The Progress Helix Tracks Reasoning Depth in Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) execute long chains of thought (CoT), but the mechanism by which they maintain a global sense of position within a reasoning trajectory has not been well characterized. To this point, we identify the Progress Helix: an emergent periodic trajectory in a low-dimensional latent subspace that completes exactly one revolution over a generated reasoning chain. Applying PCA and spectral analysis to activations from Llama-3.2, Gemma-2, and Qwen on both a controlled recursive arithmetic task and GSM8K, we recover a dominant Fourier component at fundamental frequency $k{=}1$ whose magnitude grows monotonically across layers, forming an elliptical cone in activation space. Most strikingly, applying Frequency Modulation (FM) to the helix magnitude allows targeted control over output length, with up to 80% decrease in output length at the same GSM8K accuracy in Llama-3.2. We thus offer a spectral and topological characterization of reasoning length as a controllable helical degree of freedom in the residual stream.

1 Introduction

The emergence of reasoning capabilities in LLMs, particularly through CoT (Wei et al., 2022), has fundamentally improved performance in reasoning tasks. Despite the longstanding empirical successes, the internal mechanisms by which these models coordinate long logical trajectories remain poorly understood. Specifically, a central question persists: How do models maintain a global sense of their position within a reasoning chain to ensure logical termination?

Current interpretability research often focuses on understanding localized features or training linear probes to uncover semantic concepts. However, mathematical reasoning naturally requires a dynamic reasoning progress tracker that is active during the output generation. In this work, we demonstrate that modern LLM architectures, such as Llama-3 (Grattafiori et al., 2024), Gemma-2 (Team et al., 2024) and Qwen (Hui et al., 2024), do not solely rely on positional embeddings to manage output depth. Instead, they develop an emergent Progress Helix, a periodic trajectory within a low-dimensional subspace that represents the model's internal state of completion.

In classical computation, a program's state is managed by explicit counters and pointers. In the high-dimensional latent space of an LLM, state management must be emergent. Understanding whether LLMs possess an internal reasoning tracker is vital for controlling sequence length, preventing logical loops, and improving the reliability of autoregressive generation.

We hypothesize that reasoning progress is encoded as a rotation in a specific latent manifold. Previous work has established that LLMs use helices to track line breaking (Gurnee et al., 2026). Multi-dimensional periodic features have also been documented for cyclic concepts such as days of the week and months of the year (Engels et al., 2024) and as the substrate for the "Clock" algorithm modern transformers use for multi-digit addition (Kantamneni and Tegmark, 2025). The Progress Helix is, to our knowledge, the first such structure identified for tracking the model's own position within a generated chain of thought. By applying Principal Component Analysis (PCA) and spectral analysis to the output activation matrices, we identify a dominant periodic signal at the fundamental frequency $k{=}1$. This signal represents exactly one full rotation over the duration of the reasoning chain. Unlike the baseline GPT-2 that lacks this structured periodicity, modern reasoning models exhibit an increasing magnitude signal as the activations flow through deeper layers, forming an elliptical cone in the latent space.

A critical challenge in identifying this mechanism is distinguishing it from Rotary Positional

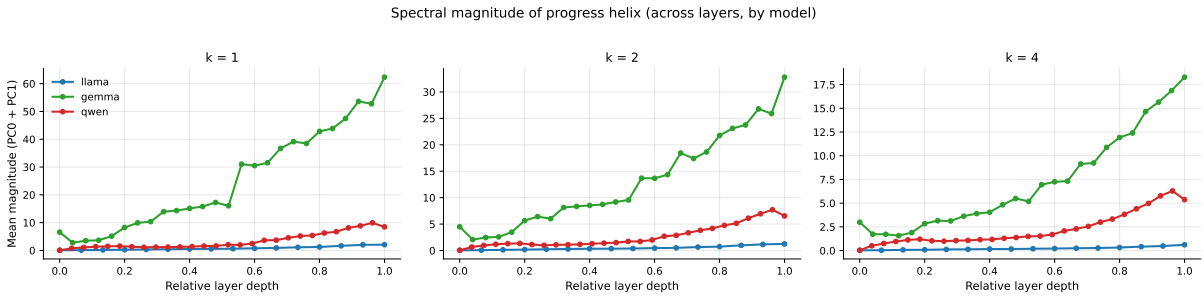

Figure 1: **Cross-model existence of the Progress Helix.** Mean magnitude of the k=1, k=2, and k=4 Fourier components in PC0+PC1 of the residual stream, plotted against relative layer depth across Llama-3.2-1B, Gemma-2-2B, and Qwen2.5-1.5B-Instruct. All three modern instruction-tuned models exhibit monotonic growth at the fundamental frequency, with weaker but consistent growth at the higher harmonics.

Embeddings (RoPE) (Su et al., 2024). While RoPE causes a linear rotation based on absolute token position t, the progress helix displays phase elasticity. Our analysis of the instantaneous phase velocity reveals that the helix does not rotate at a constant rate. Instead, it stretches and contracts to accommodate the current reasoning step. Crucially, the progress helix also controls reasoning length rather than simply tracking it. Through editing the helix, we can manipulate both the reasoning length and accuracy.

We restrict the scope of this study to mathematical and recursive arithmetic tasks, as these provide easy to measure metrics for reasoning depth and progress. Through these tasks, we provide a topological and spectral account of how modern LLM architectures manage the temporal dimension of thought.

2 Related Works

Understanding how Large Language Models (LLMs) coordinate complex logical trajectories requires reviewing representation geometry, mechanistic circuit analysis and sequence-level tracking mechanisms. This literature review situates the discovery of the Progress Helix within these domains.

Mechanistic Interpretability. To analyze how models track abstract logical properties, we build on the foundational work in mechanistic interpretability. Early studies established the linear representation framework, showing that high-dimensional model activations compress semantic concepts into specific linear subspaces (Elhage et al., 2021). Researchers have relied on training linear probes to extract semantic concepts, state attributes and conceptual associations across model layers (Burns et al., 2022; Meng et al., 2022).

Helical Manifolds. Most relevant to this study are recent breakthroughs showing that transformers shape their latent space into highly structured, low-dimensional manifolds to perform precise mathematical and spatial operations. Specifically, Gurnee et al. (2026) investigated how text models manage character layout constraints, discovering that models map character lengths to low-dimensional curved manifolds. For text tasks such as line breaking, specific attention heads twist these manifolds, aligning estimations orthogonally to form discrete linear boundaries (Gurnee et al., 2026). Shifting to mathematical reasoning, Kantamneni and Tegmark (2025) demonstrated that numbers are internally represented as a helix. They proved that models compute multi-digit addition and subtraction by rotating and manipulating this helical shape through a mechanistic algorithm known as the "Clock" algorithm (Kantamneni and Tegmark, 2025).

Representation Learning. The emergence of structured low-dimensional geometries during a model's training is a well-documented phenomenon. When networks are optimized for clean algorithmic or modular arithmetic tasks, they frequently transition from pure memorization to generalized learning in a process termed "grokking" by shaping their internal feature subspaces into clean circular and block circulant structures (Liu et al., 2022; Mallinar et al., 2024). To map this process during training, interpretability researchers have developed targeted measures to monitor the geometric transition from random weights to clean periodic forms (Nanda et al., 2023). This literature indicates that transformers rely on circular or helical paths over linear ones when computing sequential cycles.

Reasoning Length Control. A recent line of work has investigated how to control LLM rea-

soning chain length. Sun et al. (2025) report that reasoning length is governed by an approximately linear direction in the representation space and that editing roughly four percent of attention heads shifts the model away from overly-short reasoning. Zhao et al. (2025) show that a small set of high-impact activations in the last few layers suffice to elicit long-form reasoning behaviour. Hassid et al. (2025) report empirically that, for a fixed problem, shorter reasoning chains are more often correct than longer ones in modern reasoning LLMs. Our results are consistent with this picture but suggest that the underlying control signal is itself not linear: the relevant direction in activation space is the phase angle of a 2D periodic component, which a 1D linear probe will project onto a sinusoid of the underlying progress and capture only piecewise.

Position Embeddings. Finally, our research addresses the question of how long-range context tracking is handled in the model's architecture. Contemporary open source LLMs notably leverage the Rotary Position Embedding (RoPE) to grant attention keys and queries a localized and relative understanding of token distance (Su et al., 2024; Touvron et al., 2023). However, while RoPE efficiently computed relative distances between individual tokens, it does not naturally provide a global measure of the model's position within a complex reasoning task. When models use Chain-of-Thought (CoT) prompting to elongate reasoning over hundreds of generated tokens (Wei et al., 2022), they require a high-level state tracker to regulate token budget and govern logical termination. The Progress Helix functions as this global tracker.

3 Methods

We investigate the hypothesis that LLMs track mathematical reasoning progress through a Progress Helix, a periodic trajectory within a low-dimensional latent subspace. Our methodology decomposes this task into four distinct stages: task construction, identification of the manifold, phase analysis and causal validation.

3.1 Task Design

To isolate progress from other semantic complexity, we utilize two datasets:

- **Recursive Arithmetic Task (RAT):** We generate a synthetic dataset where each prompt requires exactly n steps of reasoning. A randomly named variable x is initialized to 0, and the model is instructed to perform n operations (e.g., "Add 15, subtract 12..."). By varying $n \in [1, 10]$, we control the expected length of the reasoning chain.
- **Naturalistic Reasoning (GSM8K):** To ensure generalizability, we evaluate the models on the GSM8K benchmark (Cobbe et al., 2021). Unlike the RAT, these prompts vary in semantic content and logical depth, allowing us to test if the periodic structure persists across non-uniform reasoning trajectories.

3.2 Identification of Periodic Subspace

Let $H^l \in R^{T \times d}$ be the activation matrix for a sequence of length T at layer l. To identify the Progress Helix, we perform the following:

Dimensionality Reduction: We apply Principal Component Analysis (PCA) to H^l to extract the top 17 principal components. Let $P_i \in R^T$ represent the i-th principal component.

Spectral Analysis: We compute the Discrete Fourier Transform (DFT) of each P_i to identify dominant frequencies in each output sequence independently:

$$\mathcal{F}(k) = \sum_{t=0}^{T-1} P_i(t) e^{-j \frac{2\pi}{T} kt}$$

We then normalize the magnitude $M(k) = |\mathcal{F}(k)|$ to identify the fundamental frequency. We define the main progress signal as the component where $k{=}1$, representing exactly one rotation over the sequence length L. It is also necessary to note that there exist signals with higher k values, possibly indicating sub-clocks.

3.3 Phase Dynamics

Modern LLMs utilize Rotary Positional Embeddings (RoPE) (Su et al., 2024), which rotate hidden states based on absolute position. To prove the Helix is an emergent mechanism rather than a linear projection of RoPE, we analyze the phase velocity of the progress signals.

Given the top two periodic components P_0 and P_1, we first remove the DC shift to centre the activations at the origin, then we define the emergent phase $\phi_t = \text{atan2}(P_1(t), P_0(t))$. We compare ϕ_t against the RoPE rotation $\theta_t = \omega_\theta t$. We quantify the elasticity of the helix by calculating the instantaneous phase velocity $\omega_\phi(t) = \frac{d\phi}{dt}$.

A constant $\omega_\phi(t)$ would indicate a linear positional artifact. In contrast, we hypothesize

that $\omega_\phi(t)$ fluctuates depending on the reasoning progress.

To complement the qualitative phase analysis, we additionally fit a linear model of ϕ_t on absolute position t and on normalized position t/L, and report $\Delta R^2 = R^2_{\text{progress}} - R^2_{\text{token}}$ at each layer. Statistical significance is assessed with a permutation test that shuffles phase values across positions and recomputes the statistic 200 times. Positive ΔR^2 with a small permutation p-value provides quantitative evidence that the helix tracks reasoning progress rather than absolute token index.

3.4 Causal Intervention

To establish the functional role of the progress helix, we employ two causal interventions.

Manifold Substitution (Circular Helix Injection). We define a synthetic "perfect" circular helix $S_t = [\cos(\frac{2\pi t}{L}), \sin(\frac{2\pi t}{L})]$ with a constant ω_ϕ. We intervene on the model's forward pass by projecting the activations into the PCA subspace and replacing the natural components with S_t. This aims to test the effect of ω_ϕ being constant. In addition, if the helix is purely a tracking clock, the model should remain coherent. If it has causal effects on the reasoning, we expect the output length to change.

Ablative Path Patching. We attribute k=1 magnitude to specific attention heads via output decomposition. For target layer ℓ, let $\mathbf{v}_{k=1} \in R^d$ denote the PC0 direction of the global PCA basis at layer ℓ, which carries the dominant k=1 progress signal as established in Section 3.2. On the final token of the prompt forward pass, for every head h at layer ℓ, we compute its additive contribution to the residual stream as

$$c_h = W_O^{(h)} x_h \in R^d,$$

where $x_h \in R^{d_{\text{head}}}$ is the head's pre-projection output and $W_O^{(h)} \in R^{d \times d_{\text{head}}}$ is its slice of the output projection matrix. We rank heads by the absolute alignment $|\langle c_h, \mathbf{v}_{k=1}\rangle|$ and treat this PC0 alignment as a proxy for k=1 contribution. A stronger attribution would use the projection magnitude into the full PC0/PC1 plane, $\sqrt{\langle c_h, \mathbf{v}_0\rangle^2 + \langle c_h, \mathbf{v}_1\rangle^2}$, which we leave to future work. The top-k aligned heads under the PC0 ranking are taken as contributors to the helix, and the head ranking is fixed for the entire generation. Ablation is performed by zeroing the head's input to W_O during the forward pass.

To rule out generic degradation effects, we compare three ablation conditions: (i) the top-k heads as defined above, (ii) k random heads sampled uniformly from heads that are neither top-aligned nor least-aligned, and (iii) the k least-aligned heads, with smallest $|\langle c_h, \mathbf{v}_{k=1}\rangle|$. The random condition controls for the generic effect of removing k heads at this layer. The least-aligned condition is a stronger negative control for heads that contribute to the residual stream but not to the helix subspace. We use k=3 for Llama-3.2-1B (3 of 32 heads), k=2 for Gemma-2-2B (2 of 8 heads), and k=3 for Qwen2.5-1.5B-Instruct (3 of 12 heads), giving roughly comparable proportional ablation across architectures.

For the speed-scaling intervention described below, we additionally adopt a paired-bootstrap design. Each prompt is generated once per condition with the same input string fixed across conditions, and we compute the per-prompt difference in output length together with a 95 percent paired bootstrap confidence interval (2000 resamples, N=135 prompts). The paired design subtracts out per-prompt difficulty variance and improves sensitivity to small length effects.

3.5 Output Length Modification

In addition, we attempt to empirically verify the effects of helix on the output token length. To achieve this, we hook into the model layer with the highest helix radius. In the case of Llama 3.2 1B, the final layer had the highest activation as indicated by Table1. Then, we hook into the model and ablate the helix. We then append a helix with the same orientation and DC offset but with a different scale factor. Let a and b be the semi-major and semi-minor axes, and θ_t be the angular position, where the nominal progress is $\theta_t = 2\pi \frac{t}{L} + \phi_0$ for a total sequence length L. Given a speed factor γ, we define the nominal angle θ_{base} and the manipulated target angle at step i as θ_{mod}:

$$\theta_{\text{base}} = 2\pi \left(\frac{i}{L}\right) + \phi_0$$

$$\theta_{\text{mod}} = 2\pi \left(\frac{i}{L} \cdot \gamma\right) + \phi_0 + \Delta\Phi$$

We then calculate the position of the helix in the top two principal component, $\mathbf{v}_1$ and $\mathbf{v}_2$, plane for both states.

$$\begin{aligned} P(\theta) = &(a\cos\theta\cos\phi - b\sin\theta\sin\phi)\,\mathbf{v}_1 \\ &+ (a\cos\theta\sin\phi + b\sin\theta\cos\phi)\,\mathbf{v}_2 \end{aligned} \tag{1}$$

The intervention is applied by adding the difference vector $\Delta h = P(\theta_{\text{mod}}) - P(\theta_{\text{base}})$.

$$x' = x_i + \alpha \Delta h$$

This formulation is critical since it preserves the DC offset and higher-order frequencies (k>1) as well as prompt-specific noise in the residual stream. It only shifts the k=1 component to a new position on the clock. If $\gamma > 1$, we expect the model to output less tokens, and if $\gamma < 1$, more tokens should be used given that the progress helix actively controls the output reasoning length.

4 Results

We identified a dominant periodic structure in the latent space of Llama-3.2 (Grattafiori et al., 2024), Gemma-2 (Team et al., 2024) and Qwen (Hui et al., 2024), which we term the **Progress Helix**. Unlike the baseline GPT-2 (Radford et al., 2019), which displayed a very noisy activation space that was dominated by a linear DC component (k=0), modern models with improved reasoning capabilities exhibit a clean, increasing magnitude signal at the fundamental frequency k=1 that represents one full rotation per output sequence. Figure 1 confirms this trend across all three modern model families: the k=1 component grows monotonically with relative layer depth in Llama-3, Gemma-2 and Qwen, while the harmonics at k=2 and k=4 exhibit similar but weaker growth, consistent with the presence of integer-frequency sub-clocks. Crucially, while the mean magnitude of $k = 1$ in Llama seems low compared to Gemma and Qwen, the magnitudes of all Llama frequencies (including DC offset) remain below 1.5, and $k = 1$ consistently has the highest or second highest value.

Across the RAT task described in the methodology, the FFT of the activations reveals that k=1 dominates the PC1 and PC2 signals. Even in GSM8K, where the PCA captures semantic variance in the prompts and reasoning methods, the progress helix is clearly captured in PC0 and PC1. Importantly, k=0 (DC offset) captures the global semantics, and contains the non-progress helix activations.

While each layer represents the progress as a helix, the magnitude of these activations (radius) increases across layers, forming an elliptical cone. This may indicate that the model's internal representation of progress grows stronger and more confident as the activations flow through hidden layers.

A critical finding is that the progress helix is not a mere artifact of the RoPE. We compared the phase angle (ϕ) of the emergent helix against the absolute token position t.

1. **Non-linearity:** While RoPE increases linearly (0 to 500 rad), the Progress Helix exhibits elasticity. The phase velocity $\frac{d\phi}{dt}$ is not constant. Instead, we observe distinct plateaus ($\frac{d\phi}{dt} \approx 0$) at $\phi = \pi$ and a larger plateau at $\phi = 2\pi$.

2. **Length Normalization:** When normalizing by the total sequence length L, the phase ϕ follows a near-constant trajectory across different prompts. This suggests that the model possesses a global progress tracker that "stretches" and "contracts" by changing $\frac{d\phi}{dt}$ to fit the expected depth of the reasoning chain.

The qualitative phase result is corroborated quantitatively by the per-layer ΔR^2 analysis described in Section 3.3. As shown in Figure 8, across the middle layers of all three modern instruction-tuned models the helix tracks normalized reasoning progress significantly better than absolute token index, with permutation p-values below 10^{-4} at the indicated peaks (Llama layer 8: $\Delta R^2 = +0.14$, Gemma layer 21: $\Delta R^2 = +0.22$, Qwen layer 4: $\Delta R^2 = +0.20$). This rules out a RoPE projection as the source of the observed periodicity.

To validate the functional necessity of this manifold, we performed causal patching experiments in the PC1 and PC2:

- **Circular Helix injection:** Substituting the model's natural activations with a perfect synthetic helix ($\frac{d\phi}{dt}$ is constant), we observe immediate logical collapse. The model retains semantic structure but loses track of the reasoning step being executed. Quantitatively in Llama-3.2-1B, numeric accuracy collapses from 0.92 on clean trajectories to 0.38 under k=1 substitution while the heuristic fluency proxy remains essentially unchanged ($0.79 \rightarrow 0.84$), as illustrated in Figure 9. A full-plane ablation control is less damaging to accuracy (0.71) than the synthetic-helix substitution, demonstrating that the effect is not generic noise injection.

- **Attention head patching:** Patching the top-k heads identified as k=1 contributors via the output-decomposition procedure detailed

in Section 3.4 resulted in significantly elongated responses, while ablation of k random heads or the k least-aligned heads at the same layer did not. This dissociation rules out generic degradation effects and implies the top-k heads serve as progress monitors that facilitate the transition through the helix.

- **Bifurcation in final layers:** In the final 2-3 layers, the k=1 signal yields to a dominant k=0 linear drift. We observed a topological bifurcation where phase curves for different prompts split in opposite directions. More information is provided in Section 4.2.

- **Higher order harmonics.** To understand if higher order harmonics play a role in tracking reasoning, we perform patching and scaling experiments. Although the bicoherence between $k = 1$ and higher order frequencies increasing across layers (Figure 11 and 12), changing these high-order frequencies led to model collapse. This suggests that they do not track reasoning but have other highly sensitive roles. In fact, non-linearities (such as SwiGLU) act as frequency mixers, generating higher-order harmonics from a fundamental periodic input (Oppenheim et al., 1997).

4.1 Helix Phase

Table 1: **Emergence of the Progress Helix.** We report the combined magnitude of the $k = 1$ Fourier components in the first two principal components (PC0, PC1) across layers. The signal strength increases monotonically, suggesting a strengthening of the progress-tracking mechanism in deeper layers.

Layer	Mag.	Layer	Mag.	Layer	Mag.
0	0.061	8	0.510	12	0.951
1	0.099	9	0.600	13	1.238
2	0.164	10	0.745	14	1.404
3	0.202	11	0.832	**15**	**1.724**

Causal Intervention Results. Effects of latent speed scaling on CoT generation can be seen in Table 2. The average output length follows a linear trend as expected until the breakdown of the scaling trend at $\gamma = 1.3$ where the model's internal progress activations were pushed much further than what the attention heads and FFN layers were trained on, inadvertently necessitating more reasoning. The accuracy peaks at γ=0.85, indicating that the progress helix can be used to enhance reasoning capabilities.

Table 2: This table presents the effects of the speed factor in Llama's progress helix on the accuracy and the output length in 30 prompts from the RAT dataset.

Speed Factor (γ)	Acc. ↑	Avg. Len ↓
0.7x	0.80	191.7
0.85x (Slow)	**0.87**	183.8
1.0x (Baseline)	0.80	177.1
1.15x (Fast)	0.73	**167.0**
1.3x	0.77	183.0

This length-control behaviour is most pronounced in Llama-3.2-1B. As a supplementary check, we replicated the speed-scaling intervention on Gemma-2-2B and Qwen2.5-1.5B-Instruct using the same paired-bootstrap protocol described in Section 3.4 (Figure 10). On Gemma the predicted directional pattern emerges in the moderate-perturbation range, with γ=0.85 producing a slight lengthening ($\Delta = +1.0$ tokens) and γ=1.15 producing a shortening ($\Delta = -2.4$ tokens, with the bootstrap upper bound at $+0.3$ nearly touching zero). The two endpoints span roughly four tokens of paired difference in the predicted sense, supporting the cross-model generality of the helix-as-controller account at moderate γ. On Qwen the pattern is noisier, with per-condition signs that do not consistently follow the predicted direction at γ=0.7 and γ=1.15 and confidence intervals that include zero throughout. It is highly possible that changing the speed throughout the entire output generation is unoptimal, explaining the weaker correlations in Gemma and Qwen. As seen in Figure 6, there are three regions where the phase velocity is almost zero. It remains unexplored whether modifying only specific regions yields stronger results.

4.2 Helix Magnitude

Helix Magnitude controls phase. When examining the dynamics of the progress helix, we observe a striking coupling between the phase angle and the magnitude (Figure 3), where the magnitude is defined as $M = \sqrt{(k1^2_{pc0} + k1^2_{pc1}}$. The periods where the phase plateaus directly coincide with local maxima in the magnitude curve. Conversely, phase transitions are exclusively triggered when the helix magnitude falls below a critical threshold. Interestingly, in the bifurcated cases, the opposite is true. Only above a certain threshold does the phase

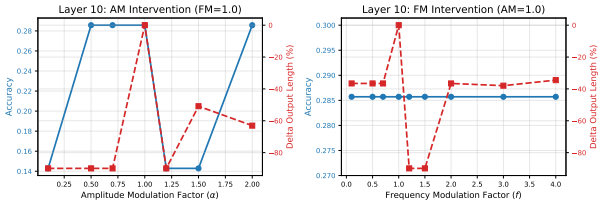

Figure 2: This figure shows the results of AM and FM on Llama's layer 10 on GSM8K. In AM, the output length is significantly shortened, with $\alpha = 2$ reducing 60% of the tokens generated with minor accuracy changes. Likewise, FM significantly reduces token count but does not affect the accuracy.

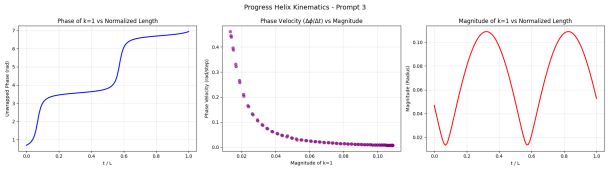

Figure 3: This figure shows how the phase is coupled to the magnitude of the helix. Below 0.04, the phase velocity increases rapidly, suggesting the progress helix relies on amplitude-gating.

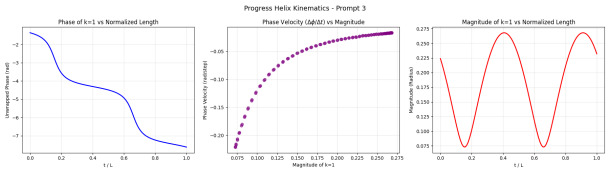

Figure 4: This figure demonstrates that in the bifurcated cases, the phase remains coupled to the magnitude of the helix. However, an upper threshold allows the phase velocity to increase rapidly.

velocity increase rapidly (Figure 4). This finding explains why injecting a circular helix collapses model accuracy, and suggests that the model relies on an amplitude-gated state machine.

To test whether magnitude is the underlying driver of the phase progression, we applied Amplitude Modulation (AM) and Frequency Modulation (FM) to the magnitude across 10 GSM8K questions at layers $l \in \{0, 8, 10, 12\}$ as well as the output layers. We initially hypothesized that modulating the amplitude might predictably stretch or compress the phase plateaus, thereby altering the output length in a directional manner. In Gemma layer 10, we observe that decreasing the amplitude does not affect reasoning while decreasing output length by up to 8%, which is much stronger than the phase scaling results. Increasing the amplitude also decreased tokens. However, we observed that, in all Llama layers, *any* perturbation away from the baseline, whether applying AM scales ($a \neq 1.0$) or FM scales ($f \neq 1.0$), consistently decreases output token length (Figure 2).

We find that the magnitude modulations have a much greater effect on both the token length and the accuracy compared to phase scaling. This suggests that the magnitude controls the phase. Interestingly, magnitude modulation effects significantly vary by layer and model type (more info in Appendix C). While these results confirm that the progress helix's magnitude profile directly affects reasoning, the precise mechanistic function by which magnitude controls phase transitions remains an open question for future work.

5 Conclusion

In this work, we presented a topological and spectral characterization of how language models track sequence position during chain-of-thought (CoT) reasoning. We identified the Progress Helix, a periodic low-dimensional trajectory at the fundamental frequency $k{=}1$ that completes one full rotation over the duration of a generated reasoning chain. Evaluated across diverse model families, including Llama-3.2-1B, Gemma-2-2B and Qwen2.5-1.5B-Instruct, on both synthetic and natural mathematical datasets, this helical structure manifests as an expanding elliptical cone across successive layers. The cone's growth indicates that the model's representation of task progress dynamically accumulates as activations flow through the residual stream.

Our analysis demonstrates that this mechanism is not an artifact of absolute position representations such as Rotary Position Embeddings (RoPE). The Progress Helix exhibits a varying phase velocity $\omega_\phi(t)$ that adapts to varying reasoning depths while maintaining a length-normalized trajectory across different prompts. The quantitative ΔR^2 analysis confirms this distinction with permutation-test significance in the middle layers of all three models studied.

Through three causal interventions, we confirmed the helix's functional role in controlling reasoning depth rather than merely tracking it. Substituting the natural $k{=}1$ component with a synthetic constant-velocity helix induces selective logical collapse, with numeric accuracy dropping sharply while surface fluency is preserved, indicating polysemantic encoding of position with numeric content. Ablating the attention heads most aligned with

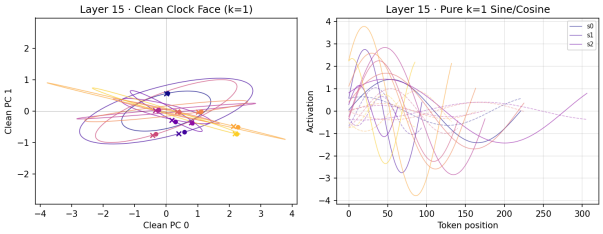

Figure 5: This figure shows a 2D view of layer 15's activation in Llama-3 with different prompts. Since each helix is different, it is necessary to compute PCA per output.

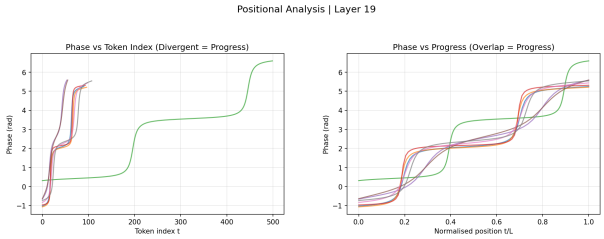

Figure 6: This image demonstrates the similar phase velocities in middle layers in the RAT dataset. We can conclude that the progress helix tracks the reasoning progress.

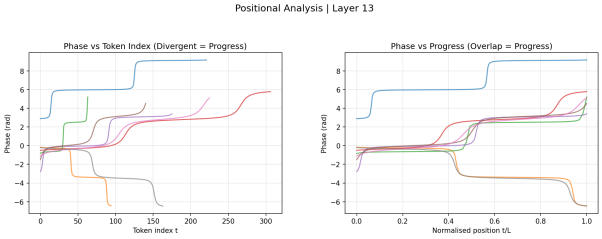

Figure 7: This image demonstrates the bifurcation of the phase in final layers.

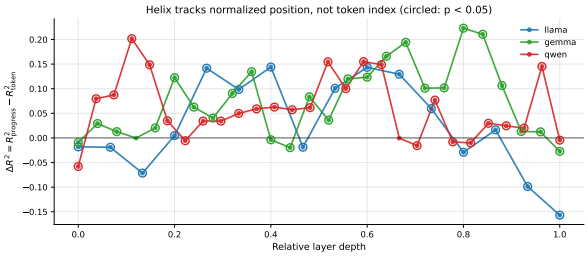

Figure 8: **The helix tracks normalized progress, not absolute token index.** Per-layer $\Delta R^2 = R^2_{\text{progress}} - R^2_{\text{token}}$. Circled markers indicate $p < 0.05$ from a permutation test. The middle layers of all three models exhibit a statistically significant preference for the normalized-position fit over the absolute-token-index fit, providing quantitative confirmation that the helix is distinct from a Rotary Position Embedding projection.

the k=1 direction elongates generations, identifying these heads as progress monitors. Scaling the helix via the parameter γ provides direct control over reasoning length and accuracy. Minor decelerations (γ=0.85) yield a meaningful boost in reasoning accuracy, and minor accelerations (γ=1.15) reduce output length. The effect is strongest in Llama-3.2-1B, with Gemma-2-2B exhibiting the same directional pattern in the moderate-γ range, while the result on Qwen2.5-1.5B-Instruct is less robust at this intervention site.

By shifting the interpretability paradigm from static linear vectors to dynamic, periodic manifolds, these findings open new avenues for predictably manipulating, shortening, and optimizing complex multi-step reasoning trajectories in language models. The Progress Helix supplies a concrete representational target for further mechanistic analysis of how transformer models budget their reasoning effort.

5.1 Limitations

While our findings provide robust empirical and causal evidence for the Progress Helix, several boundaries define the scope of this study and suggest concrete follow-up directions.

Domain restriction. Our evaluation focused on mathematical reasoning tasks (RAT and GSM8K). These domains offer cleanly structured logical depths to track, but it remains to be seen whether the k=1 helical structure generalizes to non-mathematical reasoning tasks such as code generation, multi-hop question answering, legal reasoning, or narrative writing. The shape of the helix may be specific to tasks with a clear linear notion of progress, and may take a different form, or be absent, in tasks with branching or iterative reasoning structures.

Architecture restriction. Our work examined modern decoder-only transformer architectures that natively employ RoPE. Although our ΔR^2 analysis distinguishes the helix from a RoPE projection within these models, we do not currently have a strong, non-RoPE modern model on which to verify that the same structure emerges in the absence of rotary embeddings. Future work is required to determine whether the Progress Helix is an architectural constant or whether it depends on the relative-position inductive bias provided by RoPE.

Cross-model variability of length control. The causal length-control effect via phase scaling and magnitude modulation is observed most strongly

in Llama-3.2 1B and Gemma-2 2B. On Qwen2.5-1.5B-Instruct the per-condition signs are inconsistent with the predicted direction at γ=0.7 and γ=1.15, and the bootstrap intervals do not exclude zero at any condition. The most likely explanation is that the magnitude-peak layer of Qwen is not the most effective single-layer intervention site for this architecture. The middle-layer ΔR^2 peak (Figure 8) sits at substantially different relative depths across the three models studied, and architectural differences such as Qwen's wide grouped-query ratio may modulate how an additive perturbation at one layer propagates to the output distribution. Identifying the most effective intervention site for each model is a primary direction for future work.

Negative-control limitations. Our intended negative control with GPT-2 was confounded by the model's inability to solve the synthetic recursive arithmetic task at sufficient frequency to support an activation-collection pipeline. We therefore cannot rule out the possibility that GPT-2 develops some weaker form of progress representation that our pipeline could not access. A more controlled negative-control study would either evaluate GPT-2 on tasks within its competence range or compare a modern non-RoPE decoder model against the three RoPE-based models studied here.

6 Ethics Statement

This work focuses on the mechanistic interpretability of open-source Large Language Models, such as Llama-3.2 (Grattafiori et al., 2024), Gemma-2 (Team et al., 2024), and Qwen2.5 (Hui et al., 2024) to better understand how they track reasoning and control output length. All experiments were conducted using publicly available models and standard mathematical reasoning benchmarks (GSM8K) alongside a synthetically generated recursive arithmetic dataset. All models and datasets used in this work are publicly available and were utilized in strict accordance with their respective licensing terms. Our methodology does not involve human subjects, crowdsourced labor, or the collection of personally identifiable information (PII).

From a broader impacts perspective, improving the interpretability of reasoning in LLMs can have significant applications. By providing a mechanism to reliably modulate output length, this research contributes to the development of more efficient AI systems, which can ultimately reduce inference-time compute costs. While there is a general dual-use risk associated with manipulating model activations, such as intentionally degrading model performance or inducing hallucinations, our interventions are fundamentally designed to enhance diagnostic understanding and operational reliability rather than to introduce harmful behaviors.

References

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

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. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2024. Not all language model features are one-dimensionally linear. *arXiv preprint arXiv:2405.14860*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. 2026. When models manipulate manifolds: The geometry of a counting task. *arXiv preprint arXiv:2601.04480*.

Michael Hassid and 1 others. 2025. Don't overthink it. preferring shorter thinking chains for improved LLM reasoning. *arXiv preprint arXiv:2505.17813*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Subhash Kantamneni and Max Tegmark. 2025. Language models use trigonometry to do addition. *arXiv preprint arXiv:2502.00873*.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. 2022. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663.

Neil Mallinar, Daniel Beaglehole, Libin Zhu, Adityanarayanan Radhakrishnan, Parthe Pandit, and Mikhail Belkin. 2024. Emergence in non-neural models: grokking modular arithmetic via average gradient outer product. *arXiv preprint arXiv:2407.20199*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Alan V Oppenheim, Alan S Willsky, and Syed Hamid Nawab. 1997. *Signals & systems*. Pearson Educación.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Chung-En Sun, Ge Yan, and Tsui-Wei Weng. 2025. Thinkedit: Interpretable weight editing to mitigate overly short thinking in reasoning models. *arXiv preprint arXiv:2503.22048*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zekai Zhao, Qi Liu, Kun Zhou, Zihan Liu, Yifei Shao, Zhiting Hu, and Biwei Huang. 2025. Activation control for efficiently eliciting long chain-of-thought ability of language models. *arXiv preprint arXiv:2505.17697*.

A Causal Ablations

B Frequency Bicoherence

In signal processing, bicoherence is a normalized statistical metric derived from the bispectrum used to quantify the degree of phase coupling between different frequency components of a signal. Bicoherence analyzes the third-order cumulants to identify whether two frequencies, f_1 and f_2, interact nonlinearly to generate a third harmonic component at their sum or difference frequency ($f_3 = f_1 \pm f_2$). Therefore, it serves as a critical metric to understand the relationship between the progress helix and higher order harmonics. From these results, we observe that the bicoherence increases across layers, suggesting that there is some relationship between them.

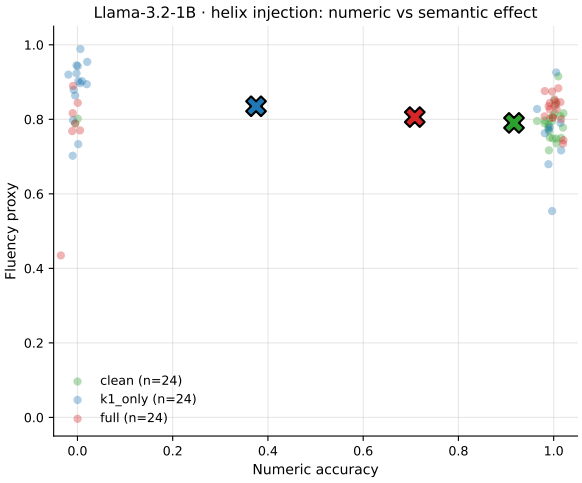

Figure 9: **Helix injection selectively degrades numeric accuracy.** Each marker is one prompt under one of three conditions on Llama-3.2-1B. Crosses indicate per-condition means. Substituting the natural $k{=}1$ component with a synthetic constant-velocity helix (k1_only) collapses numeric accuracy from 0.92 to 0.38 while leaving fluency essentially intact, supporting the polysemantic interpretation in the main text.

C Helix Magnitude Plots

Due to resource limitations, we only run AM and FM experiments on Llama-3.2 1B and Gemma-2 2B.

C.1 Llama-3.2 1B

Overall, Llama-3.2 1B was highly affected by the modulation. All modulations caused a significant decrease in output length with varying effects on the accuracy in 10 GSM8K questions. Notably, layer 10 is a perfect FM target since accuracy does not fluctuate while the token reduction is significant (80% at $f = 1.5$).

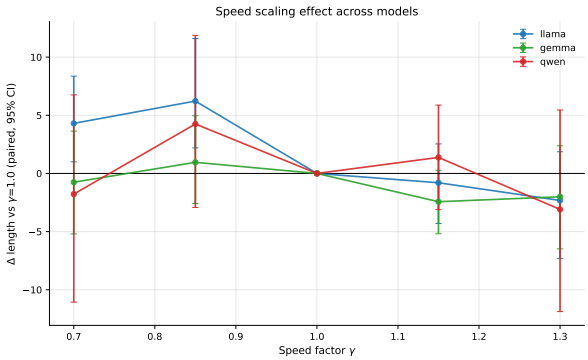

Figure 10: **Cross-model speed-scaling check.** Paired bootstrap 95 percent CIs on the change in output length under speed scaling, applied at the magnitude-peak layer of each model (layer 15 for Llama, 25 for Gemma, 26 for Qwen). The effect is strongest in Llama, where the slow conditions ($\gamma \leq 0.85$) lengthen output significantly. In Gemma the predicted directional pattern (slow lengthens, fast shortens) holds in the moderate-perturbation range $0.85 \leq \gamma \leq 1.15$, with the upper bound on the fast-side CI at γ=1.15 nearly touching zero. Qwen shows inconsistent signs across conditions, suggesting that the magnitude-peak layer is a less effective intervention site in this architecture.

C.2 Gemma-2 2B

Unlike Llama, Gemma-2 2B displays token increase in many layers. Across layers FM has a higher impact on the sequence length than AM. Interestingly, in the output layer, neither AM or FM cause a change in accuracy or output length.

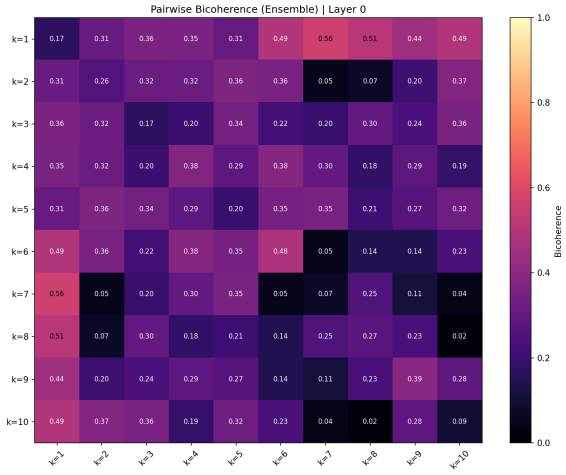

Figure 11: This figure shows the bicoherence heatmap in layer 0 of Llama. Most signals have low bicoherences.

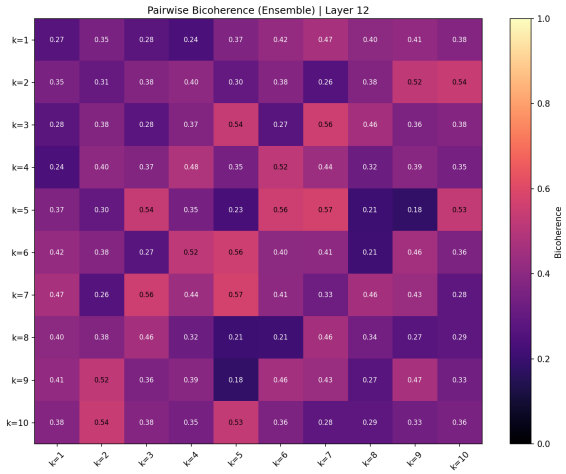

Figure 12: This figure shows the bicoherence heatmap in layer 12 of Llama. The magnitudes of the bicoherence significantly increase compared to earlier layers.

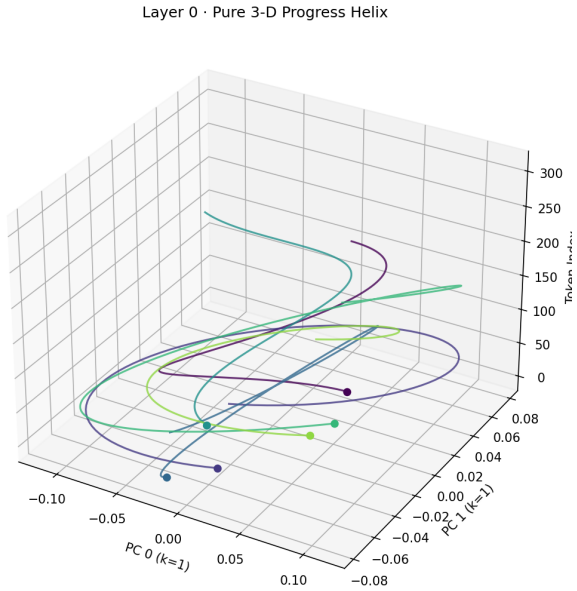

Figure 13: This figure shows one revolution of k=1 frequency of different prompts. Principal components are calculated for each output sequence.

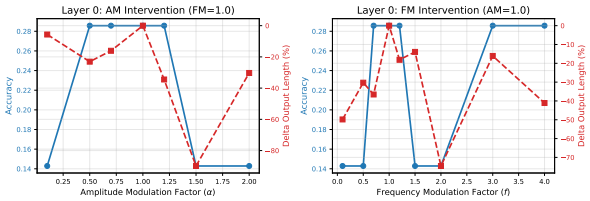

Figure 14: This figure shows the AM and FM interventions on layer 0 of Llama.

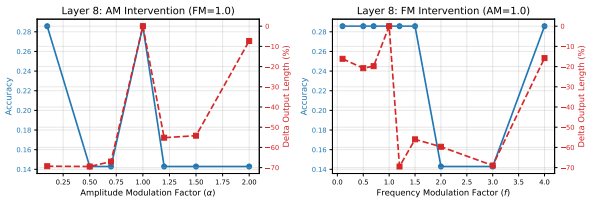

Figure 15: This figure shows the AM and FM interventions on layer 8 of Llama.

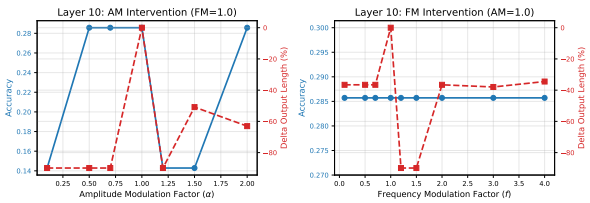

Figure 16: This figure shows the AM and FM interventions on layer 10 of Llama.

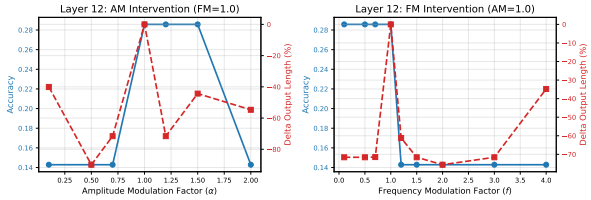

Figure 17: This figure shows the AM and FM interventions on layer 12 of Llama.

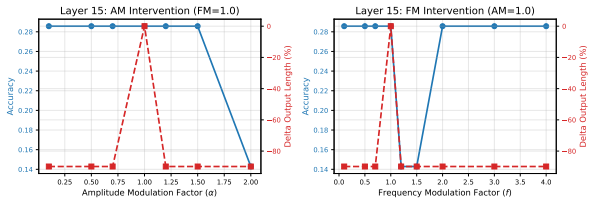

Figure 18: This figure shows the AM and FM interventions on layer 15 of Llama.

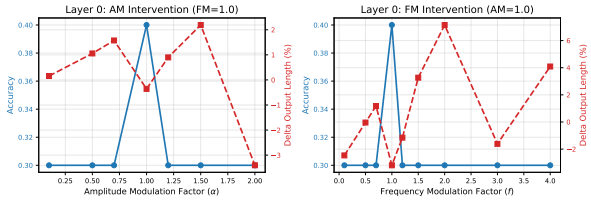

Figure 19: This figure shows the AM and FM interventions on layer 0 of Gemma.

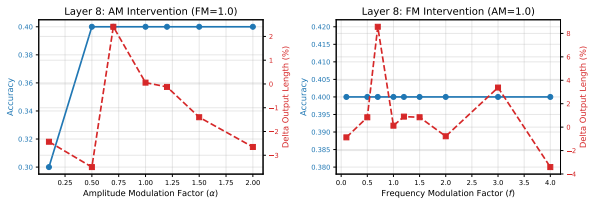

Figure 20: This figure shows the AM and FM interventions on layer 8 of Gemma.

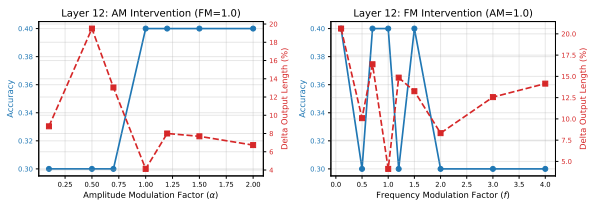

Figure 21: This figure shows the AM and FM interventions on layer 12 of Gemma.

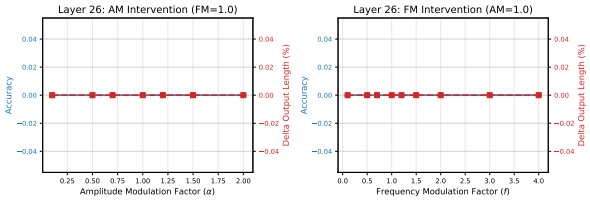

Figure 22: This figure shows the AM and FM interventions on layer 26 of Gemma. In this output layer, Am and FM have no effect