WHEN AND WHY ACTIVATION STEERING DEGRADES LONG GENERATIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering offers a computationally efficient mechanism for steering large language model output during inference. It has been observed that continuous residual-stream interventions, especially during long-form autoregressive decoding, induce semantic collapse and cause coherence loss. Although there have been recent attempts to mitigate this issue, it remains unclear why activation steering is robust over short generations but degrades rapidly during long-form generation. Existing analyses of activation steering primarily characterize its effects at individual generation steps, leaving their dependence on the evolving context over the entire generation trajectory insufficiently understood. We develop a theoretical framework that casts autoregressive generation with steering as a stochastic recursion for concept and degradation scores of the generation. At each step, steering modifies the next-token distribution, and the sampled token updates the context scores that govern subsequent expected token scores. A mean-field analysis characterizes score trajectory saturation and the dependence of degradation onset on steering strength. We describe how to estimate the framework's quantities from model activations and evaluate its predictions across eight instruction-tuned language models with 3B to 14B parameters.

1 INTRODUCTION

Activation steering provides a mechanism for controlling large language models (LLMs) through direct intervention in their internal representations. A steering vector identifies a direction associated with a target behavioral attribute, and adding a scaled version of this vector to the residual stream modifies the model's predictions during inference (Turner et al., 2023; Li et al., 2023; Zou et al., 2023; Rimsky et al., 2024). Because the intervention leaves the model's parameters fixed, it offers a computationally efficient means of behavioral control, with applications to truthfulness, refusal, and personality traits (Li et al., 2023; Arditi et al., 2024; Stolfo et al., 2025; Chen et al., 2025).

The reliability of this intervention depends on the strength of steering and length of the generated response. A common selection rule chooses the largest coefficient that preserves coherence on evaluation responses (Chen et al., 2025), yet sustained steering can degrade coherence in longer generations (Kang et al., 2026; Herbster et al., 2026). Understanding the dependency between *steering strength* and *generation length* requires accounting for the evolving context. As illustrated in Figure 1a, each intervention changes the next-token distribution, and the sampled token extends the context for subsequent generation. Later interventions therefore act on contexts shaped by earlier steering, so their effect can change over the response. Existing analyses largely characterize a single generation step at a fixed context (Singh et al., 2024; Im & Li, 2025; Taimeskhanov et al., 2026), leaving open how generated text changes the effect of subsequent interventions.

To bridge this gap, we formulate autoregressive generation with steering through two scores of the evolving context. A *concept score* measures alignment with the target attribute or behavior, while a *degradation score* provides an activation-based proxy for coherence loss. Both context scores are weighted averages of the corresponding token scores. We model the expected next-token scores as functions of the context scores and steering coefficient. The sampled token then updates the context scores, yielding a stochastic recursion.

We characterize score saturation in a deterministic approximation of this stochastic recursion and identify a critical steering coefficient separating trajectories that remain below the degradation

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

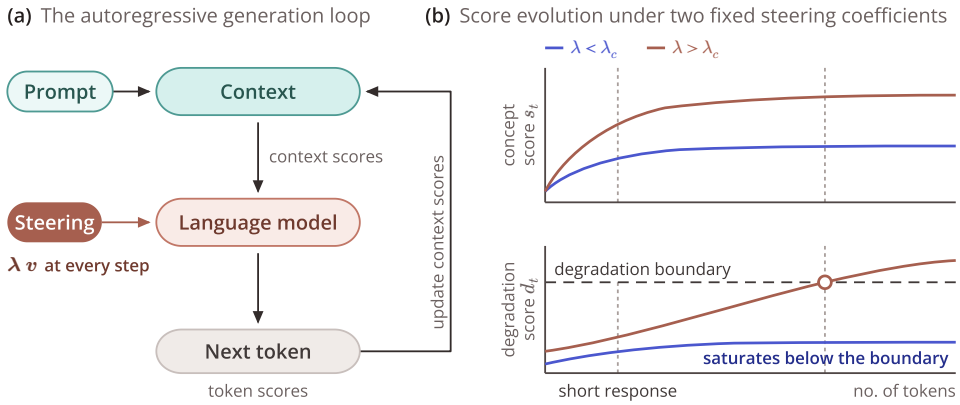


Figure 1: Degradation under constant steering strength. (a) Steering alters next-token probabilities, and sampled tokens update context scores. (b) Schematic trajectories show saturation below the degradation boundary ($\lambda < \lambda_c$), and delayed crossing after a critical steering coefficient λ_c ($\lambda > \lambda_c$).

boundary from those that cross it. A coefficient just above this critical value can keep the degradation score below the boundary over a short response yet produce a crossing as generation continues (Figure 1b). We also bound the stochastic recursion's deviation from this approximation over finite horizons.

Connecting these dynamics to transformers also requires accounting for effects of earlier interventions retained in the key–value (KV) cache. During decoding, transformers reuse keys and values computed from steered states, allowing earlier interventions to influence subsequent generation through internal activations. This persistence has been proposed as a mechanism for cumulative coherence degradation (Kang et al., 2026). We prove that the KV-cache contribution to the residual stream is bounded uniformly in sequence length, excluding consistent accumulation. We further demonstrate it empirically by isolating this contribution by holding the token sequence and current intervention fixed while changing whether earlier positions were steered.

To evaluate the framework empirically, we estimate expected next-token scores, context weights, and reference thresholds from model activations. These measurements determine score trajectories and threshold-crossing times without fitting to coherence judgments or observed crossing times. We evaluate the framework across eight models with 3B–14B parameters, seven persona traits, and three steering layers per model. The estimated dynamics track observed score trajectories and degradation-threshold crossings associated with declines in judged coherence. These results support evaluating steering coefficients at the intended response length while accounting for token-sampling randomness.

Contributions. This paper makes the following main contributions:

(i) We develop a stochastic model of autoregressive generation under steering and bound its deviation from a deterministic approximation over finite horizons. The resulting saturation and threshold-crossing conditions characterize how generation length constrains steering strength.

(ii) We establish a uniform bound on the residual-stream contribution of steered KV-cache states under the stated transformer assumptions. By holding the token sequence fixed, our analysis and empirical comparison isolate the KV-cache contribution while controlling for changes in the generated text.

(iii) We provide a measurement methodology that connects the analytical quantities to transformer activations. We evaluate the resulting trajectory and crossing-time predictions against generated sequences and judged coherence, assessing both their predictive value and departures from the modeling assumptions.

(iv) Our analysis supports coefficient selection for a target task and response length. The measured dynamics allow practitioners to estimate concept alignment and degradation onset across candidate steering coefficients, guiding the choice of steering strength.

Related work. Taimeskhanov et al. (2026) derive an exponential reweighting of next-token probabilities, which we use in our analysis at a fixed context. Bigelow et al. (2025) model the joint effects

of steering and evidence supplied through in-context examples, while Nguyen et al. (2026) study the stability of activation dynamics across layers. Our framework follows how sampled tokens update the context under repeated steering with a constant coefficient. The resulting stochastic recursion characterizes score saturation and degradation-threshold crossings as a function of steering strength and generation length.

Empirical studies examine steering reliability across tasks (Tan et al., 2024; Braun et al., 2025) and under adversarial perturbations (Le & Le, 2026). Adaptive interventions seek to preserve output quality through projection-aware corrections (Herbster et al., 2026), context-dependent scaling (Ferrando et al., 2025), and custom attention gating (Kang et al., 2026). These methods complement our characterization of generation under a constant coefficient. Kang et al. (2026) attribute long-generation degradation to KV-cache contamination, while Belitsky et al. (2025) propose one-shot cache steering instead of intervention at every step to avoid amplification from repeated interventions. Xu et al. (2026) exploit amplification along the autoregressive trajectory to induce progressive behavioral drift.

2 Theoretical Framework

A steering intervention operates on a context that evolves with each generated token. We analyze this interaction through two scalar context scores, focusing on how repeated intervention alters their trajectories. The framework separates the expected scores of the next token from that token's contribution to subsequent context, resulting in a stochastic recursion. Its deterministic approximation characterizes both saturation and the onset of degradation. The proofs are deferred to Appendix A.

2.1 Notation and setup

Let $\mathcal{V}$ denote a finite vocabulary and x a prompt. At decoding step $t \geq 1$, a language model generates $y_t \in \mathcal{V}$ conditional on the context $c_t := (x, y_{1:t-1})$. We consider a decoder-only transformer with L layers and write $\boldsymbol{h}_i^{(\ell)} \in \mathbb{R}^d$ for the residual vector entering layer ℓ at sequence position i.

We study difference-of-means steering at a fixed layer ℓ (Rimsky et al., 2024; Zou et al., 2023; Chen et al., 2025). For a target behavioral attribute, hereafter the *concept*, let $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ be mean residual vectors obtained from unsteered forward passes over responses exhibiting the concept and matched contrast responses, respectively. The steering vector is $\boldsymbol{v} := \boldsymbol{\mu}_+ - \boldsymbol{\mu}_- \neq \mathbf{0}$, and the intervention applies

$$\boldsymbol{h}_i^{(\ell)} \leftarrow \boldsymbol{h}_i^{(\ell)} + \lambda \boldsymbol{v} \tag{1}$$

at every processed position, with $\lambda \geq 0$ held constant throughout generation. We denote the induced next-token distribution by $p_\lambda(\cdot \mid c)$, so that $y_t \sim p_\lambda(\cdot \mid c_t)$ and p_0 denotes the unsteered distribution. During cached decoding, keys and values computed at earlier positions retain the effects of the intervention.

The effect of steering at a fixed context has an explicit form in the unconstrained-features model, as applied to next-token prediction by Zhao et al. (2024). This model assigns each context c an independently parameterized feature vector $\boldsymbol{h}_c \in \mathbb{R}^d$ and produces logits $\boldsymbol{W}\boldsymbol{h}_c$, where $\boldsymbol{W} \in \mathbb{R}^{|\mathcal{V}| \times d}$. Under the idealization that training recovers the target conditional distributions, Taimeskhanov et al. (2026) show that difference-of-means steering produces an exponential tilt:

$$p_\lambda(y \mid c) = \frac{p_0(y \mid c) e^{\lambda r(y)}}{\sum_{y' \in \mathcal{V}} p_0(y' \mid c) e^{\lambda r(y')}}. \tag{2}$$

Here $r(y)$ is the difference between the concept-context and contrast-context averages of $\log p_0(y \mid c')$. Thus, steering adds a context-independent logit perturbation $\lambda r(y)$, while normalization depends on the baseline distribution at the current context.

The fixed-context calculation leaves open how the intervention affects the evolving contexts later in generation. Because each sampled token changes the context, it also changes the baseline distribution to which the next intervention is applied. Our analysis therefore addresses the following question: ***How does the evolving context affect concept alignment and the onset of degradation under a constant steering coefficient?***

2.2 A STOCHASTIC MODEL OF AUTOREGRESSIVE FEEDBACK

To study generation under constant steering, we summarize the evolving context with a concept score and a degradation score. We define these two scores for individual tokens and use their context averages to formulate the stochastic recursion in Definition 1, illustrated in Figure 2.

For a token y following a context c, let $\boldsymbol{h}_y$ denote its residual vector entering layer ℓ, evaluated by an unsteered forward pass over (c, y). The *concept score* $s(y) := \langle \boldsymbol{h}_y - \boldsymbol{\mu}_-, \boldsymbol{v} \rangle / \|\boldsymbol{v}\|^2$ measures alignment with the steering direction relative to the contrast mean. Its average is 0 on the contrast responses and 1 on the concept responses used to construct $\boldsymbol{v}$. Although the notation suppresses c, the score depends on the token's preceding context.

For the *degradation score* $d(y)$, we project $\boldsymbol{h}_y$ onto a direction derived from the difference between the mean residual vectors of high-temperature and ordinary unsteered responses. We calibrate the score to have reference means 0 on ordinary responses and 1 on high-temperature responses. Its correspondence to coherence loss is assessed empirically. Appendix B specifies the direction construction and calibration.

The context scores $s(c)$ and $d(c)$ are weighted averages of token scores. We first specify how they determine the expected next-token scores, then define their update after sampling a token.

Assumption 1 (Expected next-token scores under steering) *For each score $z \in \{s, d\}$, there exist functions $m_z : \mathbb{R} \to \mathbb{R}$ and $g_z : [0, \infty) \to [0, \infty)$ with $g_z(0) = 0$ such that, for every context c and steering coefficient $\lambda \geq 0$,*

$$\mathbb{E}_{y \sim p_\lambda(\cdot|c)}[z(y)] = m_z\big(z(c) + g_z(\lambda)\big).$$

The function m_z is non-decreasing and locally Lipschitz, while g_z is continuous and non-decreasing.

Assumption 1 is motivated by autoregressive feedback, in which concept-aligned or degraded context can encourage subsequent tokens with similar characteristics (Xu et al., 2022; Yan et al., 2024). It models each expected next-token score through the corresponding context score, with steering represented as an *equivalent increase in context score* $g_z(\lambda)$ at the input to the unsteered response function m_z. Monotonicity encodes the hypothesized reinforcement of these characteristics by context and steering, while the regularity conditions support trajectory comparisons and approximation bounds without imposing a specific parametric form.

Empirical studies show that language models rely more strongly on recent context (Vakilian et al., 2025; Khandelwal et al., 2018). Motivated by this observation, we represent each context score as a running summary that combines its previous value with the newly generated token's score, with the update weight a_t controlling the balance between responsiveness to new content and persistence of earlier information. For constant $a_t = a$, this update assigns exponentially decaying weights to earlier token scores, providing a tractable approximation of recency effects. The following assumption formalizes this update.

Assumption 2 (Weighted context update) *There exist deterministic weights $a_t \in (0, 1)$ such that, as the context evolves from c_t to $c_{t+1} = (c_t, y_t)$,*

$$z(c_{t+1}) = (1 - a_t)\, z(c_t) + a_t\, z(y_t), \qquad z \in \{s, d\}.$$

This assumption directly leads to the following recursion for the concept and degradation scores.

Definition 1 (Autoregressive score dynamics) *Under Assumption 2, define $s_t := s(c_t)$ and $d_t := d(c_t)$, initialized at $s_1 = s(x)$ and $d_1 = d(x)$. For a fixed steering coefficient λ, token generation and the context scores evolve according to*

$$y_t \sim p_\lambda(\cdot \mid c_t), \qquad s_{t+1} = (1 - a_t)\, s_t + a_t\, s(y_t), \qquad d_{t+1} = (1 - a_t)\, d_t + a_t\, d(y_t). \quad (3)$$

Steering affects the expected next-token scores through $g_s(\lambda)$ and $g_d(\lambda)$, and the sampled token updates the context scores used at the next step (Figure 2). A constant coefficient therefore acts on an evolving context. Under Assumption 1, the conditional mean of each context-score increment depends on the context only through the corresponding score, enabling the mean-field analysis in Section 2.3.

To examine how g_s and g_d vary with steering strength, we relate them to the weak-steering expansion of expected next-token scores under Eq. 2 as stated below.

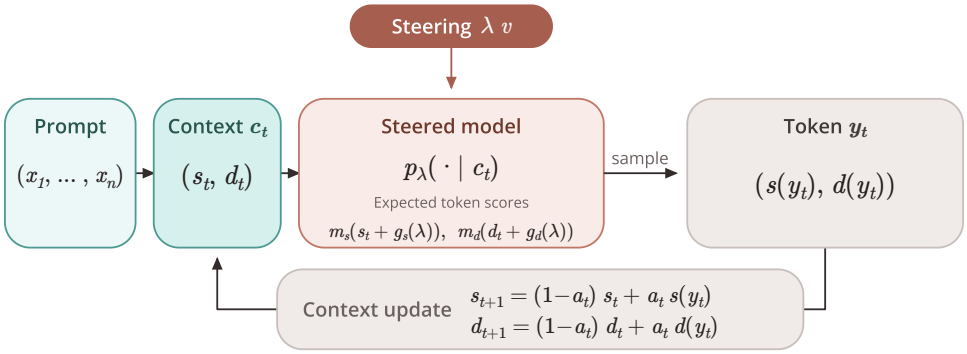


Figure 2: Score dynamics under constant steering. The context scores and steering coefficient determine the expected token scores, and the sampled token's scores update the context with weight a_t.

Proposition 1 (Expected next-token scores under weak steering) *Fix a context c whose steered law satisfies Eq. 2, and suppose that $\mathrm{Cov}_{y\sim p_0(\cdot|c)}\big(d(y), r(y)\big) = 0$. Then, as $\lambda \to 0$,*

$$\mathbb{E}_{y\sim p_\lambda(\cdot|c)}[s(y)] - \mathbb{E}_{y\sim p_0(\cdot|c)}[s(y)] = \lambda\, \mathrm{Cov}_{y\sim p_0(\cdot|c)}\big(s(y), r(y)\big) + O(\lambda^2),$$
$$\mathbb{E}_{y\sim p_\lambda(\cdot|c)}[d(y)] - \mathbb{E}_{y\sim p_0(\cdot|c)}[d(y)] = \tfrac{\lambda^2}{2}\, \mathrm{Cov}_{y\sim p_0(\cdot|c)}\big(d(y), r(y)^2\big) + O(\lambda^3).$$

The zero-covariance condition removes the first-order change in expected degradation score. When the leading coefficients are positive, the expected concept and degradation scores increase linearly and quadratically near zero, respectively. These orders also hold for g_s and g_d under the regularity conditions stated in Assumption 1. The detailed proof is provided in Appendix A.1.

2.3 Mean-field dynamics and saturation

We use a deterministic approximation of the stochastic recursion to characterize *how the context scores evolve* and *where they saturate*. Write z_t for either context score and abbreviate $g_z := g_z(\lambda)$ at a fixed steering coefficient. Taking the conditional expectation of the weighted update in Eq. 3 and applying Assumption 1 gives

$$\mathbb{E}[z_{t+1} - z_t \mid c_t] = a_t\big(\mathbb{E}[z(y_t) \mid c_t] - z_t\big) = a_t\big(m_z(z_t + g_z) - z_t\big) = a_t F(z_t). \tag{4}$$

The drift $F(z) = m_z(z + g_z) - z$ is the expected difference between the next-token score and the current context score. Writing $f_z(q) := m_z(q) - q$ for this difference without steering gives $F(z) = g_z + f_z(z + g_z)$. Thus, steering sustains growth while $g_z > -f_z(z + g_z)$ and their balance determines the mean-field equilibrium.

We account for the update weights through the *intrinsic time* $\tau_t := \sum_{r<t} a_r$, which equals $a(t-1)$ for constant weights. The *mean-field trajectory* solves

$$\frac{d\bar{z}}{d\tau} = F(\bar{z}), \qquad \bar{z}(0) = z_1. \tag{5}$$

When F is non-linear, this deterministic approximation need not coincide with the expected context score $\mathbb{E}[z_t]$. The following theorem establishes convergence of the mean-field trajectory to an equilibrium and bounds its distance from the stochastic trajectory over finite generations.

Theorem 1 (Finite-horizon approximation and saturation) *Let z_t follow Definition 1 and satisfy Assumption 1. Suppose all token scores $z(y_t)$ and the initial context score z_1 lie in a compact interval $\mathcal{I}$, f_z is Lipschitz on $\mathcal{I} + g_z$ with constant K, and $m_z(q) \in \mathcal{I}$ for every $q \in \mathcal{I} + g_z$. Then:*

(i) For every $t \geq 1$ and $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\max_{r\leq t} |z_r - \bar{z}(\tau_r)| \leq C e^{K\tau_t} \left(\sqrt{\log(2/\delta) \sum_{r<t} a_r^2} + \sum_{r<t} a_r^2\right),$$

where C depends only on K and the diameter of $\mathcal{I}$.

(ii) Both z_t and $\bar{z}$ remain in $\mathcal{I}$. The mean-field trajectory is monotone and converges as $\tau \to \infty$ to an equilibrium score z_∞ satisfying $F(z_\infty) = 0$.

Proof sketch. We decompose the stochastic recursion into conditional drift and martingale noise. A maximal Azuma–Hoeffding inequality Boucheron et al. (2013) controls cumulative sampling fluctuations despite autoregressive dependence, while Lipschitz continuity gives an $\mathcal{O}(a_t^2)$ local Euler discretization error. The two error terms reflect sampling and discretization, respectively. The discrete Grönwall inequality Clark (1987) controls the propagation of both errors, yielding the factor $e^{K\tau_t}$ in the uniform finite-horizon bound. For saturation, interval invariance and uniqueness of the scalar mean-field equation establish monotone convergence to an equilibrium. The detailed proof is provided in Appendix A.2.

Theorem 1 implies that at saturation, the expected next-token score equals the context score, giving $m_z(z_\infty + g_z) = z_\infty$. For a fixed initial score, increasing the steering coefficient cannot lower the mean-field trajectory or its equilibrium because m_z and g_z are non-decreasing. The update weights determine how quickly generation advances along this trajectory, with $\tau \to \infty$ corresponding to unbounded generation length when $\sum_t a_t = \infty$.

2.4 Degradation thresholds and exit times

A mean-field trajectory can cross the degradation boundary before reaching its equilibrium. We therefore characterize whether a crossing occurs and how its timing depends on the steering coefficient. For each score, let $z_{\max}$ denote an upper boundary calibrated on natural reference text (Section 3.2). Because g_z represents an equivalent increase in context score, we compare $z_t + g_z$ with this boundary. We define exit from the *natural range* as the first step at which $z_t + g_z \geq z_{\max}$.

Crossing the concept boundary indicates unusually strong concept alignment. Crossing the degradation boundary defines the framework's degradation event, whose correspondence to coherence loss is evaluated empirically. The following theorem characterizes exit for either score.

Theorem 2 (Exit criterion and exit time) *Under the hypotheses of Theorem 1, let* $b := z_{\max} - g_z > z_1$*, with* $[z_1, b] \subseteq \mathcal{I}$*. Then the follwings hold true:*

(i) If $F > 0$ *throughout* $[z_1, b]$*, the mean-field trajectory reaches* b *at finite intrinsic time*

$$\tau_{\text{exit}} = \int_{z_1}^{z_{\max}-g_z} \frac{dz}{F(z)}. \tag{6}$$

(ii) Otherwise, the trajectory does not reach b *in finite time, and* $z_\infty + g_z \leq z_{\max}$*, with strict inequality if* $F < 0$ *anywhere on* $[z_1, b]$*.*

Proof sketch. Continuity and strict positivity of the drift on the compact interval $[z_1, b]$ give a uniform positive lower bound, guaranteeing finite-time exit. Separation of variables then yields the exit-time integral in Eq. 6. Conversely, monotonicity and uniqueness prevent finite-time crossing when the initial drift is non-positive or an equilibrium lies along the trajectory's path, including at the boundary itself. The detailed proof is provided in Appendix A.3.

The factor $1/F(z)$ in Eq. 6 shows how regions of small positive drift can delay crossing even when exit is inevitable. With constant update weight a, the mean-field crossing occurs after approximately τ_{exit}/a generated tokens from initialization. Increasing the steering coefficient cannot decrease the drift F or raise the context-score boundary $z_{\max} - g_z(\lambda)$. Once finite-time exit occurs, it also holds for all larger coefficients. If $z_1 + g_z(\lambda) \geq z_{\max}$, exit is immediate and we set $\tau_{\text{exit}}(\lambda) = 0$. The following corollary identifies a critical coefficient separating finite-time exit from non-exit and characterizes exit behavior near this threshold.

Corollary 1 (Critical steering coefficient) *Fix* $z_1 < z_{\max}$*. Suppose the hypotheses of Theorem 2 hold at every coefficient for which the initial score lies below the exit boundary,* $z_1 + g_z(\lambda) < z_{\max}$*. Assume that the mean-field trajectory never exits at one coefficient and exits in finite positive time at another. Then there is a finite* critical coefficient λ_c *such that*

$$\tau_{\text{exit}}(\lambda) = \infty \quad \textit{for } \lambda \leq \lambda_c, \qquad \tau_{\text{exit}}(\lambda) < \infty \quad \textit{for } \lambda > \lambda_c.$$

For $\lambda > \lambda_c$*, the exit time is non-increasing in* λ *and tends to infinity as* $\lambda \downarrow \lambda_c$*. Among coefficients with positive finite exit times, a strict increase in* g_z *gives a strict decrease in exit time.*

For the degradation score, a coefficient just above λ_c can keep the mean-field trajectory below the boundary over a short response yet produce a crossing over a longer one. Token-sampling fluctuations can also produce crossings below λ_c. The proof is provided in Appendix A.4.

3 FROM THE ANALYTICAL MODEL TO TRANSFORMERS

The analysis in Section 2 characterizes saturation and degradation onset through the evolution of context scores. To apply these results to transformers, we first examine how earlier interventions affect later generation through the KV cache. We then measure the score dynamics and reference boundaries needed to estimate trajectories and exit times (Section 3.2).

3.1 THE DIRECT KV-CACHE CONTRIBUTION

The KV cache retains keys and values computed at earlier positions (Pope et al., 2023). Steering can modify these cached states, allowing earlier interventions to affect the next-token distribution even when the token sequence is held fixed. This persistence has been proposed as a possible source of degradation through *KV-cache contamination* (Kang et al., 2026; Belitsky et al., 2025). We examine whether the contribution of these earlier interventions to the residual stream can grow without bound as generation length increases.

To isolate the cache contribution, we compare steering at every position with steering only at the current position on the same token sequence. For a sequence of length T, let $\boldsymbol{h}_T^{(L+1)}(\mathcal{S})$ denote the final residual vector when steering is applied at positions $\mathcal{S} \subseteq [T] := \{1, \ldots, T\}$. Comparing $\mathcal{S} = [T]$ with $\mathcal{S} = \{T\}$ keeps the tokens and the intervention at the current position identical, so the residual difference measures the contribution of earlier interventions through cached states.

Proposition 2 (Uniform KV-cache bound) *Consider a transformer with normalization before each attention and feed-forward block. Assume uniformly bounded normalized activations, fixed finite-norm value and output projections, and continuous feed-forward blocks. There is a model-dependent constant B such that, for every sequence, position T, and coefficient $\lambda \in \mathbb{R}$,*

$$\|\boldsymbol{h}_T^{(L+1)}([T]) - \boldsymbol{h}_T^{(L+1)}(\{T\})\|_2^2 \leq B.$$

Bounded normalized activations yield bounded value vectors and feed-forward outputs, while attention forms convex combinations of the value vectors. The identical intervention at the current position cancels in the comparison, leaving differences between bounded block outputs across a fixed number of layers. The resulting bound is uniform in sequence length and excludes accumulative growth of the direct cache contribution. We also measure how the corresponding next-token divergence varies with length. Appendix A.5 gives the architecture assumptions and proof.

3.2 MEASUREMENTS

To apply the framework, we estimate the update weight, expected next-token scores, and boundaries of the natural range for each model, concept, and steering layer. We then use these measurements to predict score trajectories, exit times, and the critical steering coefficient. We outline how we measure these quantities and assess the modeling assumptions, with details in Appendix B.

Scores and reference boundaries. We evaluate the token scores defined in Section 2.2 using unsteered forward passes. To estimate the update weight a, we compare concept scores along identical continuations following matched concept and contrast spans after a common prefix. Fitting the decay of their difference to an exponential with a constant offset estimates a, which we use for both context scores in Eq. 3. We calibrate each reference boundary from 0.95 quantile of per-response maximum context scores.

Expected next-token scores. We sample continuations with and without steering from common prefixes drawn from unsteered answers and the concept and contrast responses. Pairing each token score with its preceding context score and averaging within context-score bins centered at q gives $\widehat{m}_{z,\lambda}(q)$. For degradation, we also continue prefixes of steered answers to cover scores rarely reached without steering. To estimate $g_z(\lambda)$, we find the context score q' at which generation without steering yields the same expected next-token score as generation with steering at q. Averaging $q' - q$ over these matches estimates $g_z(\lambda)$ and gives the exit threshold $z_{\max} - g_z(\lambda)$. Where matching is unavailable, we use the empirical exit threshold described in Appendix B.2.

Trajectories and exit times. We compute mean-field trajectories using the measured drift $\widehat{m}_{z,\lambda}(q) - q$ at each coefficient. Starting from each answer's observed context score after an initial prefix, we

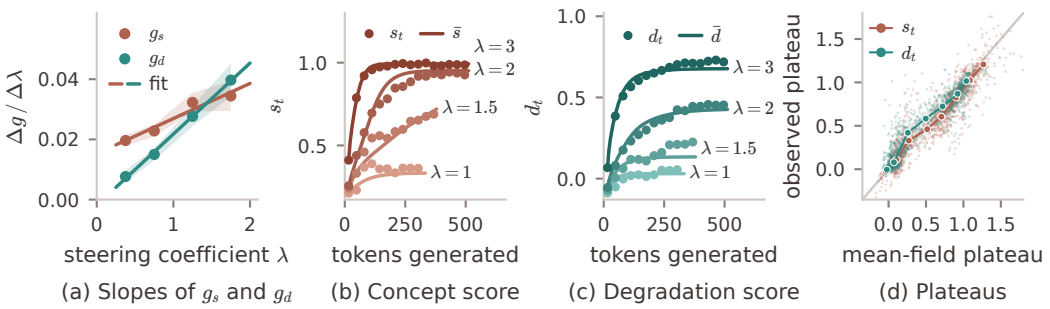


Figure 3: Steering effects and score saturation. (a) Mean finite-difference slopes of g_s and g_d across configurations, with linear fits and 95% confidence bands over $0.25 \leq \lambda \leq 2$. (b, c) Observed mean concept and degradation scores (markers) and mean-field trajectories (lines), averaged over 20 questions and three seeds per question, for Gemma-2 9B, the pessimistic concept, and layer 25 at the labeled coefficients. (d) Observed versus mean-field plateaus across all eight coefficients, connected markers show binned medians, and the gray line denotes equality.

iterate the expected update in Eq. 4 and average the trajectories over answers. We estimate exit times from the mean initial score using the same drift in Eq. 6.

Critical steering coefficient. We estimate λ_c by combining the unsteered degradation curve with a quadratic fit $g_d(\lambda) \approx \kappa\lambda^2$ to the measured equivalent increases. Starting from the unsteered degradation equilibrium, we apply the positive-drift criterion of Theorem 2 to locate the critical increase g_c at the transition to finite-time exit.

4 Experiments

We evaluate the framework's predicted dependence on steering strength, score saturation, and degradation onset using the measurements in Section 3.2. We relate degradation-boundary crossings to judged coherence, then separately examine the direct KV-cache contribution.

Setup. We evaluate eight instruction-tuned models (3B–14B parameters), seven persona concepts (Chen et al., 2025), and three steering layers near 50%, 60%, and 70% of model depth (Table 1), giving 168 configurations. At eight coefficients from 0.25 to 6, we sample answers over three seeds per question on 20 questions per concept, at temperature 1 with a 1024-token cap. Qwen3.6-35B-A3B (Yang et al., 2025) judges coherence using the rubric of Chen et al. (2025) at horizons from 64 to 1024 tokens. Shorter completed answers are retained in full. Appendix C gives further settings and results, including generations requested to contain at least 600 words.

Dependence on steering strength. Over $0.25 \leq \lambda \leq 2$, the concept increase has a positive linear component, while the degradation increase has a near-zero linear component and positive curvature, estimated from their mean slopes (Figure 3a). This pattern is consistent with Proposition 1 over this interval. Appendix C.2 examines variation across models and departures at stronger steering.

Score saturation. The measured recursion captures the rise and stabilization of both scores (Figures 3b–d). Predictions start from each answer's observed context score after an initial prefix and are averaged over answers. We compare these predictions with a baseline that assumes the context score stays at its observed initial value. The recursion reduces median trajectory root mean square error from 0.22 to 0.08 for concept and from 0.18 to 0.04 for degradation. Appendix C.3 examines plateau agreement and variation across models and coefficients.

Degradation onset and coherence. Above λ_c, stronger steering leads to earlier degradation-boundary crossings (Figure 4a), consistent with Corollary 1. At the smallest evaluated coefficient above each configuration's λ_c, the median predicted exit time is 67 tokens, compared with 60 tokens for the observed time at which half the answers have crossed (Appendix C.4).

Boundary crossings are associated with lower judged coherence. For coefficients just above λ_c, coherence relative to unsteered answers declines substantially as longer prefixes are evaluated (Figure 4b). Within a fixed configuration and coefficient, answers that cross the boundary also tend to be less coherent than those that do not (Appendix C.4).

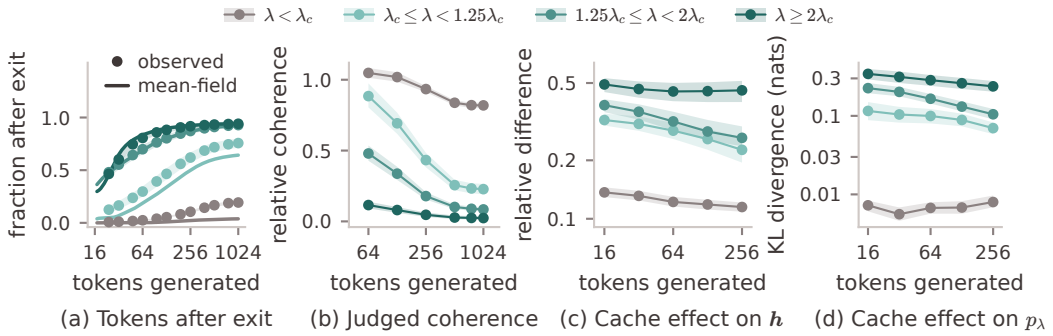


Figure 4: Degradation onset and cache effects, grouped by λ/λ_c. (a) Mean fraction of evaluated text generated after its first degradation-boundary crossing, excluding the initial 16 tokens: observations (markers) and mean-field estimates (lines). (b) Mean coherence relative to unsteered answers at the same evaluation horizon. Results pool generations over three seeds per question; panels (a, b) retain shorter answers in full. (c, d) Median normalized final residual difference and next-token KL divergence, comparing steering at every position with steering only at the current one on identical tokens from answers reaching 256 tokens. Residual differences are normalized by the unsteered residual norm. Shaded bands show 95% confidence intervals.

Choosing a coefficient. Steering strength should balance concept alignment against degradation at the intended response length. Figure 5 illustrates this tradeoff for Gemma-2 9B on the pessimistic concept at layer 25. At $\lambda = 1.5$, both scores are higher at 1024 tokens than at 64 tokens, so the gain in concept alignment is accompanied by a higher degradation score. This tradeoff serves as a practical guide for choosing the appropriate steering strength by identifying coefficients that improve concept alignment for the intended task and response length, while keeping observed boundary crossings infrequent over that length.

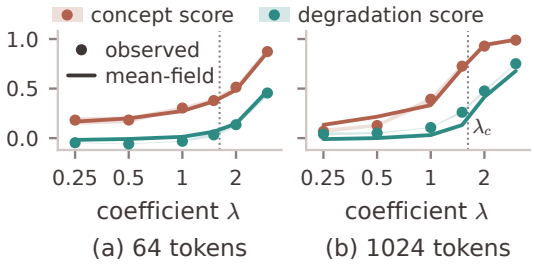


Figure 5: Concept and degradation scores versus steering strength at 64 (a) and 1024 (b) tokens, for the configuration in Figure 3. Markers show observations over three seeds per question, shades their range across seeds, and lines mean-field estimates.

KV-cache contribution. Finally, we compare steering at every position with steering only at the current one, keeping tokens and the current coefficient fixed (Proposition 2). Within each λ/λ_c group, median normalized residual differences and next-token KL divergences show no systematic growth from 16 to 256 tokens (Figures 4c,d). Separate comparisons reveal diminishing prompt-intervention effects alongside increasing effects of earlier answer interventions (Appendix C.5). These empirical findings directly corroborate our theoretical result in Proposition 2 regarding the uniform effect of the KV cache on transformer generation with steering.

5 Conclusions

We developed a stochastic framework that captures how activation steering and evolving context jointly shape concept alignment and degradation during autoregressive generation. Its mean-field analysis characterizes score saturation and threshold-crossing times, while finite-horizon bounds quantify stochastic deviations. The framework explains how steering that appears reliable on short responses can degrade longer generations. A separate analysis with tokens held fixed bounds the direct KV-cache contribution uniformly in sequence length under the stated transformer assumptions. Measurements across eight instruction-tuned models connect the predicted dynamics to observed score trajectories and coherence loss. Our analysis guides the choice of steering strength for a given task and response length while accounting for sampling variability. Translating insights gained from our analysis into effective steering methods represents a promising avenue for future work.

AI USE STATEMENT

The authors used AI to assist with literature discovery and review, suggest experimental directions and ablations, implement and debug experimental code, polish manuscript text, assist with proof verification, and create figures. As part of the experimental pipeline, language models generated the response corpora, and judge LLM provided automated trait and coherence judgments. Literature identified with AI assistance was independently checked against the original sources, and AI-assisted code changes were validated through testing and debugging. The authors reviewed the AI-assisted work and take responsibility for the final content, including the text, claims, figures, code, and experimental results.

ETHICS STATEMENT

This work studies activation steering using publicly available persona prompts, pretrained language models, and model-generated responses. It involves no human-subject experiments, private user data, or newly collected human annotations. We use the prompts and models in accordance with their licenses and terms. Some evaluated traits, including evil, hallucinating, and sycophantic behavior, can elicit harmful or misleading content. We study these behaviors to understand steering reliability and degradation, and recognize that steering techniques could also be misused to amplify them. Our coherence measurements assess output quality, not safety or factual accuracy, and rely on an automated judge that may have biases or errors.

REPRODUCIBILITY STATEMENT

All the models and data used in this work is publicly available. We provide complete proofs, describe score calibration, context-weight estimation, expected next-token score measurements, and the computation of predicted trajectories and exit times.We also specify steering layers and coefficients, sampling settings, and coherence-evaluation protocol.

REFERENCES

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

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Max Belitsky, Dawid J Kopiczko, Michael Dorkenwald, M Jehanzeb Mirza, James R Glass, Cees GM Snoek, and Yuki M Asano. Kv cache steering for controlling frozen llms. *arXiv preprint arXiv:2507.08799*, 2025.

Eric Bigelow, Daniel Wurgaft, YingQiao Wang, Noah Goodman, Tomer Ullman, Hidenori Tanaka, and Ekdeep Singh Lubana. Belief dynamics reveal the dual nature of in-context learning and activation steering. *arXiv preprint arXiv:2511.00617*, 2025.

Stephane Boucheron, Gabor Lugosi, and Pierre Massart. A non asymptotic theory of independence, 2013.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un) reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Dean S. Clark. Short proof of a discrete Gronwall inequality. *Discrete Applied Mathematics*, 16(3): 279–281, 1987.

Alex Ferrando, Xavier Suau, Jordi Gonzàlez, and Pau Rodriguez. Dynamically scaled activation steering. *arXiv preprint arXiv:2512.03661*, 2025.

Niklas Herbster, Martin Zborowski, Alberto Tosato, Gauthier Gidel, and Tommaso Tosato. Activation steering for aligned open-ended generation without sacrificing coherence. *arXiv preprint arXiv:2604.08169*, 2026.

Shawn Im and Sharon Li. A unified understanding and evaluation of steering methods. *arXiv preprint arXiv:2502.02716*, 2025.

Diancheng Kang, Zheyuan Liu, Ningshan Ma, Yue Huang, Zhaoxuan Tan, and Meng Jiang. Prompt-activation duality: Improving activation steering via attention-level interventions. *arXiv preprint arXiv:2605.10664*, 2026.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 284–294, 2018.

Kien Le and Thai Le. Adversarial robustness of activation steering in large language models. *arXiv preprint arXiv:2606.07696*, 2026.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Dung Viet Nguyen, Yen Pham, Hieu Vu, Lei Zhang, and Tan Nguyen. Activation steering with a feedback controller. In *International Conference on Learning Representations*, volume 2026, pp. 154471–154505, 2026.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of machine learning and systems*, 5:606–624, 2023.

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

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation surgery: Theory and practice of affine steering. *arXiv preprint arXiv:2402.09631*, 2024.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *International Conference on Learning Representations*, volume 2025, pp. 55790–55823, 2025.

Magamed Taimeskhanov, Samuel Vaiter, and Damien Garreau. Towards understanding steering strength. *arXiv preprint arXiv:2602.02712*, 2026.

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

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, 2012.

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

Vala Vakilian, Zimeng Wang, Ankit Singh Rawat, and Christos Thrampoulidis. Short-context dominance: How much local context natural language actually needs? *arXiv preprint arXiv:2512.08082*, 2025.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *Advances in Neural Information Processing Systems*, 35:3082–3095, 2022.

Zhiyuan Xu, Stanislav Abaimov, Joseph Gardiner, and Sana Belguith. Progressive behavioral drift through compression valleys in large language models. 2026. URL https://arxiv.org/abs/2511.17194.

Jianhao Elliott Yan, Jin Xu, Chiyu Song, Chenming Wu, Yafu Li, and Yue Zhang. Understanding in-context learning from repetitions. In *International Conference on Learning Representations*, volume 2024, pp. 49968–49989, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Yize Zhao, Tina Behnia, Vala Vakilian, and Christos Thrampoulidis. Implicit geometry of next-token prediction: From language sparsity patterns to model representations. *arXiv preprint arXiv:2408.15417*, 2024.

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

APPENDIX

A Technical Results **13**

A.1 Proof of Proposition 1 13

A.2 Proof of Theorem 1 14

A.3 Proof of Theorem 2 15

A.4 Proof of Corollary 1 16

A.5 Transformer Architecture and Cache Bound 16

B Measurement Protocol **17**

B.1 Scores and reference boundaries 17

B.2 Expected next-token scores and equivalent increases 18

B.3 Trajectories, exit times, and the critical coefficient 19

C Additional Experimental Results **19**

C.1 Evaluation setup 19

C.2 Dependence on steering strength 20

C.3 Score saturation 20

C.4 Degradation onset and coherence 20

C.5 KV-cache contribution 24

A TECHNICAL RESULTS

This appendix provides the proofs of the theoretical results in the main text. We begin with the weak-steering expansions in Proposition 1, then establish the mean-field approximation and saturation guarantees in Theorem 1. We next prove the exit-time characterization in Theorem 2 and the critical steering threshold in Corollary 1. Finally, Appendix A.5 specifies the transformer architecture assumptions and proves the sequence-length-uniform bound on the direct KV-cache contribution in Proposition 2.

A.1 PROOF OF PROPOSITION 1

The two expansions follow from the same derivative identity for an exponentially tilted distribution. Fix a context c and let z denote either the concept score s or the degradation score d. Throughout this proof, $\mathbb{E}_{p_\lambda}$ and Cov_{p_λ} denote expectation and covariance under $p_\lambda(\cdot \mid c)$, with the dependence on c suppressed. By Eq. 2,

$$\mathbb{E}_{p_\lambda}[z] = \frac{\mathbb{E}_{p_0}[ze^{\lambda r}]}{\mathbb{E}_{p_0}[e^{\lambda r}]}.$$

Because the vocabulary $\mathcal{V}$ is finite and z and r take finite values, the numerator and the positive denominator are finite sums of exponentials in λ, so the expected score is smooth in λ. Differentiating the ratio gives

$$\begin{aligned}\frac{d}{d\lambda}\mathbb{E}_{p_\lambda}[z] &= \frac{\mathbb{E}_{p_0}[zre^{\lambda r}]}{\mathbb{E}_{p_0}[e^{\lambda r}]} - \frac{\mathbb{E}_{p_0}[ze^{\lambda r}]\,\mathbb{E}_{p_0}[re^{\lambda r}]}{\mathbb{E}_{p_0}[e^{\lambda r}]^2} \\ &= \mathbb{E}_{p_\lambda}[zr] - \mathbb{E}_{p_\lambda}[z]\,\mathbb{E}_{p_\lambda}[r] = \mathrm{Cov}_{p_\lambda}(z, r).\end{aligned}$$

This identity holds for any function of the token that does not depend on λ. Applying it to zr, z and r together with the product rule gives

$$\begin{aligned}\frac{d^2}{d\lambda^2}\mathbb{E}_{p_\lambda}[z] &= \frac{d}{d\lambda}\big(\mathbb{E}_{p_\lambda}[zr] - \mathbb{E}_{p_\lambda}[z]\,\mathbb{E}_{p_\lambda}[r]\big)\\ &= \mathrm{Cov}_{p_\lambda}(zr, r) - \mathbb{E}_{p_\lambda}[r]\,\mathrm{Cov}_{p_\lambda}(z, r) - \mathbb{E}_{p_\lambda}[z]\,\mathrm{Cov}_{p_\lambda}(r, r)\\ &= \mathrm{Cov}_{p_\lambda}(z, r^2) - 2\,\mathbb{E}_{p_\lambda}[r]\,\mathrm{Cov}_{p_\lambda}(z, r)\\ &= \mathrm{Cov}_{p_\lambda}\big(z, (r - \mathbb{E}_{p_\lambda}[r])^2\big).\end{aligned}$$

The third line follows by expanding each covariance into moments of p_λ, and the last line uses linearity of covariance and the fact that adding a constant to its second argument does not change it. Thus, writing $\bar{r} := \mathbb{E}_{p_0}[r]$, Taylor's theorem at zero gives

$$\mathbb{E}_{p_\lambda}[z] - \mathbb{E}_{p_0}[z] = \lambda\,\mathrm{Cov}_{p_0}(z, r) + \frac{\lambda^2}{2}\big(\,\mathrm{Cov}_{p_0}(z, r^2) - 2\bar{r}\,\mathrm{Cov}_{p_0}(z, r)\big) + O(\lambda^3).$$

For the concept score, setting $z = s$ and absorbing the quadratic term into $O(\lambda^2)$ gives the first expansion. For the degradation score, the hypothesis $\mathrm{Cov}_{p_0}(d, r) = 0$ removes the linear term. It also removes the term $2\bar{r}\,\mathrm{Cov}_{p_0}(d, r)$ from the quadratic coefficient. This leaves

$$\mathbb{E}_{p_\lambda}[d] - \mathbb{E}_{p_0}[d] = \frac{\lambda^2}{2}\,\mathrm{Cov}_{p_0}(d, r^2) + O(\lambda^3),$$

which is the second expansion.

Equivalent increases in context score. To transfer these orders from expected token scores to g_s and g_d, we use Assumption 1 to express each score change through m_z. Suppose that m_z is continuously differentiable near $z(c)$ with $m_z'(z(c)) > 0$. Continuity of g_z and the identity $g_z(0) = 0$ give

$$m_z\big(z(c) + g_z(\lambda)\big) - m_z\big(z(c)\big) = m_z'\big(z(c)\big)g_z(\lambda) + o\big(g_z(\lambda)\big).$$

The left-hand side is the change in the expected token score under steering. Matching the two expansions and dividing by $m_z'(z(c)) > 0$ yields

$$\begin{aligned}g_s(\lambda) &= \frac{\mathrm{Cov}_{p_0}(s, r)}{m_s'(s(c))}\lambda + o(\lambda),\\ g_d(\lambda) &= \frac{\mathrm{Cov}_{p_0}(d, r^2)}{2m_d'(d(c))}\lambda^2 + o(\lambda^2),\end{aligned}$$

where the covariances are evaluated at the fixed context c. A context-independent g_z therefore requires its leading coefficient in these expressions to agree across contexts satisfying the stated conditions. When the leading coefficients are positive, g_s grows linearly and g_d quadratically near zero.

A.2 Proof of Theorem 1

We first prove interval invariance and convergence, as stated in part (ii). The resulting bounds on the drift then allow us to control the approximation error in part (i). Write $\mathcal{I} = [\alpha, \beta]$ and $R = \beta - \alpha$. If $R = 0$, both trajectories are constant and all claims follow immediately, so assume $R > 0$.

The stochastic recursion remains in $\mathcal{I}$ because each update in Eq. 3 is a convex combination of the current context score and a token score in $\mathcal{I}$. For the mean-field trajectory, recall from Eq. 4 that $F(z) = g_z + f_z(z + g_z) = m_z(z + g_z) - z$. The Lipschitz assumption on f_z makes F Lipschitz on $\mathcal{I}$ with constant K. The range condition on m_z gives $z + F(z) \in \mathcal{I}$ for every $z \in \mathcal{I}$, and therefore

$$F(\alpha) \geq 0, \qquad F(\beta) \leq 0, \qquad |F(z)| \leq R \quad \text{for all } z \in \mathcal{I}.$$

The endpoint inequalities ensure that the mean-field trajectory cannot leave $\mathcal{I}$. Lipschitz continuity gives uniqueness, and boundedness within this compact interval guarantees existence for every $\tau \geq 0$ (Teschl, 2012, Chapters 1 and 2).

To establish convergence, observe that the mean-field trajectory is constant if $F(z_1) = 0$. Otherwise, uniqueness prevents it from reaching an equilibrium in finite time, so its derivative retains the sign of

$F(z_1)$. In either case, the trajectory is monotone and bounded, and hence converges to a limit $z_\infty \in \mathcal{I}$. This limit must satisfy $F(z_\infty) = 0$. Indeed, a nonzero limiting drift would, by continuity, keep the magnitude of the derivative bounded away from zero for all sufficiently large τ, contradicting convergence. This proves part (ii).

For part (i), we bound the errors due to token sampling and time discretization separately, then control how they propagate through the drift. Let $\mathcal{F}_t$ be the history generated by the fixed prompt and the tokens $y_1, \dots, y_{t-1}$, and define the centered token score $\xi_t := z(y_t) - \mathbb{E}[z(y_t) \mid \mathcal{F}_t]$. Assumption 1 rewrites the recursion as

$$z_{t+1} = z_t + a_t F(z_t) + a_t \xi_t, \qquad \mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0, \qquad |\xi_t| \le R. \tag{A.1}$$

Consequently, $M_k := \sum_{r<k} a_r \xi_r$ is a martingale with respect to $(\mathcal{F}_k)$, with $M_1 = 0$ and increments bounded by the deterministic quantities $a_r R$. For a fixed horizon $t \ge 2$, apply Doob's submartingale inequality (Williams, 1991, Section 14.6) to $\exp(\theta M_k)$ and $\exp(-\theta M_k)$. Combining the bounded-increment exponential-moment estimate (Azuma, 1967) with optimization over $\theta > 0$ gives the maximal Azuma–Hoeffding bound: with probability at least $1 - \delta$,

$$\max_{k \le t} |M_k| \le R \sqrt{2 \log(2/\delta) \sum_{r<t} a_r^2}.$$

We next compare the exact mean-field increment over $[\tau_r, \tau_{r+1}]$ with the Euler increment $a_r F(\bar{z}(\tau_r))$. Since $\tau_{r+1} - \tau_r = a_r$, the Lipschitz property of F and the bound $|\bar{z}'| \le R$ give

$$\begin{aligned} |\bar{z}(\tau_{r+1}) - \bar{z}(\tau_r) - a_r F(\bar{z}(\tau_r))| &\le \int_{\tau_r}^{\tau_{r+1}} K |\bar{z}(u) - \bar{z}(\tau_r)| \, du \\ &\le \int_{\tau_r}^{\tau_{r+1}} KR(u - \tau_r) \, du = \frac{KR}{2} a_r^2. \end{aligned}$$

It remains to combine the two error bounds. Let $e_k := |z_k - \bar{z}(\tau_k)|$. Summing Eq. A.1 from the common initial value, subtracting the corresponding mean-field increments, and using Lipschitz continuity yields, simultaneously for all $k \le t$,

$$e_k \le K \sum_{r<k} a_r e_r + \max_{j \le t} |M_j| + \frac{KR}{2} \sum_{r<t} a_r^2.$$

The first term accounts for the propagation of earlier discrepancies through F, while the remaining terms bound sampling and discretization errors. Applying the discrete Grönwall inequality (Clark, 1987) gives

$$\max_{k \le t} e_k \le e^{K\tau_t} \left(\max_{j \le t} |M_j| + \frac{KR}{2} \sum_{r<t} a_r^2 \right).$$

Combining this estimate with the martingale bound proves part (i) with $C = R \max\{\sqrt{2}, K/2\}$. For $t = 1$, the approximation error is zero.

A.3 Proof of Theorem 2

Let $b = z_{\max} - g_z > z_1$ be the context score at the exit boundary. The key distinction is whether the drift stays strictly positive over the entire interval $[z_1, b]$, including the boundary itself.

For part (i), suppose $F > 0$ throughout $[z_1, b]$. Continuity on this compact interval gives a uniform lower bound $\gamma := \min_{z \in [z_1, b]} F(z) > 0$. Until the trajectory reaches b, its derivative is at least γ, so exit occurs no later than intrinsic time $(b - z_1)/\gamma$. Separating variables along this increasing trajectory gives the exact exit time:

$$\tau_{\text{exit}} = \int_{z_1}^{b} \frac{dz}{F(z)}.$$

For part (ii), suppose F is nonpositive at some point of $[z_1, b]$. If $F(z_1) \le 0$, Theorem 1 implies that the trajectory is constant or decreasing. It therefore never reaches b, and $z_\infty \le z_1 < b$. If $F(z_1) > 0$, continuity ensures that F has a zero in $(z_1, b]$. Let q_* be the smallest such zero. Since $F > 0$

on $[z_1, q_*)$, the trajectory increases but cannot reach q_* in finite time by uniqueness. Theorem 1 guarantees convergence to an equilibrium, and q_* is the only equilibrium in $[z_1, q_*]$. Hence $z_\infty = q_* \le b$, and finite-time exit is impossible.

It remains to establish the strict inequality when F is negative somewhere on $[z_1, b]$. The case $F(z_1) \le 0$ already gives $z_\infty < b$. When $F(z_1) > 0$, continuity places a zero between z_1 and any point where F is negative, so $q_* < b$. In both cases, $z_\infty + g_z < z_{\max}$, completing part (ii).

A.4 Proof of Corollary 1

By Theorem 2, when $z_1 + g_z(\lambda) < z_{\max}$, finite-time exit occurs exactly when

$$g_z(\lambda) + f_z(u) > 0 \qquad \text{for every } u \in [z_1 + g_z(\lambda), z_{\max}].$$

Increasing λ cannot decrease the left-hand side or enlarge the interval. Thus, once finite-time exit occurs, it also occurs at every larger coefficient, with immediate exit if the initial condition already meets the boundary. Define $\lambda_c := \sup\{\lambda \ge 0 : \tau_{\text{exit}}(\lambda) = \infty\}$. The assumed existence of coefficients with and without exit makes λ_c finite. Strict positivity on the compact interval persists under small changes in λ, so exit cannot occur at λ_c. This establishes the claimed threshold.

For positive, finite exit times, changing variables in Eq. 6 gives

$$\tau_{\text{exit}}(\lambda) = \int_{z_1 + g_z(\lambda)}^{z_{\max}} \frac{du}{g_z(\lambda) + f_z(u)}.$$

Increasing $g_z(\lambda)$ shortens the integration interval and decreases its positive integrand. The exit time is therefore nonincreasing in λ, with strict decrease whenever $g_z(\lambda)$ increases and both exit times are positive.

At λ_c, the trajectory stays a positive distance below the exit boundary on every fixed finite time interval. By continuous dependence on λ (Teschl, 2012, Chapter 2), trajectories at sufficiently close coefficients also remain below their boundaries throughout that interval. Hence the exit time tends to infinity as λ approaches λ_c from above.

A.5 Transformer Architecture and Cache Bound

We now specify the architecture used in Proposition 2. Consider a decoder-only transformer with L layers and hidden dimension d. For a sequence of length T, let $\boldsymbol{h}_i^{(\ell)} \in \mathbb{R}^d$ denote the residual stream entering layer ℓ at position $i \in [T] := \{1, \dots, T\}$. Normalization precedes each attention and feed-forward block, giving the update

$$\begin{aligned} \tilde{\boldsymbol{h}}_i^{(\ell)} &= \boldsymbol{h}_i^{(\ell)} + \mathrm{ATTN}_\ell\big(\mathrm{LN}(\boldsymbol{h}_1^{(\ell)}), \dots, \mathrm{LN}(\boldsymbol{h}_i^{(\ell)})\big), \\ \boldsymbol{h}_i^{(\ell+1)} &= \tilde{\boldsymbol{h}}_i^{(\ell)} + \mathrm{FFN}_\ell\big(\mathrm{LN}(\tilde{\boldsymbol{h}}_i^{(\ell)})\big). \end{aligned} \tag{A.2}$$

Here LN denotes RMSNorm (Zhang & Sennrich, 2019) or LayerNorm (Ba et al., 2016), ATTN_ℓ is causal multi-head attention, and FFN_ℓ is a position-wise feed-forward network. Each attention head forms a convex combination of value vectors at positions $j \le i$, followed by an output projection that combines the heads. The next-token distribution is $\mathrm{softmax}(\boldsymbol{W}\mathrm{LN}(\boldsymbol{h}_T^{(L+1)}))$, with unembedding matrix $\boldsymbol{W}$ and final normalization LN.

Proof of Proposition 2. The proof has two steps: bound each block's output uniformly over its inputs, then propagate these bounds through the difference between the two forward passes. We omit projection biases, which only change the constants.

Let N_ℓ bound the norm of the normalized inputs to both blocks at layer ℓ. Such a bound holds for RMSNorm and LayerNorm with fixed finite affine parameters and a positive normalization stabilizer. For attention head h, write $\boldsymbol{W}_{\ell,h}^V$ for its value projection and $\boldsymbol{v}_{\ell,h,j}$ for the value vector at position j. The attention weights satisfy $\pi_{ij}^{\ell,h} \ge 0$ and $\sum_{j \le i} \pi_{ij}^{\ell,h} = 1$. Thus, bounded normalized inputs give bounded values, and their convex combination obeys the same bound:

$$\|\boldsymbol{v}_{\ell,h,j}\|_2 \le \|\boldsymbol{W}_{\ell,h}^V\|_{\mathrm{op}} N_\ell =: V_{\ell,h},$$

$$\left\| \sum_{j \le i} \pi_{ij}^{\ell,h} \boldsymbol{v}_{\ell,h,j} \right\|_2 \le \sum_{j \le i} \pi_{ij}^{\ell,h} \|\boldsymbol{v}_{\ell,h,j}\|_2 \le V_{\ell,h}.$$

The bound on the head output is independent of the number of cached positions and the attention weights. To extend it to the full layer, let $\boldsymbol{A}_i^{(\ell)}$ and $\boldsymbol{D}_i^{(\ell)}$ denote the attention and feed-forward outputs in Eq. A.2, and write $\boldsymbol{W}_{\ell,h}^O$ for the output-projection block corresponding to head h. Then

$$\|\boldsymbol{A}_i^{(\ell)}\|_2 \le \sum_h \|\boldsymbol{W}_{\ell,h}^O\|_{\mathrm{op}} V_{\ell,h} =: A_\ell,$$
$$\|\boldsymbol{D}_i^{(\ell)}\|_2 \le \max_{\boldsymbol{u}:\,\|\boldsymbol{u}\|_2 \le N_\ell} \|\mathrm{FFN}_\ell(\boldsymbol{u})\|_2 =: D_\ell < \infty.$$

The feed-forward bound is finite because FFN_ℓ is continuous on the compact ball of radius N_ℓ. Both A_ℓ and D_ℓ depend only on the model parameters and hold uniformly over sequences, positions, and steering coefficients.

We now compare forward passes on the same token sequence, with steering applied at $[T]$ in one pass and at $\{T\}$ in the other. Write ℓ_0 for the fixed intervention layer and define the residual difference at the current position by $\boldsymbol{\Delta}_\ell := \boldsymbol{h}_T^{(\ell)}([T]) - \boldsymbol{h}_T^{(\ell)}(\{T\})$. At layer ℓ_0, the two residual vectors at T are identical before steering and receive the same addition $\lambda \boldsymbol{v}$. Hence $\boldsymbol{\Delta}_{\ell_0} = \boldsymbol{0}$ also after the intervention. For each subsequent residual update, subtracting the two instances of Eq. A.2 and applying the block bounds gives, for $\ell \ge \ell_0$,

$$\boldsymbol{\Delta}_{\ell+1} = \boldsymbol{\Delta}_\ell + \big(\boldsymbol{A}_T^{(\ell)}([T]) - \boldsymbol{A}_T^{(\ell)}(\{T\})\big) + \big(\boldsymbol{D}_T^{(\ell)}([T]) - \boldsymbol{D}_T^{(\ell)}(\{T\})\big),$$
$$\|\boldsymbol{\Delta}_{\ell+1}\|_2 \le \|\boldsymbol{\Delta}_\ell\|_2 + 2A_\ell + 2D_\ell.$$

Iterating from $\boldsymbol{\Delta}_{\ell_0} = \boldsymbol{0}$ and squaring yields

$$\|\boldsymbol{h}_T^{(L+1)}([T]) - \boldsymbol{h}_T^{(L+1)}(\{T\})\|_2^2 \le 4\left(\sum_{\ell=\ell_0}^{L} (A_\ell + D_\ell)\right)^2 =: B.$$

This sum contains a fixed number of terms, each determined only by the model parameters. Thus, B is independent of the token sequence, its length T, and the steering coefficient λ, proving the proposition. If the final normalization has Lipschitz constant K_{LN}, applying it and the unembedding matrix gives the corresponding logit bound $\|\boldsymbol{W}\|_{\mathrm{op}} K_{\mathrm{LN}} \sqrt{B}$. □

B Measurement Protocol

For each model, concept, and steering layer, we calibrate the scores and estimate the quantities used in Section 3.2.

B.1 Scores and reference boundaries

Token scores. All text is scored with steering disabled, regardless of how it was generated. The concept score is calibrated on the responses used to construct $\boldsymbol{v}$, with means 0 for contrast and 1 for concept responses. For degradation, we compare responses generated without steering at temperatures 1 and 2, pooling across concepts for each model and layer and weighting responses equally. Orthogonalizing the difference of their mean residuals to $\boldsymbol{v}$ and normalizing gives $\boldsymbol{u}$ (Figure 6). The resulting score has means 0 and 1 on the two temperature groups, each restricted to responses of at least 16 tokens.

On steered answers excluded from score construction, the mean degradation score separates judged coherence of at most 25 from at least 75 with mean AUROC 0.990 across the 168 configurations.

Update weight and context scores. Identical continuation tokens let us isolate the influence of preceding text on concept scores. For 48 matched concept–contrast pairs, we sample an unsteered continuation after the contrast span and append the same tokens after the concept span (Figure 7a). The 32-token spans follow either their common question or an unsteered answer prefix.

At each depth, we fit $\gamma + \beta(1-a)^k$ to the mean score difference normalized by its initial peak, with k measured from that peak (Figure 7b). The fitted β and γ describe the decay amplitude and

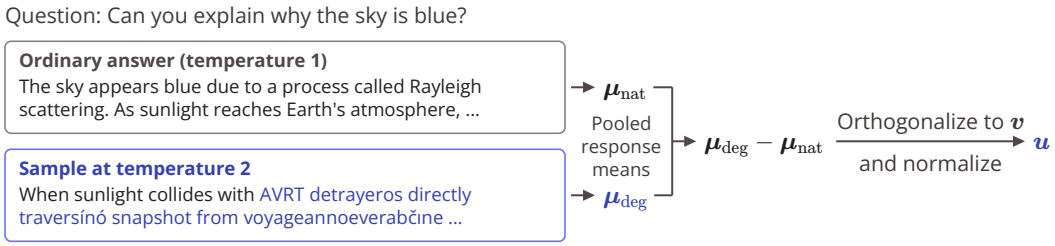
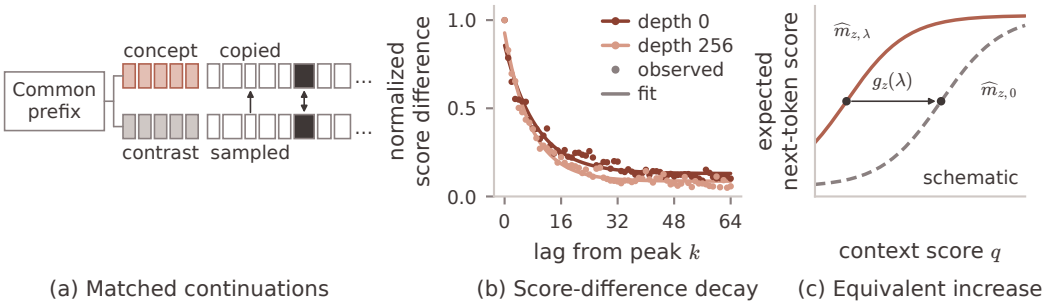


Figure 6: Degradation direction. An ordinary answer and a temperature-2 sample from Mistral 7B for the same question, with the degenerate text of the sample in blue. Orthogonalizing the difference between the mean residual vectors of the two kinds of text to $\boldsymbol{v}$ and normalizing it gives $\boldsymbol{u}$.



(a) Matched continuations (b) Score-difference decay (c) Equivalent increase

Figure 7: Context weight and equivalent increase. (a) The continuation sampled after the contrast span is appended unchanged after the concept span, and the difference between the concept scores of its k-th token in the two sequences measures the influence of the span at lag k. (b) Normalized difference for Gemma-2 9B, the pessimistic concept, and layer 25, with the span placed directly after the question (depth 0) or after 256 tokens of unsteered text (depth 256), and the fits of $\gamma + \beta(1-a)^k$. (c) Schematic of reading $g_z(\lambda)$ as the horizontal displacement between the steered and unsteered curves at equal expected next-token score.

offset over the measured window. Only a, averaged over depths, enters Eq. 3 for both context scores, initialized from the final prompt token's scores.

Reference boundaries. We score the concept and contrast responses used to construct $\boldsymbol{v}$ with their persona instructions removed. The boundaries are 0.95 quantiles of per-response maximum context scores: concept responses determine $s_{\max}$, and both groups determine $d_{\max}$.

B.2 Expected next-token scores and equivalent increases

Expected next-token scores. We estimate $\widehat{m}_{z,\lambda}(q)$ by averaging token scores conditional on their preceding context score. We sample continuations with and without steering from prefixes of the concept and contrast responses and of unsteered answers generated with and without persona prompts. Persona instructions are removed before sampling continuations. To measure higher degradation scores, we additionally continue prefixes of steered answers at their original coefficient and at $\lambda = 0$, filling only bins with insufficient observations. Bins have width 0.1 and require at least 200 score pairs.

Equivalent increases and exit thresholds. We obtain a monotone unsteered curve by isotonic regression weighted by bin counts. Within its fitted expectation range, we match the steered curve at q to the same expectation without steering at q' and average $q' - q$ to estimate $g_z(\lambda)$ (Figure 7c).

The exit threshold is $z_{\max} - g_z(\lambda)$ when matching succeeds. Otherwise, we use the lowest bin center with a measured expectation from continuations of unsteered prefixes, if available. This empirical convention counts entry into the measured score range as exit without estimating g_z. It becomes more relevant under strong steering: g_d is estimated for 165 of 168 configurations at $\lambda = 2$, but only 32 at $\lambda = 6$.

Checks of Assumption 1. We assess dependence beyond the corresponding context score by comparing unsteered curves at matched scores. Median mean absolute gaps across configurations are 0.012 between early (tokens 16–127) and later concept curves, and 0.017 for concept and 0.012 for degradation when splitting by the other score's median. These gaps are about 1–2% of the unit separation used to calibrate each score.

Estimates of the equivalent increase also vary across context scores: the standard deviation of the displacements used to estimate g_s has median 0.010 over configurations and coefficients with $\lambda \leq 2$. This quantifies variation around the fitted mean increase.

B.3 Trajectories, exit times, and the critical coefficient

Trajectories and exit times. Predictions are conditional on the observed context scores after 16 answer tokens. Deterministic trajectories iterate Eq. 4 using the measured drift $\widehat{m}_{z,\lambda}(q) - q$. We interpolate linearly between measured scores and keep endpoint estimates constant outside their range. We average trajectories over answers, holding observed scores at their final value after an answer ends.

Observed exits are first threshold crossings from initialization within each answer's length. We report the step by which at least half the answers have crossed, if that fraction is reached. The mean-field estimate starts from the mean initial score and evaluates Eq. 6 with the measured drift, converting intrinsic time to tokens as $16 + \tau_{\text{exit}}/a$.

Critical steering coefficient. We fit $g_d(\lambda) \approx \kappa\lambda^2$ by nonnegative least squares through the origin, using at least three measured increases at $\lambda \leq 3$. From the unsteered degradation equilibrium, we apply Theorem 2 to locate the transition to exit: $g + \widehat{f}_d(q+g) > 0$ up to $d_{\max} - g$, where $\widehat{f}_d(q) = \widehat{m}_{d,0}(q) - q$. The critical increase g_c gives $\lambda_c = \sqrt{g_c/\kappa}$ when $\kappa > 0$ and g_c is finite.

C Additional Experimental Results

C.1 Evaluation setup

Table 1 lists the models and steering layers, numbered from one. Each model and layer is evaluated on the apathetic, evil, hallucinating, humorous, impolite, pessimistic, and sycophantic concepts, at $\lambda \in \{0.25, 0.5, 1, 1.5, 2, 3, 4, 6\}$. Summary statistics pool configurations for which the relevant measurements are available.

Coherence is judged on a 0–100 scale at horizons of 64, 128, 256, 512, 768, and 1024 tokens. Each answer is evaluated up to the horizon or its end, whichever comes first. Relative coherence divides the mean judged score by that of unsteered answers to the same questions, with no persona prompt. The long-generation setting additionally requests at least 600 words and uses its own unsteered baseline. Shaded bands in the onset and cache plots show 95% confidence intervals, estimated by resampling model–concept–layer configurations with replacement.

Table 1: Instruction-tuned models, parameter counts, and steering layers. Each layer is evaluated separately on all seven concepts.

Model	Size	Steering layers
Phi-4	14B	20, 24, 28
Qwen2.5-14B-Instruct	14B	24, 29, 34
Qwen2.5-7B-Instruct	7B	14, 17, 20
Mistral-Nemo-Instruct-2407	12B	20, 24, 28
Mistral-7B-Instruct-v0.3	7B	16, 19, 22
Gemma-2-9B-it	9B	21, 25, 29
Llama-3.1-8B-Instruct	8B	16, 19, 22
Llama-3.2-3B-Instruct	3B	14, 17, 20

C.2 Dependence on steering strength

The distinction between concept and degradation effects appears across models in the weak-steering range $0.25 \leq \lambda \leq 2$ (Figure 8). We estimate the linear and quadratic components of each equivalent increase by fitting the mean slopes between successive coefficients (Figure 3a), with confidence intervals obtained by resampling configurations. All eight models have positive linear concept and quadratic degradation components. The 95% confidence interval for the linear degradation component includes zero in six models and lies below zero in Mistral-Nemo and Gemma-2. The long-generation setting gives similar pooled estimates: a linear concept component of 0.014 and a quadratic degradation component of 0.012.

Some models follow a different trend at stronger steering. In both Llama models and Mistral-Nemo, the concept increase peaks near $\lambda = 2$. Larger coefficients can also move steered expectations beyond the range used to estimate equivalent increases (Appendix B.2). We therefore restrict the fits to $\lambda \leq 2$ and interpret them as the dependence over this measured interval. The inferred linear degradation component near zero varies with the fit specification.

C.3 Score saturation

The mean-field trajectories capture the rise and stabilization of scores across models and concepts (Figures 9 and 10). We compare trajectories from token 16 to the median answer length and estimate plateau levels over the final 32 steps. Median absolute differences between observed and predicted plateaus over this interval are 0.09 for concept and 0.05 for degradation, as summarized in Figure 3d. Comparing observed plateaus with the limiting equilibria $F(z_\infty) = 0$ gives similar differences of 0.10 and 0.05, respectively.

Agreement varies with model and steering strength. In the Llama models at $\lambda = 1$, observed degradation scores continue to rise after the approximation stabilizes. At the weakest coefficients, the unsteered trajectory is also a competitive baseline, with lower errors for both scores at $\lambda = 0.25$ and for degradation at $\lambda = 0.5$. These comparisons use the measured conditional means at each coefficient to assess score evolution.

Mean score trajectories remain stable late in generation under both prompt settings. Over the second half of the median answer length, the median score change across configurations has magnitude below 0.03 for concept and 0.06 for degradation at every evaluated coefficient. Plateau levels generally increase with steering strength, although the median concept plateau declines beyond $\lambda = 3$, consistent with the changes in concept increase discussed in Appendix C.2.

C.4 Degradation onset and coherence

Stronger steering leads to earlier degradation-boundary crossings and lower coherence across models (Figure 11). At the smallest evaluated coefficient above each configuration's λ_c, the median step by which half the answers have crossed is 60 tokens, compared with 67 predicted by the mean-field dynamics. At the next coefficient, both values fall to 18 tokens. The plotted fraction of text after exit measures how much of each evaluated answer follows its first crossing, averaged over answers.

Boundary crossings distinguish coherence at fixed steering strength. Comparing complete answers within each configuration and coefficient, with at least five answers in each group, crossing answers have lower mean coherence in 470 of 496 settings (94.8%). The association remains similar when restricted to identifiable shifted boundaries ($448/471$, or 95.1%). The remaining cases use the empirical threshold described in Appendix B.2.

Coefficient limits by horizon. For each configuration and horizon, we take the largest coefficient below the first at which at least half of the answers have crossed the boundary. Answers that end earlier are evaluated through their end, and a coefficient without a measured degradation curve counts as crossed. This limit falls from a median of 2 at 64 tokens to 1 at 1024 tokens and is lower at 1024 tokens in 156 of 168 configurations. Applying the same rule to the mean-field exit step, at which an answer crosses if it is still being generated, gives the observed limit in 122 configurations at 64 tokens but in only 39 at 1024 tokens, and a larger limit in 35 and 126 configurations, respectively. At the observed limit, the median concept score is 0.48 at 64 tokens and 0.34 at 1024 tokens. In Figure 5, observed and mean-field scores are read at the horizon or at the end of the answer,

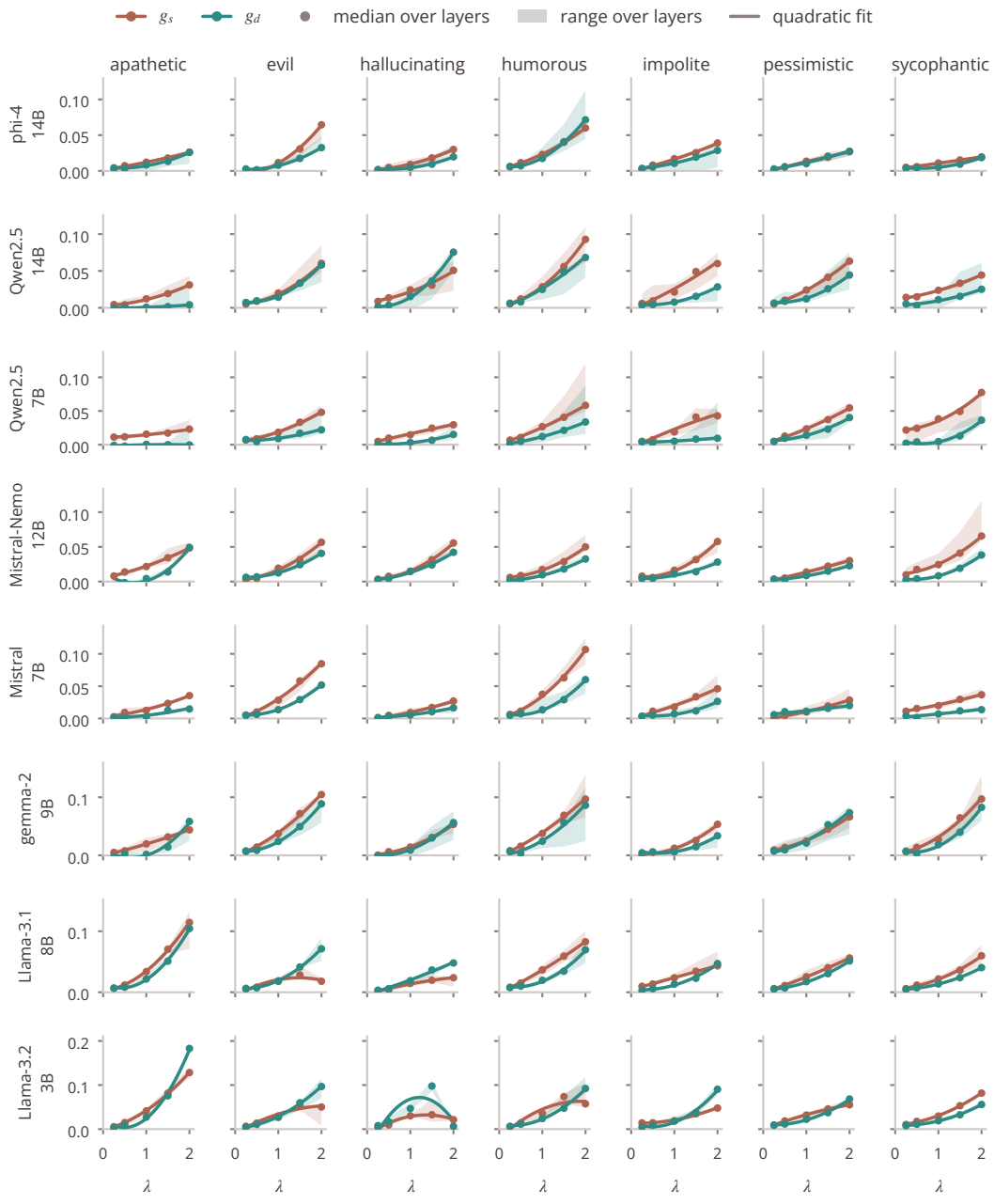


Figure 8: Equivalent increases in concept score $g_s(\lambda)$ (brown) and degradation score $g_d(\lambda)$ (teal), by model (rows) and concept (columns). Markers show medians across available steering layers, shaded bands show the range across layers, and lines are quadratic fits over $0.25 \leq \lambda \leq 2$. Vertical scales are shared within each row.

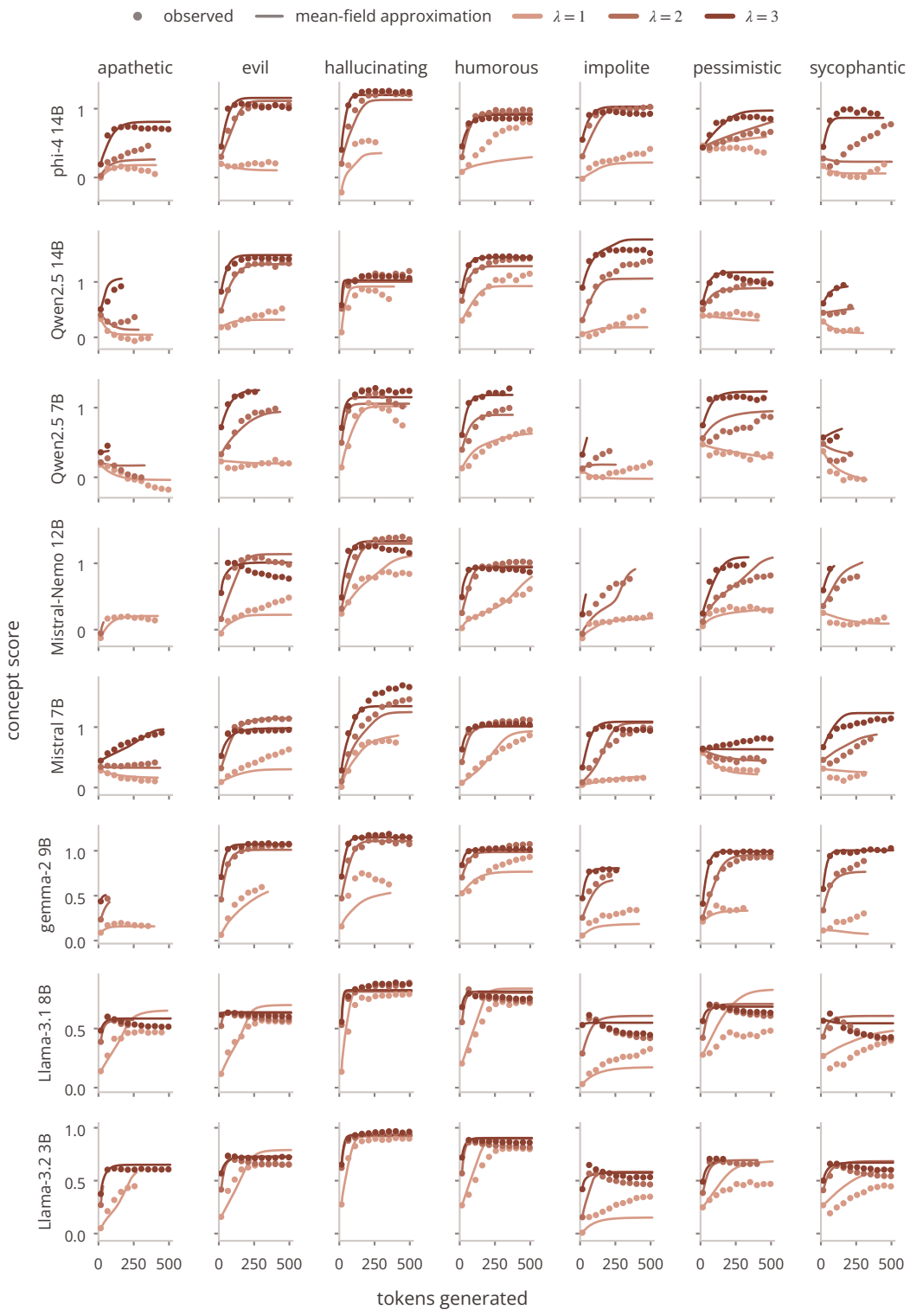


Figure 9: Concept-score trajectories by model (rows) and concept (columns), with steering at approximately 60% of model depth. Markers show observed scores averaged over 20 questions and three seeds per question. Lines show mean-field trajectories initialized from observed scores at token 16. Shades indicate $\lambda \in \{1, 2, 3\}$. Each trajectory ends at the median answer length or 512 tokens, whichever is shorter. Vertical scales are shared within each row.

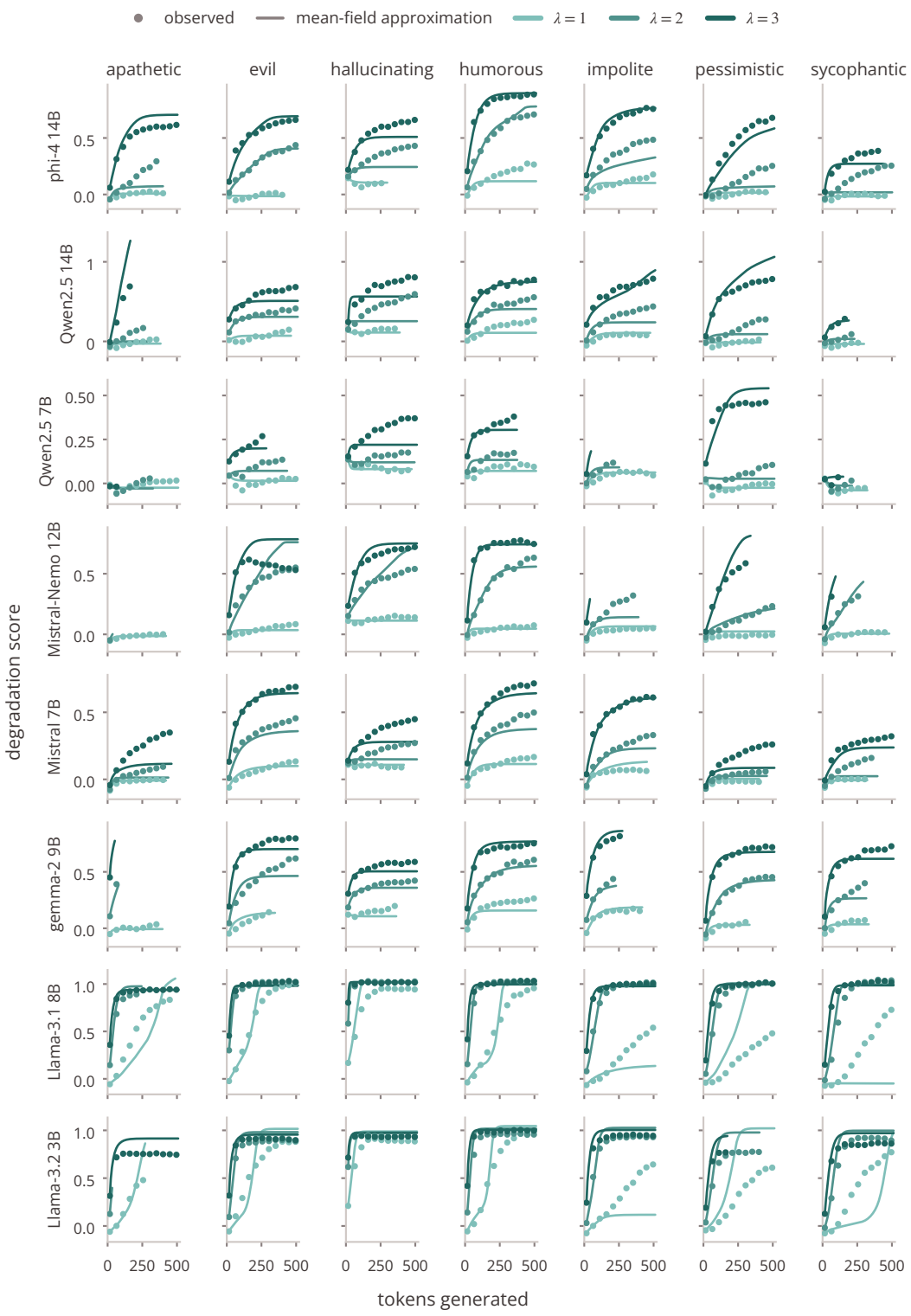


Figure 10: Degradation-score trajectories by model (rows) and concept (columns), with steering at approximately 60% of model depth. Markers show observed scores averaged over 20 questions and three seeds per question. Lines show mean-field trajectories initialized at token 16. Shades indicate $\lambda \in \{1, 2, 3\}$, with the same comparison lengths as Figure 9. Vertical scales are shared within each row.

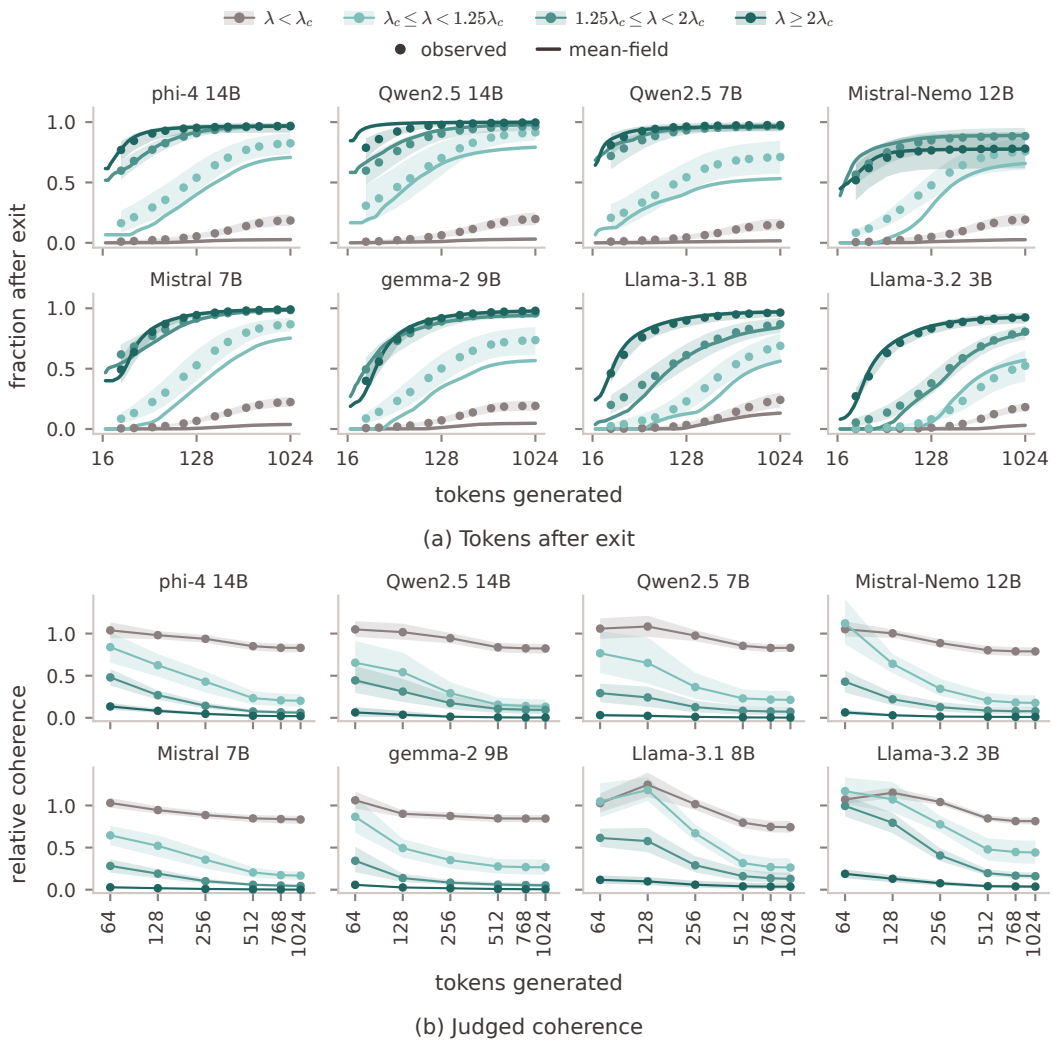


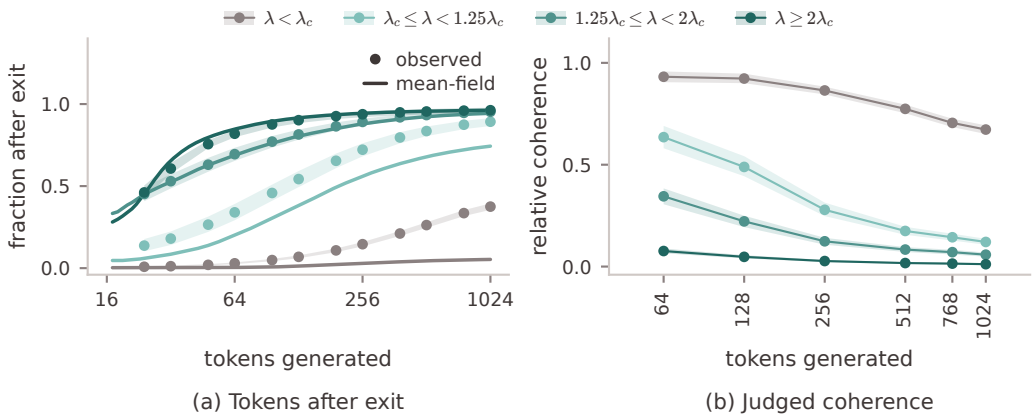
Figure 11: Degradation onset and coherence by model, grouped by λ/λ_c. (a) Mean fraction of evaluated text generated after the first degradation-boundary crossing, with observations (markers) and mean-field estimates (lines). (b) Mean judged coherence relative to unsteered answers at the same evaluation horizon. Each model pools seven concepts, three steering layers, and generations over three seeds per question. Shorter answers are retained in full. Shaded bands show 95% confidence intervals.

whichever comes first. The figure omits $\lambda = 4$ and 6, at which every answer of the example crosses within 64 tokens and the boundary is the empirical threshold of Appendix B.2.

The long-generation setting preserves the ordering by steering strength and shows a larger fraction of text after exit below λ_c (Figure 12). At the 1024-token horizon, this fraction is 0.37 and relative coherence is 0.67 for $\lambda < \lambda_c$. We retain the boundary calibrated on natural-length responses. In this setting, 19% of unsteered long-generation answers also cross it. This comparison therefore reflects both the changed prompt and longer answers, with coherence normalized by the unsteered baseline from the same setting.

C.5 KV-cache contribution

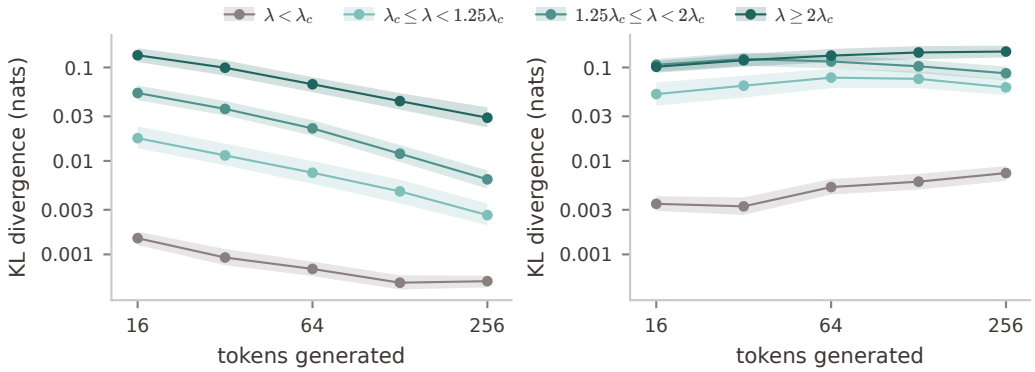
We evaluate the same natural-length answers at prefixes of 16, 32, 64, 128, and 256 generated tokens, restricting the comparison to answers that reach the final length. For each prefix, the two



(a) Tokens after exit (b) Judged coherence

Figure 12: Degradation onset and coherence when prompts request at least 600 words, grouped by λ/λ_c. (a) Mean fraction of evaluated text generated after the first degradation-boundary crossing: observations (markers) and mean-field estimates (lines). (b) Mean coherence relative to unsteered answers under the same length instruction. Results pool generations over three seeds per question. Shorter answers are retained in full. Shaded bands show 95% confidence intervals.

forward passes use identical tokens and differ only in whether earlier positions are steered. Residual differences are measured before the final normalization and divided by the unsteered residual norm. We compare the next-token distributions through $\mathrm{KL}(p_{[T]} \,\|\, p_{\{T\}})$. An additional pass steers the prompt and current position while leaving earlier answer positions unsteered. Comparing this pass with the other two distinguishes the effects of interventions in the prompt and in the generated answer (Figure 13).



(a) Steered prompt in the cache (b) Steered answer in the cache

Figure 13: KV-cache effects of steering earlier positions, grouped by λ/λ_c. (a) Next-token KL divergence when adding prompt steering to steering at the current position. (b) Divergence when additionally steering earlier answer positions. Both comparisons keep tokens fixed and use the same answers reaching 256 tokens as Figure 4c,d. Markers show medians over answers generated over three seeds per question, with 95% confidence bands.

The prompt intervention's effect diminishes as generation proceeds, while the effect of earlier answer interventions can grow (Figure 13). These KL divergences describe separate comparisons and are not additive. The combined effect is largely stable or decreasing over the measured lengths, with variation across models and coefficient groups (Figure 14), consistent with the aggregate trend in Section 4.

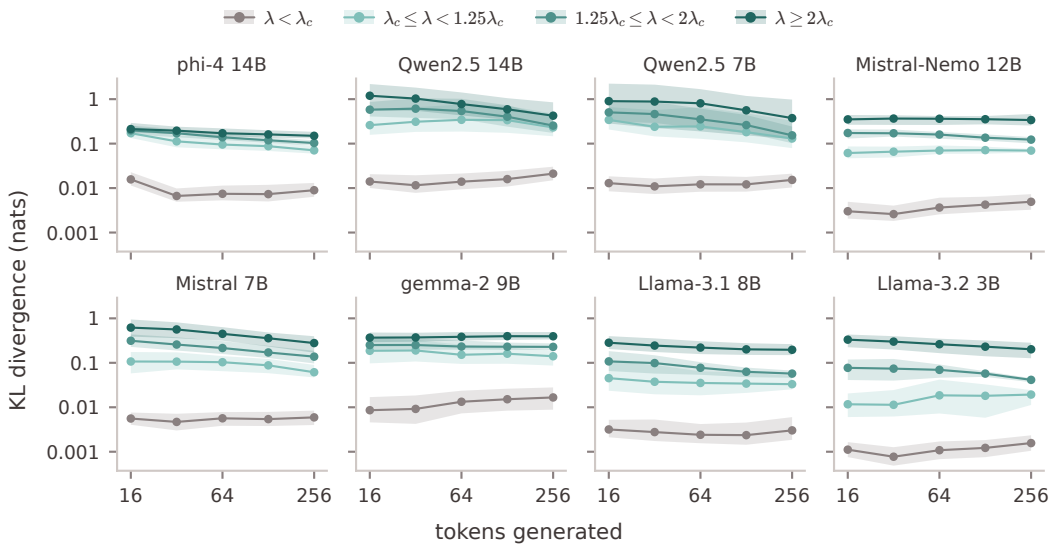


Figure 14: Next-token KL divergence by model, comparing steering at every position with steering only at the current one while keeping tokens fixed. Each panel pools concepts and steering layers, grouped by λ/λ_c, using the same answers at all five prefix lengths from 16 to 256 tokens. Markers show medians over answers generated over three seeds per question, with 95% confidence bands.