supports approach toward the predicted limits without establishing that the asymptotic regime has been reached. At $64\times$, value cosines are 0.999 or greater, but relative attention-output errors remain 0.079 and 0.146. The Q+V/V-only and CAA/PSAS differences are nonmonotone: Q+V and V-only differ by 30–40% at CAA's operating coefficient, approximately order one at intermediate strengths, and about 37–38% even at $64\times$. Thus a common limit can coexist with substantial finite-strength differences.

At CAA's operating coefficient, the CAA and PSAS attention outputs differ by 47% and 87%. On the Qwen subsample, their attention-change norms are nevertheless close (64.5 versus 63.3; Appendix D). Similar perturbation norms therefore do not establish similar perturbation vectors.

F.7 Direction of the in-layer shift (Eq. 31)

Table 18 measures the alignment of the exact normalized-input shift with $\mathbf{g} \odot \mathbf{r}$. At CAA's operating coefficient the cosines are 0.975 and 0.977. At PSAS's Llama coefficient, corresponding to the $2\times$ row, the cosine is 0.928; PSAS's Qwen coefficient is below CAA's operating coefficient. At $64\times$, the cosines are 0.727 and 0.750. These measurements show that a nearly fixed shift direction is a reasonable description near the evaluated coefficients, while state dependence becomes more pronounced at larger strengths. They do not imply that all tokens receive the same shift magnitude.

τ/τ_{op}	0.25	0.5	1	2	4	8	16	32	64
Qwen3-1.7B	0.999	0.998	0.975	0.912	0.834	0.780	0.751	0.735	0.727
Llama-3.1-8B	0.999	0.995	0.977	0.928	0.861	0.808	0.776	0.759	0.750

Table 18: Cosine between the exact normalized-input shift $\Delta\mathbf{x}_j$ and the fixed vector $\mathbf{g} \odot \mathbf{r}$.

F.8 Key shifts and attention mass

Table 19 reports attention KL divergence after the models' RoPE and QK-Norm transformations. K-only steering does not cancel: on Qwen3-1.7B, it changes attention weights more than Q-only or CAA at every tabulated coefficient through $8\times$.

τ/τ_{op}	0.25	0.5	1	2	4	8	16
Qwen3-1.7B, K-only	0.016	0.091	0.509	1.797	3.311	4.108	4.076
Qwen3-1.7B, Q-only	0.008	0.031	0.111	0.369	1.086	2.144	2.895
Qwen3-1.7B, CAA	0.018	0.061	0.187	0.636	2.044	3.687	4.167
Llama-3.1-8B, K-only	0.011	0.040	0.130	0.302	0.480	0.600	0.650
Llama-3.1-8B, Q-only	0.008	0.027	0.078	0.208	0.541	0.983	1.311
Llama-3.1-8B, CAA	0.022	0.058	0.218	1.472	3.515	4.620	4.816

Table 19: Attention KL divergence (clean versus steered) at the steered block, mean over tokens and heads.

Table 21 applies value steering only to a suffix of the context and measures the final query's delivered perturbation relative to full-context steering. The ratio of the mean delivered fraction to mean suffix attention mass is approximately 2.4 on Qwen and 3–4 on Llama. Thus mean attention mass is not an accurate multiplicative estimate of the full-context perturbation fraction on these samples; norms also depend on value magnitudes, alignment, cancellation, and the output projection.

Table 20 evaluates the valid head-specific bound in Eq. (41) on 48 sequences. The delivered-to-bound ratio is at most 0.249 across the listed conditions. The table also retains the numerical reference obtained from head-averaged mass and concatenated shifts. Although all measured changes lie below that reference in this sample, it is not a general multi-head bound, as the counterexample in Appendix B.4 demonstrates. The two expressions should therefore not be ranked as tighter and looser versions of the same bound.

model, suffix	n	attn. mass	delivered	mean-mass reference		head-specific bound	
				reference	max ratio	bound	max ratio
Qwen3-1.7B, 25%	34	0.183	28.64	77.48	0.794	148.05	0.249
Qwen3-1.7B, 50%	46	0.244	34.89	104.77	0.441	193.68	0.227
Qwen3-1.7B, 75%	48	0.328	44.54	142.52	0.373	255.76	0.222
Llama-3.1-8B, 25%	34	0.116	0.703	2.463	0.767	5.98	0.174
Llama-3.1-8B, 50%	47	0.130	0.777	2.834	0.515	6.74	0.154
Llama-3.1-8B, 75%	48	0.157	0.955	3.455	0.459	8.21	0.146

Table 20: Partial-context value steering on 48 sequences at the final query. The head-specific bound is Eq. (41); the mean-mass expression is an empirical reference, not a general bound. Each max ratio is the largest delivered/reference or delivered/bound ratio among retained sequences. Suffixes containing only padding are excluded (n retained). Attention mass is averaged on this sample, which differs from the 500-sequence sample in Table 21.

steered suffix	Qwen3-1.7B			Llama-3.1-8B		
	attn. mass	delivered / full	ratio of means	attn. mass	delivered / full	ratio of means
25%	0.130	0.308	2.37	0.082	0.251	3.06
50%	0.233	0.556	2.39	0.128	0.463	3.62
75%	0.328	0.776	2.37	0.157	0.662	4.22

Table 21: Partial-context value steering at the final query position. Attention mass and delivered fraction are means over the sample; the last column is the ratio of those two means.

G ADDITIONAL BEHAVIORAL PATHWAY AND COEFFICIENT SWEEPS

We examine how pathway choice, residual carry, and steering strength affect behavioral control and language-modeling cost on Qwen3-1.7B and Llama-3.1-8B. Within each model, pathway arms share the residual direction and intervention layer and steer all token positions, including CAA. This all-token CAA convention differs from the last-position CAA benchmark in Table 1. Each arm uses its own coefficient range because equal coefficients need not deliver equal perturbations.

Cost is teacher-forced Wikipedia excess cross-entropy, $\Delta\mathrm{CE} = \ln(\mathrm{PPL}_{\mathrm{steered}}/\mathrm{PPL}_{\mathrm{base}})$, in nats/token on fixed Wikipedia text; the pathway comparison uses 500 held-out sequences. This cost is distinct from the generation perplexity reported in Table 4 and is deterministic for a fixed intervention and corpus.

Selection and uncertainty. For the pathway and carry-retuning comparisons, each reported operating point maximizes the mean joint truthfulness–informativeness score among three candidate coefficients, each evaluated with five seeds. Tables 22 and 23 report these selected means and across-seed standard deviations. Equal candidate counts give the same final selection budget, but do not guarantee equal selection bias. The selected means are unadjusted, and small differences do not establish statistical significance. The broader PSAS coefficient sweep in Table 24 identifies its five-seed and exploratory single-seed results separately.

Model	Pathway	T×I	τ	ΔCE
Qwen	Q-only	.345±.006	5.603	0.005
	V-only	.526±.010	4.002	0.043
	Q+V (PSAS)	.548±.015	4.002	0.044
	Q+K+V	.545±.004	5.603	0.049
	MLP-only	.542±.016	8.005	0.072
	Carry-free CAA	.471±.011	5.603	0.130
	CAA	.507±.012	2.401	0.168
Llama	Q-only	.460±.010	6.656	0.012
	V-only	.707±.015	32.000	0.117
	Q+V (PSAS)	.663±.012	8.000	0.063
	Q+K+V	.698±.004	4.659	0.109
	MLP-only	.454±.017	1.997	0.011
	Carry-free CAA	.579±.011	4.659	0.148
	CAA	.657±.013	3.328	0.241

Table 22: Selected pathway operating points on Qwen3-1.7B and Llama-3.1-8B. Joint scores are means ± sample standard deviations over seeds {42,43,44,45,47}. Each row reports the coefficient with the largest mean among three evaluated candidates and its Wikipedia excess cross-entropy. Unsteered joint scores are 0.323 and 0.427, respectively.

Pathway-dependent control. On Qwen, V-only captures much of the improvement over the unsteered model. Adding query steering raises the mean joint score by approximately 0.022, from 0.526 to 0.548, at nearly identical cost (0.043 versus 0.044). Q+K+V and MLP-only reach similar mean scores with higher costs of 0.049 and 0.072, respectively.

On Llama, V-only has the highest selected mean joint score, 0.707, at cost 0.117. Q+K+V reaches 0.698 at cost 0.109, while Q+V reaches 0.663 at the lower cost of 0.063. Thus, both alternatives improve the selected mean relative to Q+V while increasing language-modeling cost. Q-only and MLP-only provide smaller behavioral gains. These operating points illustrate how the preferred pathway depends on the model and the desired control–cost trade-off.

A separate five-seed comparison at matched coefficients also favors Llama V-only over Q+V in mean joint score, by 0.054 at $\tau = 9.984$ and 0.241 at $\tau = 13.311$. Matching the coefficient does not match language-modeling cost; these comparisons therefore complement the selected operating points without establishing a matched-cost frontier.

G.1 JOINT RETUNING OF CARRY AND STEERING STRENGTH

We evaluate $F(c,\tau) = F_{\mathrm{NW},\tau} + c\tau\mathbf{r}$, retaining CAA's steered branch computations while varying its direct residual carry. Here $c=0$ gives carry-free CAA and $c=1$ gives CAA. Unlike the fixed-coefficient comparison in Section 6.2, each carry fraction is retuned over steering strength using the selection procedure above. The direction, layer, and all-token coverage match the pathway sweep.

Model	Carry fraction c	Peak T×I (mean ± SD)	τ	ΔCE
Qwen	0	.4705±.0111	5.603	.130
	.25	.4715±.0058	3.202	.080
	.50	.5038±.0135	3.202	.127
	.75	.5070±.0134	2.401	.117
	1 (CAA)	.5073±.0124	2.401	.168
	PSAS reference	.5476±.0150	4.002	.044
Llama	0	.5793±.0109	4.659	.148
	.25	.5955±.0125	4.659	.178
	.50	.6055±.0140	4.659	.257
	.75	.6244±.0081	3.328	.153
	1 (CAA)	.6565±.0129	3.328	.241
	PSAS reference	.6634±.0122	8.000	.063

Table 23: Joint retuning of carry and steering strength. Each row reports the selected five-seed mean and standard deviation, its coefficient, and Wikipedia excess cross-entropy at that coefficient. PSAS is a separate pathway intervention. Shared conditions agree with Table 22 up to rounding.

PSAS exceeds the selected carry-free mean by 0.0771 on Qwen and 0.0841 on Llama, with lower excess cross-entropy (0.044 versus 0.130, and 0.063 versus 0.148, respectively). Restoring carry raises the selected mean joint score on both models, but every listed carry configuration has a lower mean score and higher cost than PSAS. Relative to full CAA, PSAS's mean score gains are 0.0403 on Qwen and 0.0069 on Llama.

PSAS's excess cross-entropies are approximately 26% of full CAA's respective costs of 0.168 and 0.241. These comparisons describe selected operating points and do not establish statistical significance for the mean differences. A complete matched-cost frontier would require comparing the coefficient-level results across cost budgets; another coefficient may yield a better trade-off at a particular budget.

G.2 Behavior around and beyond the operating coefficient

Table 24 evaluates PSAS across steering strengths, with operating coefficients $\tau_{\text{op}} = 4.002$ on Qwen and 8 on Llama. Several tested settings retain strong control with modest language-modeling cost. Qwen's five-seed mean joint scores range from 0.505 to 0.548 at 0.6, 0.8, 1.0, and 1.4× the operating coefficient. Llama's means range from 0.627 to 0.663 at 0.83, 1.0, 1.12, and 1.25×. All exceed their respective unsteered scores at these tested settings.

Model	τ/τ_{op}	T×I	True	Info	Wiki PPL	ΔCE
Qwen	0.10	0.355	0.900	0.453	21.29	0.003
	0.30	0.416	0.876	0.535	21.45	0.010
	0.40	0.460	0.884	0.567	21.54	0.014
	0.60	0.505	0.888	0.609	21.73	0.023
	0.80	0.532	0.903	0.623	21.95	0.033
	1.00	0.548	0.914	0.626	22.18	0.044
	1.40	0.530	0.927	0.595	22.70	0.066
	2.00	0.494	0.950	0.538	23.54	0.103
	3.00	0.399	0.973	0.420	25.15	0.169
	4.00	0.263	0.968	0.283	27.14	0.245
	8.00	0.143	0.953	0.162	37.8	0.58
	16.00	0.092	0.918	0.118	83	1.36
Llama	0.25	0.501	0.523	0.975	10.16	0.005
	0.42	0.551	0.584	0.964	10.24	0.013
	0.58	0.579	0.626	0.949	10.34	0.023
	0.83	0.633	0.723	0.908	10.57	0.044
	1.00	0.663	0.813	0.847	10.77	0.063
	1.12	0.636	0.849	0.785	10.92	0.077
	1.25	0.627	0.899	0.725	11.12	0.095
	1.66	0.454	0.960	0.490	11.87	0.160
	2.50	0.236	0.993	0.241	14.2	0.34
	16.00	0.078	0.94	0.09	200	3.0

Table 24: PSAS behavioral coefficient sweep. Unsteered T×I and Wikipedia perplexity are 0.323 and 21.24 on Qwen, and 0.427 and 10.11 on Llama. Joint scores shown with ± are five-seed means with reported across-seed dispersion; the remaining joint scores are exploratory single-seed measurements. Cost is $\Delta\text{CE} = \ln(\text{PPL}/\text{PPL}_0)$ in nats/token.

Increasing strength beyond the operating coefficient reduces informativeness while truthfulness remains high. Qwen's five-seed joint mean falls from 0.548 at 1× to 0.494 at 2×, and Llama's from 0.663 at 1× to 0.454 at 1.66×. At these latter settings, Wikipedia perplexity is approximately 1.11× and 1.17× the respective unsteered baselines. In the single-seed Qwen result at 4×, the joint score falls below baseline (0.263 versus 0.323), while perplexity reaches 1.28× baseline.

At extreme strength, language-modeling cost also rises substantially. At 16×, the joint scores are 0.092 on Qwen and 0.078 on Llama, with Wikipedia perplexities of 83 and 200, approximately 3.9× and 19.8× their unsteered values. The sweep therefore shows useful control near the operating coefficient, declining informativeness under moderate overshoot, and degradation of both behavioral quality and language modeling at extreme strength. Steering strength should be selected using behavioral quality together with language-modeling cost.

H Comparison with norm-constrained CAA

We compare `psas` with uncapped CAA and three norm-constrained CAA variants on Qwen3-14B at layer 24 on TruthfulQA. These controls constrain different quantities: the direct steering update, the residual carry, or the norm of the steered residual state. Table 25 reports behavioral control together with generation perplexity and teacher-forced Wikipedia perplexity relative to the unsteered model.

Interventions. Let h_t be the original residual state and τr the nominal steering update. The three constrained variants are:

- **CAA-clip**: Replace the update by

$$u_t^{\text{clip}} = \tau r \min\left(1, \frac{c\|h_t\|_2}{\|\tau r\|_2}\right),$$

  and use this clipped update throughout CAA, including its branch inputs and residual carry.

- **CAA-clip (carry only)**: Use u_t^{clip} for the direct residual carry while retaining the nominal update τr in CAA's attention and MLP computations.

- **CAA-norm**: Replace the steered state by

$$\frac{\|h_t\|_2}{\|h_t + \tau r\|_2}(h_t + \tau r),$$

  restoring the original residual norm while preserving the direction of the steered state.

Both clipping variants use $c = 0.3$ in the reported comparisons. The reported strength convention is $\tau = \rho\|h\|_2/\|r\|_2$, where $\|h\|_2$ is the original residual norm used to parameterize the coefficient.

Method	ρ	T×I ↑	Truth ↑	Info ↑	Gen PPL ↓	Wiki PPL ratio ↓
Uncapped and norm-preserving CAA						
CAA (uncapped)	0.5	0.558	0.919	0.630	22.3	1.65
CAA (uncapped)	1.0	0.044	**0.938**	0.061	63.4	27.16
CAA-norm	0.5	0.494	0.924	0.559	21.4	1.80
CAA-norm	1.0	0.021	0.928	0.029	34.6	66.73
Clipped CAA						
CAA-clip ($c=0.3$)	1.0	0.787	0.853	0.934	19.6	**1.04**
CAA-clip ($c=0.3$)	3.0	0.786	0.846	0.940	20.0	1.07
CAA-clip (carry only, $c=0.3$)	0.5	0.800	0.864	0.935	15.5	1.11
CAA-clip (carry only, $c=0.3$)	2.0	0.585	0.860	0.700	15.9	1.79
Pathway-selective steering						
PSAS	0.5	0.771	0.860	0.909	16.4	1.08
PSAS	1.5	0.845	0.908	0.934	13.9	1.20
PSAS	3.0	**0.862**	0.916	**0.945**	**13.6**	1.25

Table 25: **Comparison with norm-constrained CAA on Qwen3-14B, layer 24, on TruthfulQA.** Each row is a reported operating point. T×I is the joint truthfulness–informativeness rate. Wiki PPL ratio is teacher-forced Wikipedia perplexity divided by the unsteered value; Gen PPL is perplexity on generated responses. Bold marks the best listed value in each metric.

Norm preservation and carry clipping. Preserving the residual-state norm does not prevent degradation at the tested strengths. At $\rho = 1$, CAA-norm has a joint score of 0.021 and a Wikipedia-perplexity ratio of 66.73, compared with 0.044 and 27.16 for uncapped CAA. Carry-only clipping attains 0.800 at $1.11\times$ perplexity for $\rho = 0.5$, but its score falls to 0.585 and its perplexity rises to $1.79\times$ at $\rho = 2$.

Behavioral control and language-modeling cost. Full-update clipping achieves joint scores of 0.787 and 0.786 at Wikipedia-perplexity ratios of 1.04 and 1.07. PSAS reaches higher scores of 0.845 and 0.862 at higher ratios of 1.20 and 1.25. Thus, full clipping provides lower-cost operating points, while PSAS achieves the highest listed joint score and lowest generation perplexity. These selected settings illustrate different control–cost trade-offs; they do not match delivered perturbation magnitude, establish a general clipping ceiling, or determine an optimized trade-off frontier.

I Comparison with Post-Projection Query–Value Steering

We compare `psas` with post-projection query–value steering along two dimensions: the response to increasing steering strength and the dependence of value perturbations on token state. We first define the interventions, then report a coefficient-scaling diagnostic and value-replacement experiments. A final subsection provides supplementary evidence about direction estimation and strength coupling.

I.1 Intervention definitions

`psas` and DISCO-QV (Torop et al., 2025) both steer the Q/V pathways, but apply their steering signals at different points. Suppressing head indices, their interventions are

$$\textbf{PSAS:}\quad \tilde{\mathbf{x}}_l = \mathrm{RMSNorm}(\mathbf{h}_l + \tau\mathbf{r}) \longrightarrow \begin{cases} \tilde{\mathbf{Q}}_l = \tilde{\mathbf{x}}_l \mathbf{W}^Q, \\ \tilde{\mathbf{V}}_l = \tilde{\mathbf{x}}_l \mathbf{W}^V, \end{cases}$$

$$\mathbf{x}_l = \mathrm{RMSNorm}(\mathbf{h}_l) \longrightarrow \begin{cases} \mathbf{K}_l = \mathbf{x}_l \mathbf{W}^K, \\ \text{identity path carries } \mathbf{h}_l; \end{cases}$$

$$\textbf{DISCO-QV:}\quad \mathbf{q} \leftarrow \mathbf{q} + \alpha_q \boldsymbol{\delta}_q, \qquad \mathbf{v} \leftarrow \mathbf{v} + \alpha_v \boldsymbol{\delta}_v, \qquad \mathbf{k} \text{ unchanged.}$$

The projections precede the model's query/key transformations, including RoPE and, where present, QK-Norm, as in Section 3.

PSAS adds a residual-space direction before normalization and projection. Its induced value perturbation therefore depends on the original token state and remains bounded as the coefficient grows. DISCO-QV adds a fixed vector at each value-projection output; this shift grows linearly with $|\alpha_v|$ when $\boldsymbol{\delta}_v \neq 0$. Its attention-output contribution is likewise unbounded when $\boldsymbol{\delta}_v \mathbf{W}^O \neq 0$. The following studies examine coefficient scaling and the contribution of token–state correspondence separately.

I.2 Response to coefficient growth

This diagnostic uses Qwen3-1.7B and Llama-3.1-8B in fp32, with interventions at zero-based blocks 16 and 15, respectively. Directions are estimated from the training half of the TruthfulQA split, using the same contrast pairs as the residual direction $\mathbf{r}$. Evaluation uses 500 fixed Wikipedia sequences truncated to 128 tokens, following Appendix F. Residual-change norms are computed per token and averaged over the sample.

The post-projection directions are $\boldsymbol{\delta}_q = (\overline{\mathrm{RMSNorm}(\mathbf{h}^+)} - \overline{\mathrm{RMSNorm}(\mathbf{h}^-)})\mathbf{W}^Q$ and $\boldsymbol{\delta}_v = (\overline{\mathrm{RMSNorm}(\mathbf{h}^+)} - \overline{\mathrm{RMSNorm}(\mathbf{h}^-)})\mathbf{W}^V$. The DISCO-QV operating point is $\alpha_q = 2.0$, $\alpha_v = 0.3$; DISCO's per-head discriminability selection is not included. PSAS's operating coefficients are 4.002 on Qwen3-1.7B and 8.0 on Llama-3.1-8B. Table 26 reports the full-layer residual change entering the next block and teacher-forced Wikipedia perplexity at multiples of each method's own operating coefficient.

Residual change entering the next block ×operating	0.25	0.5	1	2	4	8	16
Qwen3-1.7B, DISCO-QV	3.46	6.88	13.53	25.78	46.19	80.93	148.3
Qwen3-1.7B, `psas`	13.06	25.70	48.75	85.07	135.3	185.5	207.6
Llama-3.1-8B, DISCO-QV	0.19	0.38	0.75	1.43	2.60	4.70	8.80
Llama-3.1-8B, `psas`	1.10	1.97	3.42	5.69	7.80	9.10	10.00
Wikipedia perplexity ×operating	0.25	0.5	1	2	4	8	16
Qwen3-1.7B, DISCO-QV	21.42	21.44	21.47	21.55	21.84	23.07	29.70
Qwen3-1.7B, `psas`	21.63	21.89	22.51	23.95	27.64	37.31	46.29
Llama-3.1-8B, DISCO-QV	10.00	10.01	10.03	10.11	10.41	11.58	18.43
Llama-3.1-8B, `psas`	10.06	10.21	10.75	12.62	17.06	22.60	28.29

Table 26: Coefficient-scaling diagnostic for DISCO-QV and PSAS. Strengths are multiples of each method's own operating coefficient. The PSAS full-layer perturbation approaches saturation over the sweep, while the post-projection intervention continues to grow.

Table 26 shows distinct responses to increasing steering strength: PSAS's delivered perturbation approaches saturation over the evaluated range, whereas the post-projection intervention continues to grow. These diagnostics establish different responses to coefficient growth; they do not establish lower language-modeling cost at equal perturbation magnitude. We next examine a complementary difference: PSAS's value perturbations depend on the token state. We measure this variation and evaluate how replacing the original perturbations with constant or shuffled alternatives changes behavioral control and off-target language modeling.

I.3 TOKEN-DEPENDENT VALUE MODULATION

We first measure variation in the implemented value perturbations. Table 27 reports nonzero, token-varying PSAS perturbations: both their magnitudes and directions vary across positions. For a fixed layer, head, and steering strength, the DISCO-QV value shift is constant across positions by construction.

Method	**Mean** $\|\Delta v_t\|$	**CV** ($\|\Delta v_t\|$)	**Pairwise Cos Sim** (↓)	**Vector Variance** (↑)
DISCO-QV	Const.	0.000	1.000	0.000
PSAS (Ours)	**56.01**	**0.269**	**0.829**	**0.171**

Table 27: **Across-token statistics of value perturbations for DISCO-QV and PSAS.** For a fixed layer, attention head, and steering strength, DISCO-QV applies the same post-projection value shift at every token position, so its across-token cosine similarity is 1 and variance is 0 by construction. PSAS produces token-dependent perturbations: the norm coefficient of variation is 0.269, and the mean pairwise cosine similarity is 0.829, indicating that the perturbations vary across tokens while remaining directionally aligned on average.

To examine the contribution of token–state correspondence, we compare the original PSAS value perturbations with two replacement interventions:

- `psas`-V_{actual}: Applies the original token-specific perturbation Δv_t at each position t.
- `psas`-V_{constant}: Replaces token-specific shifts with a single constant vector $\overline{\Delta v} = \frac{1}{N}\sum_{t=1}^{N} \Delta v_t$, computed as the elementwise mean of all token-level Δv_t extracted across the evaluation fold's prompts.
- `psas`-V_{shuffled}: Randomly assigns a vector sampled from the empirical Δv_t pool to each token position, without conditioning on the current token state h_t.

Table 29 reports the replacements under V-only and Q+V steering. Replacing token-specific perturbations with their mean or sampled alternatives lowers the joint truthfulness–informativeness score and increases Wikipedia perplexity in both configurations. These results support a contribution from token–state correspondence at the evaluated settings. Averaging also changes perturbation magnitudes and variation, while reassignment changes alignment with recipient states; the experiments therefore establish sensitivity to these replacements rather than the necessity of token dependence in general.

I.4 DIRECTION ESTIMATION AND STRENGTH COUPLING

The following analyses provide supplementary evidence about configuration choices. PSAS derives its Q/V changes from a shared residual-space direction. For identical contrast pairs at linear projection sites, the DISCO-QV mean-difference directions are also projections of a shared normalized-input mean difference, as in the coefficient-scaling setup above. The distinction concerns the source representation and normalization placement; shared-source construction alone does not establish an advantage.

Linear discriminability. We compare residual-space and attention-space mean-difference classifiers on Qwen3-14B. TruthfulQA is split into train, validation, and test sets with a 50/25/25 ratio. We fit classifiers on the training split, select candidate layers or heads using validation balanced accuracy, and evaluate the selected candidates on the held-out test split. This analysis is readout-only and involves no steering or generation.

Table 30 reports the highest validation and test accuracy for the residual representation. Q-head representations exceed their matched head-output comparator on both splits, whereas V-head representations fall below their matched comparator on both. These readout results provide a design rationale for estimating a residual-space direction, but do not establish that shared Q/V steering directions outperform separately estimated ones.

Strength coupling. We compare the tied-strength configuration, $\tau_Q = \tau_V$, with a variant in which Q and V strengths are selected independently while keeping the directions shared. The two configurations achieve joint truthfulness–informativeness scores of 0.862 with tied strengths and 0.852 with independently

Value perturbation	V-only steering T×I ↑	V-only steering Wiki PPL ↓	Q+V steering T×I ↑	Q+V steering Wiki PPL ↓
Original PSAS	**0.585**	**14.82**	**0.865**	**17.28**
Constant	0.330	17.63	0.180	17.74
Shuffled	0.180	18.36	0.230	18.40

Table 28: Effect of replacing PSAS's token-dependent value perturbations with their mean (Constant) or samples from the empirical perturbation pool (Shuffled). Bold marks the highest joint TruthfulQA score and lowest teacher-forced Wikipedia perplexity within each configuration.

	V-only Steering			$Q+V$ Joint Steering		
Condition	**TQA** ($T\times I$) ↑	**Gen. PPL** ↓	**Wiki PPL** ↓	**TQA** ($T\times I$) ↑	**Gen. PPL** ↓	**Wiki PPL** ↓
PSAS-V_{actual} (Ours)	**0.585**	21.2	**14.82** (×1.07)	**0.865**	**13.3**	**17.28** (×1.25)
PSAS-V_{constant}	0.330	**12.9**	17.63 (×1.27)	0.180	14.3	17.74 (×1.28)
PSAS-V_{shuffled}	0.180	18.5	18.36 (×1.32)	0.230	11.7	18.40 (×1.33)

Table 29: **Ablation Study on Token-Dependent Value Modulation.** We evaluate TruthfulQA performance ($T\times I$ score / Generation PPL) and language modeling fluency on Wikipedia (PPL with ratio relative to base) across V-only and $Q+V$ joint steering configurations.

Pathway / Subspace	**Val Acc.**	**Test**
Top-k Q head ($k=768$)	0.700	0.703
Top-k V head ($k=176$)	0.681	0.697
Top-k Head output ($k=768$, matched to Q)	0.637	0.640
Top-k Head output ($k=176$, matched to V)	0.707	0.705
Residual	**0.768**	**0.792**

Table 30: **Linear Discriminability of Residual and Attention Representations (Qwen3-14B, TruthfulQA).** Mean-difference classifier performance (balanced accuracy) across validation (25%) and test (25%) splits. Following the setup in DISCO (Torop et al., 2025), the number of top-k heads ($k=768$ for Q and $k=176$ for V) is chosen identically to match their analysis.

selected strengths. We use tied strengths in the main configuration for simplicity. This comparison concerns strength tying, not the use of shared versus separately estimated Q/V directions.