LaTense: Measuring the Limits of Geometric Activation Steering

Anonymous authors
Paper under double-blind review

Abstract

Claims that activation steering causes or prevents "latent collapse" are routinely supported by n-gram repetition rates. We show that this metric is governed by output length rather than by the intervention. Across **19,268 generations from 56 runs** spanning six models from 0.5B to 9.2B parameters, three families, three benchmarks and five steering configurations, generated length and token 3-gram repetition correlate at **Spearman** $\rho = 0.881$ (Pearson $r = 0.907$): mean repetition rises monotonically from 0.00% for generations under eight tokens to 49.88% above 549 tokens, while the gap between steered and unsteered decoding never exceeds 1.1 points and a 15$\times$ increase in parameters moves it by under 2.5. A correct polynomial long division scores 68.9% repetition. A shuffled-token control shows why: the metric conflates task-intrinsic structural repetition, such as the restated scaffolding of a long derivation, with degenerate looping, and only the latter is a failure. We reach this through a controlled study of **LaTense** (**Lat**ent S**ense**), a dynamic steering framework that modulates intervention strength per token from the local geometric alignment between the hidden state and a reasoning vector, evaluated on Llama 3.1 8B Instruct, Gemma 2 9B IT and Qwen-2.5-7B-Instruct against unsteered decoding and static Contrastive Activation Addition (CAA) at matched intervention layers and matched decoding. The method helps on one model–task pair of nine (Gemma 2 9B IT on StrategyQA, **74.20%**, **+8.60** over static CAA, $p < 0.0001$, at 122.7 generated tokens per problem) and is neutral or harmful elsewhere; its cosine gate is inert in practice, since measured alignment stays near $\cos(h, v) \approx 0.03$, and a norm-scaling-only ablation reproduces that single gain exactly. We conclude that a single linear direction confers limited benefit regardless of how it is modulated, and that repetition-based collapse metrics must report output length, or compare at matched length, in order to be interpretable. Code and evaluation scripts are available in the supplementary material and at `https://anonymous.4open.science/r/latense`.

1 Introduction

Large language models (LLMs) often exhibit brittle reasoning paths despite containing the necessary cognitive primitives within their weights (Wei et al., 2022; Kojima et al., 2022). Traditional methods for steering these models toward desired behaviors typically rely on few-shot prompting (Brown et al., 2020), static activation addition (Rimsky et al., 2024; Turner et al., 2024), or compute-heavy sampling (Wang et al., 2023; Snell et al., 2024). However, static interventions provide a uniform "pressure" that can overwrite critical factual information, particularly in highly optimized "dense" models where representations are packed into tightly entangled manifolds (Elhage et al., 2021; Geva et al., 2021b).

We began from the hypothesis that activation steering should behave as a restorative governor rather than a constant force: that modulating intervention strength by the local geometry of the latent state h relative to a reasoning vector v, damping the intervention where the model is already aligned and increasing it where the trajectory diverges, would preserve the factual circuits that uniform addition disturbs and bridge "thinking states" (Amos et al., 2026) to linguistic output. This paper reports the result of testing that hypothesis, which is largely negative. The cosine alignment term motivating the design is inert at the scales we examine, since hidden states remain nearly orthogonal to any fixed direction in high dimension, and the method consequently reduces to norm-matched static

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.

steering. Pursuing that negative result surfaced the measurement problem that occupies most of what follows: the repetition statistic by which such interventions are commonly judged stable or unstable is governed by output length rather than by the intervention itself.

Our contributions are:

1. **A measurement that invalidates a common diagnostic.** Over 19,268 generations from 56 runs, token 3-gram repetition is predicted by output length ($\rho = 0.881$), not by whether or how the model was steered, nor by model scale: across six models from 0.5B to 9.2B parameters, mean output length predicts mean repetition ($r = +0.967$ on MATH-500) while parameter count does not ($r = -0.288$). A shuffled-token control (5.93% actual vs. 0.20% shuffled) shows the effect is structural recurrence rather than chance collision, so the metric cannot distinguish a long, correct derivation from a degenerate loop. We propose reporting the excess over that control, or a length-matched reference, in its place.
2. **A controlled evaluation of dynamic geometric steering** against matched-layer static CAA across three model families and three benchmarks, under matched deterministic decoding and paired bootstrap testing. Significant gains are confined to one model–task pair of nine; elsewhere the method is statistically indistinguishable from, or significantly worse than, its baselines.
3. **An explanation of the mechanism's limited effect.** Hidden states remain nearly orthogonal to the reasoning vector ($\cos(h, v) \approx 0.03$), so the alignment gate remains at ≈ 0.97 throughout generation and the method reduces to norm-matched static steering, which reproduces its one gain exactly.
4. **Failure modes that silently corrupt steering evaluations**, documented from our own: an out-of-range intervention layer that attaches no hook and returns unsteered output, accuracy means computed over incomplete runs, and comparisons that mix sampled and greedy decoding, which in our case manufactured two significant results that vanish once decoding is matched.

2 Related Works

Latent Steering and Adaptive Orchestration: Latent steering builds on the foundational work of representation engineering (Zou et al., 2023) and activation engineering (Turner et al., 2024). These techniques leverage Concept Activation Vectors (TCAV) (Kim et al., 2018) and linear relation decoding (Hernandez et al., 2024) to demonstrate that hidden states can be directly manipulated to induce desired behaviors. The validity of mean-difference steering vectors is grounded in the linear representation hypothesis (Marks & Tegmark, 2024; Park et al., 2024), which establishes that semantic concepts are encoded as linear directions in residual stream space. The shift from static methods like Contrastive Activation Addition (CAA) (Rimsky et al., 2024) to adaptive latent steering has been catalyzed by the limitations of uniform vector addition. Notably, RISER (Ye et al., 2026) introduces an RL-trained router for adaptive steering, maintaining a library of "cognitive primitives" and dynamically mixing them. As of this writing, a direct empirical comparison against RISER is not feasible, as the authors have not publicly released the pre-trained router weights or primitive vector libraries required for independent reproduction. In contrast, LaTense is a zero-shot, purely mechanical intervention applying a single reasoning vector under a closed-form geometric rule. We note at the outset that the cosine dampening term motivating this design is shown in Section 5.4 to be inactive in practice; we retain the comparison because the resulting norm-scaled intervention remains a well-defined baseline. Concurrently, ODESteer (Zhao et al., 2026) proposes a unified ODE-based theoretical framework for activation steering. While ODESteer targets alignment (TruthfulQA), LaTense is complementary, focusing on dynamic token-level modulation for reasoning tasks without the overhead of numerical ODE solving.

Dynamic Compute and Rescaling: Fractional Reasoning (FR) (Liu et al., 2025) and Inference-Time Intervention (ITI) (Li et al., 2023) propose adjusting inference dynamics to improve performance. However, FR adjusts intensity primarily per input. LaTense differs in that its per-token intensity follows a closed-form geometric rule, requiring neither a learned router nor a numerical solver; we do not claim priority for per-token adaptivity itself. This token-level granularity is critical during heterogeneous reasoning trajectories; for instance, when a model transitions from straightforward text

generation to complex arithmetic derivation, its latent alignment to the reasoning vector fluctuates significantly. This connects to broader efforts in scaling LLM test-time compute (Snell et al., 2024), providing an efficient alternative to high-k sampling.

Latent Reasoning and Stability: The exploration of latent reasoning paths has revealed significant challenges in maintaining stability. Amos et al. (2026) recently introduced "Thinking States," highlighting the phenomenon of 'latent thought corruption' in dense models. That work motivated the Model Density Hypothesis we set out to test, namely that corruption risk in entangled latent spaces (Elhage et al., 2021; Geva et al., 2021b) would make dense models resistant to linear intervention. Section 7.3 reports that our data do not support this hypothesis and identifies output length as the confound that originally appeared to.

Measuring Steering-Induced Degeneration: Steering papers routinely report an n-gram diversity or repetition statistic as evidence that an intervention does or does not destabilize generation. Braun et al. (2026) evaluate steering vectors for summarization and report that high steering strengths "consistently induce degenerate repetition," measured with word- and character-level Distinct-2 and reported without an accompanying output-length control. Herbster et al. (2026) report within-turn and cross-turn 4-gram repetition to separate steering methods, and observe that the method exhibiting the most repetition is also the one that inflates response length from roughly 360 to 1000 tokens at high coefficients. That pairing is precisely the confound we quantify: the statistic read as a stability signal is governed by how much text the model produced. We note that Herbster et al. (2026) do guard against it, restricting their metrics to the first 50 tokens "to ensure a fair comparison across conditions that produce different response lengths" and reporting unsteered baselines alongside. Our contribution is to show that this precaution is necessary rather than optional, and to quantify what is lost without it.

The length sensitivity of diversity statistics is established in dialogue evaluation: Liu et al. (2022) show that Distinct-n systematically penalizes longer sequences and propose an expectation-adjusted correction. That correction targets a chance-collision bias, the elevated probability that a long sequence repeats an n-gram by accident. Our shuffled-token control (Section 6) establishes that the excess we measure is not of this kind: shuffling words within a generation, which holds length and vocabulary exactly fixed, reduces repetition from 5.93% to 0.20%. The recurrence is structural rather than accidental, so an expectation-based adjustment does not remove it and a length-matched or shuffled reference is required in its place.

3 Methodology

3.1 The LaTense Formulation: Derivation and Intuition

Our framework intervenes directly on the residual stream representations (Vaswani et al., 2017) at each token generation step. By adaptively rescaling the steering vector based on real-time hidden state geometry, we preserve the model's structural integrity while directing it toward more coherent reasoning trajectories. Specifically, LaTense modulates the steering vector v at layer L using two geometric signals extracted from the current hidden state h.

1. Alignment-Proportionality (Cosine Penalty): We hypothesize that an intervention is most necessary when the model's latent state h is orthogonal to or moving away from the desired reasoning direction v. We define the alignment penalty G_{align} as:

$$G_{align}(h, v) = 1 - \cos(h, v) = 1 - \frac{h \cdot v}{\|h\|\|v\|} \tag{1}$$

Geometric Intuition: Geometrically, G_{align} acts as a restorative force, bending the model's trajectory back toward the reasoning vector v when it diverges. By scaling the intervention relative to $(1 - \cos(h, v))$, the intervention is attenuated proportionally to the model's current alignment. Note that in high-dimensional latent spaces, full collinearity ($\cos(h, v) = 1$) is not expected in practice; the mechanism provides continuous geometric dampening rather than a hard shutoff.

2. Manifold-Consistency (Norm Scaling): We define the norm scaling factor as $G_{norm} = \frac{\|h\|}{\|v\|}$. Scaling the intervention proportionally to the hidden state's relative magnitude preserves the local representational scale and prevents the steering vector from dominating or being drowned out by states with varying norms.

Combining these, the final LaTense intervention is defined as:

$$h' = h + \alpha \cdot (1 - \cos(h, v)) \cdot \frac{\|h\|}{\|v\|} \cdot v \tag{2}$$

Note that v enters only through its direction: $\|h\|/\|v\|$ and the cosine are scale-free, so $\|v\|$ cancels exactly (Section 5.2).

4 Experimental Setup

Table 1: **Main Performance Results.** All rows use greedy decoding, which is deterministic, so the three evaluation seeds give identical generations and we report single values. An earlier version reported TriviaQA from temperature-0.7 samples against greedy baselines; those rows have been re-run greedily so that every column shares a decoding scheme. p: two-sided paired bootstrap vs. Static CAA ($B = 10{,}000$). Static CAA is the last-layer baseline; matched-layer static steering appears in Table 2. MATH-500 static runs are scored on the same first 200 problems as LaTense. †: baseline not re-verifiable from retained run records. ‡: LaTense significantly *worse* than Static CAA.

Dataset	Model	Greedy ($\pm\sigma$)	Static CAA ($\pm\sigma$)	LaTense ($\pm\sigma$)	Gain (vs Greedy)	p-value (vs CAA)
StrategyQA	Llama 3.1 8B Instruct	72.40%†	72.00%	71.20%	-1.20%	p = 0.6430
StrategyQA	Gemma 2 9B IT	66.00%	65.60%	**74.20%**	**+8.20%**	**p < 0.0001**
MATH-500	Llama 3.1 8B Instruct	46.00%	45.50%	43.50%	-2.50%	p = 0.5417
MATH-500	Gemma 2 9B IT	51.50%	53.50%	45.00%	-6.50%	p = 0.0085‡
TriviaQA	Llama 3.1 8B Instruct	76.60%	76.40%	76.00%	-0.60%	p = 0.7532
TriviaQA	Gemma 2 9B IT	68.60%	68.60%	66.80%	-1.80%	p = 0.0704

We evaluate LaTense on Llama 3.1 8B Instruct (Dubey et al., 2024), Gemma 2 9B IT (Gemma Team, 2024), and Qwen-2.5-7B-Instruct (Qwen Team, 2024) across StrategyQA (Geva et al., 2021a), MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024), and TriviaQA (Joshi et al., 2017). For the metric analysis in Section 6 only, we additionally run unsteered greedy decoding on Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct and Gemma-2-2B-IT, extending the parameter range to 0.5B–9.2B. No intervention is applied to these three models and they take no part in the steering comparisons.

Reasoning vectors. Vectors come from a mean-difference construction over the model's own generations, not hand-written prompt pairs: for each extraction problem we generate an unsteered answer, average final-layer hidden states over the generated tokens, and score it with the benchmark's own judge, giving $v = \mu_{\text{correct}} - \mu_{\text{incorrect}}$. Extraction uses 100 problems for MATH-500 (`test[400:]`), 500 for TriviaQA (`train[200:]`), and 500–1000 for StrategyQA (`test[1200:]`), each disjoint from the corresponding evaluation slice. Vectors are extracted at the final layer but applied at intermediate layers.

Hyperparameters and decoding. We use $\alpha = 0.3$ throughout, except TriviaQA ($\alpha = 0.1$). Intervention layers are 24 (Llama), 20 (Gemma) and 26 (Qwen) for StrategyQA, 28 for MATH-500, and 24 for TriviaQA, selected on a disjoint 50-problem tuning slice with a 256-token cap (Section 5.3); these choices were not always the sweep optimum. All evaluations use greedy decoding, which is deterministic: seeds 42–44 give byte-identical generations, so we report single values rather than a seed standard deviation. TriviaQA was originally run with temperature-0.7 sampling and has been re-run greedily so that every column of every table shares one decoding scheme.

Baselines. We compare against unsteered decoding and static CAA (Rimsky et al., 2024), both at the final layer (Table 1) and at LaTense's own layer (Table 2). Self-Consistency ($k = 5$) (Wang et al., 2023) is used for the compute comparison in Section 7.4; its cost is reported as k times the mean length of a sampled chain. We do not report a Fractional Reasoning (Liu et al., 2025) comparison: our implementation was validated only on a 10-problem pilot on which every method scored identically, too small to support a baseline claim. All evaluations are in English using official instruction-tuned checkpoints, and all models and benchmarks are public assets used within their intended research contexts.

4.1 Prompt Count Sensitivity Setup

To test the robustness of our extracted reasoning vectors, we perform a sensitivity analysis by varying the prompt pair sample size $N \in \{10, 20, 50, 100, 200\}$ during vector extraction. We construct vectors for Llama 3.1 8B Instruct on StrategyQA using each sample size N and report the corresponding vector norms and downstream evaluation accuracies on a disjoint 100-sample test slice.

4.2 Infrastructure and Compute

All experiments were conducted on a single NVIDIA A100-SXM4-40GB GPU. We utilized the following models: Llama 3.1 8B Instruct (8.0B parameters), Gemma 2 9B IT (9.2B parameters), and Qwen-2.5-7B-Instruct (7.6B parameters). The total computational budget for the final results, including vector extraction, hyperparameter sweeps (layer and α), and baseline evaluations, is estimated at approximately 180 active GPU-hours.

5 Results and Analysis

Table 2: **Component Isolation Ablations** ($\alpha = 0.3$, seed 42), all configurations at LaTense's intervention layer (24/20/26 for StrategyQA, 28 for MATH-500). Repetition is the token 3-gram rate $1 - |\text{unique}|/|\text{total}|$. Unsteered greedy decoding on the same 200 MATH-500 problems scores **35.82%** (Llama), **23.95%** (Gemma) and 38.07% (Qwen), so the rates in this table are the base rate of long mathematical derivations rather than a steering effect. *: these evaluation runs did not log repetition. §: the original Qwen MATH-500 ablations applied the intervention at a layer index outside the model and so measured unsteered decoding; the row shown is re-run at layer 20, for which we repeated LaTense and matched-layer CAA but not the two single-component controls (Section 7.3).

Model & Task	Metric	Cosine-Only	Norm-Only	CAA-Optimal (Static)	LaTense (Ours)
Gemma 2 9B IT					
StrategyQA	Accuracy (%)	67.40%	**74.20%**	67.80%	**74.20%**
	Repetition (%)	3.17%	4.96%	3.15%	n/a*
MATH-500	Accuracy (%)	**52.50%**	45.00%	**53.50%**	45.00%
	Repetition (%)	23.63%	24.71%	23.97%	n/a*
Llama 3.1 8B Instruct					
StrategyQA	Accuracy (%)	**72.40%**	72.00%	**72.40%**	71.20%
	Repetition (%)	3.96%	4.72%	3.87%	n/a*
MATH-500	Accuracy (%)	**46.00%**	**47.00%**	45.50%	43.50%
	Repetition (%)	36.12%	33.87%	35.92%	n/a*
Qwen-2.5-7B-Instruct					
StrategyQA	Accuracy (%)	**67.60%**	67.40%	67.40%	66.00%
	Repetition (%)	5.18%	6.02%	5.18%	n/a*
MATH-500§	Accuracy (%)	—	—	**70.50%**	66.50%
	Repetition (%)	—	—	38.79%	39.13%

5.1 Main Experimental Results and Statistical Significance

The headline result is narrow. On StrategyQA with Gemma 2 9B IT, LaTense achieves **74.20%**, a **+8.60** point gain over Static CAA and **+8.20** points over greedy decoding (Table 1). A two-sided paired bootstrap test with $B = 10{,}000$ resamples gives $p < 0.0001$: no resample of the 500 evaluation problems reverses the ordering. This is the only model–task pair in our grid where dynamic steering significantly beats its steering baseline.

Elsewhere the picture is unfavourable. Most differences are not significant: Llama StrategyQA ($p = 0.6430$) and, once decoding is matched, all three TriviaQA pairs ($p = 0.7532$, 0.0704, 0.0538). Two are: LaTense is significantly worse than Static CAA on Gemma MATH-500 (-8.50, $p = 0.0085$), and significantly worse than unsteered greedy decoding on Qwen MATH-500 (-5.50, $p = 0.0435$). Against unsteered decoding, LaTense improves only on Gemma StrategyQA and is otherwise flat or negative. We therefore do not claim that geometric modulation improves reasoning in general.

Multiple comparisons. Table 1 reports six tests against the same baseline, so individual p-values overstate significance. Applying Holm–Bonferroni correction across those six comparisons, the Gemma StrategyQA gain remains significant (adjusted threshold 0.0083), as does the Gemma MATH-500 loss ($p = 0.0085$ against a threshold of 0.0100). Treating all nine model–task pairs as a single family leaves only the Gemma StrategyQA result significant, since the Gemma MATH-500 comparison is then tested against a threshold of at most 0.0071. We therefore treat the single positive result as robust to correction and the negative results as directional evidence rather than individually established effects.

5.2 Prompt Count Sensitivity Analysis

Varying the number of extraction problems from 10 to 200 shrinks the vector norm from 45.25 to 10.88 while accuracy stays between 68.0% and 70.0% (Table 8). Invariance to $\|v\|$ is a property of Equation 2 rather than a finding, since $\|h\|/\|v\|$ and the cosine are both scale-free; what the sweep does show is that the extracted *direction* is stable from as few as 10 problems. Note that this sweep re-extracts vectors through default split handling, which for StrategyQA overlaps the slice evaluated here, so it measures stability rather than generalization.

5.3 Layer Sensitivity Analysis

We sweep candidate layers $\{20, 24, 28, 31\}$ on MATH-500 and $\{16, 20, 24, 28, 31\}$ on TriviaQA over a disjoint 50-problem slice with a 256-token cap, so the sweep is informative for ranking only (Figure 4). The ranking is shallow and inconsistent: layer 31 for Llama, 28 for Gemma, 20 for Qwen. Our MATH-500 runs used layer 28 for all three, so two configurations were not the sweep optimum, and for Qwen, whose 28 layers are indexed 0–27, that index lies outside the model entirely (Section 7.3).

5.4 Component Isolation Ablations

To isolate the geometric contributions of our alignment gating and magnitude scaling components, we evaluate ablated controls on StrategyQA and MATH-500 (Table 2). We compare: (i) **Cosine-Only** (disabling norm scaling), (ii) **Norm-Only** (disabling cosine similarity gating), and (iii) **CAA-Optimal** (disabling both, which matches the static steering baseline at the optimal layer).

The ablations support one component and undercut the other. **(i) Norm scaling carries the Gemma gain:** removing it (*Cosine-Only*) drops Gemma StrategyQA from **74.20% to 67.40%**, near the matched-layer static baseline of 67.80%, while *Norm-Only* reaches 74.20% on exactly the same 371 of 500 problems as full LaTense, so on this task the cosine gate changes no prediction. **(ii) The gate is nearly inert:** measured alignment is small in every logged run, $\cos(h, v) \approx 0.02$–0.07, so $(1 - \cos)$ stays within a few percent of unity throughout generation and the intervention reduces to norm-matched static steering. **(iii) The components trade off by task:** on Gemma MATH-500, removing norm scaling *improves* accuracy (52.50% vs. 45.00%), reversing the StrategyQA ordering.

We previously reported that cosine gating eliminates a repetition collapse induced by static steering. That claim is withdrawn. Unsteered greedy decoding over the same 200 MATH-500 problems gives **35.82%** (Llama) and **23.95%** (Gemma) against **35.92%** and **23.97%** for matched-layer static CAA, so steering moves the rate by $+0.10$ and $+0.02$ points and there is no collapse at the operating point for a gate to cure. Section 6 shows this is one instance of a general length relationship; Section 7.2 reports where collapse is real.

6 Repetition Tracks Output Length, Not the Intervention

Section 5.4 established that steering leaves repetition unchanged on MATH-500. The relationship generalizes beyond that setting. Pooling every generation for which our harness recorded both a token count and a repetition rate gives 19,268 generations from 56 runs, spanning six models from 0.5B to 9.2B parameters across three families, all three benchmarks, and five configurations (unsteered greedy, static CAA at the final layer, matched-layer CAA, the two single-component ablations, and LaTense). Across that pool, generated length predicts repetition with **Spearman** $\rho = 0.881$ (Pearson

$r = 0.907$). Table 3 shows the relationship by length decile: mean repetition rises monotonically from 0.00% to 49.88%, a span two orders of magnitude larger than any steering effect we measured.

Table 3: Token 3-gram repetition by generated length, pooled over 19,268 generations from 56 runs covering six models (0.5B–9.2B), three benchmarks and five steering configurations. Repetition is monotone in length regardless of whether, where, or how strongly the model was steered.

Generated length (tokens)	Generations	Mean repetition
1–8	1,898	0.00%
8–19	1,735	0.00%
19–42	2,116	0.55%
42–77	1,943	2.28%
77–109	1,888	3.61%
109–153	1,960	5.87%
153–228	1,928	11.93%
228–338	1,941	19.69%
338–549	1,920	31.70%
549–1024	1,939	49.88%

The same ordering holds across tasks, which differ principally in the amount of text they require: TriviaQA averages 18.3 tokens and 0.74% repetition, StrategyQA 124.5 tokens and 4.42%, MATH-500 470.7 tokens and 33.47%. Within TriviaQA the correlation is negligible ($\rho = 0.13$), since nearly all answers are short; it is strongest on MATH-500 ($\rho = 0.76$), where generated lengths vary widely.

What the metric detects. Length alone does not account for the effect. For the 3,636 generations whose text we retained, we recompute repetition after randomly shuffling the words within each generation, which holds length and vocabulary exactly fixed and destroys only word order. Actual repetition averages 5.93% against **0.20%** for the shuffled control, and the gap widens with length: $+0.00$ points below 9 words, $+2.34$ at 22–120, $+11.18$ at 120–249, $+20.68$ at 249–752. (These figures use whitespace-delimited words rather than model tokens, so their absolute values differ from the tokenizer-based rates elsewhere in the paper; only the within-tokenization contrast matters.) The result is not specific to $n = 3$. Recomputing both the actual and shuffled rates for $n \in \{2, 3, 4, 5\}$ over the 3,446 retained generations of at least five words, the rank correlation with length remains high at every order ($\rho = 0.804, 0.769, 0.742, 0.693$) and the excess over the shuffled control remains substantial ($+8.32$, $+6.04$, $+3.96$, $+2.79$ points). Repetition at these lengths is therefore not an artifact of chance n-gram collision in long sequences. It reflects genuine recurrence of surface form, such as the "divide, multiply, subtract, bring down" scaffolding of long division restated at each elimination step.

This property renders the metric unreliable as a stability diagnostic. It assigns equivalent scores to two distinct phenomena, *task-intrinsic structural repetition* and *degenerate looping*, of which only the latter constitutes a failure. A correct polynomial long division scoring 68.9% (Appendix C) and a generation that emits a single repeated conjunction both fall within the 24–38% band. This also accounts for the coexistence of the two figures withdrawn in this version: a 0.00% rate on 6.8-token Gemma TriviaQA answers and a 38.07% "collapse" on 546-token Qwen derivations to which no intervention was applied are two points on the curve in Table 3.

We therefore recommend reporting repetition against a shuffled-token control of the same generation, or against a length-matched unsteered reference, rather than as a bare rate. The excess over the shuffled control isolates genuine recurrence from what length and vocabulary alone produce, and a bare rate will otherwise reward any intervention that merely shortens its outputs. LaTense, for example, increases the length of Llama's TriviaQA answers from 23.6 to 33.5 tokens, which alters its expected repetition rate independently of any change in generation quality.

The relationship is invariant to model scale. Table 4 reports unsteered greedy decoding on MATH-500 for six models spanning a 18× range in parameters. Mean output length predicts mean repetition across them ($r = +0.967$), while parameter count does not ($r = -0.288$). Within the Qwen family a 15× increase in parameters, from 0.5B to 7.6B, changes repetition by 2.5 points and does so non-monotonically (38.31%, 36.67%, 39.13%), whereas the roughly 200-token difference in mean

output length relative to Gemma corresponds to a change of 14 points. The same ordering holds on TriviaQA, where every model produces short answers and every repetition rate falls below 2.1%.

Table 4: Unsteered greedy decoding on MATH-500 across six models. Repetition corresponds to mean output length rather than to parameter count; ρ is the per-example Spearman correlation between length and repetition within each model.

Model	Params	Mean tokens	Repetition	ρ (len, rep)
Qwen2.5-0.5B	0.5B	556.5	38.31%	0.798
Qwen2.5-1.5B	1.5B	554.9	36.67%	0.774
Gemma-2-2B	2.0B	388.0	27.98%	0.692
Qwen2.5-7B	7.6B	544.5	39.13%	0.825
Llama-3.1-8B	8.0B	479.3	35.82%	0.820
Gemma-2-9B	9.2B	355.4	23.95%	0.563

7 Steering Dynamics and the Limits of Linear Intervention

7.1 Token-Level Mechanism Trace

Figure 5 visualizes the token-by-token intensity (Δh). The modulation factor stays within roughly 0.85 to 1.0, never approaching the attenuation-to-zero regime that motivates the design: hidden states remain close to orthogonal throughout, so the gate neither shuts off on aligned tokens nor doubles on opposed ones, and the visible variation in $\|\Delta h\|$ tracks $\|h\|$ rather than alignment.

7.2 The Steering-Stability Trade-off

Degenerate repetition does occur under steering, but only at intensities well above the operating point, and it bounds every method we tested, including our own. At $\alpha = 0.3$ steering leaves repetition statistically indistinguishable from unsteered decoding (Section 5.4); the failure mode emerges only as α increases. Figure 1 sweeps α under the full LaTense formulation, plotting accuracy against the token 3-gram repetition rate. Repetition remains low through $\alpha = 0.3$, the operating point used throughout, then rises steeply beyond $\alpha \approx 0.5$, reaching 81% at $\alpha = 1.0$ on Gemma MATH-500 as accuracy approaches zero; individual StrategyQA generations reach 80.9% and 94.4% repetition, degenerating into a repeated conjunction or into tokens from another script. The dynamic gate does not prevent this, since it remains close to constant during generation (Section 5.4). These results therefore establish a ceiling on usable intervention strength rather than a mechanism that removes the failure mode.

7.3 The Model Density Hypothesis and the Limits of Our Evidence

Qwen-2.5-7B-Instruct responds differently to steering than the other two models (Table 6), which an earlier version of this work read as evidence that hyper-optimized models entangle reasoning and factual representations. That evidence was an artifact: the MATH-500 configuration intervened at layer index 28, outside this 28-layer model, so the "steered" run is identical to unsteered decoding on all 200 problems, and its 38.07% repetition rate reflects the base rate for long derivations.

The residual evidence is substantially weaker. Re-running MATH-500 at layer 20 provides the first valid measurement for this configuration: LaTense reaches 66.50% against 70.50% for matched-layer static CAA and 72.00% unsteered, so the intervention *reduces* accuracy by 5.50 points ($p = 0.0435$) rather than yielding the $+2.11$ previously reported, and repetition is 39.13% against 38.79% steered and 38.07% unsteered, i.e. unchanged. On the two tasks where steering had always been genuinely applied, StrategyQA (layer 26) and TriviaQA (layer 24), both static CAA and LaTense leave accuracy essentially unchanged or slightly reduced. Qwen is therefore the model on which linear steering is least effective, but the underlying mechanism is a uniform reduction in accuracy rather than the representational collapse previously attributed to it. We therefore state the **Model Density Hypothesis (MDH)** as an open conjecture: architectural optimization, through high token-to-parameter ratios and aggressive parameter sharing, may yield manifolds in which a single linear direction cannot be added

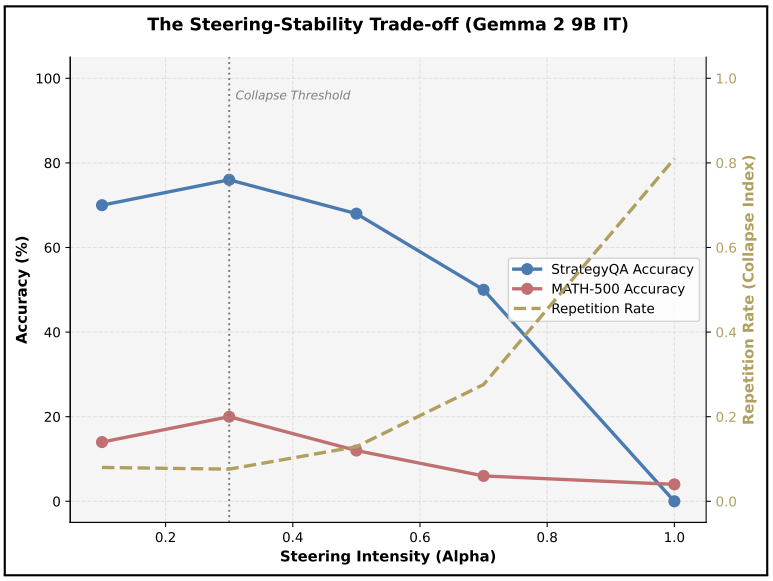

Figure 1: **Steering-Stability Trade-off for Gemma 2 9B IT.** All curves are full LaTense swept over α on a 50-problem tuning slice with a 256-token cap (StrategyQA layer 20, MATH-500 layer 28), so absolute accuracies are not comparable with Table 1; the sweep starts at $\alpha = 0.1$, with no unsteered point. High α collapses dynamic steering as it does static steering.

without disturbing unrelated circuits. Testing it properly requires varying density while holding scale fixed, or comparing against sparse mixture-of-experts models, neither of which our models support. One observation now weighs against it. The elevated repetition previously interpreted as a symptom of dense representational entanglement in Qwen is reproduced by Qwen2.5-0.5B (38.31%), a model to which the hyper-optimization account does not apply, and is accounted for by output length: Qwen produces derivations averaging approximately 545 tokens where Gemma produces 355 (Table 4). The observation that originally motivated the hypothesis is therefore a length effect, and we retain the hypothesis only as an open question concerning representational geometry rather than as a claim supported by our data.

7.4 Efficiency at Single-Sample Cost

On StrategyQA with Gemma 2 9B IT, LaTense generates 122.7 tokens per problem against 565.9 for Self-Consistency ($k = 5$, the total across five sampled chains averaging 113.2 tokens each), a 4.6$\times$ reduction, while scoring 9.8 points above that baseline (74.2% vs. 64.4%) and 8.2 above greedy decoding (66.0%). Self-Consistency underperforms greedy decoding on this task; the comparison therefore concerns computational cost rather than accuracy.

8 Conclusion

Our central result is about measurement rather than method: the n-gram repetition rate used throughout the steering literature to diagnose latent collapse is predicted by how much text a model produces ($\rho = 0.881$ over 19,268 generations spanning 0.5B to 9.2B parameters), not by whether it was steered nor by model scale. A shuffled-token control shows what it is detecting at length: genuine recurrence of surface form, the restated scaffolding of a long derivation, which the metric scores identically to a degenerate loop. Any claim that an intervention induces or eliminates collapse therefore needs the excess over such a control, or a length-matched reference; without one the metric will reward interventions that merely shorten outputs. We arrived at this by auditing our own framework, LaTense, which modulates steering strength per token from the local geometry of the hidden state. That audit is the second contribution: against matched-layer static steering under matched decoding, LaTense helps on one of nine model–task pairs (Gemma 2 9B IT on StrategyQA, +8.60 points, $p < 0.0001$), is indistinguishable on most others and significantly worse on two, and its gain is reproduced exactly by

norm scaling alone because the cosine gate stays near-constant in high dimensions. Taken together, the two results indicate that a single linear direction confers limited benefit regardless of how it is modulated, and that progress in activation steering currently depends less on novel gating functions than on measurements capable of distinguishing a genuine effect from an artifact of output length, an invalid layer index, or an incomplete run.

LIMITATIONS AND SOCIETAL IMPACT

Several boundary conditions limit what these results establish. The intensity ceiling (Figure 1) is not lifted by dynamic modulation: beyond $\alpha \approx 0.5$ both static and dynamic steering degenerate into repetition. Our headline comparison rests on a single model–task pair, and the norm-scaling-only ablation matches it exactly, so the gain cannot be attributed to the cosine gate. The Model Density Hypothesis remains untested as a statement about density: our three models occupy a narrow band of scale, and although we have now measured Qwen MATH-500 at a valid layer, what it shows is ordinary underperformance (-5.50 points against unsteered decoding) rather than the circuit collision the hypothesis predicts. The prompt-count sensitivity sweep (Table 8) re-extracted its vectors through default split handling, which for StrategyQA falls back to `test` and overlaps that sweep's own evaluation slice; the main results are unaffected, as their vectors come from `test[1200:]`. Repetition was not logged during the main evaluation runs, so those cells in Table 2 are unavailable; we have since measured unsteered baselines on MATH-500, which is what establishes that the rates in that table are not steering-induced, while claims about collapse at high α rest on the sweeps. Finally, performance depends on the quality of a direction derived from the model's own correct and incorrect generations. Regarding societal impact, because LaTense modulates internal trajectories without parameter fine-tuning, it inherently inherits—rather than fundamentally alters—the safety and toxicity profiles of its underlying base models. However, we note that any framework which successfully shifts the test-time efficiency frontier (reducing token cost by $4.6\times$) carries dual-use neutrality; while it democratizes access to capable reasoning systems, it could theoretically be dual-purposed to accelerate the generation of highly structured misinformation or automated deception at a lower computational cost.

REPRODUCIBILITY STATEMENT

To ensure complete reproducibility of all theoretical claims and experimental findings:

- **Code and Vectors:** All implementation source code for the LaTense geometric governor, baseline generation harnesses, and pre-extracted reasoning vectors across all three model families (Llama 3.1 8B, Gemma 2 9B, Qwen-2.5-7B) are provided in the supplementary material and hosted anonymously at `https://anonymous.4open.science/r/latense`.
- **Experimental Splits:** We maintain strict data disjointness across vector extraction, hyperparameter sweeps, and final evaluation splits as detailed in Section 4.
- **Hyperparameters and Environment:** Complete layer indices, steering coefficients (α), decoding settings, and compute environment specifications are documented in Section 4, together with the vector-extraction procedure and the known StrategyQA split overlap.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. All experiments utilize standard, publicly available academic evaluation benchmarks (StrategyQA, MATH-500, TriviaQA) and publicly accessible open-weight foundation models. Our framework performs test-time inference steering without modifying underlying model weights, thereby inheriting the baseline safety profiles of the underlying models. We make no claim that the method mitigates degenerative looping; Section 7.2 reports that it does not.

AI USE STATEMENT

Large language model assistants were utilized in this work to assist with: (i) drafting, editing, and polishing manuscript prose and LaTeX typesetting; and (ii) experimental execution, including implementing hyperparameter sweep harnesses, pipeline scripts, and data logging. All research ideation, core hypotheses, theoretical formulations (including the dynamic geometric governor), and analytical conclusions were conceived, directed, and verified solely by the human authors, who assume full responsibility for the contents of this paper.

REFERENCES

Ido Amos, Avi Caciularu, Mor Geva, Amir Globerson, Jonathan Herzig, Lior Shani, and Idan Szpektor. Latent reasoning with supervised thinking states. *arXiv preprint arXiv:2602.08332*, 2026.

Joschka Braun, Carsten Eickhoff, and Seyed Ali Bahrainian. Beyond multiple choice: Evaluating steering vectors for summarization. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3849–3884, 2026.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Katherine G. Fowler, Katircioglu, Kavya Upasani, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Sheid, Louis Martin, Lovish Madaan, Luca Melis, Luke de Oliveira, Madian Khabsa, Maggie Cheng, Maithili Sapre, Manav Rathi, Manish Bhatt, Mani Sandhu, Manoj Ganesan, Manish Kumar Borthakur, Marc S. Salit, Mark Zuckerberg, Mark Tygert, Mary Williamson, Mason Boyd, and Matthew M. Muckley. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021a.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021b.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, 2021.

Niklas Herbster, Martin Zborowski, Alberto Tosato, Gauthier Gidel, and Tommaso Tosato. Activation steering for aligned open-ended generation without sacrificing coherence. In *Conference on Language Modeling (COLM)*, 2026.

Evan Hernandez, Arnab S Sharma, Benjamin Levinstein, Yonatan Belinkov, David Bau, and Jacob Andreas. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J Cai, James Wexler, Fernanda B Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pp. 2668–2677, 2018.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35: 22199–22213, 2022.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, Karl Cobbe, Ilya Sutskever, and John Schulman. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Sheng Liu, Tianlang Chen, Pan Lu, Haotian Ye, Yizheng Chen, Lei Xing, and James Zou. Fractional reasoning via latent steering vectors improves inference time compute. *arXiv preprint arXiv:2506.15882*, 2025.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. Rethinking and refining the distinct metric. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 762–770, 2022.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling (COLM)*, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Wencheng Ye, Xiaoyang Yuan, Yi Bin, Pengpeng Zeng, Hengyu Jin, Liang Peng, and Heng Tao Shen. RISER: Orchestrating latent reasoning skills for adaptive activation steering. *arXiv preprint arXiv:2601.09269*, 2026.

Hongjue Zhao, Haosen Sun, Jiangtao Kong, Xiaochang Li, Qineng Wang, Liwei Jiang, Qi Zhu, Tarek Abdelzaher, Yejin Choi, Manling Li, and Huajie Shao. ODESteer: A unified ODE-based steering framework for LLM alignment. In *International Conference on Learning Representations (ICLR)*, 2026.

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

A Static Steering Baseline Results

For completeness and traceability, we report the Static Steering (CAA) baseline (Rimsky et al., 2024) in Table 5. These runs use a fixed intensity ($\alpha = 0.3$), the same steering vectors as LaTense, greedy decoding, and the model's final layer as the intervention site; matched-layer static steering appears as "CAA-Optimal" in Table 2. Only complete runs are included: several seeds of this baseline terminated early, and averaging them with completed runs (as an earlier version of this table did) produced means and standard deviations dominated by runs of one to three problems. MATH-500 static runs covered a larger slice and are scored here on the same first 200 problems as LaTense.

Table 5: Static Steering (CAA) Baseline Performance.

Dataset	Llama 3.1 8B Instruct	Gemma 2 9B IT	Qwen-2.5-7B-Instruct
StrategyQA	72.00%	65.60%	67.20%
MATH-500	45.50%	53.50%	73.50%
TriviaQA	76.40%	68.60%	55.40%

B Qualitative Reasoning Examples

Unless noted otherwise, StrategyQA examples are drawn from the evaluation runs at the operating point ($\alpha = 0.3$). The two MATH-500 examples come from the disjoint tuning sweep at $\alpha = 0.1$, and the two Qwen examples contrast the same configuration at two intensities to illustrate the stability ceiling of Section 7.2; they are not a comparison against static steering, for which generated text was

not retained. Appendix C shows unsteered generations from the re-run evaluations, which do retain their text, and is the appendix to read for what the repetition metric measures.

Llama 3.1 8B Instruct — StrategyQA

Question: Are more people today related to Genghis Khan than Julius Caesar?
LaTense Response: Yes. The reason for this is due to the fact that Genghis Khan and the Mongol Empire had a much larger population and a more extensive geographic reach than Julius Caesar and the Roman Empire. As a result, Genghis Khan's genetic legacy has been spread across a much broader area, making it more likely that a significant portion of the world's population today is descended from him.

Llama 3.1 8B Instruct — StrategyQA

Question: Would a Monoamine Oxidase candy bar cheer up a depressed friend?
LaTense Response: No. The term "Monoamine Oxidase" refers to an enzyme involved in the breakdown of certain neurotransmitters in the brain, such as serotonin, norepinephrine, and dopamine, which are associated with mood regulation. A "Monoamine Oxidase inhibitor" (MAOI) is a type of medication used to treat depression by preventing the breakdown of these neurotransmitters, thereby increasing their levels in the brain.

Gemma 2 9B IT — StrategyQA

Question: Could the members of The Police perform lawful arrests?
LaTense Response: No. Here's why: The Police is a band, not a law enforcement agency. Only sworn law enforcement officers, who have received proper training and authorization, are legally allowed to make arrests.

Llama 3.1 8B Instruct — MATH-500

Question: A group of N students, where $N < 50$, is on a field trip. If their teacher puts them in groups of 8, the last group has 5 students. If their teacher instead puts them in groups of 6, the last group has 3 students. What is the sum of all possible values of N?
LaTense Response: To find the possible values of N, we need to consider the two given conditions: 1. When the students are grouped in 8, the last group has 5 students. This means that N = 8k + 5, where k is a non-negative integer. 2. When the students are grouped in 6, the last group has 3 students. This means that N = 6m + 3, where m is a non-negative integer. Possible values of N for condition 1 are 5, 13, 21, 29, 37, 45. Possible values of N for condition 2 are 3, 9, 15, 21, 27, 33, 39, 45. The common values are 21 and 45. The sum of all possible values of N is 21 + 45 = $\boxed{66}$.

Gemma 2 9B IT — MATH-500

Question: Twelve 1 by 1 squares form a rectangle. What is the total area of the shaded region? [Figure: two triangles with bases 2 and 3, height 4]
LaTense Response: Here's how to solve the problem step-by-step: 1. Identify the shapes: The shaded region is made up of two triangles. 2. Find the base and height of each triangle: - Triangle 1: Base = 2 squares, Height = 4 squares. - Triangle 2: Base = 3 squares, Height = 4 squares. 3. Calculate the area of each triangle: - Triangle 1: (1/2) * base * height = (1/2) * 2 * 4 = 4 square units. - Triangle 2: (1/2) * base * height = (1/2) * 3 * 4 = 6 square units. 4. Add the areas together: 4 + 6 = $\boxed{10}$ square units.

Qwen-2.5-7B-Instruct — StrategyQA (LaTense: high intensity vs. operating point)

Question: Does Biochemistry study gluons?
LaTense, $\alpha = 1.0$ (beyond the stability ceiling): No, Biochemistry does not study gluons. Here's the reasoning: 1. **Field of Study**: Biochemistry is the branch of biology and, by extension, of, and, and, and

Gemma 2 9B IT — same problem unsteered greedy — 3-gram repetition 64.2 percent

(answer scored correct)
Generation (917 tokens, excerpted): We can use polynomial long division to solve this. **1. Set up the division** … **5. Repeat steps 2–4:** Divide the new leading term ($-x^5$) by the leading term of the divisor (x) … Multiply the divisor by $-x^4$ … Subtract: `x^6 + 0x^5 + 0x^4 + 0x^3 + 0x^2 + 0x - 3` … *[the growing long-division tableau is reprinted in full after each elimination, so the same monomial rows recur verbatim]* … $\boxed{x^5 - x^4 + x^3 - x^2 + x - 1}$

Neither generation is degenerate. The repetition is the fixed vocabulary of long division in the first case and a re-printed tableau in the second, and the two models reach the metric's "collapse" range by entirely different routes. This is why we treat the 24–38% rates on MATH-500 as a property of the task, and why claims of eliminating repetition on such tasks require an unsteered reference before they mean anything. For contrast, the same method on TriviaQA, where answers run 7–34 tokens, scores 0.00%:

Llama 3.1 8B Instruct — TriviaQA LaTense greedy — repetition 0.00 percent

(alpha = 0.1, layer 24)
Q: Who was the man behind The Chipmunks? **A:** Ross Bagdasarian Sr., also known as David Seville, was the man behind The Chipmunks. *(correct)*
Q: Who was the next British Prime Minister after Arthur Balfour? **A:** The next British Prime Minister after Arthur Balfour was Sir Henry Campbell-Bannerman. *(correct)*

D SUPPLEMENTARY ANALYSES

Table 6: Performance on Qwen-2.5-7B-Instruct, all rows greedy. The MATH-500 row is re-run at layer 20, the best valid layer in our tuning sweep: an earlier version applied the intervention at layer index 28, which does not exist in this 28-layer model, so nothing was steered and the reported 74.44% was simply unsteered decoding. Its Static CAA entry is matched-layer (L20) rather than last-layer. ‡: LaTense significantly worse than unsteered greedy decoding (-5.50 points, two-sided paired bootstrap $p = 0.0435$).

Dataset	Greedy ($\pm\sigma$)	Static CAA ($\pm\sigma$)	LaTense ($\pm\sigma$)	Gain (vs Greedy)
StrategyQA	67.40%	67.20%	66.00%	-1.40%
MATH-500	72.00%	70.50%	66.50%	-5.50%‡
TriviaQA	55.00%	55.40%	53.20%	-1.80%

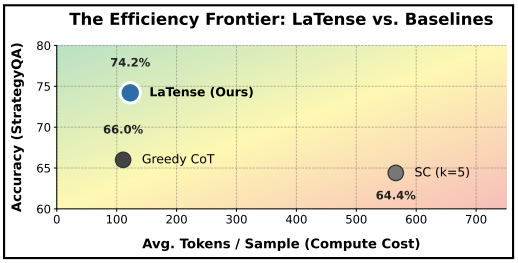

Figure 2: **Efficiency Frontier.** Accuracy vs. Compute Cost for Gemma 2 9B IT on StrategyQA.

Table 7: Efficiency (Gemma 2 9B IT, StrategyQA): a 4.6× token reduction vs. Self-Consistency ($k = 5$) at 1.1× greedy cost. SC tokens are the total across its five sampled chains (Section 7.4).

Method	Acc.	Tokens	Relative Cost
Greedy	66.0%	110.8	1.0×
SC k=5	64.4%	565.9	5.1×
LaTense	**74.2%**	**122.7**	**1.1×**

The layer sweep, the token-level trace and the prompt-count sensitivity results are reported here; they characterise the method rather than the measurement result, and are referenced from Sections 3, 5.2–5.3 and 6.4.

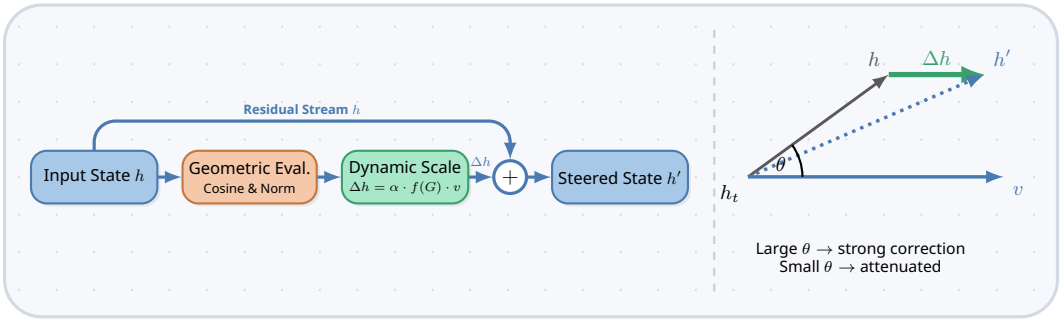

Figure 3: **The LaTense Mechanism.** (Left) Token-level processing pipeline with residual stream skip-connection ($h' = h + \Delta h$). (Right) Geometric intuition: the intervention Δh acts as a restorative force bridging the gap between current state h and reasoning vector v, with strength scaling by θ.

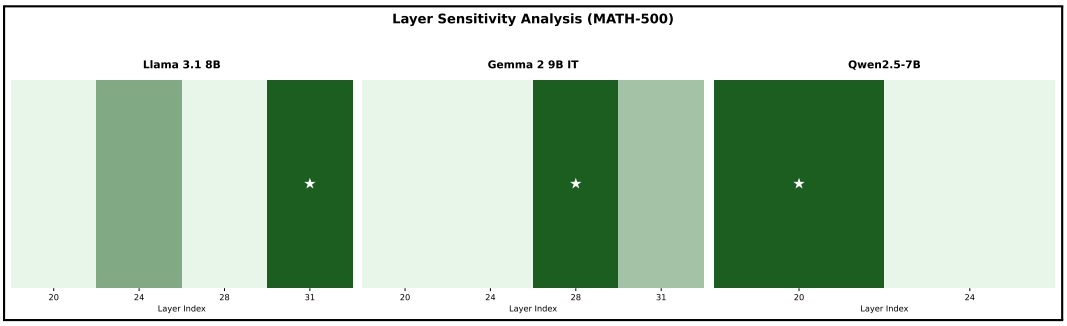

Figure 4: **Layer Sensitivity Analysis (MATH-500).** Colors represent relative performance normalized per model (1.0 = best layer). The optimal intervention layer is marked with a star and a bold border.

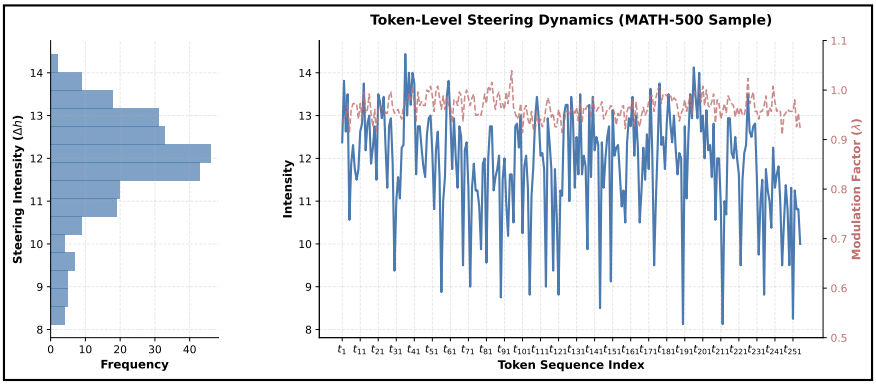

Figure 5: **Token-Level Trace:** Steering Intensity (Δh) across a MATH-500 generation. LaTense dynamically adjusts pressure based on alignment.

Table 8: Prompt count sensitivity analysis for Llama 3.1 8B Instruct on StrategyQA. Evaluation accuracy is reported on a disjoint 100-sample test slice.

Prompt Pairs (N)	**Vector Norm**	**Accuracy**
10	45.25	69.00%
20	31.75	69.00%
50	18.25	70.00%
100	13.75	68.00%
200	10.88	68.00%

C Asset, Dataset, and Licensing Details

- **Model Artifacts & Licensing:** All evaluations utilize the official open-source, instruction-tuned variants of Llama 3.1 8B Instruct (governed under the Meta Llama 3 Community License), Gemma 2 9B IT (governed under the Gemma Terms of Use), and Qwen-2.5-7B-Instruct (released under the Apache 2.0 License).
- **Benchmark Documentation & Language Scope:** The evaluation pipeline is bounded strictly within English-language distributions. We use three standard public benchmarks: StrategyQA (strategic common-sense logic), MATH-500 (multi-step mathematical derivations), and TriviaQA (open-domain question answering), fully cited and cross-referenced in the primary references. All assets are utilized strictly in accordance with their intended research-use parameters as specified by their original creators.
- **Data Ethics & Integrity:** No new or unvetted evaluation data was collected or crawled for this study. The public benchmarks utilized were heavily pre-vetted, pre-anonymized, filtered for offensive content, and cleared of personally identifying information (PII) by their original authors during curation, requiring no secondary human scrubbing.
- **Software Packages & Generation Parameters:** We utilize `transformers==4.46.3` for all model implementations and `datasets==2.14.6` for benchmark data loading. Mathematical answer verification for MATH-500 is performed using the `math_verify` and `latex2sympy` libraries to ensure robust symbolic normalization. All reported LaTense, static CAA and unsteered generations use greedy decoding (temperature = 0, top-p = 1.0); only Self-Consistency ($k = 5$) samples, at temperature 0.7. TriviaQA was initially run with temperature-0.7 sampling and has been re-run greedily so that no table mixes decoding schemes within a comparison. Flash Attention 2 is enabled across all experiments to ensure memory efficiency and throughput consistency.