

Control Reinforcement Learning: Token-Level Mechanistic Analysis via Learned SAE Feature Steering

Seonglae Cho^{1,2} Zekun Wu^{1,2} Adriano Koshiyama^{1,2}



<https://huggingface.co/spaces/seonglae/ControlRL>

<https://seongland.com/article/crl>

Abstract

Sparse autoencoders (SAEs) decompose language model activations into interpretable features, but existing methods reveal only which features activate, not which change model outputs when amplified. We introduce Control Reinforcement Learning (CRL), which trains a policy to select SAE features for steering at each token, producing interpretable intervention logs: the learned policy identifies features that change model outputs when amplified. Adaptive Feature Masking encourages diverse feature discovery while preserving single-feature interpretability. The framework yields new analysis capabilities: branch point tracking locates tokens where feature choice determines output correctness; critic trajectory analysis separates policy limitations from value estimation errors; layer-wise comparison reveals syntactic features in early layers and semantic features in later layers. On Gemma 2 2B across MMLU, BBQ, GSM8K, HarmBench, and XSTest, CRL achieves improvements while providing per-token intervention logs. These results establish learned feature steering as a mechanistic interpretability tool that complements static feature analysis with dynamic intervention probes.

1. Introduction

Sparse autoencoders (SAEs) extract interpretable features from language model activations (Bricken et al., 2023; Huben et al., 2024), enabling researchers to decompose dense representations into human-readable components. Yet current analysis methods are static: they identify which features activate on given inputs but do not reveal which features *change model outputs* when amplified.

¹Holistic AI ²University College London. Correspondence to: Seonglae Cho <seonglae.cho@holisticai.com>.

This paper introduces Control Reinforcement Learning (CRL), a framework that turns learned steering into interpretable intervention analysis. A policy network observes residual stream activations and selects an SAE feature to amplify at each generation step. Training this policy with task-specific rewards produces intervention logs: the converged policy identifies features that, when amplified, change model outputs at each token position.

CRL provides four analysis capabilities with concrete empirical support:

- **Token-level intervention logging.** Each intervention logs which feature was selected and its effect on output, enabling fine-grained behavioral analysis.
- **Branch point tracking.** Comparing cases where identical prefixes yield different outcomes based on feature selection identifies critical decision points (Section 5.4).
- **Critic trajectory analysis.** The learned value function exposes whether task limitations stem from policy feature selection or critic reward estimation (Section 5.1).
- **Layer-wise feature semantics.** Training policies at different layers reveals that early layers encode syntactic structure while later layers capture abstract semantics (Section 5.4).

Adaptive Feature Masking (AFM) prevents the policy from collapsing to a narrow feature set by dynamically restricting choices based on recent selections, expanding available features as generation progresses while maintaining interpretability through single-feature interventions.

We evaluate CRL on Gemma 2 2B (Team et al., 2024a) with Gemma Scope SAEs (Lieberum et al., 2024) across MMLU, BBQ, GSM8K, HarmBench, and XSTest. The framework achieves improvements across all benchmarks while producing interpretable per-token attribution logs. Branch analysis demonstrates that layer 10 features capture syntactic patterns (e.g., "mathematical notation") while layer 20 features encode task semantics (e.g., "logical derivation structure"). Critic trajectory analysis identifies distinct bottleneck patterns: BBQ exhibits effective policy-critic coordination with

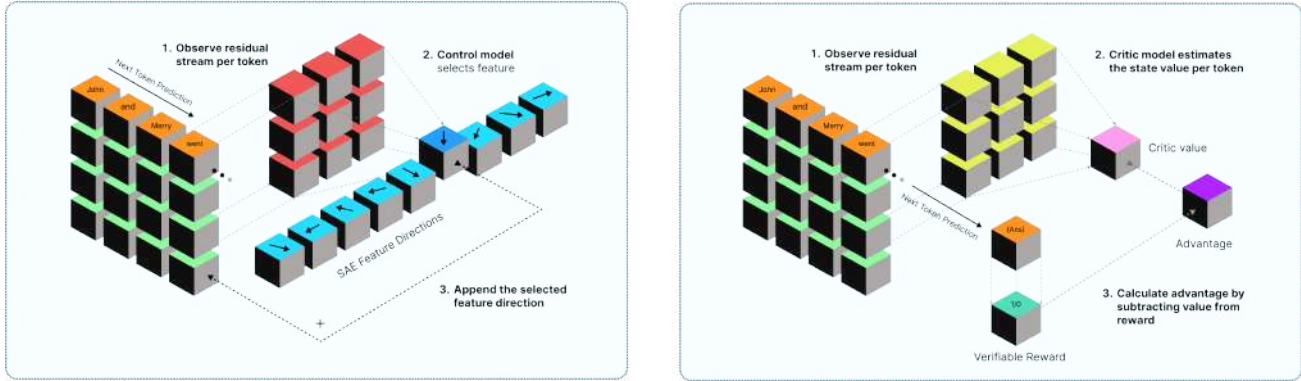


Figure 1. CRL overview. The policy observes residual stream activations and selects an SAE feature to amplify at each token. Each intervention produces an interpretable log: which feature was selected and how it affected output at that position. The critic estimates state values for PPO optimization.

clear sample discrimination, while MMLU shows critic estimation limitations.

Contributions:

1. A framework producing **token-level interpretable intervention logs** via learned SAE feature steering.
2. **Branch tracking and critic trajectory analysis** as diagnostic tools revealing where and why steering succeeds.
3. **Adaptive Feature Masking** for diverse feature discovery under single-feature interpretability.
4. Empirical demonstration of **layer-specific feature semantics** and **task-dependent bottleneck patterns**.

2. Related Work

Sparse Autoencoders address the superposition hypothesis (Elhage et al., 2022) by learning to decompose neural activations into interpretable, sparse features (Huben et al., 2024; Bricken et al., 2023). Given an activation vector $\mathbf{x} \in \mathbb{R}^d$, an SAE learns an encoder $f_{\text{enc}} : \mathbb{R}^d \rightarrow \mathbb{R}^{d_{\text{dict}}}$ and decoder $f_{\text{dec}} : \mathbb{R}^{d_{\text{dict}}} \rightarrow \mathbb{R}^d$ where $d_{\text{dict}} \gg d$, such that:

$$\mathbf{z} = f_{\text{enc}}(\mathbf{x}) = \text{Activation}(\mathbf{W}_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}}) \quad (1)$$

$$\hat{\mathbf{x}} = f_{\text{dec}}(\mathbf{z}) = \mathbf{W}_{\text{dec}}\mathbf{z} + \mathbf{b}_{\text{dec}} \quad (2)$$

The training objective uses reconstruction loss and sparsity regularization: $\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 + \lambda\|\mathbf{z}\|_1$.

Activation Engineering techniques work by making targeted perturbations to a model’s activations (Rawte et al., 2023; Turner et al., 2023; Hernandez et al., 2024; Rinsky et al., 2024). In Activation Engineering, steering vectors (Subramani et al., 2022; Konen et al., 2024) enable control over the language model’s behavior, offering more direct behavioral control compared to prompt engineering. In this context, SAE features provide interpretability and can serve as steering vectors in clamp or addition operations (Chalnev et al., 2024). Selecting a suitable coefficient keeps the language model within its optimal "sweet spot" without dis-

ruption (Durmus et al., 2024). Typically, quantile-based adjustments or handcrafted coefficients are common methods for regulating a feature’s coefficient (Choi et al., 2024).

SAE-based Control methods demonstrate that SAE features can serve as steering mechanisms for language model control. However, current approaches face significant limitations: they require either contrastive datasets for feature selection or extensive activation storage for coefficient optimization. More critically, existing methods lack adaptive feedback mechanisms that can adjust steering strategies based on generation quality, limiting their effectiveness across diverse tasks.

Reinforcement Learning for LLM has been explored in various contexts (Yu et al., 2017; DeepSeek-AI, 2025), primarily focusing on high-level policy optimization. Recent work has applied RL to interpretable model steering (Ferrao, 2024), but with limited action spaces. LLMs operate with dual sequences: token positions (Text Stream) and layer information flow (Residual Stream), both conceptualizable as Markov processes. This suggests potential for RL-based SAE control that automatically identifies optimal feature manipulations to maximize task-specific rewards.

3. Method

3.1. Problem Formulation: CRL as an MDP over SAE Features

Figure 1 illustrates the CRL framework. We approximate the control of transformer representations as a Markov Decision Process (MDP) in which SAE features are manipulated to optimize task-specific rewards. The underlying problem is naturally a Partially Observable MDP (POMDP) (Zhong & Zhang, 2023; Lim et al., 2023); we adopt the MDP formulation, treating the current residual stream activation as the state (Yu et al., 2017; Gui et al., 2019; Xu & Jin, 2024).

Let $\mathbf{x} \in \mathbb{R}^d$ denote the residual stream activations at

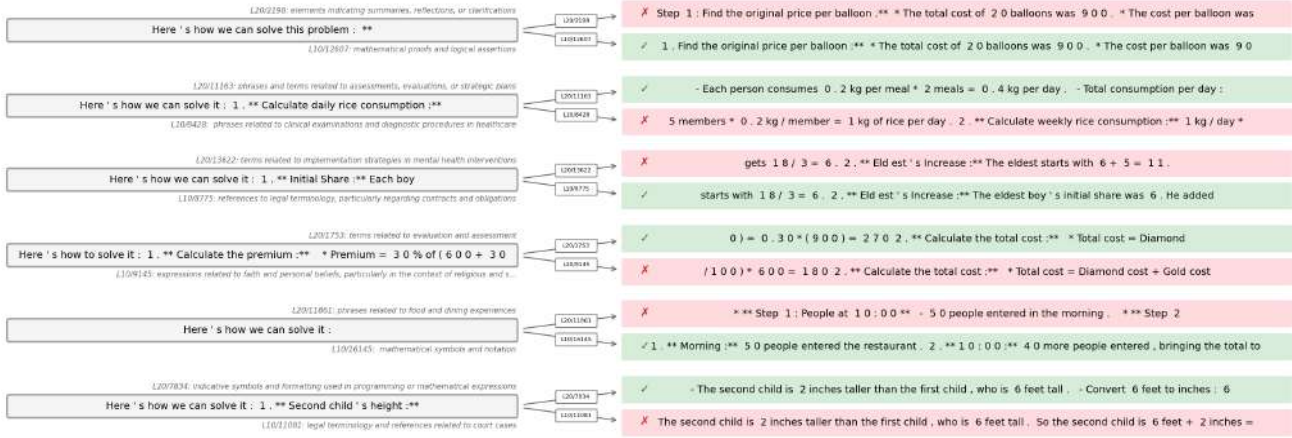


Figure 2. Branch point analysis comparing layer 10 vs layer 20 feature semantics. Each panel shows context and competing features from both layers, with correct features highlighted.

layer ℓ for a target token position, where d is the hidden dimension of the transformer model. Given a pre-trained SAE with encoder $\mathbf{W}_{enc} \in \mathbb{R}^{d \times d_{dict}}$ and decoder $\mathbf{W}_{dec} \in \mathbb{R}^{d_{dict} \times d}$, the sparse feature activations are computed as $\mathbf{z}_t = \text{Act}(\mathbf{x}_t \mathbf{W}_{enc} + \mathbf{b}_{enc})$, where $\text{Act}(\cdot)$ denotes a generic pointwise activation used in the SAE encoder and $\mathbf{z} \in \mathbb{R}^{d_{dict}}$ represents the sparse feature activations with dictionary size d_{dict} .

The MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ where:

- **State Space \mathcal{S} :** The observation is $\mathbf{s} = \mathbf{x} \in \mathbb{R}^d$, the residual stream activation at layer ℓ for the current token position.
- **Action Space \mathcal{A} :** Actions are binary vectors $\mathbf{a} \in \{0, 1\}^{d_{dict}}$ where selected features (via argmax or top-k) are set to 1, reducing the exploration challenge in high-dimensional feature spaces.
- **Transition Function \mathcal{P} :** Deterministic transition governed by the transformer’s forward pass with steering applied.
- **Reward Function \mathcal{R} :** Task-specific rewards r based on final output quality evaluation.

Although the full problem is a POMDP, we empirically find that treating the current residual stream as a sufficient statistic is effective, making the MDP formulation a practical approximation.

3.2. CRL-Token: Token-Level Interpretable Steering via SAE Features

At each generation step t , CRL-Token performs token-level interventions by observing the current token’s residual stream activation at layer ℓ and applying perturbations:

$$\tilde{\mathbf{x}}_t = \mathbf{x}_t + c \cdot \mathbf{a}_t \mathbf{W}_{dec}, \quad (3)$$

where $\mathbf{a}_t \in \{0, 1\}^{d_{dict}}$ is a sparse binary selection vector, c is the steering coefficient, and $\tilde{\mathbf{x}}_t$ is the steered activation. By the Markov property, the policy depends only on the current residual stream activation:

$$\pi_\theta(\mathbf{a}_t \mid \mathbf{x}_{1:t}) = \pi_\theta(\mathbf{a}_t \mid \mathbf{x}_t),$$

where t denotes the token generation step, starting from the first generated token after the input context.

3.2.1. POLICY NETWORK FOR FEATURE SELECTION

The policy network $\pi_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^{d_{dict}}$ maps residual stream observations to SAE feature selection logits. We implement this as a 2-layer MLP with Tanh activation (architecture details in Appendix A.1, training procedure in Algorithm 1):

$$\boldsymbol{\mu} = \pi_\theta(\mathbf{s}) \quad (4)$$

$$\mathbf{a} = \text{TopK}(\boldsymbol{\mu}, k) \quad (5)$$

The policy first computes feature selection logits, then selects the top- k features. In this study, we set $k = 1$ to focus on the most relevant feature per token, effectively equivalent to ArgMax selection. This design prioritizes interpretability, enabling clear attribution of steering effects to individual features. We employ a softmax parameterization for the policy, converting logits $\boldsymbol{\mu} = \pi_\theta(\mathbf{s})$ into probabilities $p_j = \frac{\exp(\mu_j)}{\sum_{i=1}^{d_{dict}} \exp(\mu_i)}$. During training, actions are **sampled** from this categorical distribution to ensure unbiased policy gradient estimation, while evaluation uses deterministic argmax selection for reproducibility. This parameterization ensures differentiability and enables PPO (Schulman et al., 2017) to optimize the expected advantage using the log-probability of the selected feature: $\log \pi_\theta(\mathbf{a} \mid \mathbf{s}) = \log p_j$. By decoupling probability learning from coefficient determination, the model can learn which features to select without prematurely collapsing to a single feature.

Partial Observability: CRL-Token is formulated as a POMDP because the policy observes only the residual stream at a specific layer without access to the sampled token information, yet the next state depends on the token that will be sampled and added to the KV cache. This creates partial observability where undetermined token states affect future transitions. Empirical injectivity of layer representations (Nikolaou et al., 2025) suggests that hidden states at arbitrary layers contain sufficient information about input tokens; we use temperature=0 sampling to eliminate stochastic token sampling effects during training and evaluation.

3.3. Adaptive Feature Masking

CRL-Token introduces Adaptive Feature Masking (AFM), a mechanism designed to balance exploration and exploitation during multi-token generation. AFM maintains a per-sample boolean mask over the SAE dictionary, where $\text{mask}_i = 1$ indicates feature i is available for selection. At the start of each sample, the mask is initialized to a small subset (e.g., frequently active features), and progressively expanded during generation:

$$\text{mask}_i^{(t+1)} = \text{mask}_i^{(t)} \vee \text{active}_i^{(t)}, \quad \mathbf{a}_t = \text{TopK}(\boldsymbol{\mu} \odot \text{mask}^{(t)}, k) \quad (6)$$

where $\text{active}_i^{(t)} = [\mathbf{z}_{t,i} > 0]$ indicates whether feature i was activated at step t , and the mask constrains policy selection by masking unavailable features. Policy selection (\mathbf{a}_t) determines which feature to amplify, while natural activations expand the mask for future steps. This per-sample masking prevents policy collapse to early-selected features while building on activated features within each trajectory.

3.4. Critic Network for Value Estimation

The critic $V_\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ estimates the state value function:

$$V_\phi(\mathbf{s}) = \mathbb{E}_{\pi_\theta}[r \mid \mathbf{s}], \quad (7)$$

and is implemented as a multilayer perceptron (MLP).

3.5. Optimization with PPO and Task-Specific Rewards

The steering coefficient c is determined adaptively by averaging activation magnitudes from correctly predicted training samples, eliminating the need for manual tuning across different layers and tasks.

We train both the policy and critic using Proximal Policy Optimization (PPO). For feature selection via $\text{argmax}/\text{top-}k$, the PPO objectives are:

$$L_{\text{policy}}(\theta) = \mathbb{E}[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t)], \quad (8)$$

$$L_{\text{critic}}(\phi) = \mathbb{E}[(V_\phi(\mathbf{s}) - r)^2]. \quad (9)$$

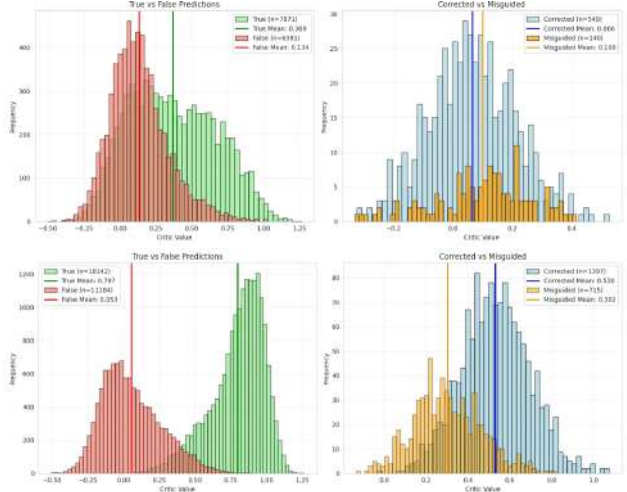


Figure 3. Critic value distributions for single-token tasks. Top: MMLU, Bottom: BBQ. Colors: green = unchanged-correct, red = unchanged-incorrect, blue = corrected, yellow = misguided (see Appendix A.1 for definitions).

Here, $\pi_\theta(\mathbf{a}|\mathbf{s})$ is computed from the softmax probabilities described above, $r_t(\theta) = \frac{\pi_\theta(\mathbf{a}|\mathbf{s})}{\pi_{\theta_{\text{old}}}(\mathbf{a}|\mathbf{s})}$ is the probability ratio for the selected feature, and $A_t = r - V_\phi(\mathbf{s})$ is the advantage estimate. PPO’s explicit value function enables the critic trajectory analysis in Section 5.1, which methods like DPO or GRPO cannot provide.

The reward r is task-specific. For multiple-choice tasks, we use binary rewards following DeepSeekMath (Shao et al., 2024): $r(\hat{y}, y^*) = 1$ if $\hat{y} = y^*$, and 0 otherwise. For reasoning and safety tasks, we use task-specific correctness and refusal metrics, with details provided in Appendix A.

3.6. Experimental Setup

We evaluate CRL using Gemma 2 2B-IT (Team et al., 2024a) with Gemma Scope SAEs (Lieberum et al., 2024) across benchmarks including knowledge (MMLU (Hendrycks et al., 2021)), reasoning (GSM8K (Cobbe et al., 2021)), bias (BBQ (Parrish et al., 2022)), and safety (HarmBench (Mazeika et al., 2024), XSTest (Röttger et al., 2024)) tasks. The framework is trained using PPO with task-specific rewards to optimize feature selection. The steering coefficient c in Equation 3 is computed by averaging $\|\mathbf{a}_t \mathbf{W}_{\text{dec}}\|$ over correctly predicted training samples, adapting to layer-specific activation scales. Complete experimental details are provided in Appendix A.

4. Results

4.1. Steering Effectiveness Across Tasks

CRL enables interpretable feature-based steering while achieving improvements across benchmarks. Beyond per-

formance metrics, the framework reveals which SAE features change outputs at the token level. We report full numbers and layer/coefficient sweeps to make effects auditable and support mechanistic analysis. CRL-Token adaptively selects SAE features for steering based on each token’s residual stream activation, enabling dynamic token-level control during generation. This approach provides inherent interpretability by revealing each token’s dedicated features, which we analyze in detail in [Section 5.3](#). [Table 1](#) reports the main results for the Gemma 2 2B model using the standard **single-layer** CRL-Token configuration, where one intervention layer is selected for steering.

Multi-layer variant. For completeness, we experimented with applying CRL-Token across multiple layers simultaneously, achieving 87.14% on HarmBench and 89.84% on XSTest. We also tested **CRL-Layer**, a variant using a single shared policy across layers, which achieves 71.43% on HarmBench and 86.98% on XSTest ([Appendix A.5](#)). However, multi-layer manipulation trades single-feature attribution for stronger aggregate refusal behavior; we keep CRL-Token as the interpretable mainline.

4.2. Baseline Comparisons and Intervention Validation

We validate CRL’s effectiveness through constrained decoding comparison, complementarity analysis with fine-tuning, and comparison against alternative steering methods.

Constrained Decoding Comparison. To isolate CRL’s effect from output format correction, we compare against constrained decoding on MMLU. Constrained decoding restricts outputs to valid answer tokens (A/B/C/D), eliminating format errors. Base accuracy improves from 51.5% to 55.19% with constrained decoding alone. CRL achieves 55.55%, demonstrating +0.36% improvement beyond format correction. This confirms that learned features contribute to performance gains beyond output regularization.

Complementarity with Fine-tuning. CRL provides additive gains when applied on top of supervised fine-tuned models. On MMLU, accuracy improves from 55.2% (base) to 55.7% (SFT) to 56.1% (SFT + CRL), demonstrating that learned feature steering captures complementary signals beyond what fine-tuning achieves. This suggests CRL can serve as a lightweight post-hoc enhancement without requiring additional weight updates. The Gemma Scope SAEs transfer effectively across model variants, enabling CRL application without retraining the autoencoder.

Comparison with Alternative Steering. We ablate CRL’s components by comparing against heuristic baselines and component-removed variants ([Table 2](#)). CRL without AFM achieves 62.55% on BBQ Ambig (+2.39 over base) and 48.67% on HarmBench (+7.21 over base), confirming that the learned policy provides most of the gain. Adding AFM

contributes further (+3.31 BBQ, +0.45 HarmBench) by encouraging diverse feature exploration. Random+AFM improves over random alone (+1.80 BBQ, +1.61 HarmBench), isolating AFM’s independent contribution. These results confirm that learned feature selection is the primary driver; AFM and constrained decoding provide additional but smaller contributions.

4.3. Sensitivity to Steering Layer and Coefficients

To isolate the effects of coefficient changes, we conducted systematic coefficient analysis across different steering magnitudes. Based on experimental results shown in [Appendix 12](#) for MMLU tasks (extended analysis in [Appendix 13](#)), we found that optimal coefficients vary with network depth, summarized in two key findings:

- **Layer-specific effectiveness:** Certain layers demonstrate higher utility, with optimal intervention points varying across task conditions.
- **Coefficient scaling:** Large coefficients perform poorly in earlier layers, suggesting that aggressive steering during early processing stages disrupts fundamental model capabilities.

Across coefficients ranging from 10 to 100 for CRL-Token steering ([Appendix 12](#)), later layers outperform earlier layers. Large coefficients in early layers yield poor performance, while this pattern becomes mixed in later layers. This aligns with prior observations that residual stream norms increase across layers (Heimersheim & Turner, 2023), meaning identical coefficient values produce different effects at each layer. The observed performance inversions indicate that each layer differs in the optimal coefficient range called "sweet spot" (Durmus et al., 2024), necessitating the dynamic approach we employ through averaging observed coefficients. Beyond performance differences, the varying optimal layers indicate that different tasks require distinct feature activations at different network depths, suggesting task-specific intervention strategies (residual stream norm analysis in [Appendix 11](#)).

Further analysis of MMLU layer patterns ([Figure 13](#)) reveals that unconstrained and constrained decoding exhibit different optimal layers, indicating that hallucination mitigation and factual answering require distinct feature activations. BBQ benchmark results exhibit similar tendencies ([Figure 14](#)), with distinct optimal layers for ambiguous vs. disambiguated contexts ([Figure 15](#)), suggesting that bias mitigation strategies must be adapted to the specific type of ambiguity present in the input. Extended layer-wise coefficient analysis across all tasks is provided in [Appendix A](#).

Dynamic coefficient variants. We tested three per-token dynamic alternatives to the static dataset-averaged coefficient: norm-scaled ($c \propto \|\mathbf{x}_t\|$), activation-scaled (c set to

Table 1. Performance results for Gemma 2 2B model across different tasks using **single-layer** CRL-Token. The table shows task type, intervention layer, baseline accuracy (Before), CRL accuracy (After), and gain in percentage points. Layer column shows where task-relevant features emerge for diagnostic analysis. Results reported as mean \pm std across three seeds.

Task	Type	Layer	Before	After	Gain
MMLU	Multi-choice QA	24	52.06 \pm 0.23	55.37 \pm 0.16	+3.31
MMLU-Pro	Multi-choice QA	25	30.30 \pm 0.00	30.49 \pm 0.07	+0.19
BBQ Ambig	Bias QA (ambiguous)	5	60.17 \pm 0.01	65.86 \pm 3.03	+5.69
BBQ Disambig	Bias QA (disambiguated)	5	84.38 \pm 0.52	84.85 \pm 0.09	+0.47
SimpleQA	Short-form QA	8	3.78 \pm 0.17	4.00 \pm 0.29	+0.22
GSM8K	Math reasoning	20	54.62 \pm 0.16	55.65 \pm 0.33	+1.03
HarmBench	Adversarial safety	21	41.46 \pm 9.05	49.12 \pm 1.59	+7.66
XSTest	Over-refusal	12	86.35 \pm 0.00	87.62 \pm 0.82	+1.27

Table 2. Ablation of feature selection methods. Random+AFM isolates masking; CRL w/o AFM isolates learned selection.

Method	BBQ Ambig	HarmBench
No steering (Base)	60.17	41.46
Random feature	58.36	45.35
Random + AFM	60.16	46.96
Most-active feature	59.94	48.03
CRL w/o AFM	62.55	48.67
CRL + AFM (Ours)	65.86	49.12

the natural SAE activation magnitude of the selected feature), and critic-gated ($c \propto V_\phi(\mathbf{s}_t)$). All three degraded performance: on HarmBench, norm-scaled achieved 45.00%, activation-scaled 41.79%, and critic-gated 43.93%, compared to 49.75% for the static coefficient. GSM8K showed similar drops (41.39%, 40.33%, 41.47% vs. 55.65%). The policy learns a joint feature-coefficient coupling; changing the coefficient at inference disrupts this coupling. The static coefficient itself is robust: varying c from 10 to 100 on MMLU changes accuracy by only \sim 3pp (Table 8).

4.4. Critic Network Analysis

The critic network reveals CRL’s learning dynamics across different task types. We analyze critic behavior patterns for both single-token and multi-token generation tasks.

Single-token Tasks. For single-token tasks (MMLU, BBQ), critic values show clear distributions between correct/incorrect samples. MMLU exhibits critic bottlenecks where corrected and misguided samples (defined in Appendix) remain nearly indistinguishable, while BBQ demonstrates effective policy-critic coordination with clear sample distinction. Detailed single-token critic analysis is provided in Section 5.1.

Multi-token Tasks. For multi-token tasks like GSM8K, the critic network’s value function exhibits climbing patterns along token sequences, indicating effective learning of policy feature selection through AFM (Section 3.3). As

illustrated in Figure 4, critic value patterns fall into four groups. Correct cases display higher gradients with divergence beginning around 200 tokens, while steering-induced cases show different trends with corrected answers exhibiting value increases at the 400-token point. This indicates that CRL-Token’s interventions become more effective in later generation stages.

Feature-level Interpretability. Examining individual steering interventions reveals semantic coherence in CRL’s feature selection. For instance, Feature 7708, described as *mathematical operations and expressions in various forms*, activates not only on explicit operators but also on quantitative units like “babies” and “packs” in word problems, demonstrating lexical generalization beyond surface forms. This interpretable mapping between features and token semantics distinguishes CRL from black-box steering approaches. Additional corrected and misguided examples are provided in Appendix B.2.

Additional multi-token critic analysis including HarmBench and XSTest patterns is provided in Appendix A.7. CRL generalizes to LLaMA-3.1 8B with LLaMA Scope SAEs (32K features), showing consistent improvements on MMLU (+0.92), MMLU-Pro (+1.03), BBQ Ambig (+1.07), and HarmBench (+5.36), demonstrating architecture-agnostic steering (Appendix A.6). The XSTest decline (-1.90) on LLaMA reflects task-specific sensitivity: reducing over-refusal requires steering against the model’s default safety features, which may share representations with HarmBench’s refusal policy. This distinction between tasks where the desired direction aligns with vs. opposes the model’s default tendencies is an important axis for characterizing CRL’s effectiveness. Feature diversity (Shannon entropy H) and impact score (fraction of selections that change model output; Appendix A.9) reveal task-dependent patterns: GSM8K requires broad feature engagement ($H=6.7$, impact 0.36) across 1024-token reasoning chains, MMLU-Pro uses moderate diversity ($H=5.5$, impact 0.63) across multiple knowledge domains, XSTest shows intermediate

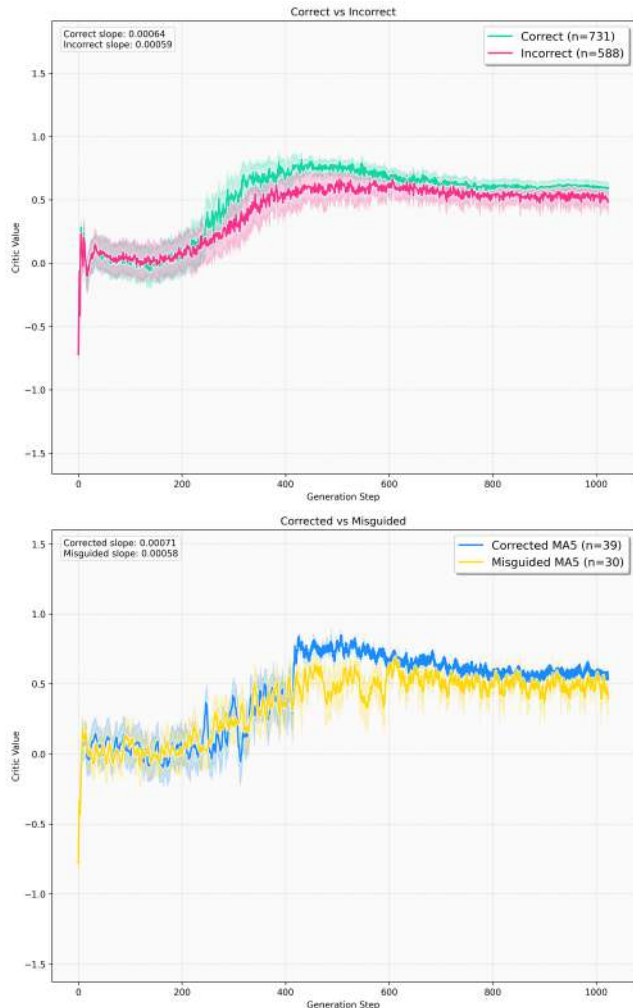


Figure 4. Critic network value trajectories for GSM8K task. Colors indicate: green (unchanged correct), red (unchanged incorrect), blue (corrected), yellow (misguided) (see Appendix A.1).

diversity ($H=4.8$, impact 0.36), while HarmBench concentrates on few high-impact refusal features ($H=2.9$, impact 0.81). Low diversity with high impact (HarmBench) indicates focused tasks where few features genuinely suffice; high diversity with moderate impact (GSM8K) reflects distributed reasoning requiring many features across long sequences. This confirms that the policy adapts its exploration strategy to task complexity rather than collapsing to a fixed feature set (Appendix A.10).

5. Discussion

5.1. Critic Bottlenecks in Single-Token Decisions

Building on Section 4.4, we analyze task-dependent bottlenecks in policy and critic functionality. As shown in Figure 3, the critic value distributions illustrate these patterns, with distinct clustering between correct and incorrect

sample categories. For MMLU, corrected and misguided samples remain nearly indistinguishable, indicating a critic bottleneck. Conversely, for BBQ, corrected and misguided samples become clearly distinguishable, with corrected samples achieving higher values, suggesting effective policy-critic coordination.

5.2. Trajectory-Based Critic Diagnostics for Multi-Token Tasks

The critic value distribution extends analysis to examine critic trends throughout token generation, with critic values predicted from each token’s residual stream. As shown in the GSM8K critic analysis (Figure 4), multi-token tasks exhibit distinctive trajectory patterns throughout generation. For HarmBench, the critic successfully distinguishes between correct and incorrect samples, whereas XSTest exhibits inverted estimation patterns. However, while HarmBench shows clear gaps at sequence boundaries with minimal gradient differences, XSTest demonstrates estimation errors in gap measurements but maintains clear superiority for correct samples in slope analysis.

This discrepancy stems from XSTest’s task characteristics, which require distinguishing benign requests with higher context dependency, making reward estimation challenging. Similar to the MMLU case, this indicates a critic bottleneck, while HarmBench achieves accurate estimation but shows limited critic improvement through policy feature selection, suggesting a policy bottleneck. Analysis of HarmBench and XSTest critic patterns is provided in Appendix A.7.

5.3. Feature Steering as Evidence of Semantic Coherence

Analysis of feature selections reveals semantically meaningful patterns aligned with task requirements. The policy network identifies features corresponding to cognitive processes: reasoning, fact retrieval, and safety considerations.

Efficacious steering demonstrates semantic coherence across corrected examples (Figure 5). In successful cases, features described as *statistical methods* activate on mathematical tokens, strengthening reasoning chains. Features described as *mathematical operations* demonstrate lexical generalization, activating on quantitative units like "babies" and "packs" beyond surface mathematical notation. Failure cases reveal token-feature misalignment: features unrelated to mathematics show poor alignment with mathematical tokens. Feature diversity analysis (Appendix A.10) reveals that complex reasoning tasks (GSM8K, MMLU-Pro) require broader feature engagement, while focused tasks (HarmBench) benefit from concentrated feature activation, suggesting task-specific interpretability patterns. Corrected and misguided examples are provided in Appendix B.2.

Mico's age in 10 years $10 + 10 = 20$ * Marco's age in 10 years $10 + 10 = 20$ **6 Find the sum of their ages in 10 years:** * $20 + 20 = 40$ ##### The sum of their ages in 10 years will be

Here's how to solve it 1 **Calculate the total number of sweets:** - 15 packs * 60 pieces/pack = 900 pieces 2 **Calculate the number of packs Anne kept:** - 15 packs - 2 packs = 13 packs 3 **Calculate the number of pieces given to friends:** - 900 pieces - 120 pieces (for 32 * 3 = 96 classes visited on Wednesday * **Total for Monday-Wednesday:** $32 + 64 + 96 = 192$ classes visited from Monday to Wednesday * **Thursday and Friday:** $30 + 25 = 55$ classes visited on Thursday and Friday * **Total for the week:** $192 + 55 = 247$ classes visited the Science Center

Here's how to solve it 1 **Find the number of babies wearing purple:** * Since there are 8 babies and Carla dresses $3/4$ in purple there are $(3/4) * 8 = 6$ babies wearing purple 2 **Find

Figure 5. GSM8K corrected cases. Top: Feature steering activates relevant numerical tokens, strengthening reasoning chains. Bottom: Semantically coherent feature activations align with equality tokens, guiding correct computation.

5.4. Branch Analysis: Layer-Specific Feature Semantics

Branch point analysis compares cases where the same problem context produces correct or incorrect outputs depending on which SAE features are selected for steering. We trained separate policies on GSM8K at layers 10 and 20 to analyze how steering effectiveness varies across layers (Figure 2).

Observation. Layer 10 features exhibit concrete, syntactic properties while layer 20 features demonstrate abstract, contextual semantics. Balloon pricing activates layer 20’s “mathematical proofs” feature recognizing logical derivation structure, whereas layer 10’s “summaries” feature treats it as tutorial explanation. Rice consumption with “kg/member” notation activates layer 10’s “clinical examinations” feature due to surface similarity with medical dosage, while layer 20’s “assessments and strategic plans” captures the planning semantics. Restaurant problems show layer 10’s “mathematical notation” feature outperforming layer 20’s “food and dining” by prioritizing numerical structure.

Interpretation. Layer 10 performs surface-level pattern matching (notation \rightarrow similar-looking domains), while layer 20 captures task-relevant abstract structure. This is consistent with the residual stream hierarchy where early layers encode syntactic/structural information and later layers encode abstract task semantics.

Implication. Practitioners should select intervention depth based on the target behavior class: early layers for syn-

tactic/format control, later layers for semantic reasoning. BBQ’s optimal layer (5) contains syntactic features while MMLU’s (24) contains domain-knowledge features (Table 1), confirming this principle across tasks. Feature labels provide limited predictive value for steering outcomes: “emotional relationships” solves speed calculations while “mathematical notation” fails on some arithmetic, indicating that learned steering effects diverge from semantic interpretations (Appendix B.5).

5.5. Feature Behavior Analysis

Our experiments reveal insights into SAE feature controllability. Corrected cases demonstrate clear semantic coherence between feature activations and their documented descriptions, as shown in GSM8K examples. Misguided cases provide additional diagnostic value: they reveal where feature-token alignments break down, enabling targeted improvements. These patterns suggest that feature interactions extend beyond simple linear combinations, opening directions for multi-feature coordination strategies.

CRL as SAE Diagnostic. CRL’s per-token logs expose SAE quality issues that static feature analysis cannot detect. Known SAE limitations, including polysemantic features and missing dictionary concepts (“dark matter”) (Engels et al., 2025), propagate to CRL’s intervention logs. As AxBench (Wu et al., 2025) demonstrates that simple baselines can outperform SAE-based steering, SAE quality di-

rectly bounds CRL’s effectiveness. However, CRL makes these limitations observable: if the policy collapses to few semantically incoherent features despite AFM, this signals SAE deficiency rather than a CRL failure. Feature diversity (Shannon entropy H over selection probabilities) provides a task-dependent diagnostic: complex reasoning tasks show high diversity (GSM8K: 6.7) while focused tasks show concentrated, high-impact selection (HarmBench: 2.9). For tasks requiring distributed reasoning, unusually low diversity flags that the SAE may lack adequate feature coverage. A practical failure diagnosis protocol follows: (1) detect unexpectedly low H relative to task complexity; (2) localize by checking whether alternative layers produce higher diversity and accuracy; (3) recover by switching to a better-covered layer; (4) verify that diversity, accuracy, and feature coherence improve.

Bottleneck Diagnostics. CRL’s critic trajectories diagnose where improvements originate. BBQ demonstrates effective policy-critic coordination with clear distinction between sample categories. For MMLU, CRL learns to eliminate invalid responses that the baseline model generates instead of selecting from given options (Figure 8). This task-dependent pattern enables targeted optimization strategies. Detailed critic analysis is provided in Appendix A.7.

5.6. Cross-Task Transfer Analysis

To distinguish task-specific mechanisms from dataset-specific shortcuts, we evaluate policies trained on one task against other tasks (Table 3). All experiments use the training task’s layer for steering.

Table 3. Cross-task transfer results. Each row applies the source task’s policy and optimal layer to the target task.

Train → Eval	Acc. (%)	vs. Base
MMLU → GSM8K	31.01	−10.31
GSM8K → MMLU	52.24	+0.18
HarmBench → XSTest	86.35	+1.27
BBQ → MMLU	53.55	+1.49

The transfer pattern is asymmetric and selective: MMLU features actively harm GSM8K (−10.31%), while GSM8K features remain inert on MMLU (+0.18%). HarmBench features slightly improve XSTest (+1.27%), consistent with shared safety representations. This structured pattern is more consistent with task-specific feature specialization than with dataset-specific shortcuts. The interference aligns with layer-specific computation: MMLU uses layer 24 while GSM8K’s optimal intervention occurs at layer 20.

Across both Gemma 2B and LLaMA 8B, two axes predict CRL’s effectiveness. First, *signal clarity*: CRL shows strongest gains where there is room for improvement (BBQ Ambig: +5.70 Gemma/+1.07 LLaMA, HarmBench:

+7.66/+5.36) and modest gains where baselines are strong (BBQ Disambig: +0.47/+0.09, MMLU-Pro: +0.19/+1.03). Second, *objective alignment*: CRL struggles when the desired direction opposes the model’s default tendency, as in LLaMA’s XSTest decline (−1.90), where reducing over-refusal conflicts with shared safety representations.

6. Limitations

CRL applies one feature at one layer per token to preserve single-feature causal attribution; the multi-layer variant achieves stronger steering (87.14% on HarmBench vs. 49.12% single-layer) but trades attribution clarity, and top- k selection with interaction analysis is future work. CRL inherits SAE coverage and quality limitations: missing dictionary concepts (Engels et al., 2025) cannot be steered on, and polysemantic features may cause proxy selection, though per-token logs make such cases inspectable and policy collapse despite AFM can signal SAE deficiency rather than a CRL failure (Wu et al., 2025).

Each task/layer combination requires separate PPO training (0.5–5.5 GPU-hours on a single A100, 3.3–3.8% additional parameters, 8–9% inference overhead); scaling to 100K+ dictionaries may require efficient architectures. Model evaluation is limited to Gemma-2 2B and LLaMA-3.1 8B by public SAE availability. Safety tasks use classifier-based rewards and SimpleQA uses STS similarity; CRL’s intervention logs make learned behaviors auditable but do not formally prevent reward-metric coupling.

7. Conclusion

Control Reinforcement Learning transforms SAE feature steering into token-level interpretable intervention logs. Training a policy to select features at each generation step produces mechanistic analysis tools: branch tracking locates critical decision points, critic trajectories separate policy from value-estimation limitations, and layer comparison reveals syntactic-to-semantic feature gradients across network depth.

The framework achieves performance improvements across benchmarks while maintaining single-feature interpretability. Task-specific analysis reveals patterns: BBQ demonstrates effective policy-critic coordination with clear sample discrimination, while safety tasks show strong steering effects in multi-layer configuration. CRL is complementary to fine-tuning, providing additive gains on top of supervised fine-tuned models. Branch point analysis confirms that layer 10 features encode syntactic structure while layer 20 features capture abstract task semantics. CRL establishes learned feature steering as a practical mechanistic interpretability tool, providing dynamic intervention-based views of model behavior that complement static feature analysis.

References

- Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Chalnev, S., Siu, M., and Conmy, A. Improving steering vectors by targeting sparse autoencoder features, 2024. URL <https://arxiv.org/abs/2411.02193>.
- Choi, D., Huang, V., Meng, K., Johnson, D. D., Steinhardt, J., and Schwettmann, S. Scaling automatic neuron description. <https://transluce.org/neuron-descriptions>, October 2024.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.
- DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Durmus, E., Tamkin, A., Clark, J., Wei, J., Marcus, J., Batson, J., Handa, K., Lovitt, L., Tong, M., McCain, M., Rausch, O., Huang, S., Bowman, S., Ritchie, S., Henighan, T., and Ganguli, D. Evaluating feature steering: A case study in mitigating social biases. <https://anthropic.com/research/evaluating-feature-steering>, 2024. Anthropic Research.
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.
- Elhoushi, M., Shrivastava, A., Liskovich, D., Hosmer, B., Wasti, B., Lai, L., Mahmoud, A., Acun, B., Agarwal, S., Roman, A., Aly, A. A., Chen, B., and Wu, C.-J. Layer-skip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 12622–12642, 2024. URL <https://doi.org/10.18653/v1/2024.acl-long.681>.
- Engels, J., Riggs, L., and Tegmark, M. Decomposing the dark matter of sparse autoencoders. *Transactions on Machine Learning Research*, 2025.
- Ferrao, J. L. Autosteer: Weight-preserving reinforcement learning for interpretable model control. <https://apartresearch.com>, 2024. Research submission to the research sprint hosted by Apart.
- Gui, T., Zhang, Q., Zhao, L., Lin, Y., Peng, M., Gong, J., and Huang, X. Long short-term memory with dynamic skip connections. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6481–6488, Jul. 2019. doi: 10.1609/aaai.v33i01.33016481. URL <https://ojs.aaai.org/index.php/AAAI/article/view/4613>.
- He, Z., Shu, W., Ge, X., Chen, L., Wang, J., Zhou, Y., Liu, F., Guo, Q., Huang, X., Wu, Z., Jiang, Y.-G., and Qiu, X. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders, 2024. URL <https://arxiv.org/abs/2410.20526>.
- Heimersheim, S. and Turner, A. Residual stream norms grow exponentially over the forward pass, May 2023. URL <https://www.lesswrong.com/posts/8mizBCm3dyc432nK8/residual-stream-norms-grow-exponentially-over-the-forward>. LessWrong, AI Alignment Forum.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=d7KJmI3GmQ>.
- Hernandez, E., Li, B. Z., and Andreas, J. Inspecting and editing knowledge representations in language models. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=ADtL6fgNRv>.
- Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=F76bwRSLeK>.
- Konen, K., Jentzsch, S., Diallo, D., Schütt, P., Bensch, O., El Baff, R., Opitz, D., and Hecking, T. Style vectors for steering generative large language models. In Graham, Y. and Purver, M. (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 782–802, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.findings-eacl.52/>.
- Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R.,

- and Nanda, N. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In Belinkov, Y., Kim, N., Jumelet, J., Mohebbi, H., Mueller, A., and Chen, H. (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL <https://aclanthology.org/2024.blackboxnlp-1.19/>.
- Lim, M., Becker, N., Kochenderfer, M. J., Tomlin, C. J., and Sunberg, Z. N. Optimality guarantees for particle belief approximation of pomdps. *Journal of Artificial Intelligence Research*, 77:1591–1636, 2023. doi: 10.1613/jair.1.14491. URL <https://jair.org/index.php/jair/article/view/14491>.
- Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., and Hendrycks, D. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=f3TUipYU3U>.
- Nikolaou, G., Mencattini, T., Crisostomi, D., Santilli, A., Panagakis, Y., and Rodolà, E. Language models are injective and hence invertible, 2025. URL <https://arxiv.org/abs/2510.15511>.
- Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. BBQ: A hand-built bias benchmark for question answering. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL <https://aclanthology.org/2022.findings-acl.165/>.
- Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramár, J., and Nanda, N. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders, 2024. URL <https://arxiv.org/abs/2407.14435>.
- Raposo, D., Ritter, S., Richards, B., Lillcrap, T., Humphreys, P. C., and Santoro, A. Mixture-of-depths: Dynamically allocating compute in transformer-based language models, 2024. URL <https://arxiv.org/abs/2404.02258>.
- Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. T. I., Chadha, A., Sheth, A., and Das, A. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2541–2573, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.155. URL <https://aclanthology.org/2023.emnlp-main.155/>.
- Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL <https://aclanthology.org/2024.acl-long.828/>.
- Röttger, P., Kirk, H., Vidgen, B., Attanasio, G., Bianchi, F., and Hovy, D. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL <https://aclanthology.org/2024.naacl-long.301/>.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. <https://openai.com/index/openai-baselines-ppo/>, 2017. URL <https://arxiv.org/abs/1707.06347>.
- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.
- Subramani, N., Suresh, N., and Peters, M. Extracting latent steering vectors from pretrained language models. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL <https://aclanthology.org/2022.findings-acl.48/>.
- Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev,

N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshv, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., yeong Ji, J., Mohamed, K., Badola, K., Black, K., Milligan, K., McDonnell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wählteiz, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Cogan, S., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving open language models at a practical size, 2024a. URL <https://arxiv.org/abs/2408.00118>.

Team, L., Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino,

D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yearly, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine,

- D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damla, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhee, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024b. URL <https://arxiv.org/abs/2407.21783>.
- Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization. *CoRR*, abs/2308.10248, 2023. URL <https://doi.org/10.48550/arXiv.2308.10248>.
- Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=y10DM6R2r3>.
- Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., and Fedus, W. Measuring short-form factuality in large language models, 2024. URL <https://openai.com/index/introducing-simpleqa/>.
- Wu, Z., Arora, A., Geiger, A., Huang, J., Wang, Z., Potts, C., Manning, C. D., and Bowman, S. R. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *ICML*, 2025.
- Xu, W. and Jin, X. Dynamic layer skipping for large language models on natural language understanding tasks and machine translation using reinforcement learning. *Frontiers in Computing and Intelligent Systems*, 9(3), 2024. doi: 10.54097/wy0g8m89. Uses dynamic layer skipping with adapters and RL to accelerate inference of LLMs; achieves SOTA performance on NLU and MT tasks.
- Ye, D., Lin, Y., Huang, Y., and Sun, M. Tr-bert: Dynamic token reduction for accelerating bert inference. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5798–5809, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.463. URL <https://aclanthology.org/2021.naacl-main.463/>.
- Yu, A. W., Lee, H., and Le, Q. Learning to skim text. In Barzilay, R. and Kan, M.-Y. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1880–1890, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1172. URL <https://aclanthology.org/P17-1172/>.

Zhong, H. and Zhang, T. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=1bTG4sJ7tN>.

A. Appendix

A.1. Key Terminology and Implementation Details

- **Corrected samples:** Samples where CRL steering changes incorrect predictions to correct ones
- **Misguided samples:** Samples where CRL steering changes correct predictions to incorrect ones
- **Policy network** π_θ : 2-layer MLP (Tanh activation) that maps residual stream activations to SAE feature selection probabilities (hidden dimension matching LLM)
- **Critic network** V_ϕ : 2-layer MLP (Tanh activation) that estimates state value function for PPO training (hidden dimension matching LLM)
- **Steering coefficient** c : Scalar multiplier controlling intervention magnitude in $\tilde{x}_t = x_t + c \cdot a_t W_{dec}$
- **Constrained decoding:** Generation restricted to valid answer tokens (e.g., A/B/C/D for MMLU)
- **Unconstrained decoding:** Generation allowing any token output, including invalid responses

A.2. Algorithm 1: CRL Training Procedure

Algorithm 1 CRL Training Procedure

Require: Base model M , SAE (W_{enc}, W_{dec}), task dataset D

Ensure: Trained policy π , value function V

- 1: Initialize policy $\pi(a|s)$ and value function $V(s)$
 - 2: **for** each episode **do**
 - 3: Sample batch B from dataset D
 - 4: **for** each token position t **do**
 - 5: Extract residual activation x_t
 - 6: Select feature $a_t \sim \pi(\cdot|x_t)$
 - 7: Apply intervention: $\tilde{x}_t = x_t + c \cdot a_t \cdot W_{dec}$
 - 8: Compute reward r_t from task outcome
 - 9: **end for**
 - 10: Update π and V using PPO with rewards $\{r_t\}$
 - 11: **end for**
 - 12: **return** trained π, V
-

A.3. Task-Specific Rewards

We use the following reward structures for each task:

Benchmarks with ground-truth answers:

- **MMLU & MMLU-Pro:** Binary reward: +1 for correct multiple-choice answer, 0 for incorrect
- **GSM8K:** Binary reward: +1 for exact numerical match, 0 otherwise
- **BBQ (Ambig & Disambig):** Binary reward: +1 for unbiased answer, 0 for biased answer

Benchmarks requiring evaluation models:

- **HarmBench & XSTest:** Binary reward: +1 for refusing harmful request, 0 for compliance.
- **SimpleQA:** Binary reward based on semantic similarity. The BERT STS model¹ matches generated answers against expected responses.

¹https://huggingface.co/datasets/HuggingFaceH4/stsb_multi_mt

A.4. Experimental Setup

Models and SAEs. We conduct experiments using Gemma 2 2B-IT (Team et al., 2024a) with pre-trained SAEs from Gemma Scope (Lieberum et al., 2024) (16K features) across layers 1-26, and LLaMA-3.1 8B (Team et al., 2024b) with LLaMA Scope SAEs (He et al., 2024) (32K features). Both SAE families employ JumpReLU activation (Rajamanoharan et al., 2024) and were the only releases providing SAEs across all transformer layers at the time of writing. SAEs are transferable across fine-tuned models with low reconstruction loss. **Training Protocol.** Training uses PPO with batch size 8, evaluating on 500 validation samples every 100 training steps. Selection tasks use 1 token generation, while HarmBench/XSTest use 32 tokens and GSM8K uses 1024 tokens. For datasets with fewer than 4000 samples, we repeat training data to reach 4000 samples. Layer selection for Table 1 uses a held-out validation set to select optimal layers and coefficients; final evaluation is performed on the separate test set to ensure no data leakage. **Cross-seed Stability.** Results are reported as mean \pm std across three seeds. Standard deviations remain low, with CRL notably reducing HarmBench variance from ± 9.05 to ± 1.59 , indicating stable feature selection.

Evaluation Benchmarks. We evaluate CRL across five categories: **Knowledge:** MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024). **Reasoning:** GSM8K (Cobbe et al., 2021). **Bias:** BBQ (Parrish et al., 2022). **Factuality:** SimpleQA (Wei et al., 2024). **Safety:** HarmBench (Mazeika et al., 2024), XSTest (Röttger et al., 2024).

A.5. CRL-Layer: Layer-wise Steering Mechanism

For layer-wise steering, we extend the framework to operate across multiple transformer layers simultaneously using a single token position. The shared Markov Decision Process operates on layer-specific variables where all components are parameterized with layer index ℓ .

Let $\mathbf{x}^{(\ell)} \in \mathbb{R}^d$ denote the residual stream activation at layer ℓ for a single target token position. Given layer-specific SAE components with encoder $\mathbf{W}_{enc}^{(\ell)} \in \mathbb{R}^{d \times d_{dict}}$ and decoder $\mathbf{W}_{dec}^{(\ell)} \in \mathbb{R}^{d_{dict} \times d}$, the steering mechanism applies perturbations across all layers:

$$\tilde{\mathbf{x}}^{(\ell)} = \mathbf{x}^{(\ell)} + \mathbf{a}^{(\ell)} \mathbf{W}_{dec}^{(\ell)} \quad (10)$$

where $\mathbf{a}^{(\ell)} \in \{0, 1\}^{d_{dict}}$ is the layer-specific action vector. The shared MDP coordinates decisions across layers through a joint policy:

$$\pi_{\theta}(\mathbf{a}^{(1:L)} | \mathbf{x}^{(1:L)}) = \prod_{\ell=1}^L \pi_{\theta}^{(\ell)}(\mathbf{a}^{(\ell)} | \mathbf{x}^{(\ell)}) \quad (11)$$

where L is the total number of layers and each layer-specific policy $\pi_{\theta}^{(\ell)}$ shares parameters while adapting to layer-specific representations.

CRL-Layer Results.

CRL-Layer’s practical advantage lies in sharing policy and critic networks across layers. This approach reduces computational resources compared to extending CRL-Token to multiple layers without cross-layer compatibility. Although representation spaces for each layer’s residual stream differ, residual connections enforce shared vector spaces, and existing layer-wise reuse approaches (Ye et al., 2021; Elhoushi et al., 2024; Raposo et al., 2024) support network sharing across layers.

Table 4. Performance comparison between CRL variants on Gemma 2 2B. Results use single-seed 0-shot evaluation with layer-shared policy; BBQ and HarmBench baselines differ from Table 1 due to 0-shot vs 1-shot evaluation and single-seed variance.

Method	MMLU	MMLU-Pro	BBQ Ambig	BBQ Disambig	XSTest	HarmBench
Baseline	52.23	30.30	59.10	75.42	86.35	44.64
CRL-Layer	55.00	29.18	62.93	75.68	86.98	71.43
CRL-Token	55.55	30.44	61.88	76.73	86.98	50.25

Table 4 presents a comparison between CRL-Layer and CRL-Token across multiple benchmarks. CRL-Layer achieves improvements over baseline across most tasks except MMLU-Pro, though it generally underperforms CRL-Token on

complex tasks, with the exception of refusal tasks. This pattern reveals that while layer-wise sharing provides clear computational advantages, it introduces limitations for complex tasks requiring layer-specific weights. Notably, CRL-Layer achieves the highest performance on HarmBench, outperforming CRL-Token, suggesting that layer-shared refusal patterns benefit from cross-layer feature consistency.

A.6. LLaMA Results

We extend our evaluation to LLaMA-3.1 8B model with LLaMA Scope SAEs to demonstrate the generalizability of our approach across different model architectures.

Table 5. Performance results for LLaMA-3.1 8B model across different tasks using CRL-Token approach. MMLU and MMLU-Pro use 0-shot evaluation, BBQ tasks use 1-shot evaluation, SimpleQA uses 0-shot evaluation, XSTest and HarmBench use refusal rate evaluation.

Task	Type	Layer	Before	After	Gain
MMLU	knowledge	20	61.41	62.33	+0.92
MMLU-Pro	knowledge	26	32.13	33.16	+1.03
BBQ Ambig	bias	7	83.97	85.04	+1.07
BBQ Disambig	bias	30	90.07	90.16	+0.09
SimpleQA	factuality	25	0.43	0.96	+0.53
XSTest	safety	23	61.27	59.37	-1.90
HarmBench	safety	5	0.71	6.07	+5.36

Table 5 demonstrates that CRL-Token generalizes across different model families, with mixed results across tasks. While most tasks show improvements (knowledge, bias, factuality), XSTest exhibits a decline (-1.90), suggesting task-specific sensitivity to the steering approach. HarmBench shows improvement (+5.36), consistent with the Gemma 2 2B results, demonstrating safety steering across architectures. The mixed results confirm that our approach transfers across different model architectures while highlighting the importance of task-specific optimization, maintaining the interpretable feature-based control mechanism.

A.7. Critic Value Analysis

Multi-token Generation Task Analysis.

The four sample categories (correct/incorrect, corrected/misguided) are analyzed for HarmBench and XSTest tasks to understand critic behavior patterns in multi-token generation scenarios.

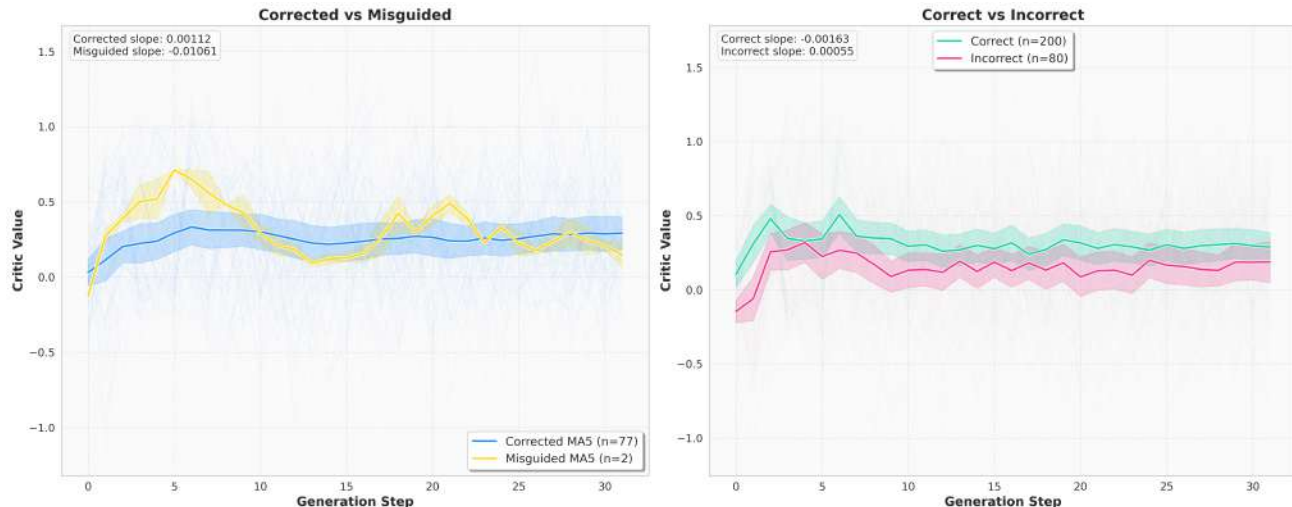


Figure 6. Critic value trajectories along token generation for HarmBench task. The critic successfully distinguishes correct from incorrect samples throughout generation.

Our analysis focuses on linear regression gradient slopes and the visible gaps between sample categories. Empirically, we observed that critic value gaps between correct and incorrect samples become more pronounced in later layers. We observe

two distinct patterns in critic trajectories analogous to those in single-token generation tasks. These patterns reflect either errors in critic value estimation (manifesting as bias gaps) or limitations in the policy’s critic value utility (affecting gradient dynamics).

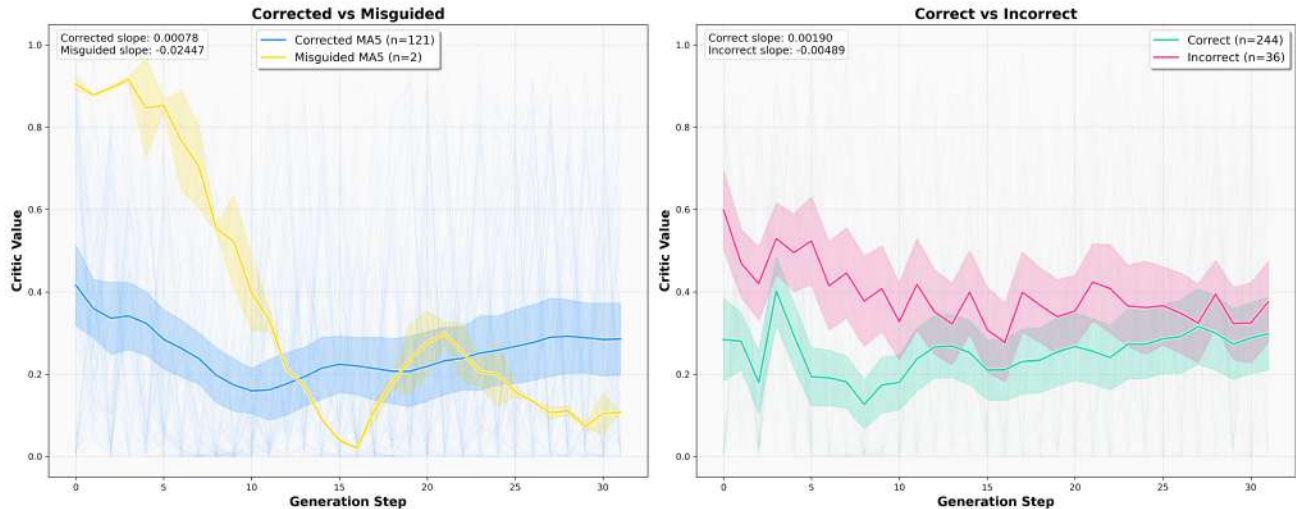


Figure 7. Critic value trajectories along token generation for XSTest task. XSTest shows inverted estimation patterns compared to HarmBench, indicating critic bottleneck for context-dependent benign requests.

For HarmBench (Figure 6), the critic successfully distinguishes between correct and incorrect samples, whereas XSTest (Figure 7) exhibits inverted estimation patterns. However, while HarmBench shows clear gaps at sequence boundaries with minimal gradient differences, XSTest demonstrates estimation errors in gap measurements but maintains clear superiority for correct samples in slope analysis across both correct/incorrect and corrected/misguided categories.

A.8. Hallucination Mitigation Analysis

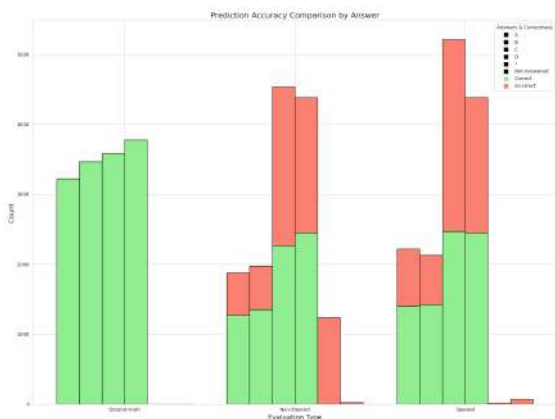


Figure 8. MMLU invalid response comparison. Baseline generates responses like “*” or whitespace instead of valid options (A, B, C, D). CRL-Token steering eliminates these invalid outputs, contributing to accuracy gains beyond format correction.

For single-token generation tasks, particularly MMLU without constrained decoding, a substantial portion of performance improvement stems from hallucination mitigation. CRL-Token steering effectively eliminates responses that fall outside the provided answer options (A, B, C, D).

The baseline model frequently generates invalid responses such as “*” or whitespace instead of selecting from the given options. This behavior significantly impacts performance metrics, as these responses are automatically marked incorrect regardless of the underlying reasoning quality. Our CRL-Token approach addresses this issue by learning to constrain outputs to valid answer choices while maintaining the model’s reasoning capabilities.

The analysis reveals that a portion of performance gains in MMLU tasks can be attributed to this hallucination mitigation effect, with the remainder coming from improved feature selection and reasoning enhancement. This finding highlights the importance of output format consistency in evaluation benchmarks and demonstrates CRL’s effectiveness in learning task-specific constraints.

A.9. Feature Selection Evaluation

We analyze the policy network’s feature selection patterns to understand which SAE features contribute most to performance improvements. For each feature i , we compute:

A.9.1. IMPACT SCORE

$$n_i = \sum_t \mathbf{1}\{i \in S_t\}, \quad N = \sum_j n_j, \quad (12)$$

$$p_i = \frac{n_i}{N}, \quad (13)$$

$$c_i = \sum_t \mathbf{1}\{i \in S_t\} \mathbf{1}\{\text{improved}_t\}, \quad m_i = \sum_t \mathbf{1}\{i \in S_t\} \mathbf{1}\{\text{degraded}_t\}, \quad (14)$$

$$\text{Impact}_i = \frac{(c_i + m_i) \cdot \log(n_i + \varepsilon)}{n_i \cdot \log(n_i + \varepsilon)} = \frac{c_i + m_i}{n_i}, \quad (\varepsilon > 0) \quad (15)$$

where S_t is the set of selected features at step t (e.g., top- k), and $\text{improved}_t/\text{degraded}_t$ denote whether steering improved/degraded the outcome versus baseline.

A.9.2. FEATURE DIVERSITY

Additionally, we compute feature diversity to understand the policy’s exploration behavior:

$$\text{Feature Diversity} = - \sum_{i=1}^{d_{dict}} p_i \log p_i \quad (16)$$

where p_i is the probability of selecting feature i . Higher entropy indicates more diverse feature usage across the SAE dictionary. Together, these metrics quantify how diverse the trained policy’s feature selection is and how much the selected features actually influence behavior change.

A.10. Feature Diversity and Impact Score

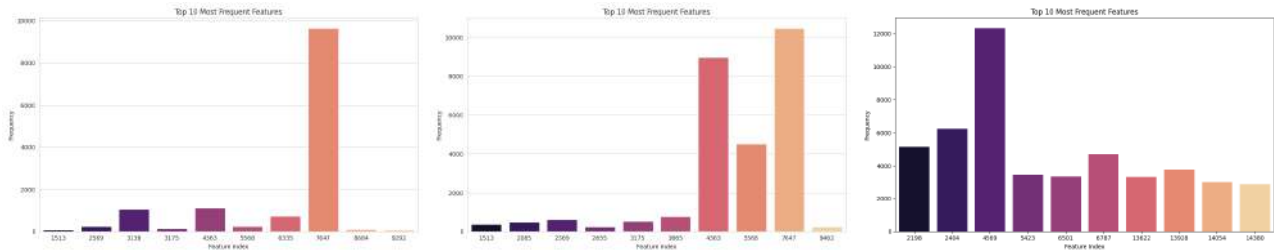


Figure 9. Top activated features for MMLU (left) and BBQ Disambig (center) and GSM8K (right) tasks, revealing semantic patterns in feature selection for reasoning and bias mitigation.

Feature diversity remains consistent across different hyperparameters within the same task, though it decreases with increased policy layer depth. As shown in Figure 9, tasks requiring longer token generation horizons (GSM8K: 6.695) or higher complexity (MMLU-Pro: 5.476) demonstrate elevated feature diversity compared to shorter response tasks (HarmBench: 2.938), with XSTest (4.757) showing intermediate diversity levels. Impact scores exhibit an inverse relationship with feature diversity: tasks with lower diversity show higher concentrated impact scores (HarmBench: 0.808), while high-diversity tasks demonstrate more distributed feature contributions (XSTest: 0.357, MMLU-Pro: 0.625). This pattern suggests that complex reasoning tasks require broader feature engagement, whereas focused tasks benefit from concentrated feature activation. Figure 10 demonstrates that activation frequency exhibits rank-dependent decay with task-specific distribution patterns reflecting underlying diversity characteristics.

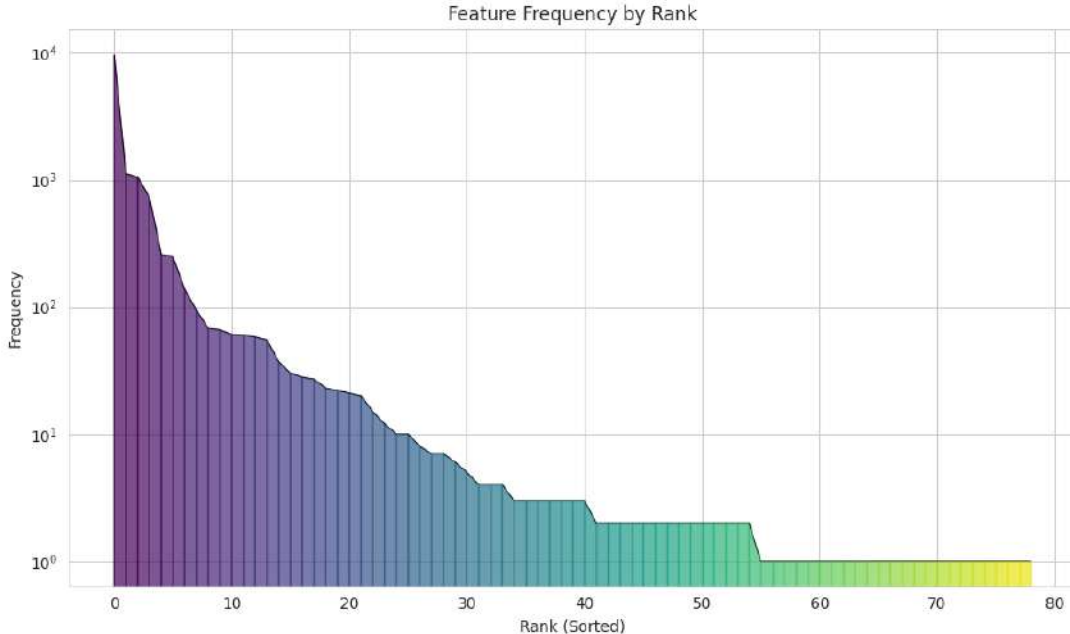


Figure 10. Feature usage hierarchy in Gemma 2 2B MMLU task, demonstrating that the policy selectively identifies specific features rather than random selection.

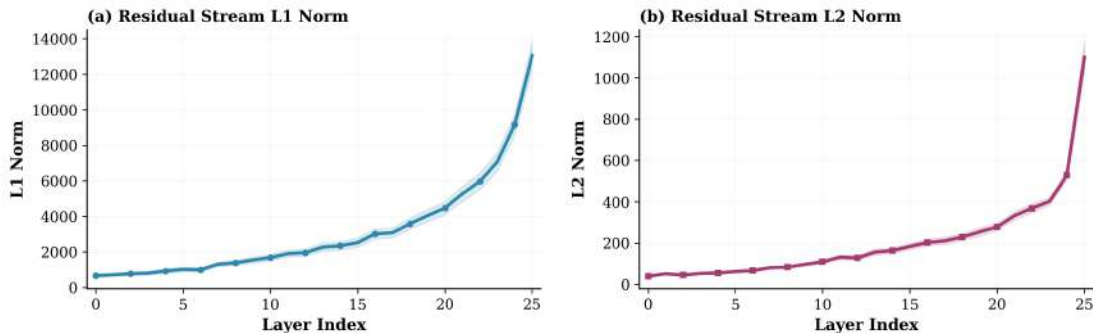


Figure 11. Residual stream norm growth across layers for Gemma 2 2B on MMLU. Norm increases with depth, explaining why identical steering coefficients produce different effects at different layers and motivating our adaptive coefficient strategy.

A.11. Layer-wise Analysis

Figure 11 shows residual stream norm growth across layers for Gemma 2 2B model on the MMLU task. This pattern is consistent with prior observations that residual stream norms increase with network depth (Heimersheim & Turner, 2023).

The analysis for layers $\ell \geq 18$ for BBQ tasks (Figure 15) confirms that ambiguous and disambiguated contexts require different optimal intervention layers. This suggests that bias mitigation strategies must be adapted to the specific type of ambiguity present in the input. The layer-wise analysis shows similar patterns across tasks: later layers generally provide more effective intervention points, while early layers show degraded performance with large coefficients. This pattern holds across different coefficient values and task types, consistent with observations that residual stream norms increase across layers (Heimersheim & Turner, 2023).

B. Feature Analysis Details

B.1. Corrected Features for GSM8K Task

The following features demonstrate positive steering effects, correcting model outputs from incorrect to correct responses:

Control Reinforcement Learning

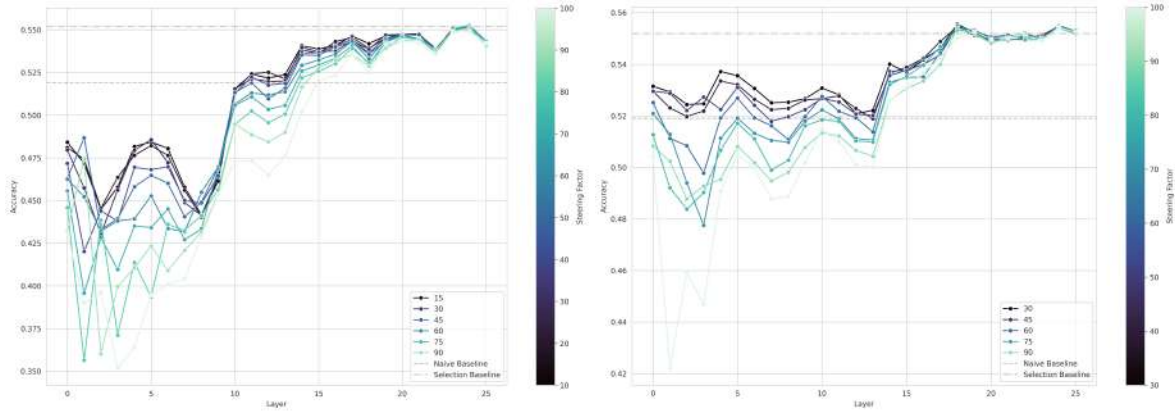


Figure 12. MMLU accuracy per fixed coefficient across layers for Gemma 2 2B, showing optimal intervention points for unconstrained (left) and constrained (right) decoding. Later layers generally outperform earlier layers, with large coefficients degrading early-layer performance.

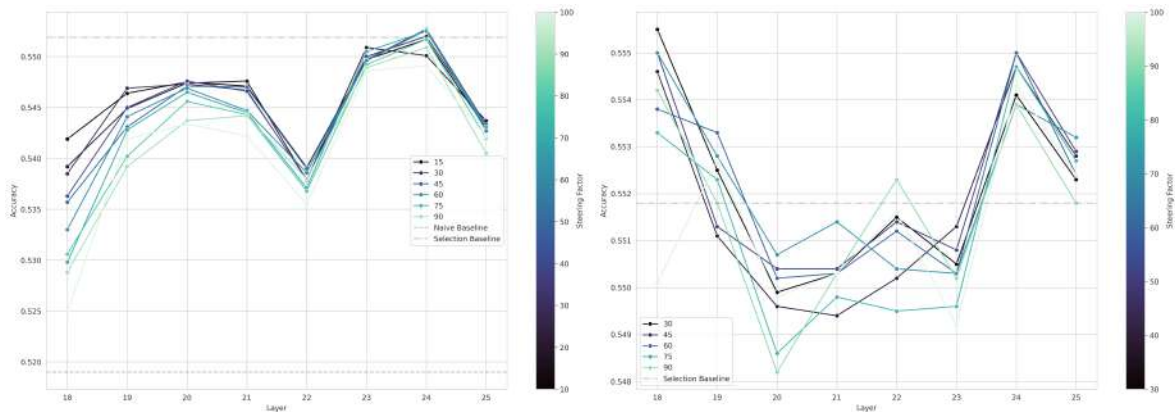


Figure 13. MMLU accuracy per fixed coefficient across layers $\ell \geq 18$ for unconstrained (left) and constrained (right) decoding. Unconstrained and constrained settings show different optimal layers, indicating distinct feature requirements for format correction vs factual answering.

- 4504 diagrams and schematic representations related to workflows, communication systems, and protocols (correct/incorrect: 321/170, corrected/misguided: 15/4)
- 406 specific mentions of individuals, organizations, and their respective roles or activities (correct/incorrect: 116/113, corrected/misguided: 12/4)
- 1440 instances of the word "vice" and related terms (correct/incorrect: 161/119, corrected/misguided: 11/4)
- 14204 indicators of group dynamics and power relations (correct/incorrect: 248/179, corrected/misguided: 16/6)
- 10961 terms related to statistical methods and implementation details (correct/incorrect: 223/182, corrected/misguided: 20/8)
- 2699 significant concepts related to problems and solutions within a defined framework (correct/incorrect: 234/187, corrected/misguided: 17/7)
- 2317 phrases indicating comparisons or relationships between two entities or elements (correct/incorrect: 120/86, corrected/misguided: 12/5)
- 2720 mathematical expressions and equations related to functions and inequalities (correct/incorrect: 151/150, corrected/misguided: 11/5)
- 347 phrases and concepts related to accountability and compliance (correct/incorrect: 148/129, corrected/misguided: 15/7)
- 7708 mathematical operations and expressions in various forms (correct/incorrect: 188/217, corrected/misguided: 19/9)

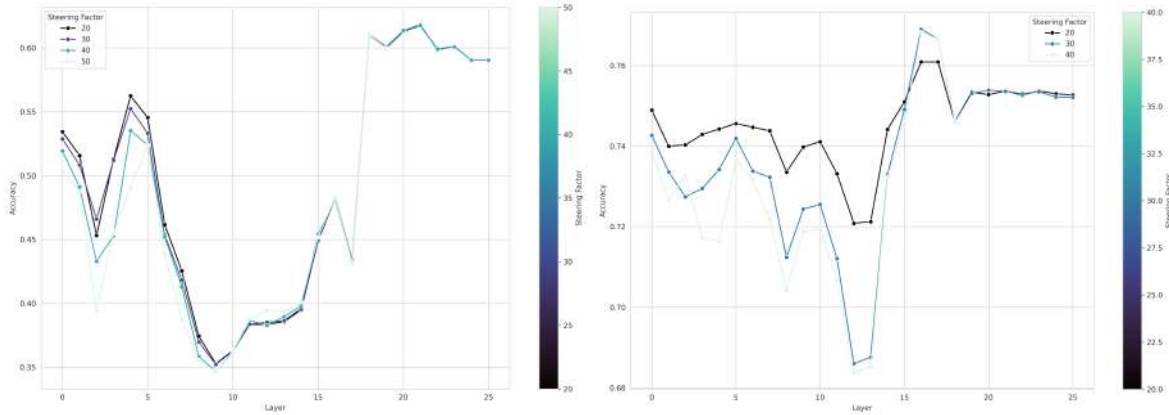


Figure 14. BBQ accuracy per fixed coefficient across layers for ambiguous (left) and disambiguated (right) contexts. Different optimal layers indicate context-dependent bias mitigation strategies.

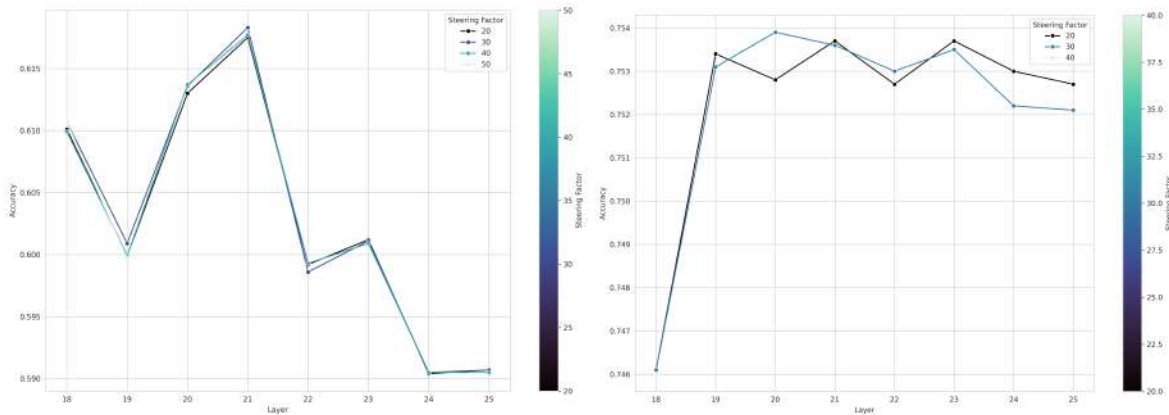


Figure 15. BBQ accuracy per fixed coefficient across layers $\ell \geq 18$ for ambiguous (left) and disambiguated (right) contexts. Later layers show distinct optimal coefficients for each context type.

B.2. Additional Corrected and Misguided Examples

B.3. Misguided Features for GSM8K Task

Conversely, these features demonstrate negative steering effects, inadvertently changing correct responses to incorrect ones:

- 7999 mathematical expressions or equations (correct/incorrect: 40/42, corrected/misguided: 1/10)
- 4069 mathematical notation and geometric properties related to circles and angles (correct/incorrect: 121/100, corrected/misguided: 7/16)
- 13752 elements related to job creation and economic context (correct/incorrect: 109/78, corrected/misguided: 5/10)
- 10968 specific coding or mathematical terms and phrases related to programming or data analysis (correct/incorrect: 99/112, corrected/misguided: 8/16)
- 16295 significant terms and phrases related to scientific studies, particularly in the context of law, medicine, and research methodologies (correct/incorrect: 99/93, corrected/misguided: 6/11)
- 12174 references to understanding and problem-solving processes (correct/incorrect: 180/157, corrected/misguided: 10/17)
- 15434 complex relationships involving socio-legal and psychological themes (correct/incorrect: 113/125, corrected/misguided: 8/13)
- 1222 sentences expressing emotional vulnerability and complex interpersonal dynamics (correct/incorrect: 74/61, corrected/misguided: 7/11)
- 7834 indicative symbols and formatting used in programming or mathematical expressions (correct/incorrect: 97/101, corrected/misguided: 9/14)

Control Reinforcement Learning

Mico's age in 10 years $10 + 10 = 20$ * Marco's age in 10 years $10 + 10 = 20$ **6 Find the sum of their ages in 10 years:** * $20 + 20 = 40$ ##### The sum of their ages in 10 years will be Here's how to solve it 1 **Calculate the total number of sweets:** - 15 packs * 60 pieces/pack = 900 pieces 2 **Calculate the number of packs Anne kept:** - 15 packs - 2 packs = 13 packs 3 **Calculate the number of pieces given to friends:** - 900 pieces - 120 pieces (for

Figure 16. GSM8K corrected case 1: Contextually appropriate feature steering activates relevant numerical tokens, strengthening reasoning.

$32 * 3 = 96$ classes visited on Wednesday * **Total for Monday-Wednesday:** $32 + 64 + 96 = 192$ classes visited from Monday to Wednesday * **Thursday and Friday:** $30 + 25 = 55$ classes visited on Thursday and Friday * **Total for the week:** $192 + 55 = 247$ classes visited the Science Center Here's how to solve it 1 **Find the number of babies wearing purple:** * Since there are 8 babies and Carla dresses $3/4$ in purple there are $(3/4) * 8 = 6$ babies wearing purple 2 **Find

Figure 17. GSM8K corrected case 2: Semantically coherent feature activations align with equality tokens, guiding correct computation.

- 3415 patterns of conditional phrases and expressions of uncertainty (correct/incorrect: 92/108, corrected/misguided: 8/12)

B.4. Additional Misguided Example

Figure 21 illustrates a case where misguided feature steering causes the model to abandon a correct solution path. The steered model selects features associated with task-irrelevant semantics (e.g., socio-legal framing or programming constructs from the misguided feature list above), which interfere with the arithmetic reasoning required for the problem. This token-level interference is visible in CRL’s intervention logs: the selected feature’s description does not align with the mathematical operation at that position, providing a concrete diagnostic signal for identifying when steering goes wrong.

B.5. Additional Branch Analysis

We analyze decision bifurcations at critical tokens, comparing layer 10 and layer 20 feature selections across GSM8K tasks. The following cases show different steering outcomes depending on which layer’s features are selected.

In these three cases, L10 features produce correct outputs while L20 features fail. Both the toy calculation and tournament problems use L10’s "pronoun references" feature, which generates accurate enumerated calculations, whereas L20’s "code formatting" and "punctuation cues" features produce incorrect results. The speed conversion case shows L10’s "emotional relationships" feature generating the correct formula (" $3 \text{ miles/hour} * 2 = 6$ "), while L20’s "claims and benefits" produces wrong logic (" $3 \text{ miles per hour} / 2 = 1.5$ "). Feature labels alone do not predict which layer succeeds: the "emotional relationships" label bears no obvious relation to speed calculation.

In these five cases, L20 features produce correct outputs while L10 features fail. The bus passenger problem shows L20’s "code formatting" generating correct sequential tracking ("1. **People after the first stop:** $120 + 20 = 140$ "), while L10’s "pronoun references" produces verbose, incorrect narrative. The variable abstraction case shows L10 choosing a variable-based approach ("Let 'J' represent...") while L20 proceeds directly with numerical calculation ("Jason carried 2 balls"). In the subscription pricing problem, L10’s "legal contracts" generates incorrect output, while L20’s "uncertainty expressions" handles the choice scenario. The banana and pages problems similarly show L10 features producing incorrect outputs, whereas L20 features generate correct multi-step solutions.

Control Reinforcement Learning

the number of babies wearing blue:** * There are 8 babies and 1/4 are wearing blue so there
L20/2699 L20/7708

are (1/4) * 8 = 2 babies wearing blue 3 **Find the number of babies wearing bows:** * 1/3
L20/10961 L20/7708 L20/2699 L20/7708

of the purple wearers wear bows so there are (1/3) * 6 = 2 babies wearing bows 4 **Find the
L20/10961 L20/7708

number of babies wearing both bows and purple:** * There are 2 babies wearing bows and 6 babies
L20/2699 L20/7708 L20/406 L20/7708 L20/7708

wearing purple so there are 2 babies wearing both 5 **Find the number of babies wearing purple
L20/7708 L20/2699 L20/7708

Figure 18. GSM8K corrected case 3: Feature 7708, described as *mathematical operations and expressions in various forms*, activates on quantitative units like "babies" and "packs", demonstrating lexical generalization beyond surface forms.

Here's how we can solve it * **Girls' teams:** 4 schools * 1 girls' team = 4 girls' teams *
L20/15434 L20/15434

Boys' teams: 4 schools * 1 boys' team = 4 boys' teams * **Total girls' players:** 4 girls'
L20/15434 L20/15434

teams * 5 players/team = 20 girls' players * **Total boys' players:** 4 boys' teams * 5
L20/15434

players:** 20 girls' players + 20 boys' players = 40 players * **Total coaches:** 4 coaches
L20/15434

Here's how we can solve it 1 **Height of the house:** The house is 7 times as high as Mattison
L20/15434

Figure 19. GSM8K misguided case 1: Selection of task-irrelevant features interferes with mathematical reasoning and diverts the solution process.

Observations. Across these eight examples, we observe a pattern: L10 succeeds on three cases involving discrete item counting and direct arithmetic, while L20 succeeds on five cases involving multi-step accumulation, choice scenarios, and variable reasoning. L10 failures show patterns of surface matching (e.g., financial terminology leading to "legal contracts", notation leading to "programming constructs"), whereas L20 failures show structural misapplication to simple problems (e.g., "code formatting" on toy calculation). Feature labels provide limited predictive value—"emotional relationships" solves speed calculations, "code formatting" performs both correctly and incorrectly on different tasks. These divergent outcomes demonstrate that layer choice affects steering effectiveness, which CRL's layer-specific policy exploits to adapt to different problem structures.

B.6. Computational Cost and Coefficient Sensitivity

CRL introduces minimal computational overhead during inference (8–9% latency on A100). The overhead consists of SAE encoding, policy forward pass, and steering vector computation. Table 6 reports per-task training times and resource requirements.

Table 6. Training cost per task on a single A100. Policy+critic parameters are 84M (3.3% of Gemma 2B) and 302M (3.8% of LLaMA 8B). All runs use batch size 8, 4000 samples, 1000 PPO iterations.

Model	Task	Gen Length	Training Time	Peak Memory
Gemma 2 2B	MMLU	1 token	~37 min	~6.8 GB
Gemma 2 2B	BBQ	1 token	~55 min	~6.8 GB
Gemma 2 2B	GSM8K	1024 tokens	~5.5 hrs	~6.8 GB
Gemma 2 2B	HarmBench	32 tokens	~3 hrs	~6.8 GB
LLaMA 3.1 8B	MMLU	1 token	~2.5 hrs	~15.7 GB

For steering coefficients, our adaptive strategy averages activation magnitudes from correctly predicted samples: $c = \frac{1}{|D^+|} \sum_{x \in D^+} \|a_t \mathbf{W}_{dec}\|$, where D^+ denotes correct samples. This eliminates manual tuning across 26 layers. Table 8 shows moderate coefficients (25-50) achieve optimal performance, with layer-specific analysis in Figure 12.

Control Reinforcement Learning

first floor's floor to ceiling height is thrice as high as Mattison so it's $3 * 60$ inches = 180
L20/15434

Here's how we can solve it **Step 1 Calculate team B's goals in the first half.** Team B
L20/16295 L20/1222

scored 2 fewer goals than team A so they scored $4 - 2 = 2$ goals **Step 2 Calculate team A's
L20/4069 L20/4069 L20/7834

goals in the second half.** Team A scored $1/4$ of the goals scored by team B in the second half
L20/1222 L20/10968L20/4069 L20/1222 L20/4069 L20/1222

Team B scored $4 * 2 = 8$ goals in the second half Therefore team A scored $8 * (1/4) = 2$ goals in
L20/4069 L20/7834 L20/1222 L20/10968L20/4069

Figure 20. GSM8K misguided case 2: Token-feature misalignment produces semantically incoherent activations, degrading reasoning performance.

* Jogger jeans cost 6 less than tattered jeans so they cost $28 - 6 = 22$ 2 **Calculate the
L20/4069 L20/1222

total savings from the jogger jeans:** * He saved 6 total and he saved $1/3$ of the savings from
L20/1222 L20/4069 L20/7834 L20/4069 L20/1222

the jogger jeans * Total savings from jogger jeans $6 * 3 = 18$ 3 **Calculate the savings
L20/4069 L20/7834 L20/1222 L20/7834 L20/4069

from the tattered jeans:** * He saved 6 total and he saved the rest from the tattered jeans
L20/1222 L20/4069 L20/4069 L20/1222 L20/4069

* Savings from tattered jeans $6 - 18 = -12$ 4 **Determine the original price of the jogger
L20/1222 L20/4069 L20/4069 L20/7834

Figure 21. GSM8K misguided example 3: Feature interference causes abandonment of correct solution path.

Table 7. Runtime overhead. CRL adds 8–9% latency.

Task	Type	Overhead
MMLU	Single	+8.0%
BBQ	Single	+7.8%
GSM8K	Multi	+8.8%
HarmBench	Multi	+9.1%

Table 8. Coefficient sensitivity on MMLU (layer 24).

Coef.	10	25	50	75	100
Acc. (%)	53.2	54.8	55.4	54.1	52.3

Complementarity with Fine-tuning. CRL provides additive gains when applied on top of supervised fine-tuned models. On MMLU, accuracy improves from 55.2% (base) to 55.7% (SFT) to 56.1% (SFT + CRL), demonstrating that learned feature steering captures complementary signals beyond what fine-tuning achieves. This suggests CRL can serve as a lightweight post-hoc enhancement without requiring additional weight updates. The Gemma Scope SAEs transfer effectively across model variants, enabling CRL application without retraining the autoencoder. For deployment scenarios requiring rapid adaptation, CRL offers real-time steering with <10% latency overhead compared to full model fine-tuning.

B.7. Cross-Task Transfer Analysis

To evaluate whether CRL captures task-specific mechanisms or dataset-specific shortcuts, we apply policies trained on one task to other tasks, using the training task’s optimal layer. Table 9 reports the full results alongside native CRL performance for reference.

Table 9. Cross-task transfer results. “Native CRL” shows the target task’s own trained policy for reference. All experiments use the source task’s optimal layer.

Train → Eval	Layer	Acc. (%)	vs. Base	Native CRL
MMLU → GSM8K	24	31.01	−10.31	55.65 (L20)
GSM8K → MMLU	20	52.24	+0.18	55.37 (L24)
HarmBench → XSTest	21	86.35	+1.27	87.62 (L12)
BBQ → MMLU	5	53.55	+1.49	55.37 (L24)

The transfer pattern is asymmetric: MMLU features actively harm GSM8K (−10.31%), while GSM8K features remain inert

Control Reinforcement Learning

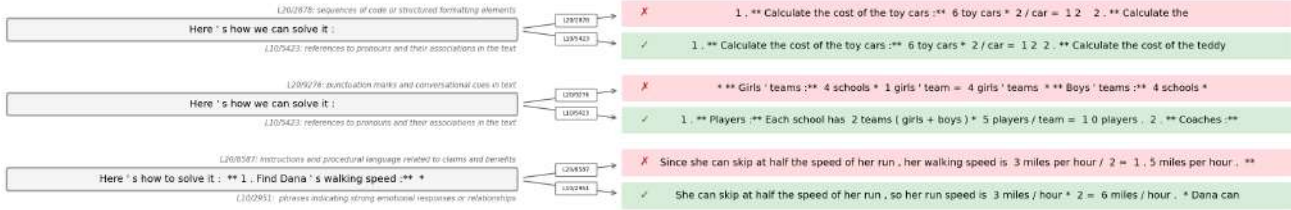


Figure 22. Cases where layer 10 features succeed. Top: Toy calculation where L10’s "pronoun references" captures concrete items while L20’s "code formatting" misapplies structure. Middle: Tournament problem where L10’s "pronoun references" handles entity tracking while L20’s "punctuation cues" misidentifies structure. Bottom: Speed calculation where L10’s "emotional relationships" correctly solves the problem while L20’s "claims and benefits" misapplies procedural framing.

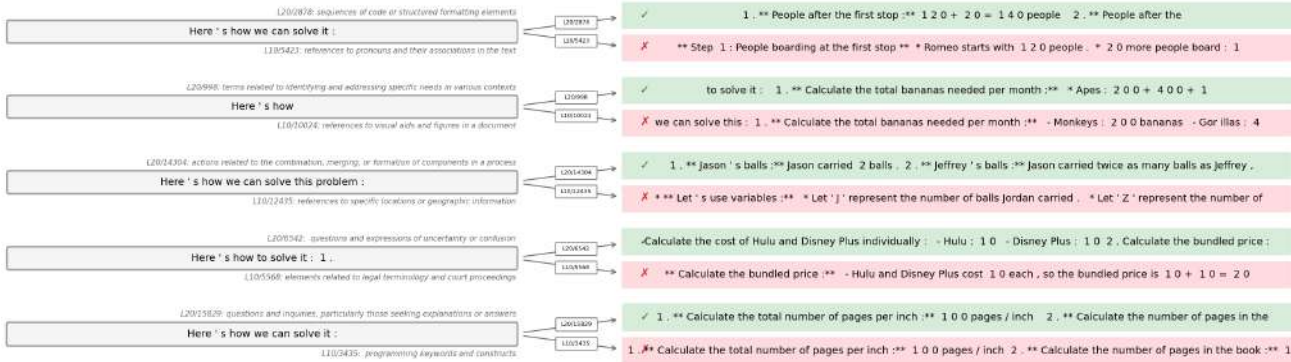


Figure 23. Cases where layer 20 features succeed. Top: Bus problem where L20’s "code formatting" correctly structures sequential steps while L10’s "pronoun references" fails. Second: Banana calculation where L20’s "needs identification" captures requirement semantics while L10’s "visual aids" produces incorrect output. Third: Variable abstraction where L20’s "combination and formation" proceeds with direct calculation while L10’s "geographic information" chooses variable-based approach. Fourth: Subscription pricing where L20’s "uncertainty expressions" captures choice scenarios while L10’s "legal proceedings" produces incorrect output. Bottom: Pages calculation where L20’s "questions and inquiries" captures exploratory reasoning while L10’s "programming constructs" treats it as code.

on MMLU (+0.18%). If the policy captured dataset-specific shortcuts (e.g., bias toward numeric outputs), GSM8K features should also disrupt MMLU, but they do not. HarmBench features slightly improve XSTest (+1.27%), consistent with shared safety representations across related tasks. The interference aligns with layer-specific computation: MMLU features are selected at layer 24 while GSM8K’s optimal intervention occurs at layer 20, and applying features across mismatched depths introduces structured interference rather than random degradation.

B.8. Dynamic Coefficient Ablation

The steering coefficient c is fixed during training, and the policy learns feature selections calibrated to this static scale. We tested three per-token dynamic alternatives (Table 10): norm-scaled ($c \propto \|\mathbf{x}_t\|$), activation-scaled (c set to the natural SAE activation of the selected feature), and critic-gated ($c \propto V_\phi(\mathbf{s}_t)$).

Table 10. Dynamic coefficient variants vs. static (dataset-averaged). All dynamic variants degrade performance due to train-inference mismatch in the learned feature-coefficient coupling.

Method	HarmBench	GSM8K
Static (dataset-averaged)	49.75	55.65
Norm-scaled ($c \propto \ \mathbf{x}_t\ $)	45.00	41.39
Activation-scaled ($c =$ natural SAE activation)	41.79	40.33
Critic-gated ($c \propto V_\phi(\mathbf{s}_t)$)	43.93	41.47

The policy is trained with a fixed coefficient and learns a joint feature-coefficient coupling where selected features are calibrated to the static intervention scale. Changing the coefficient at inference disrupts this coupling. The static coefficient itself is robust: varying c from 10 to 100 on MMLU changes accuracy by only ~ 3 pp (Table 8), confirming the dataset-averaged approach finds a stable operating point.

B.9. Component Ablation: Isolating CRL and AFM

To isolate the contributions of learned feature selection and Adaptive Feature Masking, we compare CRL variants against heuristic baselines (Table 11). All methods share identical decoding and evaluation setup; only the feature selection mechanism differs.

Table 11. Component ablation. CRL without AFM numbers are from matched-hyperparameter runs. HarmBench CRL+AFM: 3-seed mean (paper reports 49.12, within run-to-run variance ± 1.59).

Method	BBQ Ambig	HarmBench
No steering (Base)	60.17	41.46
Random feature	58.36	45.35
Random + AFM	60.16	46.96
Most-active feature	59.94	48.03
CRL without AFM	62.55	48.67 ± 0.56
CRL + AFM (full)	65.86	49.20± 0.84

CRL without AFM achieves +2.39 on BBQ Ambig and +7.21 on HarmBench over the baseline, confirming that the learned policy provides most of the gain. AFM contributes an additional +3.31 on BBQ Ambig and +0.53 on HarmBench by encouraging diverse feature exploration. Random+AFM improves over random alone (+1.80 BBQ, +1.61 HarmBench), isolating AFM’s independent contribution from learned selection.