# SOFTSPHERE: UNDERSTANDING AND MITIGATING COLLAPSE IN LATENT REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Latent reasoning feeds continuous representations back into a language model, but repeated feedback can reduce their diversity. We study and characterize this representation collapse. Our analysis shows how contracting feedback suppresses trajectory differences and how nondegenerate token sampling can replenish variation. Inspired by these observations, we propose SoftSphere, which projects each soft token onto a sphere whose radius matches the mean vocabulary embedding norm. This correction preserves direction and introduces no trainable parameters or auxiliary loss. We establish guarantees for preserving temporal effective rank and slowing state alignment. We evaluate SoftSphere on both token-anchored and noise-anchored latent reasoning modes across a variety of RL tasks, which show consistent improvements in both final reward and diversity. SoftSphere also induces fewer repetitions in the reasoning chain and greater trajectory effective rank. These findings support controlling geometry as a simple way to improve the usefulness of RL rollouts in latent reasoning.

## 1 INTRODUCTION

Latent reasoning allows language models carry intermediate computation through continuous representations. Soft-token methods feed mixtures of token embeddings back into the model. Two common modes support exploration during reinforcement learning (RL): *token-anchored* feedback blends soft tokens with sampled-token embeddings (Yue et al., 2025b), while *noise-anchored* feedback injects noise into mixture weights or embeddings (Zheng et al., 2025; Butt et al., 2025). Their benefit depends on how the representations evolve: additional computation can be wasted if successive steps or sampled attempts become too similar.

We study *representation collapse* in this feedback process. Our measurements distinguish three symptoms: additional thinking steps add little representational variation, small perturbations fade over repeated forward passes, and diversity across sampled generations can vanish during RL (Section 2). We quantify variation along a trajectory using covariance effective rank, which measures how broadly that variation is spread across directions. Terminal-reward RL provides only indirect supervision of these dynamics, since latent steps receive no supervised next-token targets.

The intuition is a feedback loop: if each pass dampens differences between states, feeding its output back repeatedly compounds that damping. Our analysis formalizes this loss of variation under uniform contraction (Section 3). Sampling discrete tokens can reintroduce differences at each step when the sampling distributions remain nonconcentrated; deterministic mixture feedback lacks this source of variation, while stochastic latent methods can supply it through noise. This explains a possible mechanism for loss of diversity, distinct from saturation of temporal effective rank.

We propose SoftSphere, a simple correction that projects each nonzero soft token onto a sphere, $\Pi_R(s_t) = Rs_t/\|s_t\|_2$, with radius $R$ equal to the mean norm of vocabulary embeddings. The projection is compatible with both token-anchored and noise-anchored feedback, preserves the token's direction, and adds no trainable parameters or auxiliary loss. Intuitively, when successive states shrink, later directions contribute less to the trajectory's variation. Restoring their magnitude prevents this radial decay from suppressing the available directional variation. For radially contracting updates with radius-independent direction dynamics and sufficient angular diversity, we prove that projection preserves a fixed fraction of the available angular dimension, while unprojected effective



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.

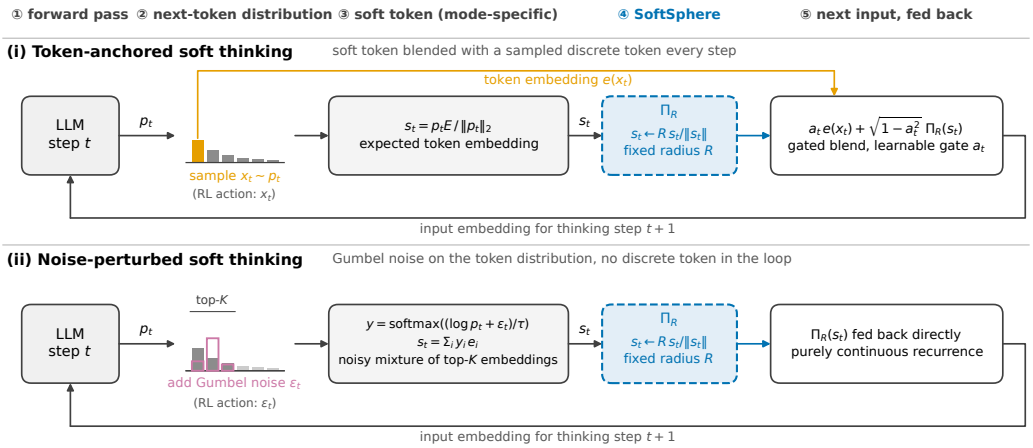


Figure 1: **Where SoftSphere enters the latent-thinking loop.** Each thinking step maps the next-token distribution $p_t$ to a soft token $s_t$ that is fed back as the next input. The two rollouts differ only in step ③: **(i) token-anchored:** $s_t$ is the expected token embedding, gated with the embedding of a sampled token $x_t$ (the RL action); **(ii) noise-anchored:** Gumbel noise $\epsilon_t$ (the RL action) re-weights the top-$K$ support and the mixture is fed back directly. SoftSphere inserts $\Pi_R(s_t) = R\, s_t/\|s_t\|$ at the same point in both, fixing the norm while leaving the direction and the RL action unchanged.

rank has a dimension-independent bound. We also show that projection slows alignment between states (Section 4).

We evaluate SoftSphere in both modes on Qwen2.5-1.5B-Instruct and Llama-3.2-1B-Instruct across mathematical reasoning, code generation, text-to-SQL, and instruction following. Across model families and tasks, projection improves pass@8 by 1.63 percentage points on average, compared with 0.48 points at pass@1 (Table 5). Indeed, the larger gain at $k = 8$ indicates that additional attempts become more useful. Projection also increases temporal effective rank in paired rollout probes. These results support preserving variation as a practical way to improve latent reasoning. Our main contributions are:

- **Characterizing collapse.** We identify temporal rank saturation, fading input sensitivity, and loss of rollout diversity, and analyze how contraction and token sampling affect latent variation.
- **Provable mitigation.** We introduce SoftSphere, a fixed-radius projection with no added parameters, and establish conditional guarantees for preserving temporal effective rank and slowing state alignment.
- **Experimental validation.** We evaluate both latent reasoning modes across two model families and five tasks, showing improved sampling gains and higher effective rank and less repetition.

## 2 Representation Collapse in Latent Reasoning

The collapse phenomenon is especially prominent for modeling during RL, where unlike supervised training, there is no direct supervision for the next generation step to counteract any representation contraction. Only a reward can be offered at the end of generation, and the reward is often sparse.

Furthermore, latent modeling can be especially susceptible to collapse during RL, where there is either no or little signal from discretization into the token space, which can interrupt collapse. We discuss this further in Section 3.

Figure 2 summarizes three observables: diversity across thinking steps, sensitivity to small changes in latent inputs, and diversity across generations of the same prompt. All measurements concern $D$-dimensional soft-token embeddings across complete forward passes, rather than individual layers. In the HRPO and pure-latent implementations, the soft token is $s_t = \sum_v p_t(v) e_v / \|p_t\|_2$; it is fed back directly in pure-soft mode and through the gated blend in HRPO (Section 5.2). Covariance diagnostics use the stored soft component, before the HRPO blend.

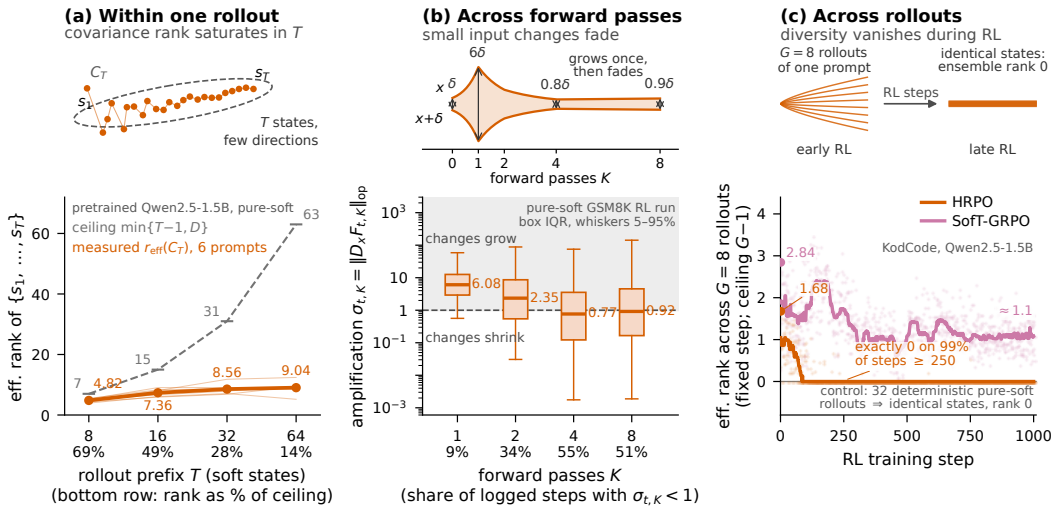


Figure 2: **Three modes of representation collapse in soft-token reasoning. (a)** Within one rollout, the effective rank $r_{\text{eff}}(C_T) = (\text{Tr}C_T)^2/\text{Tr}(C_T^2)$ of the covariance of the first $T$ soft states saturates near 9 while its ceiling $\min\{T-1, D\}$ grows from 7 to 63 (pretrained Qwen2.5-1.5B-Instruct, pure-soft, GSM8K prompts; faint lines: individual prompts). **(b)** The largest local amplification $\sigma_{t,K}$ of a small input change is estimated by power iteration with context fixed. Recorded mean estimates have medians 6.08 after one pass and 0.77, 0.92 after $K{=}4, 8$; boxes: interquartile range, whiskers: 5–95% over 1000 logged training steps of a pure-soft GSM8K run; bottom row: share of steps with $\sigma_{t,K} < 1$. **(c)** Across $G{=}8$ rollouts of one prompt at a fixed thinking step, the ensemble effective rank, averaged over prompts and valid thinking steps, collapses during RL: HRPO falls from 1.68 to exactly 0 on 99% of steps after step 250, SofT-GRPO from 2.84 to ≈1 (dots: logged values; lines: 25-step rolling median). Deterministic pure-soft rollouts have zero ensemble variance by construction.

These observations distinguish limited variation within a rollout from loss of diversity across rollouts; local sensitivity estimates do not establish uniform contraction. Effective rank is reported as zero when variance is numerically negligible. Appendix B provides metric definitions, probe protocols, and detailed results.

## 3 Mechanisms of Collapse and the Role of Discretization

Rank collapse has been studied across transformer depth (Dong et al., 2021; Noci et al., 2022) and in bootstrapped value learning (Kumar et al., 2021). These settings motivate our analysis but differ from temporal covariance saturation under latent feedback and policy-gradient RL. We hypothesize that terminal, and often sparse, rewards leave contracting latent dynamics insufficiently constrained because intermediate steps lack supervised next-token targets.

**Continuous feedback compounds contraction.** Consider an idealized autonomous recurrence $h_{t+1} = F(h_t)$ on a closed invariant subset of $\mathbb{R}^D$, with $\|F(u) - F(v)\| \le a\|u - v\|$ for all states and $0 < a < 1$. For an ensemble of trajectories with covariance $\Sigma_t$,

$$\|h_T^{(g)} - h_T^{(g')}\| \le a^T \|h_0^{(g)} - h_0^{(g')}\|, \qquad \text{Tr}\Sigma_T \le a^{2T}\text{Tr}\Sigma_0. \tag{3.1}$$

Trajectories approach a fixed point, so later steps contribute diminishing deviations to temporal covariance. This gives a possible mechanism for saturation, not a prediction of effective rank: participation ratio is scale invariant. The uniform contraction assumption also abstracts away the growing transformer context and is stronger than the local attenuation estimated in Section 2.

**Token sampling can counteract collapse.** Replacing sampled tokens with continuous mean feedback removes a source of stochastic variation, making latent trajectories more susceptible to representation collapse under contracting dynamics.

To formalize this distinction, compare mean feedback $\mu(p) = \sum_i p_i e_i$ with sampled-token feedback $e_I$, where $I \sim p$. For arbitrary token embeddings, sampling introduces conditional variance

$$\mathbb{E}_{I\sim p}\|e_I - \mu(p)\|^2 = \frac{1}{2}\sum_{i,j} p_i p_j \|e_i - e_j\|^2. \tag{3.2}$$

Assume distinct embeddings are separated by at least $\delta > 0$ and all reachable next-token distributions satisfy $\|p\|_2^2 \le \kappa < 1$. For two rollouts whose next embeddings $X_{t+1}, Y_{t+1}$ are sampled independently conditional on their joint history, Cauchy–Schwarz bounds the collision probability by $\kappa$, yielding

$$\mathbb{E}\|X_{t+1} - Y_{t+1}\|^2 \ge \delta^2(1-\kappa). \tag{3.3}$$

Nonconcentrated sampling can thus counteract multiplicative loss of ensemble spread, without guaranteeing temporal rank growth or retention of initial-state information.

## 4 Mitigating Collapse with SoftSphere

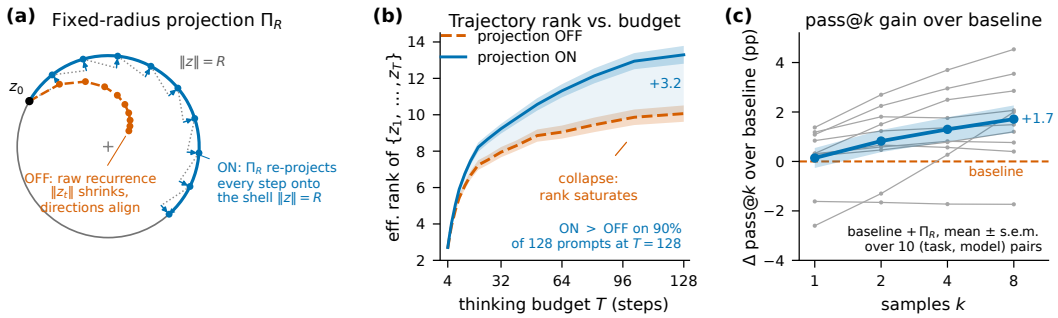


Figure 3: **SoftSphere stabilizes soft-token magnitude.** **(a)** Fixed-radius projection preserves each soft token's direction. **(b)** Projection sustains higher temporal effective rank as the thinking budget grows. **(c)** Average pass@$k$ gains over ten task–model pairs increase with the number of sampled attempts; individual effects vary.

This motivates us to study the role of discrete-token structure. First, we conduct a simple experiment: replace the model-sampled token in HRPO's gated blend with a uniformly *random* vocabulary token at each thinking step. On GSM8K with Qwen2.5-1.5B-Instruct, random-token feedback achieves 61.8% pass@8, compared with 34.7% for pure latent feedback and 66.8% for model-sampled tokens. These results suggest that discrete-token feedback supplies useful structure, even when the token does not encode contextual information. Further details and results can be found in Appendix D.

Motivated by this observation, we test whether constraining soft-token magnitude mitigates the covariance saturation observed in Section 2. We propose SoftSphere, which projects each soft token onto a sphere whose radius equals the mean norm of the current vocabulary embeddings:

$$R = \frac{1}{V}\sum_{v=1}^{V}\|e_v\|_2, \qquad \widetilde{s}_t = \Pi_R(s_t) := R\frac{s_t}{\|s_t\|_2}. \tag{4.1}$$

For nonzero states, projection fixes the norm while preserving the current direction. It acts on the soft component before the HRPO gated blend and directly on the feedback input in pure-soft and SofT-GRPO rollouts. We recompute $R$ once per rollout batch; the numerical implementation is specified in Appendix A.2.

**Provable benefits.** We compare noiseless rollouts $h_{t+1} = F_t(h_t)$ and $\widetilde{h}_{t+1} = \Pi_R(F_t(\widetilde{h}_t))$, initialized at the same point on the sphere of radius $R$. The following informal theorems summarize two guarantees; exact statements, assumptions, constants, and proofs appear in Appendix F.

**Theorem A** (Informal: projection increases effective rank)**.** *Suppose each update multiplies the state norm by a time-dependent factor $a_t \in (0, a]$, where $a < 1$, and its direction update is independent*

Table 1: Temporal effective rank with and without projection for pretrained Qwen2.5-1.5B-Instruct on six GSM8K test prompts (pure-soft feedback, temperature 0.7, float32). Budgets use prefixes of 64-step rollouts. The final column reports the paired improvement (mean ± standard error across prompts).

| Steps $T$ | Unprojected | Projected | Improvement |
|---|---|---|---|
| 8 | 4.82 | 5.15 | $0.34 \pm 0.05$ |
| 16 | 7.36 | 8.15 | $0.80 \pm 0.15$ |
| 32 | 8.56 | 9.93 | $1.37 \pm 0.25$ |
| 64 | 9.04 | 11.63 | $2.60 \pm 0.69$ |

*of radius. Assume the covariance of the shared unit directions has rank $d \geq 1$, with the ratio of its smallest to largest positive eigenvalue at least $\eta \in (0, 1]$. For $T \geq 2/(1-a)$, the temporal covariance effective ranks satisfy*

$$r_{\text{raw}} \leq K, \qquad r_{\text{projected}} \geq \eta d. \tag{4.2}$$

*Here $K$ depends only on $a$, independently of dimension and rollout length. In particular, projection strictly increases effective rank whenever $\eta d > K$.*

Radial contraction makes early states dominate covariance; projection restores equal magnitudes without creating new directions. The proof bounds the raw covariance's trace relative to its leading eigenvalue and identifies the projected covariance as $R^2$ times the covariance of the shared unit directions (Theorem A).

**Theorem B** (Informal: projection slows temporal alignment)**.** *Let $F(h) = ah + (1-a)h_\star$, where $0 < a < 1$ and $0 < \|h_\star\|_2 < R$. Assume the common initial state makes an acute, nonzero angle with $h_\star$. For any fixed lag $\ell \geq 1$, let $\rho_t$ and $\widetilde{\rho}_t$ denote the cosine similarities between states at steps $t$ and $t + \ell$ without and with projection, respectively. There exists $q \in (a, 1)$ such that, as $t \to \infty$,*

$$1 - \rho_t = \Theta(a^{2t}), \qquad 1 - \widetilde{\rho}_t = \Theta(q^{2t}). \tag{4.3}$$

*Thus projection strictly slows temporal alignment.*

Maintaining radius $R$ reduces the relative pull toward $h_\star$, slowing but not preventing alignment. The proof linearizes the angular dynamics and uses $1 - \cos\theta \sim \theta^2/2$ to convert angular decay factors $a$ and $q > a$ into the stated cosine-similarity rates.

Theorem B gives the exact rates and also characterizes the mean cosine similarity over a growing prefix: its gap from one decays as $1/T$ in both cases, with a larger leading constant under projection.

**Projection increases temporal effective rank.** We compare paired rollouts with fixed model weights on the same six GSM8K prompts used in Section 2. Table 1 reports effective rank with and without projection, together with the mean paired difference and its standard error across prompts. Projection increases effective rank on all six prompts at every tested budget. The mean improvement grows from $0.34 \pm 0.05$ at $T = 8$ to $2.60 \pm 0.69$ at $T = 64$, supporting reduced covariance saturation over these horizons.

The correction is modest in magnitude in the measured training runs: mean projection gains $R/\|s_t\|_2$ are 1.037 for Qwen on Numina and 1.025 for Llama on KodCode. These gains quantify the per-step correction along the projected trajectories.

## 5 Experimental setup

### 5.1 Datasets and tasks

**GSM8K.** GSM8K evaluates multi-step arithmetic reasoning through grade-school word problems (Cobbe et al., 2021). We train on GSM8k-Aug, a synthetic augmentation with structured arithmetic solution steps (Deng et al., 2024), and use final-answer correctness as the reinforcement

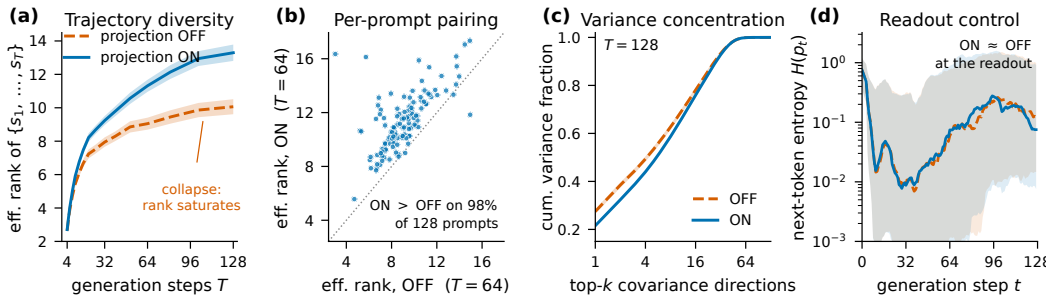


Figure 4: **token evolution during generation. Soft-token generation evolution with SoftSphere $\Pi_R$ ON vs. OFF** (Qwen2.5-1.5B-Instruct, 128 GSM8K prompts, soft-only recurrence, temperature 0.7). **(a) Trajectory diversity.** Effective rank of the across-step covariance of the soft tokens: without projection the rank saturates near 10 of a possible 127– the trajectory stops exploring new directions, while projection sustains growth. **(b) Per-prompt pairing.** Paired per-prompt ranks at the production thinking budget $T$=64: ON exceeds OFF on 97.7% of prompts ($+2.27 \pm 0.16$). **(c) Variance concentration.** The raw recurrence concentrates trajectory variance on fewer covariance directions. **(d) Readout distribution.** The per-step next-token entropy is unchanged by projection–$\Pi_R$ improves the latent geometry without perturbing the emitted token distributions.

learning (RL) reward. **Numina.** We use NuminaMath-CoT, a mathematical reasoning corpus spanning school-level exercises and competition problems, with chain-of-thought solutions (Li et al., 2024). RL rewards correctness of the extracted final answer. **Instruction following.** We use prompts that combine up to five automatically verifiable output constraints, including length, keyword, and formatting requirements (Pyatkin et al., 2025). The RL reward is the fraction of constraints satisfied. **KodCode.** We study Python code generation from natural-language specifications using the easy-difficulty subset of KodCode-Light-RL-10K (Xu et al., 2025). A binary RL reward indicates whether the generated program passes all associated unit tests. **WikiSQL.** WikiSQL pairs natural-language questions with SQL queries over individual Wikipedia tables (Zhong et al., 2017). Given a question, table title when available, and column names and types, the model generates the canonical SQL query; the RL reward is exact string match after case and whitespace normalization and removal of trailing semicolons.

### 5.2 Training pipeline

We fine-tune Qwen2.5-1.5B-Instruct and Llama-3.2-1B-Instruct directly with RL, without supervised warm-up. We compare token-anchored (HRPO) and noise-anchored (SofT-GRPO) reasoning with SoftSphere (Equation (4.1)) enabled or disabled, holding the initial checkpoint, data, seed, reward, and optimization settings fixed within each pair.

**Latent rollouts.** Given the next-token distribution $p_t$, HRPO (Yue et al., 2025b) forms a probability-weighted embedding mixture, normalizes it by $\|p_t\|_2$, and blends it with a sampled-token embedding through learned gates. SofT-GRPO (Zheng et al., 2025) directly feeds back a convex embedding mixture with Gumbel-perturbed weights. Both modes switch to discrete answer generation at the task's answer marker; SofT-GRPO enforces a 128-step reasoning cap.

**Policy optimization.** We optimize clipped GRPO (Shao et al., 2024) using $G = 8$ completions per prompt, group-relative verifier-reward advantages, clipping width 0.2, one update per rollout batch, and no reference-policy KL penalty. Updates replay stored soft tokens and reasoning masks as fixed inputs, recomputing policy scores and HRPO's gated blends. Gradients update the policy and active gates; generated soft states and their projection remain detached. HRPO uses categorical token log-probabilities; SofT-GRPO uses a clipped, masked mean Gumbel log-density estimator on the stored support for reasoning steps and categorical log-probabilities for answers.

Appendix A gives exact rollout formulas, implementation settings, and evaluation details; Appendix C reports projection ablations.

Table 2: **Pass@$k$ across tasks and rollout mechanisms, Qwen2.5-1.5B-Instruct.** pass@1 and pass@8 (%, $n$=8 samples, temperature 0.8, top-$p$ 0.95) on 1,500-problem held-out splits (GSM8K: full 1,319-problem test set) for GRPO (discrete-token baseline), HRPO (gated latent reasoning), and SofT-GRPO, with the fixed-radius projection $\Pi_R$ OFF and ON ( shaded : SoftSphere, denoted $\Delta_{\Pi_R}$). $\Delta$ columns give the ON − OFF difference for each pair (green: projection helps, orange: hurts). **Bold** marks the best method per row and metric. †RL training collapsed (near-zero-accuracy policy). "–": not evaluated in this configuration.

| | GRPO | | HRPO | | HRPO+$\Pi_R$ | | $\Delta_{\Pi_R}$ | | SofT-GRPO | | SofT-GRPO+$\Pi_R$ | | $\Delta_{\Pi_R}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task | @1 | @8 | @1 | @8 | @1 | @8 | @1 | @8 | @1 | @8 | @1 | @8 | @1 | @8 |
| GSM8K | 37.4 | 70.2 | 37.2 | 69.7 | **39.2** | **73.1** | +2.0 | +3.4 | 35.6 | 65.3 | 37.1 | 67.2 | +1.5 | +1.9 |
| NuminaMath | 22.5 | 45.1 | 23.7 | 45.6 | **24.4** | **47.4** | +0.7 | +1.8 | 14.6 | 38.3 | 17.2 | 41.4 | +2.6 | +3.1 |
| KodCode | 34.6 | 60.4 | 39.3 | 64.9 | 40.1 | **65.9** | +0.8 | +1.0 | 31.5 | 60.2 | **41.2** | 65.2 | +9.7 | +5.0 |
| WikiSQL | 64.5 | 70.2 | 66.4 | 71.1 | 64.8 | 69.3 | −1.6 | −1.8 | **68.1** | **72.4** | 67.5 | 71.6 | −0.6 | −0.8 |
| IF | 25.1 | 27.4 | 25.8 | 28.4 | 27.2 | 32.9 | +1.4 | +4.5 | 30.3 | 35.3 | **31.3** | **37.7** | +1.0 | +2.4 |

Table 3: **Pass@$k$ across tasks and rollout mechanisms, Llama-3.2-1B-Instruct.** Same setup and notation as Table 2.

| | GRPO | | HRPO | | HRPO+$\Pi_R$ | | $\Delta_{\Pi_R}$ | | SofT-GRPO | | SofT-GRPO+$\Pi_R$ | | $\Delta_{\Pi_R}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task | @1 | @8 | @1 | @8 | @1 | @8 | @1 | @8 | @1 | @8 | @1 | @8 | @1 | @8 |
| GSM8K-Aug | 67.9 | **89.3** | 69.5 | 86.5 | **70.7** | 88.5 | +1.2 | +2.0 | 57.2 | 72.4 | 53.9 | 69.1 | −3.3 | −3.3 |
| NuminaMath | 21.7 | **44.5** | 19.7 | 40.0 | 20.7 | 43.5 | +1.0 | +3.5 | 19.73 | 40.84 | **22.1** | 44.7 | +2.4 | +3.9 |
| KodCode | 44.5 | 64.3 | 44.4 | 64.2 | 44.6 | **64.6** | +0.2 | +0.4 | 44.5 | 64.1 | **45.1** | 63.4 | +0.6 | −0.7 |
| WikiSQL | 71.5 | **77.1** | 70.5 | 73.3 | 70.7 | 74.5 | +0.2 | +1.2 | **72.3** | 75.8 | 70.6 | 73.8 | −1.7 | −2.0 |
| IF | 26.3 | 29.7 | 28.1 | 35.7 | 25.5 | 37.7 | −2.6 | +2.0 | 23.5 | 24.5 | **33.6** | **41.1** | +10.1 | +16.6 |

# 6 SoftSphere improves rollout reward and diversity

## 6.1 Reward Improvements

Manifold projection generally yields accuracy gains in both latent reasoning modes (Tables 2–4). Note that WikiSQL shows no consistent benefit, plausibly reflecting its short reasoning chains: GRPO models average just 6 thought tokens, versus 287 on NuminaMath, leaving less scope for projection to help.

## 6.2 Diversity Gains

We quantify the value of additional attempts using the *sampling gain*, $g_k = \text{pass@}k - \text{pass@}1$, where pass@$k$ measures success in at least one of $k$ samples. Across 16 matched HRPO run pairs, SoftSphere improves pass@8 by 1.63 percentage points on average, about 3.4 times the 0.48-point gain at pass@1. Sampling gain increases by 1.15 points on average. The pass@8 and sampling-gain improvements are statistically significant; the pass@1 change is not (Table 5). For the 10 HRPO pairs evaluated with 16 samples (Table 4), pass@16 improves by 2.04 points, while the 1.33-point increase in $g_{16}$ is not statistically significant.

Effects vary across tasks and models (Tables 2 and 3). Mathematics and code benefit substantially from multiple attempts. On instruction following, Llama HRPO gains 2.0 points at pass@8 despite losing 2.6 points at pass@1, illustrating improved coverage without higher single-sample accuracy. WikiSQL has smaller sampling gains, and Qwen HRPO declines at both metrics.

These findings are consistent with the conditional guarantees for preserving temporal effective rank and slowing state alignment (Theorems A and B): retaining variation may help stochastic rollouts explore alternative solutions. The guarantees concern within-rollout geometry, however, and do not imply diversity across generations or higher pass@$k$.

Table 4: **Effect of the fixed-radius projection $\Pi_R$.** pass@16 (%) for HRPO and SofT-GRPO with the projection OFF and ON ( shaded : SoftSphere), matched by model, task, and training seed. $\Delta$ columns give the ON − OFF difference (green: projection helps, orange: hurts) with $p$-values from exact McNemar tests on discordant problems. "–": run collapsed or not evaluated.

| Task | HRPO | HRPO+$\Pi_R$ | $\Delta_{\Pi_R}$ ($p$) | SofT-GRPO | SofT-GRPO+$\Pi_R$ | $\Delta_{\Pi_R}$ ($p$) |
|---|---|---|---|---|---|---|
| *Llama-3.2-1B-Instruct* | | | | | | |
| GSM8K-Aug | 89.5 | 91.2 | +1.7 (.022) | 75.8 | 73.0 | −2.8 (.017) |
| IF | 37.3 | 40.2 | +2.9 (.007) | 24.5 | 42.4 | +17.9 (<.001) |
| KodCode | 65.5 | 69.4 | +3.9 (<.001) | 70.7 | 71.4 | +0.7 (.064) |
| NuminaMath | 47.1 | 49.7 | +2.6 (.020) | 42.7 | 46.1 | +3.4 (.011) |
| WikiSQL | 74.0 | 74.9 | +0.9 (.219) | 76.3 | 75.1 | −1.2 (.056) |
| *Qwen2.5-1.5B-Instruct* | | | | | | |
| GSM8K | 77.3 | 79.5 | +2.1 (.056) | 74.5 | 74.7 | +0.2 (.893) |
| IF | 29.9 | 33.9 | +4.0 (<.001) | 36.7 | 39.3 | +2.6 (.003) |
| KodCode | 69.8 | 71.7 | +1.9 (.042) | 67.5 | 72.7 | +5.2 (<.001) |
| NuminaMath | 50.1 | 52.0 | +1.9 (.080) | 44.2 | 47.5 | +3.3 (.009) |
| WikiSQL | 72.0 | 70.5 | −1.5 (.014) | 73.2 | 72.5 | −0.7 (.324) |

Table 5: **Paired effect of the SoftSphere projection $\Pi_R$ on best-of-$n$ sampling.** Differences (ON − OFF, percentage points) for HRPO pairs matched by model, task, and training seed, evaluated at temperature 0.8. Gain is pass@$k$ − pass@1, with both estimated from the same $n$ samples. The $n$=16 block summarizes the HRPO pairs in Table 4. Statistics use unrounded scores; $p$-values are two-sided Wilcoxon signed-rank tests across pairs within each block.

| Metric | Mean $\Delta$ | Median $\Delta$ | Wilcoxon $p$ |
|---|---|---|---|
| pass@1 | +0.48 | +0.65 | 0.23 |
| pass@8 | +1.63 | +2.03 | **0.011** |
| pass@16 | +2.04 | +2.00 | **0.006** |

### 6.3 Trajectory analysis

We measure within-rollout repetition using rep-4, the fraction of token 4-grams that repeat an earlier occurrence. On 128 held-out NuminaMath problems with Qwen2.5-1.5B-Instruct, HRPO exhibits significantly more repetition than SoftSphere (Table 6). Projection reduces rep-4 by 0.95 percentage points (paired 95% bootstrap CI: $[-1.85, -0.02]$), despite longer average completions. Removing latent feedback yields no significant change in repetition relative to HRPO. Appendix E provides the metric definition and experimental protocol.

Figure 5 shows a matched pair of completions: HRPO repeats an incorrect derivative calculation, while SoftSphere answers correctly using 42% fewer tokens. The pair was selected for legibility among cases with a large rep-4 gap and does not establish the behavior's prevalence.

## 7 Related Work

**Latent reasoning.** Implicit CoT internalizes reasoning through a curriculum (Deng et al., 2024), Coconut feeds back hidden states (Hao et al., 2024), and Soft Thinking uses probability-weighted embeddings without further training (Zhang et al., 2025). RL extensions introduce embedding noise (Butt et al., 2025), gated discrete–continuous feedback (HRPO) (Yue et al., 2025b), or Gumbel-reparameterized mixtures (SofT-GRPO) (Zheng et al., 2025). We add fixed-radius projection to the latter two mechanisms, retaining their reward objectives.

**Collapse in latent reasoning.** Soft thinking can be dominated by its highest-probability token (Wu et al., 2025); SeLaR counters this with entropy-gated feedback and directional correction (Fu & Luo, 2026). SofT-GRPO attributes training collapse with diffuse soft inputs to distribution shift (Zheng et al., 2025). Most closely, WIDER diagnoses temporal rank collapse and adds spectral regularization to supervised and distilled implicit-CoT training (Hao & Yang, 2026). We distinguish temporal

Table 6: **Verbatim repetition and completion length.** Results for Qwen1.5B NuminaMath on 128 held-out problems. Tokens denotes mean completion length. Projection reduces repetition despite longer completions.

| Run | rep-4 | Δ rep-4 (95% CI) | Tokens |
|---|---|---|---|
| HRPO | 0.2951 | — | 524.5 |
| HRPO $+\Pi_R$ (SoftSphere) | **0.2856** | $+0.0095$ $[-0.0185, -0.0002]$ | 549.0 |
| no latent feedback | 0.2941 | $+0.0010$ $[-0.0097, +0.0078]$ | 573.0 |

**Prompt.** Find the 2017th derivative of $\cos x$. A: $-\sin x$; B: $-\cos x$; C: $\sin x$; D: $\cos x$.
**HRPO** (rep-4 $= 0.55$): *So,* $f_1(x) = -\cos x$. … *Now, we find the second derivative of* $f_2(x)$ … [derivative block repeats 4×] … $f_{2017}(x) = f_1(x) = -\cos x$. #### **B** ×
**HRPO +** $\Pi_R$ (rep-4 $= 0.23$): *We observe that the sequence repeats every 4 derivatives* … $2017 \equiv 1 \pmod 4$, *therefore* $f_{2017}(x) = f_1(x) = -\sin x$. #### **A** ✓

Figure 5: **What rep-4 measures.** Abridged prompt and completion excerpts for the same held-out NuminaMath problem and seed.

covariance saturation, perturbation attenuation, and loss of diversity across generations under RL. SoftSphere controls magnitude while preserving directions, without an auxiliary loss; its rank guarantee assumes sufficient angular diversity.

**Representation collapse and recurrent dynamics.** Related degeneracies include token uniformity across depth in pure attention (Dong et al., 2021) and feature-rank loss in bootstrapped value learning (Kumar et al., 2021). STARS regularizes Jacobian spectral radii and varies training loop counts to stabilize depth-recurrent models (Yang et al., 2026). We analyze autoregressive soft-token feedback under policy gradients, distinguishing ensemble contraction, temporal rank saturation, and loss of initial-state information; convergence need not imply ineffective reasoning.

**Hyperspherical normalization.** nGPT normalizes parameter vectors and activations to unit norm throughout the transformer, including activation normalization at inference (Loshchilov et al., 2025). SoftSphere projects soft-token feedback during training rollouts and evaluation, with radius matched to vocabulary embeddings. HRPO's $1/\|p_t\|_2$ rescaling (Yue et al., 2025b) does not fix embedding norms. We control these directly, with conditional guarantees for temporal effective rank and alignment (Section 4).

**Overthinking.** Extended explicit reasoning can repeat strategies with little benefit (Chen et al., 2024) or lose accuracy as chains lengthen, motivating effort control (Yang et al., 2025). SoftSphere changes feedback dynamics without a length penalty or stopping rule. Its reduced repetition despite longer completions separates repetition from length (Section 6.3).

**RL exploration and sample coverage.** DAPO counters entropy collapse through asymmetric clipping (Yu et al., 2025), while RL can improve pass@1 yet reduce pass@$k$ at large $k$ (Yue et al., 2025a). Soft-token RL also reports larger gains with multiple attempts (Butt et al., 2025; Zheng et al., 2025). Our geometry correction complements policy-level exploration: its larger aggregate pass@8 gains support better coverage, while temporal rank, policy entropy, and diversity across generations remain distinct.

## 8 Conclusion

We characterized representation collapse in latent reasoning and introduced SoftSphere, a fixed-radius projection of soft-token feedback. Our analysis provides conditional guarantees for preserving temporal effective rank and slowing alignment. Experiments show higher temporal rank, less repetition, and stronger aggregate HRPO gains at pass@8 than pass@1. These findings suggest that preserving latent variation can make additional sampled attempts more useful.

## REFERENCES


Natasha Butt, Ariel Kwiatkowski, Ismail Labiad, Julia Kempe, and Yann Ollivier. Soft tokens, hard truths, 2025. URL https://arxiv.org/abs/2509.19170.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? on the overthinking of o1-like LLMs, 2024. URL https://arxiv.org/abs/2412.21187.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit CoT to implicit CoT: Learning to internalize CoT step by step, 2024. URL https://arxiv.org/abs/2405.14838. Dataset: https://huggingface.co/datasets/zen-E/GSM8k-Aug.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803. PMLR, 2021. URL https://proceedings.mlr.press/v139/dong21a.html.

Renyu Fu and Guibo Luo. SeLaR: Selective latent reasoning in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7073–7087. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.320. URL https://aclanthology.org/2026.acl-long.320/.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space, 2024. URL https://arxiv.org/abs/2412.06769.

Yuwen Hao and Menglin Yang. Think wider: Mitigating latent rank collapse in implicit chain-of-thought reasoning, 2026. URL https://arxiv.org/abs/2609.07406.

Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=O9bnihsFfXU.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath. Hugging Face dataset repository, 2024. URL https://huggingface.co/datasets/AI-MO/NuminaMath-CoT.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. nGPT: Normalized transformer with representation learning on the hypersphere. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=se4vjm7h4E.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27198–27211, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ae0cba715b60c4052359b3d52a2cff7f-Abstract-Conference.html.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1645. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/46499a0622ecf568b72d17b61e45dbd5-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track. Dataset: https://huggingface.co/datasets/allenai/IF_multi_constraints_upto5.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Junhong Wu, Jinliang Lu, Zixuan Ren, Gangqiang Hu, Zhi Wu, Dai Dai, and Hua Wu. LLMs are single-threaded reasoners: Demystifying the working mechanism of soft thinking, 2025. URL https://arxiv.org/abs/2508.03440.

Zhangchen Xu, Yang Liu, Yueqin Yin, Mingyuan Zhou, and Radha Poovendran. KodCode: A diverse, challenging, and verifiable synthetic dataset for coding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6980–7008, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.365. URL https://aclanthology.org/2025.findings-acl.365/. Dataset: https://huggingface.co/datasets/KodCode/KodCode-Light-RL-10K.

Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2502.18080.

Xiao-Wen Yang, Ziyu Han, Xi-Hua Zhang, Wen-Da Wei, Jie-Jing Shao, Lan-Zhe Guo, and Yu-Feng Li. Stabilizing recurrent dynamics for test-time scalable latent reasoning in looped language models, 2026. URL https://arxiv.org/abs/2605.26733. ICML 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, 2025a. URL https://arxiv.org/abs/2504.13837.

Zhenrui Yue, Bowen Jin, Huimin Zeng, Honglei Zhuang, Zhen Qin, Jinsung Yoon, Lanyu Shang, Jiawei Han, and Dong Wang. Hybrid latent reasoning via reinforcement learning, 2025b. URL https://arxiv.org/abs/2505.18454.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of LLMs in continuous concept space, 2025. URL https://arxiv.org/abs/2505.15778.

Zhi Zheng, Yu Gu, Wei Liu, Yee Whye Teh, and Wee Sun Lee. SofT-GRPO: Surpassing discrete-token LLM reinforcement learning via gumbel-reparameterized soft-thinking policy optimization, 2025. URL https://arxiv.org/abs/2511.06411.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2SQL: Generating structured queries from natural language using reinforcement learning, 2017. URL https://arxiv.org/abs/1709.00103.

# A Further training and evaluation details

## A.1 Training details

We finetune the full model with AdamW, using a base learning rate of $2 \times 10^{-6}$ for GSM8k-Aug, Numina, and instruction following, and $10^{-6}$ for KodCode and WikiSQL. HRPO gate matrices and retention parameters use separate learning rates of $10^{-4}$ and $10^{-3}$, respectively. Training uses a cosine learning-rate schedule with $10\%$ warm-up, weight decay $0.1$ on decayed parameter groups, and gradient-norm clipping at $0.1$. Computation uses bfloat16 mixed precision with float32 master weights and optimizer states. The projection setting is retained when generating evaluation rollouts.

## A.2 Rollout and optimization details

**Token-anchored feedback.** Let $e_v \in \mathbb{R}^D$ be token $v$'s input embedding and $p_t$ the next-token distribution conditioned on the prompt and preceding discrete and latent inputs. HRPO samples $y_t \sim p_t$ and constructs

$$s_t = \frac{\sum_{v=1}^{V} p_t(v) e_v}{\|p_t\|_2}. \tag{A.1}$$

The next input is the elementwise blend $a_t \odot e_{y_t} + \sqrt{1 - a_t^2} \odot (i_t \odot \widetilde{s}_t)$, where retention and input gates $a_t, i_t$ are computed from $e_{y_t}$, and $\widetilde{s}_t = \Pi_R(s_t)$ with projection enabled, otherwise $s_t$. Retention is initialized near one to favor discrete embeddings. Rollouts use temperature $0.7$ and the full vocabulary distribution.

**Noise-anchored feedback.** SofT-GRPO uses temperature 1, top-$k = 5$, and top-$p = 0.95$ filtering. We select the five largest entries of the filtered distribution as $S_t$, renormalize their probabilities, and form

$$w_t(v) = \frac{\exp\big((\log(p_t(v) + 10^{-6}) + \xi_{t,v})/\tau_g\big)}{\sum_{u \in S_t} \exp\big((\log(p_t(u) + 10^{-6}) + \xi_{t,u})/\tau_g\big)}, \qquad s_t = \sum_{v \in S_t} w_t(v) e_v, \tag{A.2}$$

with $\tau_g = 0.1$ and independent standard Gumbel samples $\xi_{t,v}$ clipped to $[-1.5, 3]$. The next input is $\widetilde{s}_t$ directly, without gates or $\|p_t\|_2$ normalization.

**Answer generation.** Prompts use the model's chat template and task-specific instructions separating reasoning from the final response. The answer marker is `####` for mathematics, instruction following, and SQL, and the opening Python code fence for KodCode. If an unfinished SofT-GRPO rollout reaches $128$ reasoning steps without the marker, we insert it to start discrete answer decoding. Rollouts may transition earlier or terminate at an end-of-sequence token. The task verifier scores the final response; completions without exactly one answer marker receive zero reward.

**Numerical projection.** The normalization denominator floors the soft-token root-mean-square (RMS) magnitude at $10^{-8}$; the rescaling factor is uncapped.

**Advantages and loss averaging.** For the $G = 8$ verifier rewards per prompt, we compute $A_g = (R_g - \overline{R})/(\mathrm{std}(R_1, \ldots, R_G) + 10^{-4})$. The clipped GRPO objective is averaged first over positions within each completion, then over completions. Replay and gradient handling follow Section 5.2.

## A.3 Evaluation protocol

**Held-out sets.** For NuminaMath, KodCode, WikiSQL and the IFEval-style set we hold out a fixed subset of the training split (seed 42, chosen before training, so evaluation rows are the complement of the trained rows); KodCode is first restricted to its `easy` rows. GSM8K uses its test split. Rows without a parseable gold answer (e.g. NuminaMath proofs without `\boxed{}`) are dropped identically at training and evaluation time.

**Generation.** Checkpoints are evaluated with the exact rollout they were trained with: the latent-thinking settings (answer marker, thinking budget, rollout mode, projection and its parameters) are recovered from the saved training arguments and the gate parameters are re-attached, so the same soft-token loop is used at test time; token-level GRPO baselines use standard sampling. Prompts use the training system prompt and chat template. For every problem we draw $n{=}8$ completions (16 where noted) at temperature $0.8$, top-$p$ $0.95$, with a fixed sampling seed and a per-dataset generation limit (512 tokens for GSM8K and WikiSQL, 768 for KodCode, 1024 for NuminaMath and IFEval-style).

**Scoring.** The prediction is the text after the model's answer marker and is judged by the same functions as the training reward: exact match after numeric normalisation (GSM8K); `\boxed{}` extraction with LaTeX normalisation and a guarded symbolic-equivalence check (NuminaMath); execution of the generated program against the problem's unit tests (KodCode); normalised logical-form match with the gold query (WikiSQL); and IFEval constraint checkers, requiring all constraints to hold (IFEval-style).

**Metric.** From the per-problem correct count $c$ out of $n$ samples we report the unbiased estimator pass@$k = 1 - \binom{n-c}{k} / \binom{n}{k}$ (Chen et al., 2021), averaged over problems. Because all runs are evaluated on the same problems with the same sampling seed, differences between runs are computed *paired* per problem, which removes problem-difficulty variance from the standard errors we report.

## B Additional collapse diagnostics

**Covariance effective rank saturates across generation steps.** For the first $T$ soft states of one rollout, define

$$\bar{s} = \frac{1}{T}\sum_{t=1}^{T} s_t, \qquad C_T = \frac{1}{T-1}\sum_{t=1}^{T}(s_t - \bar{s})(s_t - \bar{s})^\top, \qquad r_{\text{eff}}(C_T) = \frac{(\mathrm{Tr}C_T)^2}{\mathrm{Tr}(C_T^2)}. \tag{B.1}$$

This participation ratio has ceiling $\min\{T-1, D\}$; the implementation reports zero at numerically vanishing variance. We use *covariance saturation* to describe a leveling-off of effective rank as $T$ grows: additional states contribute diminishing spectral diversity despite the increasing rank ceiling. In a pure-soft probe of pretrained Qwen2.5-1.5B-Instruct on six GSM8K test prompts, doubling the rollout prefix from 32 to 64 steps raises mean effective rank only from $8.56$ to $9.04$, while its ceiling increases from 31 to 63 (unprojected results in Table 1).

**Small input changes can fade over repeated reasoning steps.** In a pure-soft GSM8K run, the recorded mean amplification estimates have medians of $6.08$, $2.35$, $0.77$, and $0.92$ after $K = 1, 2, 4, 8$ forward passes, respectively. Factors above one allow small input changes to grow; factors below one indicate shrinkage. These results suggest that repeated feedback can dampen differences despite initial amplification.

We measure the largest local output change per unit change in the input embedding:

$$\sigma_{t,K} = \|D_x F_{t,K}(x)\|_{\text{op}}, \tag{B.2}$$

where $F_{t,K}(x)$ is the soft output after $K$ passes from input $x$. We estimate this factor by power iteration, holding the preceding context fixed and, for gated probes, fixing sampled token identities. These local, averaged estimates do not establish contraction across all states.

**Diversity across generations can vanish during RL.** We separately compute covariance across generations of the same prompt at a fixed thinking step. Its effective rank is at most $G - 1$ for $G$ valid generations; training logs average over prompts and valid thinking steps. As a control, 32 deterministic pure-soft generations per prompt yield zero measured ensemble variance. With stochastic feedback, collapse also appears during optimization: in a Qwen KodCode HRPO run with $G = 8$, logged effective rank falls from $1.68$ to numerical zero by approximately step $250$ and remains near zero through step $1000$. A SofT-GRPO run falls from $2.84$ to approximately $1$ late in training. These observations separate saturation within a rollout from loss of diversity across rollouts.

Table 7: Projection variants on NuminaMath. pass@$k$ in %; paired per-problem standard errors are 0.5 pp (pass@1) and 1.1 pp (pass@8). Diagnostics: effective rank, trace and norm of the thinking-state covariance (last 20% of training). Only the hard, fixed-radius projection raises the rank and pass@8.

| Variant | $\alpha$ | pass@1 | pass@8 | eff. rank | trace | $\|s\|/R$ |
|---|---|---|---|---|---|---|
| No projection | – | 23.8 | 45.6 | 19.9 | 0.82 | 0.99 |
| SOFTSPHERE $\Pi_R$ | 1 | 24.4 | **47.4** | **22.8** | **0.85** | 1.01 |
| Partial normalisation | 0.5 | 23.4 | 44.2 | 21.2 | 0.83 | 1.00 |
| Partial normalisation | 0.25 | 23.8 | 44.4 | 19.9 | 0.81 | 0.99 |
| Centred, token-dep. $R_t$ | 1 | **24.9** | 45.5 | 19.4 | 0.78 | 0.97 |
| Centred, token-dep. $R_t$ | 0.5 | 23.4 | 45.3 | 19.0 | 0.77 | 0.96 |

## C ABLATIONS: VARIANTS OF THE PROJECTION

**Setup.** All runs share the token-anchored recipe on NuminaMath (Qwen2.5-1.5B-Instruct, lr $2 \times 10^{-6}$, 500 steps, seed 5267) and are evaluated on the 1439-problem held-out split with $n$=8 samples at $T$=0.8. Training-time diagnostics are averages over the last 20% of steps of the across-time covariance of the thinking states of a rollout: participation-ratio effective rank, trace, and state norm (relative to $R$). A safety cap on the projection gain ($10\times$) was present on the variant runs but never activated (capped fraction 0 throughout), so the runs differ only through the maps below.

**Variants.** SOFTSPHERE is the hard projection $\Pi_R(s) = R\,s/\|s\|$. *Partial normalisation* relaxes the constraint, $\Pi_R^\alpha(s) = s\,(R/\|s\|)^\alpha$ with $\alpha \in \{0.25, 0.5\}$ ($\alpha$=1 recovers $\Pi_R$, $\alpha$=0 disables it); under a per-step radial contraction $a$ the norm settles at $r_\infty = R\,a^{(1-\alpha)/\alpha}$, i.e. it is bounded away from zero but remains a free, step-dependent variable. *Centred projection* re-normalises around the embedding mean $\mu$ with a token-dependent radius, $\tilde{\Pi}(m) = \mu + R_t\,(m-\mu)/\|m-\mu\|$, $R_t^2 = \sum_i p_{t,i}\|e_i - \mu\|^2$, applied to the raw mixture $m = \sum_i p_{t,i} e_i$ in place of the $1/\|p_t\|_2$ normalisation. It is exact for one-hot $p_t$ (returns $e_i$), undoes only the shrinkage caused by averaging toward the centre, and was motivated by the fact that real embedding tables have $\mu \neq 0$, so that normalising $s$ rather than $s - \mu$ mixes the mean direction into the radius. We also test the combination (centred, $\alpha$=0.5).

**Results.** Table 7 shows a single pattern: the diversity gain and the pass@8 gain appear together, and only for the hard fixed-radius map. SOFTSPHERE raises the effective rank by $+3.0$ over no projection (with *fewer* thinking states, 477 vs. 549, so this is not a length effect) and pass@8 by $+1.8$ pp. Softening the constraint removes the gain gradually: the rank advantage falls to $+1.4$ at $\alpha$=0.5 and to 0 at $\alpha$=0.25, tracking the shrinking work the projection does (mean per-step gain $1.037 \to 1.017 \to 1.008$), and pass@8 drops 3 pp below SOFTSPHERE ($\approx$ 3 SE) to at or below the unprojected baseline. Centring fares no better: its rank is not above baseline (19.4 vs. 19.9), its states are *smaller* and *less spread* than even the unprojected ones (norm 0.97, trace 0.78), and pass@8 returns to the baseline value; its $+0.5$ pp pass@1 edge over SOFTSPHERE is one standard error. Adding partial normalisation to centring is worst on every diagnostic.

**Why the variants are worse.** Both relaxations re-open the radial degree of freedom that causes collapse. With $\alpha < 1$ the norm is again a dynamical variable that the recurrence contracts every step, so trajectories spread no more than without projection. The centred map preserves "legitimate" token norms by construction, but that is precisely the wrong target: its radius $R_t$ follows the entropy and support of $p_t$, so the norm carries step-dependent information and can shrink with it. What breaks the contraction is pinning every soft state to the *same* radius, independently of $p_t$; neither variant does this, and neither delivers the effect. The two conventions for the fixed-radius map (Euclidean vs. RMS) are the same projection up to the scale constant and are therefore not ablated.

**Takeaway.** The projection should be hard ($\alpha$=1) and its radius fixed; theoretically motivated refinements that keep the norm faithful to the token embeddings trade away the mechanism that produces the gains. Differences of $\lesssim 1$ pp at pass@1 are within the seed-to-seed spread of this recipe and should not be read as a ranking.

## D FURTHER DETAILS ON THE RANDOM TOKEN EXPERIMENT

To separate the *information* carried by the discrete token from its *stochastic re-separation* effect, we replace the sampled token $x_t$ in the gated blend by a token drawn uniformly at random from the vocabulary at every thinking step (thinking capped at 40 steps, since a random token essentially never emits the answer marker; the control is HRPO under the same cap). On GSM8K with Qwen2.5-1.5B-Instruct, pure latent thinking is deterministic and its rollouts coincide, so pass@8 $\approx$ pass@1 (34.7 vs. 33.0). Mixing in a token that carries *no information* already recovers most of the lost diversity: pass@8 rises to 61.8, closing $84\%$ of the gap to the model-sampled token at the same budget (66.8) and $73\%$ of the gap to uncapped HRPO (71.7); it also beats isotropic Gaussian noise at matched budget ($+3.3$ pass@1, $+2.0$ pass@8), and training-set accuracy is indistinguishable from the sampled-token run (0.46 vs. 0.45). The token's content matters mainly for single-shot accuracy (test pass@1 20.2 vs. 30.0). Hence most of the benefit of discrete supervision at $k > 1$ is re-separation of collapsed rollouts, not the token identity—the component that a deterministic projection cannot supply.

## E ADDITIONAL TRAJECTORY ANALYSIS

**Metric and protocol.** We quantify verbatim self-repetition with the repeated-4-gram rate,

$$\text{rep-4} = 1 - \frac{|\{\,(y_i, \ldots, y_{i+3})\,\}|}{\#\{\,i\,\}}, \tag{E.1}$$

where $i$ ranges over all valid 4-gram starting positions in a completion. The metric counts occurrences beyond the first appearance of each distinct 4-gram; higher values indicate more repetition.

We compare three Qwen2.5-1.5B-Instruct NuminaMath runs at seed 5267: HRPO with projection disabled, HRPO with projection enabled, and a discrete-token control without latent feedback. Each run generates 4 samples for each of 128 held-out problems (512 completions per run), using temperature 0.8, top-$p$ 0.95, and at most 1,024 new tokens. We average statistics within each problem before computing paired differences against HRPO. Confidence intervals use 20,000 paired bootstrap resamples over problems.

**Detailed results.** Table 6 in the main text reports repetition, paired differences, and mean completion length. Projection lowers rep-4 from 0.2951 to 0.2856 while increasing mean length from 524.5 to 549.0 tokens. The control without latent feedback has rep-4 of 0.2941, with a confidence interval for its difference from HRPO that includes zero.

## F FORMAL GUARANTEES FOR MANIFOLD PROJECTION

For the theoretical results, use the population-normalized temporal covariance

$$\widehat{C}_T(z) = \frac{1}{T}\sum_{t=0}^{T-1}(z_t - \bar{z})(z_t - \bar{z})^\top, \qquad \bar{z} = \frac{1}{T}\sum_{t=0}^{T-1} z_t. \tag{F.1}$$

The change from the empirical divisor $T-1$ leaves effective rank unchanged. For a positive-semidefinite matrix $C$, order its eigenvalues as $\lambda_1(C) \geq \cdots \geq \lambda_D(C) \geq 0$. Write $N_\varepsilon(C) = \#\{k : \lambda_k(C) > \varepsilon\}$ and $\log_+(x) = \max\{\log x, 0\}$.

### F.1 COVARIANCE-RANK SEPARATION

**Theorem A** (Covariance-rank separation under projection)**.** *Let $R > 0$ and $u_0 \in \mathbb{S}^{D-1}$. Suppose $F_t(ru) = a_t r G_t(u)$ for $r > 0$, where $G_t : \mathbb{S}^{D-1} \to \mathbb{S}^{D-1}$ and $0 < a_t \leq a < 1$. From $h_0 = \widetilde{h}_0 = Ru_0$, iterate $h_{t+1} = F_t(h_t)$ and $\widetilde{h}_{t+1} = \Pi_R(F_t(\widetilde{h}_t))$. Let $u_{t+1} = G_t(u_t)$, $\mu_T = \widehat{C}_T(u)$, and $d = \operatorname{rank}(\mu_T) \geq 1$. Assume*

$$T \geq \frac{2}{1-a}, \qquad \frac{c}{d} \leq \lambda_d(\mu_T) \leq \lambda_1(\mu_T) \leq \frac{\bar{c}}{d}, \qquad 0 < c \leq \bar{c}. \tag{F.2}$$

*Then*

$$r_{\text{eff}}(\widehat{C}_T(h)) \le \frac{16}{(1-a^2)^2}, \qquad r_{\text{eff}}(\widehat{C}_T(\widetilde{h})) \ge \frac{c}{\bar{c}}d, \tag{F.3}$$

$$Tr\widehat{C}_T(h) \le \frac{R^2}{T(1-a^2)}, \qquad Tr\widehat{C}_T(\widetilde{h}) \ge cR^2. \tag{F.4}$$

*At threshold* $\varepsilon_\star = cR^2/(2d)$*,*

$$N_{\varepsilon_\star}(\widehat{C}_T(h)) \le \left\lceil \frac{\log_+\!\big(2d/[cT(1-a^2)]\big)}{2\log(1/a)} \right\rceil, \qquad N_{\varepsilon_\star}(\widehat{C}_T(\widetilde{h})) = d. \tag{F.5}$$

*In particular, projection strictly increases effective rank whenever* $(c/\bar{c})d > 16/(1-a^2)^2$.

The informal statement in Section 4 uses $K = 16/(1-a^2)^2$ and $\eta = c/\bar{c}$.

*Proof of Theorem A.* For a sequence $z_0, \dots, z_{T-1}$, write

$$\bar{z} := \frac{1}{T}\sum_{t=0}^{T-1} z_t, \qquad \widehat{C}_T(z) := \frac{1}{T}\sum_{t=0}^{T-1}(z_t - \bar{z})(z_t - \bar{z})^\top.$$

Under the assumed factorization $F_t(ru) = a_t r G_t(u)$, where $\|G_t(u)\| = 1$ and $0 < a_t \le a < 1$, define

$$A_0 = 1, \qquad A_t = \prod_{s=0}^{t-1} a_s, \qquad u_{t+1} = G_t(u_t).$$

Starting both recurrences at $Ru_0$, induction gives $h_t = RA_t u_t$ and $\widetilde{h}_t = Ru_t$, with $A_t \le a^t$. Indeed, the unprojected update multiplies the radius by $a_t$, whereas projection restores radius $R$ without changing the angular update.

First consider the unprojected trajectory. Its uncentered second moment and centered covariance satisfy

$$M_T := \frac{1}{T}\sum_{t=0}^{T-1} h_t h_t^\top, \qquad \widehat{C}_T(h) = M_T - \bar{h}\bar{h}^\top \preceq M_T.$$

For $0 \le k \le D-1$, split $M_T = P_k + Q_k$ into summands with $t < k$ and $t \ge k$, respectively, restricting both sums to $0 \le t < T$. Then $\operatorname{rank}(P_k) \le k$, and eigenvalue monotonicity and Weyl's inequality imply

$$\begin{aligned}\lambda_{k+1}\big(\widehat{C}_T(h)\big) &\le \lambda_{k+1}(M_T) \le \|Q_k\|_{\text{op}} \\ &\le \frac{1}{T}\sum_{t=k}^{T-1}\|h_t\|^2 \le \frac{R^2}{T}\sum_{t=k}^{\infty} a^{2t} = \Phi_T a^{2k}, \qquad \Phi_T := \frac{R^2}{T(1-a^2)}.\end{aligned}$$

Empty sums are zero, so the argument also covers $k \ge T$. Similarly,

$$\text{Tr}\widehat{C}_T(h) \le \text{Tr}M_T \le \frac{R^2}{T}\sum_{t=0}^{\infty} a^{2t} = \Phi_T.$$

For the threshold $\varepsilon_\star = cR^2/(2d)$, an eigenvalue $\lambda_{k+1}(\widehat{C}_T(h)) > \varepsilon_\star$ requires $\Phi_T a^{2k} > \varepsilon_\star$. Counting the admissible nonnegative integers $k$ therefore gives

$$N_{\varepsilon_\star}\big(\widehat{C}_T(h)\big) \le \left\lceil \frac{\log_+\!\big(2d/[cT(1-a^2)]\big)}{2\log(1/a)} \right\rceil, \qquad \log_+(x) := \max\{\log x, 0\}.$$

To bound participation ratio, the assumption $T \ge 2/(1-a)$ yields

$$\|\bar{h}\| \le \frac{R}{T}\sum_{t=0}^{T-1} a^t \le \frac{R}{T(1-a)} \le \frac{R}{2}.$$

The $t = 0$ summand of the covariance then implies

$$\lambda_1\big(\widehat{C}_T(h)\big) \geq \frac{1}{T}\|h_0 - \bar{h}\|^2 \geq \frac{R^2}{4T}.$$

For every nonzero positive-semidefinite matrix $C$, $\|C\|_F^2 \geq \lambda_1(C)^2$. Consequently,

$$r_{\text{eff}}\big(\widehat{C}_T(h)\big) \leq \left(\frac{\mathrm{Tr}\widehat{C}_T(h)}{\lambda_1(\widehat{C}_T(h))}\right)^2 \leq \frac{16}{(1-a^2)^2}.$$

For the projected trajectory, let $\mu_T := \widehat{C}_T(u)$. Since $\widetilde{h}_t = Ru_t$,

$$\widehat{C}_T(\widetilde{h}) = R^2\mu_T.$$

By hypothesis, $\mu_T$ has rank $d \geq 1$ and its positive eigenvalues lie between $c/d$ and $\bar{c}/d$. It follows that

$$\begin{aligned}
\mathrm{Tr}\widehat{C}_T(\widetilde{h}) &\geq cR^2, \\
\lambda_d\big(\widehat{C}_T(\widetilde{h})\big) &\geq \frac{cR^2}{d} > \varepsilon_\star, \\
\lambda_1\big(\widehat{C}_T(\widetilde{h})\big) &\leq \frac{\bar{c}R^2}{d}.
\end{aligned}$$

The first inequality proves the projected trace bound, and the second, together with rank $d$, gives $N_{\varepsilon_\star}(\widehat{C}_T(\widetilde{h})) = d$. Finally, $\sum_j \lambda_j(C)^2 \leq \lambda_1(C)\mathrm{Tr}C$ for every positive-semidefinite $C$, so

$$r_{\text{eff}}\big(\widehat{C}_T(\widetilde{h})\big) \geq \frac{\mathrm{Tr}\widehat{C}_T(\widetilde{h})}{\lambda_1(\widehat{C}_T(\widetilde{h}))} \geq \frac{c}{\bar{c}}\,d.$$

□

### F.2 Temporal alignment under projection

**Theorem B** (Projection slows temporal alignment)**.** *Let $D \geq 2$, $\|u\| = 1$, $0 < a < 1$, and $0 < r_\star < R$. For $F(h) = ah + (1-a)r_\star u$, compare the unprojected and projected recurrences above, initialized at a common $h_0$ with $\|h_0\| = R$ and $0 < \angle(h_0, u) < \pi/2$. Define*

$$q = \frac{aR}{aR + (1-a)r_\star} \in (a, 1), \qquad \rho_{s,t}(z) = \frac{\langle z_s, z_t\rangle}{\|z_s\|\|z_t\|}. \tag{F.6}$$

*For every fixed lag $\ell \geq 1$, there are constants $c_\ell, \widetilde{c}_\ell > 0$ such that, as $t \to \infty$,*

$$1 - \rho_{t,t+\ell}(h) \sim c_\ell a^{2t}, \qquad 1 - \rho_{t,t+\ell}(\widetilde{h}) \sim \widetilde{c}_\ell q^{2t}. \tag{F.7}$$

*For every fixed window length $m \geq 2$, the normalized token Gram matrices $(\rho_{t+i,t+j})_{i,j=0}^{m-1}$ approach the all-ones matrix with Frobenius errors $\Theta(a^{2t})$ and $\Theta(q^{2t})$, respectively. For the growing-prefix mean $\bar{\rho}_T(z) = \frac{2}{T(T-1)}\sum_{0\leq s<t<T}\rho_{s,t}(z)$, there instead exist $0 < B_{\text{off}} < B_{\text{on}} < \infty$ such that*

$$1 - \bar{\rho}_T(h) \sim \frac{2B_{\text{off}}}{T}, \qquad 1 - \bar{\rho}_T(\widetilde{h}) \sim \frac{2B_{\text{on}}}{T}. \tag{F.8}$$

*Proof of Theorem B.* Let $\theta_0 = \angle(h_0, u)$ and write

$$h_0 = R(\cos\theta_0\, u + \sin\theta_0\, v), \qquad v \perp u, \quad \|v\| = 1.$$

Both trajectories remain in the plane spanned by $u, v$, with positive coefficients along these vectors. In particular, all states and pre-projection outputs are nonzero.

**Angular asymptotics.** The unprojected recurrence has the explicit solution

$$h_t = r_\star u + a^t(h_0 - r_\star u).$$

Consequently, its angle $\theta_t = \angle(h_t, u)$ satisfies

$$\theta_t = \arctan\left(\frac{a^t R \sin\theta_0}{r_\star + a^t(R\cos\theta_0 - r_\star)}\right) = La^t + O(a^{2t}), \qquad L = \frac{R\sin\theta_0}{r_\star} > 0. \tag{F.9}$$

Set $b = (1-a)r_\star$. The projected trajectory has norm $R$ and angle $\widetilde{\theta}_t = \angle(\widetilde{h}_t, u)$ obeying

$$\widetilde{\theta}_{t+1} = g(\widetilde{\theta}_t), \qquad g(\theta) = \arctan\left(\frac{aR\sin\theta}{aR\cos\theta + b}\right). \tag{F.10}$$

For $0 < \theta < \pi/2$, $0 < g(\theta) < \theta$. Thus $\widetilde{\theta}_t$ decreases to zero: continuity would make any positive limit a fixed point of $g$. Taylor expansion gives

$$g(\theta) = q\theta + O(\theta^3), \qquad q = \frac{aR}{aR+b} \in (a, 1).$$

Choose $q < q_+ < 1$. For all sufficiently large $t$, $\widetilde{\theta}_{t+1} \le q_+\widetilde{\theta}_t$, so $\sum_t \widetilde{\theta}_t^2 < \infty$. Since

$$\frac{\widetilde{\theta}_{t+1}}{q\widetilde{\theta}_t} = 1 + O(\widetilde{\theta}_t^2),$$

the logarithms of these positive ratios form an absolutely convergent tail series. Their product converges to a finite positive limit, yielding

$$\widetilde{\theta}_t = \widetilde{L}q^t(1 + o(1)) \qquad \text{for some } \widetilde{L} > 0. \tag{F.11}$$

**Fixed-lag and fixed-window correlations.** The common oriented plane gives $\rho_{s,t}(h) = \cos(\theta_s - \theta_t)$, with the analogous identity for the projected trajectory. For fixed $\ell \ge 1$,

$$\theta_t - \theta_{t+\ell} = L(1-a^\ell)a^t + O(a^{2t}),$$
$$\widetilde{\theta}_t - \widetilde{\theta}_{t+\ell} = \widetilde{L}(1-q^\ell)q^t(1+o(1)).$$

Using $1 - \cos x \sim x^2/2$ proves the claimed equivalents with $c_\ell = \frac{1}{2}L^2(1-a^\ell)^2$ and $\widetilde{c}_\ell = \frac{1}{2}\widetilde{L}^2(1-q^\ell)^2$. Both constants are positive. For a fixed window and $0 \le i < j < m$,

$$1 - \rho_{t+i,t+j}(h) \sim \tfrac{1}{2}L^2(a^i - a^j)^2 a^{2t},$$

and similarly with $L, a$ replaced by $\widetilde{L}, q$. Summing squared entrywise deficits and taking the square root gives the two Frobenius rates. Because $q > a$, the projected decay exponent $-2\log q$ is strictly smaller than $-2\log a$.

**Growing-prefix correlations.** For either trajectory, let $\vartheta_t$ denote its angle and set

$$B_T = \sum_{t=0}^{T-1}(1-\cos\vartheta_t), \qquad S_T = \sum_{t=0}^{T-1}\sin\vartheta_t.$$

The angular asymptotics give $B_T \to B \in (0,\infty)$ and $S_T \to S < \infty$. For normalized states $v_t = \cos\vartheta_t\, u + \sin\vartheta_t\, v$,

$$\bar{\rho}_T = \frac{\|\sum_{t<T} v_t\|^2 - T}{T(T-1)}, \qquad \sum_{t<T} v_t = (T - B_T)u + S_T v.$$

It follows that

$$1 - \bar{\rho}_T = \frac{2TB_T - B_T^2 - S_T^2}{T(T-1)} \sim \frac{2B}{T}.$$

Applying this identity to both trajectories gives the prefix asymptotics, with

$$B_{\text{off}} = \sum_{t\ge 0}(1-\cos\theta_t), \qquad B_{\text{on}} = \sum_{t\ge 0}(1-\cos\widetilde{\theta}_t).$$

To compare these constants, the explicit unprojected solution implies $\|h_t\| \le a^t R + (1-a^t) r_\star < R$ for every $t \ge 1$. For a state of radius $r > 0$ and angle $0 < \theta < \pi/2$, one affine update produces angle

$$H(r,\theta) = \arctan\left(\frac{ar\sin\theta}{ar\cos\theta + b}\right).$$

Both partial derivatives are strictly positive on this domain. The initial angles and first updated angles coincide, but all later unprojected radii are smaller than $R$. Induction therefore gives $\widetilde{\theta}_t > \theta_t$ for every $t \ge 2$. Since $1 - \cos\theta$ is strictly increasing on the acute interval, summing yields $B_{\mathrm{on}} > B_{\mathrm{off}} > 0$. □

## G Additional results on token discretization

### Contraction across thinking steps with and without token discretization

There are three distinct notions of representation collapse in a discretized reasoning process. The distance between two continuous latent trajectories may shrink, the laws of two token chains may forget their initial tokens, and two independently sampled token rollouts may lose sample diversity. For deterministic dynamics, pathwise separation and sample diversity reduce to the same observable. After sampling, rollout diversity and forgetting of initial conditions separate: a token chain can forget its initial condition while two independent realizations remain a constant distance apart. The results below therefore state the observable being contracted in each case.

We use *contraction factor* for a number in $[0,1]$ that multiplies a distance in one step, and *decay exponent* for its negative logarithm. Thus a factor $a \in (0,1)$ corresponds to the positive exponent $\log(1/a)$.

### Setup

Let $V \ge 2$ and suppose that the token embeddings $e_1, \dots, e_V \in \mathbb{R}^D$, where $D \ge V - 1$, form a centered regular simplex of radius $R$:

$$\sum_{i=1}^{V} e_i = 0, \qquad \|e_i\| = R, \qquad \langle e_i, e_j \rangle = -\frac{R^2}{V-1} \quad (i \ne j). \tag{G.1}$$

Every pair of distinct embeddings is then separated by

$$\delta_{\mathrm{tok}} := \|e_i - e_j\| = R\sqrt{\frac{2V}{V-1}}, \qquad i \ne j. \tag{G.2}$$

Write $\mathcal{K} := \mathrm{conv}\{e_1, \dots, e_V\}$. Its diameter is $\delta_{\mathrm{tok}}$. The regular-simplex assumption is an idealized geometry: it gives exact formulas below, but necessarily requires $V \le D + 1$.

Without token discretization, $G \ge 2$ representations evolve through the same sequence of differentiable forward maps,

$$h_{t+1}^{(g)} = F_t\big(h_t^{(g)}\big), \qquad g = 1, \dots, G, \tag{G.3}$$

where $F_t : \mathcal{K} \to \mathcal{K}$ extends differentiably to a neighborhood of $\mathcal{K}$. Define the ensemble diameter, mean, and covariance at step $t$ by

$$\mathcal{D}_t := \max_{g,g'} \big\| h_t^{(g)} - h_t^{(g')} \big\|, \tag{G.4}$$

$$m_G(h_t) := \frac{1}{G}\sum_{g=1}^{G} h_t^{(g)}, \qquad \widehat{C}_G(h_t) := \frac{1}{G}\sum_{g=1}^{G} \big(h_t^{(g)} - m_G(h_t)\big)\big(h_t^{(g)} - m_G(h_t)\big)^\top. \tag{G.5}$$

We repeatedly use the pairwise identity

$$\mathrm{Tr}\widehat{C}_G(h_t) = \frac{1}{2G^2}\sum_{g,g'=1}^{G} \big\| h_t^{(g)} - h_t^{(g')} \big\|^2. \tag{G.6}$$

With discretization, a forward pass from token $i$ induces a distribution over the next token. In the memoryless abstraction this is a stochastic matrix $P_t$,

$$P_t(i,j) := \mathbb{P}\{I_{t+1} = j \mid I_t = i\}, \qquad X_t := e_{I_t}. \tag{G.7}$$

For an autoregressive transformer, the effective state also contains the context and KV cache. The total-variation part of the Markov analysis below extends after taking that full context as the state and defining its Dobrushin coefficient accordingly. The regular-simplex Wasserstein identity continues to concern the last-token embeddings and does not automatically extend to distances between full contexts. The independent-sampling result will be stated directly for context-dependent next-token distributions.

A common soft-thinking construction produces a next-token distribution $p_t(z)$ and feeds back its mean embedding

$$M(p) := \sum_{i=1}^{V} p_i e_i, \qquad F_t(z) = M\big(p_t(z)\big). \tag{G.8}$$

Explicit token reasoning instead samples $I_{t+1} \sim p_t(X_t)$ and feeds back $e_{I_{t+1}}$. Hence the two recurrences use the same distribution-valued decoder rule, differing in mean feedback versus sampling and re-embedding.

NO DISCRETIZATION: MULTIPLICATIVE CONTRACTION

For $0 \le s \le t$, set

$$L_t := \sup_{z \in \mathcal{K}} \|DF_t(z)\|_{\mathrm{op}}, \qquad \Gamma_{s,t} := \prod_{u=s}^{t-1} L_u, \qquad \Gamma_{t,t} := 1. \tag{G.9}$$

**Theorem A** (Continuous latent states contract multiplicatively)**.** *Assume $0 < L_t < \infty$ for every $t$. Then the following statements hold.*

1. *Every pair of latent trajectories satisfies*

$$\big\|h_T^{(g)} - h_T^{(g')}\big\| \le \Gamma_{0,T}\big\|h_0^{(g)} - h_0^{(g')}\big\|, \qquad \mathcal{D}_T \le \Gamma_{0,T}\mathcal{D}_0. \tag{G.10}$$

2. *The ensemble covariance trace obeys*

$$\mathit{Tr}\widehat{C}_G(h_T) \le \Gamma_{0,T}^2 \mathit{Tr}\widehat{C}_G(h_0). \tag{G.11}$$

3. *Suppose the uniform transverse Lyapunov bound is negative:*

$$\lambda_\perp := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log L_t = -x_{\mathrm{lat}} < 0. \tag{G.12}$$

*For every $\varepsilon \in (0, x_{\mathrm{lat}})$ there is a finite $C_\varepsilon \ge 1$ such that, for all $T$,*

$$\mathcal{D}_T \le C_\varepsilon e^{-(x_{\mathrm{lat}}-\varepsilon)T}\mathcal{D}_0, \tag{G.13}$$

$$\mathit{Tr}\widehat{C}_G(h_T) \le C_\varepsilon^2 e^{-2(x_{\mathrm{lat}}-\varepsilon)T}\mathit{Tr}\widehat{C}_G(h_0). \tag{G.14}$$

*Thus distance and root-mean-square spread have certified decay exponent $x_{\mathrm{lat}}$, while variance has twice that exponent.*

4. *Under the stronger per-step condition $L_t \le a < 1$, the finite-time bounds sharpen to*

$$\boxed{\mathcal{D}_T \le a^T \mathcal{D}_0, \qquad \mathit{Tr}\widehat{C}_G(h_T) \le a^{2T}\mathit{Tr}\widehat{C}_G(h_0)}. \tag{G.15}$$

*This certifies distance-decay exponent at least $\log(1/a)$ and variance-decay exponent at least $2\log(1/a)$. If in addition $F_t \equiv F$ is autonomous, then $F$ has a unique fixed point $h_\star \in \mathcal{K}$ and*

$$\big\|h_T^{(g)} - h_\star\big\| \le a^T\big\|h_0^{(g)} - h_\star\big\| \le \delta_{\mathrm{tok}} a^T. \tag{G.16}$$

*Proof.* Because $\mathcal{K}$ is convex, the line segment between $u, w \in \mathcal{K}$ lies in $\mathcal{K}$. The fundamental theorem of calculus gives

$$F_t(u) - F_t(w) = \int_0^1 DF_t\big(w + s(u-w)\big)(u-w)\, ds,$$

and hence $\|F_t(u) - F_t(w)\| \le L_t\|u - w\|$. Iterating this inequality proves (G.10). Applying its squared version to every summand in (G.6) proves (G.11).

Fix $\varepsilon \in (0, x_{\mathrm{lat}})$. The definition of the limsup in (G.12) implies that, for all sufficiently large $T$,

$$\sum_{t=0}^{T-1} \log L_t \le -(x_{\mathrm{lat}} - \varepsilon)T.$$

The finitely many remaining horizons can be absorbed into a constant $C_\varepsilon$, so $\Gamma_{0,T} \le C_\varepsilon e^{-(x_{\mathrm{lat}}-\varepsilon)T}$ for every $T$. Substitution proves (G.13) and (G.14). The uniform bounds in (G.15) follow by $\Gamma_{0,T} \le a^T$. Finally, in the autonomous case, $\mathcal{K}$ is compact and hence complete, and $F : \mathcal{K} \to \mathcal{K}$ is an $a$-contraction. The Banach fixed-point theorem gives the unique $h_\star$, and iteration of the one-step bound gives (G.16). □

The Lyapunov hypothesis in Theorem A is uniform over the region containing the ensemble. A negative exponent estimated in one direction along one trajectory does not by itself imply the global diameter bound. Conversely, the theorem gives an upper bound on distances; an exact exponent requires a matching lower bound. For example, if

$$e^{\mu}\|u - w\| \le \|F_t(u) - F_t(w)\| \le e^{\lambda}\|u - w\|, \qquad \mu \le \lambda < 0, \tag{G.17}$$

then iteration sandwiches every nonzero pairwise distance between $e^{\mu T}$ and $e^{\lambda T}$ times its initial value. Its decay exponent therefore lies in $[-\lambda, -\mu]$.

For an autonomous map satisfying the last part of the theorem, the entire simplex is $\varepsilon$-indistinguishable after

$$T_\varepsilon := \max\left\{0, \left\lceil \frac{\log(\delta_{\mathrm{tok}}/\varepsilon)}{\log(1/a)} \right\rceil\right\} \tag{G.18}$$

steps. An $L_f$-Lipschitz readout has output diameter at most $L_f\varepsilon$ from that step onward.

### What mixing and discretization do in one step

**Lemma B** (Simplex mixing geometry)**.** *For probability vectors $p, q \in \mathbb{R}^V$ and the mean-embedding map in (G.8),*

$$\|M(p) - M(q)\|^2 = R^2 \frac{V}{V-1}\|p - q\|_2^2, \tag{G.19}$$

$$\|M(p)\|^2 = R^2 \frac{V\|p\|_2^2 - 1}{V-1}. \tag{G.20}$$

*If $I \sim p$, then the mean-square displacement created by sampling and re-embedding is*

$$\mathbb{E}\big[\|e_I - M(p)\|^2\big] = R^2 \frac{V(1 - \|p\|_2^2)}{V-1} = \frac{\delta_{\mathrm{tok}}^2}{2}\big(1 - \|p\|_2^2\big). \tag{G.21}$$

*Proof.* If $v \in \mathbb{R}^V$ satisfies $\sum_i v_i = 0$, then

$$\begin{aligned}\left\|\sum_i v_i e_i\right\|^2 &= R^2 \sum_i v_i^2 - \frac{R^2}{V-1}\sum_{i \ne j} v_i v_j \\ &= R^2\left(1 + \frac{1}{V-1}\right)\sum_i v_i^2,\end{aligned}$$

because $\sum_{i\ne j} v_i v_j = -\sum_i v_i^2$. Taking $v = p - q$ proves (G.19). Since $M(p) = \sum_i (p_i - 1/V)e_i$ and $\sum_i (p_i - 1/V)^2 = \|p\|_2^2 - 1/V$, the same calculation proves (G.20). Finally, $M(p) = \mathbb{E}[e_I]$ and $\|e_I\| = R$, so

$$\mathbb{E}\|e_I - M(p)\|^2 = \mathbb{E}\|e_I\|^2 - \|M(p)\|^2.$$

Substituting (G.20) and (G.2) proves (G.21). □

Writing $H_2(p) := -\log\|p\|_2^2$, the quantity $\|p\|_2^2 = e^{-H_2(p)}$ is the collision probability. Thus decoder entropy pulls a soft mean embedding toward the center by (G.20), while token sampling restores a vertex-scale state and injects the displacement in (G.21). The mixing identity alone does not determine the iterated token rate; that rate depends on the induced token transition kernel.

**Lemma C** (Bounded discretization errors accumulate additively)**.** *Let $Q_t : F_t(\mathcal{K}) \to \mathcal{K}$ be any deterministic discretizer satisfying*

$$\sup_{z \in F_t(\mathcal{K})} \|Q_t(z) - z\| \le \epsilon_t, \tag{G.22}$$

*and, from an initial state in $\mathcal{K}$, let $\widehat{h}_{t+1} = Q_t(F_t(\widehat{h}_t))$. Two discretized trajectories satisfy*

$$\left\|\widehat{h}_T - \widehat{h}'_T\right\| \le \Gamma_{0,T}\|\widehat{h}_0 - \widehat{h}'_0\| + 2\sum_{s=0}^{T-1} \epsilon_s \Gamma_{s+1,T}. \tag{G.23}$$

*If $h_{t+1} = F_t(h_t)$, then*

$$\|\widehat{h}_T - h_T\| \le \Gamma_{0,T}\|\widehat{h}_0 - h_0\| + \sum_{s=0}^{T-1} \epsilon_s \Gamma_{s+1,T}. \tag{G.24}$$

*In particular, if $L_t \le a < 1$ and $\epsilon_t \le \epsilon$, then the two additive terms are at most $2\epsilon(1-a^T)/(1-a)$ and $\epsilon(1-a^T)/(1-a)$, respectively.*

*Proof.* The triangle inequality and (G.22) give

$$\|\widehat{h}_{t+1} - \widehat{h}'_{t+1}\| \le L_t\|\widehat{h}_t - \widehat{h}'_t\| + 2\epsilon_t,$$
$$\|\widehat{h}_{t+1} - h_{t+1}\| \le L_t\|\widehat{h}_t - h_t\| + \epsilon_t.$$

Unrolling the two scalar recurrences proves (G.23) and (G.24). The uniform case follows by summing a geometric series. □

Lemma C is a stability bound, not a universal slower contraction theorem. Nearest-token maps are discontinuous at their cell boundaries, and different token transition rules can produce qualitatively different rates, as shown next.

**Proposition D** (The latent rate does not determine the greedy-token rate)**.** *Let*

$$Q(z) \in \operatorname*{argmin}_{e \in \{e_1,\dots,e_V\}} \|z - e\|$$

*be nearest-token discretization, with any fixed rule for ties. There are contractive maps for which $Q \circ F$ preserves all pairwise token distances, and contractive maps for which $Q \circ F$ collapses every token to one token in a single step. Specifically:*

1. *For $F(z) = az$ with $0 < a < 1$,*

$$\|F^T(e_i) - F^T(e_j)\| = a^T\delta_{\mathrm{tok}}, \qquad (Q \circ F)^T(e_i) = e_i. \tag{G.25}$$

2. *For $F(z) = (1-b)e_1 + bz$ with $0 < b < 1/2$, latent distances contract by $b$ per step, but*

$$Q(F(e_i)) = e_1 \qquad \text{for every } i. \tag{G.26}$$

*Consequently, a negative latent Lyapunov exponent and simplex geometry alone imply no nontrivial universal rate after greedy tokenization.*

*Proof.* For the first map, $F^T(e_i) = a^T e_i$. Moreover, $\langle ae_i, e_i\rangle = aR^2$ whereas $\langle ae_i, e_j\rangle = -aR^2/(V-1)$ for $j \ne i$. Because all vertices have the same norm, nearest-vertex decoding is equivalent to maximizing this inner product. Hence $Q(ae_i) = e_i$, proving (G.25).

For the second map, $\|F(u) - F(w)\| = b\|u - w\|$ and $F(\mathcal{K}) \subseteq \mathcal{K}$. It remains to identify the nearest vertex. If $i \ne 1$, then

$$\langle F(e_i), e_1 - e_i\rangle = (1-2b)R^2\frac{V}{V-1} > 0.$$

For $k \notin \{1, i\}$,

$$\langle F(e_i), e_1 - e_k \rangle = (1-b)R^2 \frac{V}{V-1} > 0,$$

where the $be_i$ contribution cancels because $e_i$ has the same inner product with $e_1$ and $e_k$. Thus $e_1$ has strictly larger inner product with $F(e_i)$ than every other vertex and is its unique nearest token. The case $i = 1$ is immediate, proving (G.26). □

TOKEN DISCRETIZATION: CONTRACTION OF INITIAL-STATE MEMORY

For probability measures $\mu, \nu$ on the $V$ tokens, write

$$\mathrm{TV}(\mu, \nu) := \frac{1}{2} \sum_{i=1}^{V} |\mu_i - \nu_i|.$$

The appropriate worst-case coefficient for a stochastic token transition is its Dobrushin coefficient

$$\eta_t := \max_{i,j} \mathrm{TV}\big(P_t(i,\cdot), P_t(j,\cdot)\big) \in [0,1]. \tag{G.27}$$

**Theorem E** (Token-channel memory contracts at the Dobrushin rate)**.** *Let $\mu_{t+1} = \mu_t P_t$ and $\nu_{t+1} = \nu_t P_t$ be the laws of two time-inhomogeneous token chains. Then*

$$\mathrm{TV}(\mu_T, \nu_T) \le \left( \prod_{t=0}^{T-1} \eta_t \right) \mathrm{TV}(\mu_0, \nu_0). \tag{G.28}$$

*Moreover, $\eta_t$ is the exact worst-case one-step contraction coefficient:*

$$\eta_t = \sup_{\mu \neq \nu} \frac{\mathrm{TV}(\mu P_t, \nu P_t)}{\mathrm{TV}(\mu, \nu)}. \tag{G.29}$$

*Let $W_1$ denote Wasserstein distance with the Euclidean cost between token embeddings. Under (G.1),*

$$W_1(\mu, \nu) = \delta_{\mathrm{tok}} \, \mathrm{TV}(\mu, \nu), \tag{G.30}$$

*so (G.28) also holds with* TV *replaced by $W_1$. In particular, if $0 < \eta < 1$ and $\eta_t \le \eta$ for every $t$, then*

$$\boxed{W_1(\mu_T, \nu_T) \le \eta^T W_1(\mu_0, \nu_0), \qquad x_{\mathrm{tok}}^{\mathrm{mem}} \ge -\log \eta}. \tag{G.31}$$

*Here $x_{\mathrm{tok}}^{\mathrm{mem}}$ denotes the certified decay exponent for dependence on the initial token. More generally, if all $\eta_t > 0$ and*

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \log \eta_t = -x_{\mathrm{tok}}^{\mathrm{mem}} < 0,$$

*then for every $\varepsilon \in (0, x_{\mathrm{tok}}^{\mathrm{mem}})$ there is $C_\varepsilon < \infty$ such that*

$$\mathrm{TV}(\mu_T, \nu_T) \le C_\varepsilon e^{-(x_{\mathrm{tok}}^{\mathrm{mem}} - \varepsilon)T} \, \mathrm{TV}(\mu_0, \nu_0), \tag{G.32}$$

$$W_1(\mu_T, \nu_T) \le C_\varepsilon e^{-(x_{\mathrm{tok}}^{\mathrm{mem}} - \varepsilon)T} W_1(\mu_0, \nu_0). \tag{G.33}$$

*Proof.* It suffices to prove the one-step bound. Couple $I \sim \mu$ and $J \sim \nu$ so that

$$\mathbb{P}\{I \neq J\} = \mathrm{TV}(\mu, \nu),$$

which is possible by maximal coupling. Conditional on $I = J = i$, use the same draw from $P_t(i,\cdot)$ for both next states. Conditional on $(I, J) = (i, j)$ with $i \neq j$, maximally couple the two rows $P_t(i,\cdot)$ and $P_t(j,\cdot)$. The conditional probability that the next states disagree is then $\mathrm{TV}(P_t(i,\cdot), P_t(j,\cdot)) \le \eta_t$. Consequently this construction couples $\mu P_t$ and $\nu P_t$ with disagreement probability at most $\eta_t \, \mathrm{TV}(\mu, \nu)$. The coupling characterization of total variation therefore gives

$$\mathrm{TV}(\mu P_t, \nu P_t) \le \eta_t \, \mathrm{TV}(\mu, \nu).$$

Iteration proves (G.28). Taking $\mu$ and $\nu$ to be point masses at a pair of tokens attaining the maximum in (G.27) proves the reverse inequality required for (G.29).

For (G.30), an optimal transport plan leaves the common mass $\sum_i \min\{\mu_i, \nu_i\} = 1 - \mathrm{TV}(\mu, \nu)$ on the diagonal and moves the remaining mass $\mathrm{TV}(\mu, \nu)$ between distinct vertices. Every such move costs exactly $\delta_{\mathrm{tok}}$. Thus the transport cost is exactly $\delta_{\mathrm{tok}} \, \mathrm{TV}(\mu, \nu)$. The uniform and asymptotic conclusions follow by bounding the product of the $\eta_t$ in the same way as the product of the $L_t$ in Theorem A. □

Theorem E shows precisely what additional object is needed after discretization: the continuous Jacobian factors $L_t$ are replaced by the channel coefficients $\eta_t$. Neither $\eta_t$ nor its asymptotic product is determined by the latent Lyapunov exponent, as Proposition D already shows in the deterministic limit.

### INDEPENDENT SAMPLED ROLLOUTS: A DIVERSITY FLOOR

At step $t$, let $q_t(\cdot \mid c)$ denote the next-token distribution at a reachable context $c \in \mathcal{C}_t$, and define

$$\kappa_t := \sup_{c \in \mathcal{C}_t} \big\| q_t(\cdot \mid c) \big\|_2^2, \qquad \kappa := \sup_{t \geq 0} \kappa_t, \qquad H_2^{\min} := -\log \kappa. \tag{G.34}$$

Here $\|q\|_2^2$ is the probability that two independent samples from $q$ collide, and $H_2^{\min}$ denotes the infimum Rényi-2 entropy over reachable steps and contexts; the infimum need not be attained.

**Theorem F** (Independent token sampling prevents total sample collapse)**.** *Let $X_t = e_{I_t}$ and $Y_t = e_{J_t}$ be two token rollouts whose next-token draws are independent conditional on their joint pair of histories. Then, for every $t \geq 0$,*

$$\mathbb{E}\big[\|X_{t+1} - Y_{t+1}\|\big] \geq \delta_{\mathrm{tok}}(1 - \kappa_t), \tag{G.35}$$

$$\mathbb{E}\big[\|X_{t+1} - Y_{t+1}\|^2\big] \geq \delta_{\mathrm{tok}}^2(1 - \kappa_t). \tag{G.36}$$

*These bounds hold even when the two current histories are identical.*

*More generally, consider $G \geq 2$ rollouts $X_t^{(1)}, \ldots, X_t^{(G)}$ sampled mutually independently conditional on their joint history at each pass, write $\mathbf{X}_t := (X_t^{(1)}, \ldots, X_t^{(G)})$, and form their cross-generation covariance as in (G.5). Then*

$$\mathbb{E}\big[Tr\widehat{C}_G(\mathbf{X}_{t+1})\big] \geq \frac{G-1}{2G}\, \delta_{\mathrm{tok}}^2(1 - \kappa_t). \tag{G.37}$$

*If $\kappa < 1$, the expected sample distance is bounded away from zero uniformly in the number of thinking steps. In particular, its asymptotic decay exponent is zero:*

$$-\lim_{T \to \infty} \frac{1}{T} \log \left( \frac{\mathbb{E}\|X_T - Y_T\|}{\delta_{\mathrm{tok}}} \right) = 0. \tag{G.38}$$

*Proof.* Condition on a pair of histories with contexts $c, c' \in \mathcal{C}_t$. Conditional independence and Cauchy–Schwarz give

$$\begin{aligned} \mathbb{P}\{X_{t+1} = Y_{t+1} \mid c, c'\} &= \sum_{i=1}^{V} q_t(i \mid c) q_t(i \mid c') \\ &\leq \|q_t(\cdot \mid c)\|_2 \|q_t(\cdot \mid c')\|_2 \leq \kappa_t. \end{aligned}$$

Under regular-simplex geometry,

$$\|X_{t+1} - Y_{t+1}\| = \delta_{\mathrm{tok}} \mathbf{1}\{X_{t+1} \neq Y_{t+1}\},$$

and the analogous identity holds after squaring. Taking expectations proves (G.35) and (G.36).

For the covariance claim, apply the pairwise identity (G.6) to the random token ensemble. Each of the $G(G-1)$ ordered pairs of distinct generations has expected squared distance at least $\delta_{\mathrm{tok}}^2(1 - \kappa_t)$; the $G$ diagonal pairs contribute zero. This proves (G.37). Finally, when $\kappa < 1$, the ratio inside the logarithm in (G.38) lies between $1 - \kappa$ and 1 for $T \geq 1$. Dividing its bounded logarithm by $T$ proves the limit. □

The collision parameter $\kappa_t$ and the Dobrushin coefficient $\eta_t$ measure different things. For example, if every transition row is uniform, then $\eta_t = 0$: the chain forgets its initial token in one step. At the same time $\kappa_t = 1/V$, so two independent next-token samples disagree with probability $1 - 1/V$. The memory has collapsed, but the independently sampled rollouts remain diverse.

If $\kappa < 1$, then under the uniform latent bound $L_t \le a < 1$, two continuous trajectories that begin at simplex vertices have distance at most $\delta_{\mathrm{tok}} a^T$. The independent-token floor is no smaller than this upper bound once

$$T \ge T_\times := \left\lceil \frac{\log\big(1/(1-\kappa)\big)}{\log(1/a)} \right\rceil. \tag{G.39}$$

Thus any uniform nondegenerate sampling rule eventually yields more expected sample separation than the contracting latent recurrence. This comparison is between pathwise latent distance and independent-rollout sample diversity, not between their memories of the initial state.

**Corollary G** (Absorbing-coalescence benchmark)**.** *Suppose the token recurrence is memoryless. Start two chains at distinct tokens, sample their transitions independently while they are distinct, and after their first meeting use the same transition draw so that they remain equal. If*

$$\tau := \inf\{t \ge 1 : X_t = Y_t\},$$

*then, for every integer $T \ge 0$,*

$$\mathbb{P}\{\tau > T\} \ge \prod_{t=0}^{T-1} (1-\kappa_t) \ge (1-\kappa)^T, \qquad \mathbb{E}\|X_T - Y_T\| \ge \delta_{\mathrm{tok}} (1-\kappa)^T. \tag{G.40}$$

*If $\kappa < 1$, (G.40) bounds this coupling's collapse exponent above by $-\log(1-\kappa)$. Compared with a latent per-step factor $a$, the certified token separation decays strictly more slowly when*

$$1 - \kappa > a \quad \Longleftrightarrow \quad H_2^{\min} > \log \frac{1}{1-a}. \tag{G.41}$$

*Proof.* On $\{\tau > t\}$, condition on the two distinct current tokens. The same collision calculation used in Theorem F shows that their independent next draws meet with probability at most $\kappa_t$. Therefore

$$\mathbb{P}\{\tau > t+1\} \ge (1-\kappa_t)\mathbb{P}\{\tau > t\}.$$

Induction proves the survival bounds. Because the chains are synchronized after meeting, $\|X_T - Y_T\| = \delta_{\mathrm{tok}} \mathbf{1}\{\tau > T\}$, proving the distance bound. Finally, $1-\kappa > a$ is equivalent to $e^{-H_2^{\min}} < 1-a$, which gives (G.41). □

The memoryless hypothesis in Corollary G is essential: in a transformer, equal current tokens need not mean equal histories or equal KV caches. One may instead define coalescence on the full effective state. The corollary is a lower survival bound for an explicit coupling, not the Dobrushin contraction rate of an arbitrary token chain.

### An exact regular-simplex softmax model

The next result uses the pre-decoder continuous-state abstraction from the reference model, rather than the mean-feedback construction in (G.8). The unquantized recurrence feeds back $F(z) = az$ directly; the tokenized recurrence instead applies a specified softmax decoder to $F(e_i)$ and re-embeds its sample. In this model the two multiplicative rates are directly comparable, and decoding confidence can make token-channel memory contract much more slowly than continuous latent distance.

**Theorem H** (Exact latent and token-memory rates)**.** *Let the continuous forward map be*

$$F(z) = az, \qquad 0 < a < 1,$$

*and, after a forward pass from token $e_i$, sample the next token according to*

$$P(i,j) := \frac{\exp\big(\beta\langle a e_i, e_j\rangle\big)}{\sum_{k=1}^{V} \exp\big(\beta\langle a e_i, e_k\rangle\big)}, \qquad \beta > 0, \tag{G.42}$$

*where $\beta$ is the inverse temperature, or decoding confidence. Define*

$$\zeta := \frac{\beta a R^2 V}{V-1}, \qquad \eta := \frac{e^\zeta - 1}{e^\zeta + V - 1} \in (0,1). \tag{G.43}$$

*Then*

$$P = \eta I_V + (1-\eta)\frac{\mathbf{1}\mathbf{1}^\top}{V}, \qquad P^T = \eta^T\left(I_V - \frac{\mathbf{1}\mathbf{1}^\top}{V}\right) + \frac{\mathbf{1}\mathbf{1}^\top}{V}. \tag{G.44}$$

*For $i \neq j$, the continuous latent distance, token-law distance, and conditional mean-embedding distance are exactly*

$$\|F^T(e_i) - F^T(e_j)\| = \delta_{\mathrm{tok}} a^T, \tag{G.45}$$

$$\mathrm{TV}\big(P^T(i,\cdot), P^T(j,\cdot)\big) = \eta^T, \tag{G.46}$$

$$\big\|\bar{e}_T(i) - \bar{e}_T(j)\big\| = \delta_{\mathrm{tok}}\eta^T, \qquad \bar{e}_T(i) := \mathbb{E}[e_{I_T} \mid I_0 = i]. \tag{G.47}$$

*Equivalently, the two positive decay exponents are*

$$\boxed{x_{\mathrm{lat}} = \log\frac{1}{a}, \qquad x_{\mathrm{tok}}^{\mathrm{mem}} = \log\frac{1}{\eta}}. \tag{G.48}$$

*Token-channel memory contracts strictly more slowly than continuous latent distance precisely when $\eta > a$, or equivalently*

$$\boxed{\beta > \frac{V-1}{aR^2V}\log\left(\frac{1+a(V-1)}{1-a}\right)}. \tag{G.49}$$

*As $\zeta \to \infty$,*

$$x_{\mathrm{tok}}^{\mathrm{mem}} = Ve^{-\zeta} + O(e^{-2\zeta}), \tag{G.50}$$

*and hence the token-memory exponent tends to zero in the hard-decoding limit.*

*Proof.* The simplex inner products imply that the diagonal logit in (G.42) is $\beta a R^2$, while every off-diagonal logit is $-\beta a R^2/(V-1)$. Factoring the off-diagonal exponential out of the denominator gives

$$P(i,i) = \frac{e^\zeta}{e^\zeta + V - 1}, \qquad P(i,j) = \frac{1}{e^\zeta + V - 1} \quad (i \neq j).$$

These entries are exactly those of the first matrix in (G.44). The matrices $I_V - \mathbf{1}\mathbf{1}^\top/V$ and $\mathbf{1}\mathbf{1}^\top/V$ are complementary orthogonal projections, so taking the $T$-th power proves the second identity.

The continuous identity (G.45) follows from $F^T(z) = a^T z$. Subtracting rows $i$ and $j$ of the second matrix in (G.44) gives $\eta^T(\boldsymbol{e}_i - \boldsymbol{e}_j)$, where $\boldsymbol{e}_i$ denotes the $i$-th coordinate vector. Its total variation norm is $\eta^T$, proving (G.46). Because the token embeddings are centered,

$$\bar{e}_T(i) = \sum_{j=1}^{V} P^T(i,j)e_j = \eta^T e_i.$$

Subtracting two such means proves (G.47) and hence (G.48).

The tokenized exponent is smaller precisely when $\eta > a$. Substitution of (G.43) and rearrangement give

$$e^\zeta > \frac{1+a(V-1)}{1-a},$$

which is equivalent to (G.49). Finally,

$$1 - \eta = \frac{V}{e^\zeta + V - 1} = Ve^{-\zeta} + O(e^{-2\zeta}),$$

and $-\log\eta = (1-\eta) + O((1-\eta)^2)$ proves (G.50). □

The exact memory contraction in Theorem H still does not imply contraction between independent samples. Indeed, if $I_T$ and $J_T$ are sampled independently from rows $i \neq j$ of $P^T$, then

$$\mathbb{P}\{I_T = J_T\} = \frac{1-\eta^{2T}}{V}, \qquad \mathbb{E}\|e_{I_T} - e_{J_T}\| = \delta_{\mathrm{tok}}\left(1 - \frac{1}{V} + \frac{\eta^{2T}}{V}\right). \tag{G.51}$$

Thus the conditional means converge together at factor $\eta$, while independent sample distance converges to the positive value $\delta_{\mathrm{tok}}(1 - 1/V)$.

For deterministic nearest-token decoding, the same model satisfies $Q(ae_i) = e_i$, as in Proposition D; its token contraction factor is exactly one even though the continuous factor is $a < 1$. At the opposite extreme, a constant decoder has token contraction factor zero. Sampling produces intermediate memory factors only after the transition kernel has been specified.

### RATE SUMMARY

The contraction statements can be summarized as follows. A factor is a per-step multiplier; its corresponding positive exponent is the negative logarithm of that factor.

| Observable | Uniform factor or bound | Decay exponent |
|---|---|---|
| Continuous pairwise distance | $a$ | $\log(1/a)$ |
| Continuous covariance trace | $a^2$ | $2\log(1/a)$ |
| Token initial-state memory | $\eta$ | $\log(1/\eta)$ |
| Independent token samples | floor $\delta_{\mathrm{tok}}(1-\kappa)$, $\kappa < 1$ | $0$ |
| Absorbing-coalescence benchmark | at least $1-\kappa$, $\kappa < 1$ | at most $\log \frac{1}{1-\kappa}$ |

The token-memory factor $\eta$ and the collision bound $\kappa$ are not interchangeable. The former compares different transition rows and controls forgetting of initial conditions; the latter controls agreement of independent samples and therefore their diversity. For general, non-simplex embeddings with minimum and maximum distinct-token distances $\delta_{\min}$ and $\delta_{\max}$, the exact identity (G.30) is replaced by

$$\delta_{\min}\,\mathrm{TV}(\mu,\nu) \le W_1(\mu,\nu) \le \delta_{\max}\,\mathrm{TV}(\mu,\nu),$$

provided $\delta_{\min} > 0$. The qualitative distinction between continuous multiplicative contraction, token-channel mixing, and independent-sampling diversity remains unchanged.