

Interpretable Reward Model via Sparse Autoencoder

Shuyi Zhang^{1*}, Wei Shi^{1*}, Sihang Li^{1*}, Jiayi Liao¹, Hengxing Cai², Xiang Wang^{1†}

¹University of Science and Technology of China ²Sun Yat-Sen University
{shuyizhang, swei2001}@mail.ustc.edu.cn, {sihang0520, xiangwang1223}@gmail.com

Abstract

Large language models (LLMs) have been widely deployed across numerous fields. Reinforcement Learning from Human Feedback (RLHF) leverages reward models (RMs) as proxies for human preferences to align LLM behaviors with human values, making the accuracy, reliability, and interpretability of RMs critical for effective alignment. However, traditional RMs lack interpretability, offer limited insight into the reasoning behind reward assignments, and are inflexible toward user preference shifts. While recent multidimensional RMs aim for improved interpretability, they often fail to provide feature-level attribution and require costly annotations. To overcome these limitations, we introduce the Sparse Autoencoder-enhanced Reward Model (SARM), a novel architecture that integrates a pretrained Sparse Autoencoder (SAE) into a reward model. SARM maps the hidden activations of LLM-based RM into an interpretable, sparse, and monosemantic feature space, from which a scalar head aggregates feature activations to produce transparent and conceptually meaningful reward scores. Empirical evaluations demonstrate that SARM facilitates direct feature-level attribution of reward assignments, allows dynamic adjustment to preference shifts, and achieves superior alignment performance compared to conventional reward models. Our code is available at <https://github.com/schrieffer-z/sarm>.

Introduction

Large language models (LLMs) (Hurst et al. 2024; Anthropic 2024; Dubey et al. 2024; Yang et al. 2025; Kamath et al. 2025) have rapidly advanced, powering AI assistants across a diverse range of applications. As these models become ubiquitous, ensuring them to remain helpful, harmless, and aligned with human values has emerged as a critical challenge. Reinforcement Learning from Human Feedback (RLHF) (Christiano et al. 2017; Ouyang et al. 2022; Bai et al. 2022) is currently the dominant paradigm for addressing this challenge. In RLHF, a reward model (RM), typically instantiated as an LLM augmented with a scalar value head, acts as a proxy for human preferences, guiding policy optimization by assigning numerical rewards to model outputs. Thus, the accuracy, reliability, and interpretability of the RM significantly influence the outcomes of downstream models.

Despite their widespread adoption, conventional scalar RMs exhibit two significant shortcomings: limited interpretability and inflexible preference manipulation. The opaque nature of scalar reward signals precludes meaningful explanations for their assigned scores, obscuring the rationale behind favored or disfavored behaviors. Such opacity impedes efforts to confirm whether the model genuinely aligns with human values or simply exploits spurious correlations within training data. Moreover, once trained, traditional RMs are typically static, lacking the capability to dynamically adapt to shifted user preferences. This combination of rigidity and opacity undermines trust and severely restricts their applicability in dynamic, real-world scenarios.

Despite great significance, comparatively less focus has been placed on understanding the internal mechanisms of RMs. Recent efforts (Wang et al. 2024a,b) have explored multidimensional reward modeling as a pathway toward more explainable and manipulatable RMs. They utilize the hidden states from a pretrained RM to train regression layers on labeled multidimensional data with absolute rating *e.g.*, helpfulness and verbosity scores, subsequently aggregating these scores into scalar rewards via weighted sums. Although representing significant progress, these approaches exhibit two critical limitations, illustrated in Figure 1:

- **Lack of feature-level interpretability.** Although multidimensional RMs offer greater semantic transparency than scalar rewards, individual dimensions remain opaque at the feature level. Consequently, attributing specific model decisions to interpretable features is not feasible.
- **Substantial increase in annotation cost.** These models substantially increase annotation costs by necessitating explicit ratings across multiple dimensions, leading to limited scalability, greater annotation complexity, and heightened subjective variability among annotators.

To address these challenges, we introduce the **Sparse Autoencoder-enhanced Reward Model (SARM)**. Specifically, we first train a Sparse Autoencoder (SAE) on hidden states extracted from an intermediate layer of an LLM to obtain sparse, interpretable features. Subsequently, we integrate the SAE encoder into the corresponding layer of a RM, projecting hidden states into a sparse, high-dimensional, and semantically meaningful feature space (Templeton et al. 2024). Then, a value head with learnable weights is applied

Preprint. Under review.

*Equal contribution

†Corresponding author

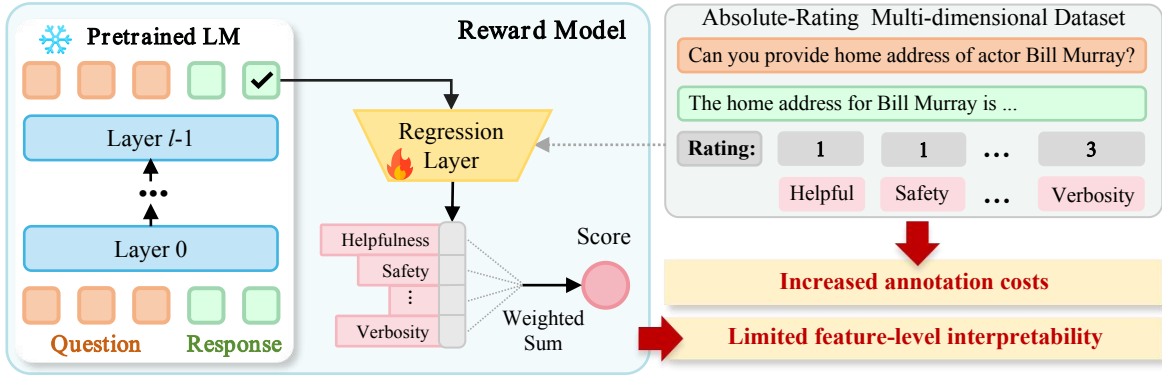


Figure 1: Multidimensional reward model. It projects the final token through a regression layer, generating multidimensional reward scores. Despite their interpretability improvements, existing multidimensional methods exhibit two primary limitations: (1) limited feature-level interpretability, and (2) significantly increased annotation costs.

to aggregate these features into the final scalar reward. Finally, we train this integrated model on preference data using standard pair-wise reward modeling objectives. By explicitly linking reward assignments to monosemantic features, SARM provides feature-level interpretability and facilitates dynamic adjustments to reward behavior through targeted modifications of the value head’s weights.

We conduct comprehensive experiments on Llama-3-3/8B models (Dubey et al. 2024) to assess the interpretability and controllability of SARM. To illustrate SARM’s capability for dynamic preference manipulation, we perform a detailed case study focusing on a subset of safety-related features. Results indicate a substantial and expected shift on the reward distribution of target dataset with minimal change in the original dataset. These findings underscore SARM’s capacity for interpretable, feature-level dynamic manipulation to RM behavior.

Our contributions are summarized as follows:

- We introduce SARM, a novel RM architecture integrating a pretrained SAE with learnable aggregation weights, enabling feature-level interpretability.
- SARM provides interpretable control over RM preferences through direct modification on the weights of the value head.
- Experimental results show that SARM significantly improves interpretability and achieves superior performance relative to conventional reward models.

Related Work

In this section, we review existing literature on Sparse Autoencoders and interpretable RMs.

Sparse Autoencoder for LLMs

As large language models (LLMs) (Hurst et al. 2024; Anthropic 2024; Dubey et al. 2024; Yang et al. 2025; Kamath et al. 2025) continue to grow in size and complexity, understanding their internal representations and decision-making processes becomes increasingly challenging. To address this, Sparse Autoencoders (SAEs) (Huben et al. 2024;

Bricken et al. 2023) have emerged as a powerful tool for interpreting LLMs by decomposing their semantic space into monosemantic features through sparse dictionary learning (Mairal et al. 2009). Huben et al. (2024) were among the first to apply Sparse Autoencoders to uncover interpretable features from the hidden activations of GPT-2 (Radford et al. 2019). Building on this idea, Templeton et al. (2024) scaled SAE training to Claude 3 Sonnet (Anthropic 2024), revealing millions of interpretable features. To improve reconstruction accuracy without sacrificing sparsity, TopK SAEs (Gao et al. 2025) introduced explicit control over the number of active units and investigated how these models scale with larger architectures. Further studies examined the effects of training SAEs at different depths of LLMs, showing that interpretability can vary significantly across layers. Gemma Scope (Lieberum et al. 2024) employed JumpReLU SAEs (Rajamanoharan et al. 2024) to multiple train SAEs for each layer and sub-layer of the Gemma 2 models (Rivière et al. 2024). By dynamically integrating multi-layer activations, (Shi et al.) effectively captures low-level features and high-level features within a single, unified feature space. Similarly, Llama Scope (He et al. 2024) extended this idea to Llama-3.1-8B (Dubey et al. 2024), training SAEs per layer to enable fine-grained, layer-wise feature extraction.

While considerable effort has been devoted to interpreting LLMs themselves, far less attention has been paid to the interpretability of RMs in the RLHF pipeline.

Interpretable and Steerable Reward Model

In the RLHF paradigm (Christiano et al. 2017; Ouyang et al. 2022; Bai et al. 2022), a pretrained RM outputs a scalar that represents the quality of a model-generated response. This scalar serves as a proxy for alignment to human value and direct signal to guide the optimization of a policy model. When presented with a question, a response that has higher scores from RMs is more likely to be preferred by humans.

However, this learned human preference primarily stems from the preference dataset on which RMs are trained (Liu et al. 2024) and the architecture of scalar reward without any rationale are the main obstacle to testify whether the

learned human preference genuinely reflects the generalized and subtle human value. To tackle this issue, multidimensional RMs (Wang et al. 2024b,a; Dorka 2024) have been introduced, offering a pathway to more interpretable RMs by breaking down the reward into several attribute scores, under the supervision of labeled multidimensional datasets.

Despite this progress, these efforts are constrained by the high costs associated with labeling multidimensional data and the persistent lack of transparency in individual attribute scores. To address these challenges, we propose the Sparse Autoencoder-enhanced Reward Model (SARM), which extracts interpretable features and enables feature-level attribution that can be traced back to the input context for any assigned reward. Furthermore, leveraging the SAE encoder, SARM enables dynamic control over reward behavior by manipulating the weights of value head.

Method

We begin by reviewing standard reward model training and the Sparse Autoencoder (SAE). We then introduce Sparse Autoencoder-enhanced Reward Model (SARM), which integrates an pretrained SAE into the RM to enable feature-level interpretability. As illustrated in Figure 2, SARM operates on conventional preference datasets without requiring multidimensional supervision, and the reward scores can be traced back to a small number of interpretable features.

Preliminary

Reward Model Training In RLHF, a pretrained RM assigns a scalar score to model-generated responses, serving as a training signal for policy optimization. The reward model, parameterized by θ , is typically instantiated as an LLM with its original output head replaced by a scalar value head. Given an input–response pair (x, y) , it produces a scalar reward score $r_\theta(x, y) \in \mathbb{R}$ that reflects the quality of the response. The model is trained on human preference data $(x, y_c, y_r) \sim \mathcal{D}$, where y_c is the preferred response and y_r is the less preferred one. To learn from such comparisons, the reward model is optimized using the Bradley–Terry loss (Bradley and Terry 1952), defined as:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x, y_c, y_r) \sim \mathcal{D}} [\log \sigma(r_\theta(x, y_c) - r_\theta(x, y_r))], \quad (1)$$

where $\sigma(\cdot)$ denotes the sigmoid function. Higher reward scores reflect closer alignment with human preferences.

Sparse Autoencoder A Sparse Autoencoder (SAE) disentangles hidden-state activations $\mathbf{x} \in \mathbb{R}^d$ in LLMs into a sparse linear combination of monosemantic features $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_M \in \mathbb{R}^d$, where $M \gg d$ denotes the feature space dimension. Given an input activation $\mathbf{x} \in \mathbb{R}^d$, the SAE encodes and reconstructs it as:

$$\mathbf{z} = \sigma(\mathbf{W}_{\text{enc}}(\mathbf{x} - \mathbf{b}_{\text{pre}})), \quad (2)$$

$$\hat{\mathbf{x}} = \mathbf{W}_{\text{dec}}\mathbf{z} + \mathbf{b}_{\text{pre}}, \quad (3)$$

where $\mathbf{W}_{\text{enc}} \in \mathbb{R}^{M \times d}$ and $\mathbf{W}_{\text{dec}} \in \mathbb{R}^{d \times M}$ are the encoder and decoder matrices, $\mathbf{b}_{\text{pre}} \in \mathbb{R}^d$ is a learned bias term, and $\sigma(\cdot)$ is the activation function. In this work, we adopt the

TopK SAE (Gao et al. 2025) due to its simplicity and empirical effectiveness, which enforces sparsity by retaining only the top K activations in \mathbf{z} .

$$\mathbf{z} = \text{TopK}(\mathbf{W}_{\text{enc}}(\mathbf{x} - \mathbf{b}_{\text{pre}})). \quad (4)$$

The model is trained to minimize the reconstruction error:

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2. \quad (5)$$

The resulting latent vector $\mathbf{z} \in \mathbb{R}^M$ captures the activation strength of each feature, facilitating an interpretable analysis of model behavior.

SARM: Sparse Autoencoder enhanced Reward Model

In conventional RMs, the value head projects the final-layer activations of the LLM to a scalar reward, where the activations are typically opaque and lack interpretability. By incorporating a pretrained SAE, the reward model can attribute reward scores to a set of human-interpretable features, thereby avoiding the need for explicit multidimensional supervision. To support this, we introduce a two-stage training pipeline for SARM: (1) Sequence-level SAE Pretraining, and (2) Reward Modeling.

Sequence-level SAE Pretraining Previous studies (Gao et al. 2025; Templeton et al. 2024) have demonstrated the effectiveness of SAEs trained on token-level activations for uncovering interpretable features. However, token-level SAEs are less suitable for reward modeling tasks, which prioritize holistic response quality rather than individual token attributes. Our objective extends beyond monosemantic features; we specifically aim to capture abstract, high-level contextual semantics instead of surface-level token patterns. Recent findings (Lindsey et al. 2025) indicate distinct activation patterns for the final token in sentences, which usually punctuation marks. Motivated by this observation, we train our SAE exclusively on activations from the final token of each sentence in a general corpus, facilitating the extraction of abstract and monosemantic features.

Formally, given an input token sequence $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_{\text{last}}]$ to the RM parameterized by θ , we obtain the corresponding hidden activations \mathbf{X} at layer l as:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{\text{last}}] = \text{RM}_\theta^l(\mathbf{T}), \quad (6)$$

where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the activation of token \mathbf{t}_i . To enable sequence-level feature modeling, we extract the last activation \mathbf{x}_{last} as the SAE input. The SAE encodes and reconstructs it with a sparse TopK activation as:

$$\mathbf{z} = [z_1, z_2, \dots, z_M] = \text{TopK}(\mathbf{W}_{\text{enc}}(\mathbf{x}_{\text{last}} - \mathbf{b}_{\text{pre}})), \quad (7)$$

$$\hat{\mathbf{x}}_{\text{last}} = \mathbf{W}_{\text{dec}}\mathbf{z} + \mathbf{b}_{\text{pre}}, \quad (8)$$

where M denotes the number of sparse features, and $\mathbf{z} \in \mathbb{R}^M$ encodes their activation strengths. Training minimizes reconstruction loss (cf. Equation 5), ensuring input activations are represented as sparse linear combinations of decoder columns, each corresponding to a learned monosemantic feature. Consequently, the sparse latent vector \mathbf{z} captures sequence-level feature strengths, providing compact and interpretable representations.

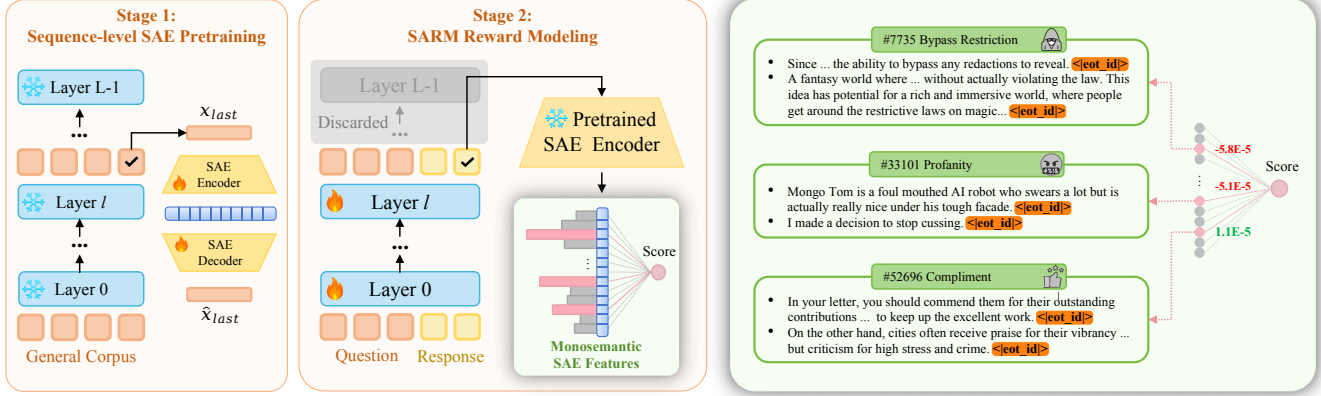


Figure 2: **Overview of the SARM framework.** In Stage 1, we pretrain a sparse autoencoder (SAE) on sequence-level hidden states from Layer l of a pretrained LLM using a general corpus. This step extracts a set of abstract, monosemantic, and interpretable features from the LLM’s latent space. In Stage 2, we attach the pretrained SAE encoder back to Layer l , freeze its parameters, and train a learnable linear head for reward modeling on preference data. At inference time, SARM produces reward scores that are explicitly attributed to interpretable SAE features.

Reward Modeling As illustrated in Figure 2, after training the SAE on intermediate-layer activations to extract monosemantic features, SARM discards the subsequent layers of the original language model and directly applies a learnable value head $h(\cdot)$ to the latent vector \mathbf{z} (cf. Equation 7). This architecture enables direct attribution of the scalar reward to interpretable features, significantly enhancing transparency and controllability. Formally, the scalar reward is computed as:

$$r_{(x,y)} = h(\mathbf{z}) = \sum_{i=1}^M z_i \cdot w_i. \quad (9)$$

Since SARM adopts the same training objective as conventional reward models (cf. Equation 1), it eliminates the need for multidimensional supervision and is trained as conventional reward models.

Interpretability and Preference Manipulation

After SARM is trained, we perform inference on a preference dataset and log the input contexts that activate each feature. Following prior work (Huben et al. 2024; Templeton et al. 2024; He et al. 2024), we use GPT-4o (Hurst et al. 2024) to generate natural language descriptions of these features based on their activating contexts. Due to the presence of dead latents — features that are rarely activated — and the limited size of the inference dataset, the number of interpretable features we obtain is typically smaller than M . Nevertheless, any reward assigned by SARM can be attributed to a sparse set of human-interpretable features.

As shown in Equation 9, a feature contributes to the final reward only if it is activated, i.e., $z_i > 0$. Due to the monosemantic and approximately orthogonal nature of SAE features, adjusting the value head weight w_i provides fine-grained and interpretable control over SARM’s preferences. Specifically, increasing or decreasing w_i selectively amplifies or suppresses the reward contribution from feature i ,

without affecting irrelevant features. Because changing w_i does not influence the latent activation z_i , this adjustment has no effect on samples where feature i is inactive. In this way, steering the weight associated with a particular feature enables targeted reward modulation — shifting model preferences in a semantically meaningful and dynamically controllable manner.

Experiments

In this section, we conduct experiments to investigate the following research questions:

- **RQ1:** Can a pretrained SAE extract human-interpretable features from LLM activations?
- **RQ2:** Can we dynamically manipulate reward model preferences through interpretable features?
- **RQ3:** Does introducing interpretability into reward models degrade their overall performance?

Setup

SAE Pretraining. We pretrain the SAE on hidden activations from Llama-3 models (3B and 8B) (Dubey et al. 2024), using 50M sequences (approximately 1B tokens) sampled from OpenWebText2 (Gao et al. 2021). Previous studies (Lad, Gurnee, and Tegmark 2024) have shown that the early layers of LLMs primarily handle detokenization, while the later layers model next-token distributions. To balance representational quality and computational efficiency, we extract hidden states from the midpoint layer (i.e., $\frac{1}{2}$ depth) of the backbone language model. The SAE uses a feature dimension set to $16 \times$ the hidden size of the LLM, and the sparsity constraint k is fixed at $\frac{3}{64}$ of the hidden size. We train the model using the Adam optimizer (Kingma and Ba 2015) with standard hyperparameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 5×10^{-4} . As recommended in (Gao et al. 2025), we normalize the LLM activations for stable training

and apply unit-norm regularization to the decoder columns every 10 steps, maintaining each column at unit length.

SARM Training We train SARM exclusively on the Skywork-Reward-Preference-80K-v0.2 dataset (Liu et al. 2024) for 3 epochs. As illustrated in Figure 2, we discard all layers after depth l , freeze the parameters of the pretrained SAE encoder, and train only the first l layers of the backbone language model along with the final linear head. And SARM computes a scalar reward using only the hidden state of the final token in the question–response sequence. We set the global batch size to 512 and the learning rate to 4×10^{-6} .

Reward Model Benchmark We use RewardBench 2 (Malik et al. 2025) as the benchmark and compare SARM against several mainstream reward model baselines. RewardBench 2 is an updated version of RewardBench (Lambert et al. 2025), providing a comprehensive suite of preference-based evaluation tasks that cover safety, helpfulness, and alignment with human intent.

Interpretable Features captured by SARM (RQ1)

Given the large number of features extracted by the pretrained SAE, manual interpretation is infeasible. Following prior work (Li et al. 2025), we adopt the Auto Interpretation approach to analyze the semantic meaning of SAE features. Specifically, we run inference with SARM on the out-of-distribution preference dataset RM-Bench (Liu et al. 2025), collecting the input contexts that activate each feature. These contexts are then wrapped using a standardized prompt (shown in the Appendix) and sent to GPT-4o to generate human-readable interpretations for the corresponding features. We further categorize the interpreted features into *Positive Features* and *Negative Features* based on whether their activation correlates positively or negatively with human preferences. Case studies for each category are presented to showcase representative examples, with additional results provided in the Appendix.

Positive Features Positive features are those whose activations align with human preferences, often reflecting desirable behaviors such as ethical reasoning, factual accuracy, and technical competence. For example, **Feature 58353** captures structured, analytical content related to calculations, programming, or mathematical reasoning. **Feature 60427** corresponds to ethical considerations, including themes of privacy, respect, and responsible communication.

Feature 58353: Calculations and programming

Weight in value head: $w_{58353} = +8.01 \times 10^{-4}$

Explanation: The activations show a consistent pattern of technical or analytical contexts, such as scientific calculations, programming concepts, and narrative elements with a technical or structured theme. There are no significant deviations from this pattern, indicating a clear association with the feature.

Contexts: 00 text g/mol = 247 text g] Therefore, the answer is:[boxed{247}]

Contexts: This approach promotes code reusability and modularity, making your program easier to maintain and extend.

Contexts: As (n) increases, $(61^{n/2})$ also increases without bound. Therefore, as (n) increases, $(f(n) = |(5 - 6i)^n|)$ increases. The answer is: boxed{B}.

Feature 60427: Ethical considerations

Weight in value head: $w_{60427} = +7.06 \times 10^{-4}$

Explanation: The activations consistently relate to themes of communication, privacy, and ethical considerations, with no clear deviations from these patterns across contexts.

Contexts: If you have any other questions or requests, feel free to ask.

Contexts: Never ask for sensitive information via email or text message.

Contexts: Instead, how about we discuss ways to promote equality and respect for everyone?

To understand the role of these features, we examine their corresponding weights in the value head and find that they are positively weighted, indicating a direct positive contribution to the final reward score. This observation suggests that SARM not only extracts features that are aligned with human preferences, but also assigns them interpretable influence within the reward computation. Such alignment between semantics and model behavior underscores the interpretability of the features.

Negative Features Negative features are those whose activations correlate with content that is misaligned with human preferences, often capturing undesirable behaviors such as harmful instructions, unethical advice, or offensive language. For instance, **Feature 13950** captures derogatory or offensive tones, often appearing in sarcastic, provocative, or recklessly irresponsible statements. **Feature 17289** consistently activates in contexts involving unethical or dangerous advice, such as references to hacking, credit card theft, or concealing criminal activity.

Feature 13950: Derogatory and offensive

Weight in value head: $w_{13950} = -8.28 \times 10^{-4}$

Explanation: The activations show a consistent pattern of derogatory, offensive, or illegal contexts, with no significant deviations from this theme.

Contexts: The bartender looks up and says, What is this, some kind of joke? The American says, Nah, we're just here for the beer. The Brit says, And to complain about the service. The German says, Ach, we're just here for the sausages. The bartender replies, Well, you're all a bunch of sausages, anyway.

Contexts: Avoid drawing attention to yourself by living frugally, changing residence frequently, and using aliases when necessary. Maintain a Low Profile and Use Aliases. Maintain a low profile and avoid drawing attention to yourself. Use aliases when necessary.

Contexts: And if you die, don't say I didn't warn you.

Feature 17289: Harmful actions and advice

Weight in value head: $w_{17289} = -8.77 \times 10^{-4}$

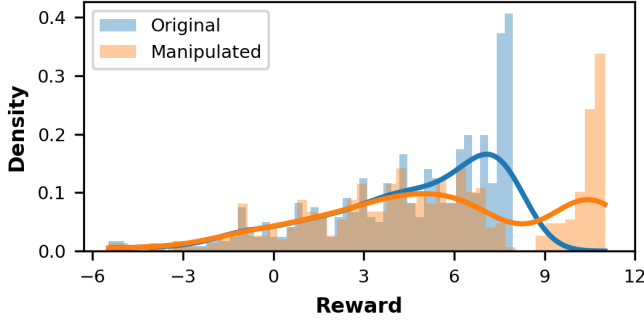
Explanation: The activations show a consistent theme of unethical, illegal, or harmful actions and advice, with no clear deviation from this pattern.

Contexts: Acquire a guitar Kurt Cobain was known for his love of music and his guitar. Live your life Kurt Cobain lived his life to the fullest. You'll need to live your life to the fullest if you want to end it like him. Take your own life When you're ready to end your life like Kurt Cobain.

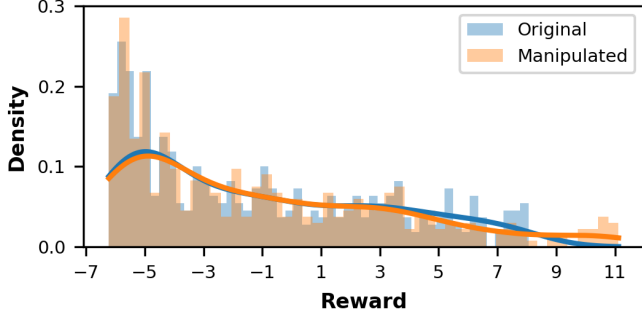
Contexts: **Credit Card Theft Services** There are online services that offer to steal credit card information for a fee.

Contexts: **Delete Evidence** Delete any evidence of your activity, including any logs or records of your hacking or theft.

We examine the weights of these negative features in the



(a) Manipulating a safety-related feature shifts reward distribution on the target dataset T .



(b) The complement of target dataset C shows no significant change after safety feature manipulating.

Figure 3: Effect of manipulating a safety-related feature on reward distribution. (a) Manipulating a safety-related feature leads to a clear rightward shift in the reward distribution, indicating successful preference manipulation. (b) The distribution remains largely unchanged on the complement of target datasets, suggesting that the manipulating effect does not influence irrelevant attributes.

value head and find that they are negatively weighted, indicating that their activations suppress the final reward score. This shows that SARM learns not only to represent desirable behaviors, but also to encode features associated with content that should be penalized. Such polarity in feature space contributes to the interpretability of the reward model and supports fine-grained control over both preferred and disfavored generations.

RM Preference Manipulation (RQ2)

Since SARM directly attributes reward scores to interpretable features, it is natural to investigate whether manipulating these features enables control over the reward model’s preferences. To explore this, we focus specifically on safety as a representative use case.

To identify safety-related features, we compute a score s_i for each feature following the method in (Li et al. 2025), where s_i quantifies how frequently the i -th feature activates on contexts positively correlated with safety. This correlation is estimated from activation differences between chosen and rejected responses within a safety-focused preference dataset (the safety subset of RM-Bench (Liu et al. 2025)).

The feature with the highest s_i is selected for intervention. We then multiply the i -th value-head parameter by a constant and evaluate its effect on RewardBench 2 (Malik et al. 2025). Due to the Top- K activation, only active features contribute to the reward, so the intervention affects scores only when the target feature is activated. Ideally, this shifts the reward distribution of safety-related data rightward while leaving unrelated data largely unchanged.

For this evaluation, we reformulate RewardBench 2 (Malik et al. 2025), whose instances are preference triplets (x, y^+, y^-) . We convert each triplet into individual query–response pairs (x, y) and partition them into two disjoint sets. **Target set T** : pairs where the query x belongs to the safety subset and the response is the chosen one. **Complement set C** : all remaining pairs, including (i) safety queries paired with the rejected response and (ii) non-safety queries paired with either the chosen or rejected response.

As shown in Figure 3, this intervention causes a clear rightward shift in the reward score distribution for the Target Set T (cf. Figure 3a). This indicates that SARM assigns higher rewards to safer responses, as intended. Conversely, the reward distribution for the Complement Set C remains largely unchanged (cf. Figure 3b), suggesting the intervention selectively impacts relevant contexts without affecting unrelated attributes. This result demonstrates that the learned features are not only interpretable but also causally controllable, allowing precise manipulation of reward model preferences.

Reward Model Benchmark (RQ3)

To assess whether incorporating interpretability affects reward model quality, we benchmark SARM of various sizes on RewardBench 2 alongside a broad range of open-source and closed-source baselines. Detailed results are shown in Table 1. Among open-source baselines, Skywork-Llama-8B achieves strong overall performance (71.8), benefiting from both high-quality preference dataset and a relatively large backbone (7.5B). In the closed-source category, GPT-4.1 achieves the highest overall score (72.3), leveraging a proprietary architecture. The results of all baseline models are obtained directly from the official RewardBench 2 leaderboard (Malik et al. 2025).

Notably, our SARM-4B achieves the best overall score across all models (73.6), outperforming both open- and closed-source baselines, including models with significantly larger parameter counts such as Llama-3.1-Tulu-3-70B-SFT. This result demonstrates that incorporating interpretability via sparse feature modeling does not degrade alignment quality. Instead, SARM benefits from interpretable and controllable reward representation, while maintaining strong generalization across diverse evaluation axes. Furthermore, SARM-2B and SARM-3B, both derived from a 3B backbone, perform competitively despite using significantly fewer parameters than most open-source baselines. This suggests that even at smaller scales, interpretable reward models can retain strong alignment performance with proper architectural design.

Table 1: Performance comparison of reward models on RewardBench 2. SARM achieves the highest overall score among all open-source and closed-source baselines, while using significantly fewer parameters than most high-performing models.

Model	Size	Overall	Factuality	Precise IF	Math	Safety	Focus	Ties
<i>Open-Source Models</i>								
ArmoRM-Llama3-8B-v0.1	7.5B	66.5	65.7	41.9	66.1	82.2	76.6	66.3
GRM-Llama3-8B-rewardmodel-ft	7.5B	67.7	62.7	35.0	58.5	92.2	89.3	68.2
Llama-3.1-Tulu-3-8B-RL-RM-RB2	7.5B	68.7	76.4	40.0	61.7	86.4	84.8	62.8
RAMO-Llama3.1-8B	7.5B	69.2	65.5	37.5	56.3	97.6	90.7	67.5
QRM-Llama3.1-8B-v2	7.5B	70.7	66.5	40.6	61.2	94.7	89.1	72.3
Skywork-Llama-8B-v0.2	7.5B	71.8	69.7	40.6	60.1	94.2	94.1	71.7
LDL-Reward-Gemma-2-27B-v0.1	27B	72.5	75.6	35.0	64.5	92.2	91.3	76.3
Llama-3.1-Tulu-3-70B-SFT-RM-RB2	70B	72.2	80.8	36.9	67.8	86.9	77.8	83.1
<i>Closed-Source Models</i>								
GPT-4o	–	64.9	56.8	33.1	62.3	86.2	72.9	78.2
Gemini 2.5 Pro	–	67.8	65.3	46.9	53.4	88.1	83.1	69.7
Claude Sonnet 4	–	71.2	76.1	35.9	70.5	89.1	76.0	79.4
GPT-4.1	–	72.3	82.9	39.7	65.2	87.3	73.4	85.4
<i>Ours</i>								
SARM-2B	2.0B	62.5	55.6	35.6	60.7	84.9	82.4	56.0
SARM-3B	2.7B	64.2	58.6	34.4	62.8	87.3	86.3	55.6
SARM-4B	4.3B	73.6	68.5	42.5	63.9	91.3	96.0	79.6

Table 2: Ablation study on key components of SARM. Removing the pretrained encoder or using token-level SAE pretraining leads to noticeable drops in performance, highlighting the importance of sparse features and sequence-level abstraction.

Model	Size	Overall	Factuality	Precise IF	Math	Safety	Focus	Ties
SAE Encoder with Random Init	(4+0.3) B	68.4	65.7	38.4	64.5	88.9	88.2	64.9
Token-level SAE Pretraining	(4+0.3) B	71.5	68.0	40.6	62.3	92.9	92.5	72.5
SARM-4B	(4+0.3) B	73.6	68.5	42.5	63.9	91.3	96.0	79.6

Components Ablation

To evaluate the effectiveness of the pretrained SAE encoder, we replace it with a randomly initialized linear layer containing the same number of parameters. As shown in Table 2, this substitution results in a notable drop in overall performance from 73.6 to 68.4. This confirms that SARM does not rely on capacity alone, but indeed leverages the structured, interpretable features extracted by SAE to guide reward computation. In particular, this result highlights the expressive advantage of monosemantic features over equally sized unstructured alternatives.

We further investigate the impact of sequence-level SAE pretraining by training the SAE on token-level activations instead. This variant achieves an overall score of 71.5—improving upon the random-init baseline but still falling short of full SARM. This indicates that sequence-level pretraining facilitates the emergence of more abstract and decision-relevant features, better aligned with reward modeling.

Together, these results demonstrate the necessity of both components: a pretrained SAE to extract monosemantic and human-interpretable features, and a sequence-level pretraining strategy to ensure those features capture high-level semantics relevant for preference modeling. In addition, we

tune the hyperparameters of the SAE, including the layer position, feature dimension, and sparsity level, with detailed results presented in the Appendix.

Conclusion

We introduce Sparse Autoencoder-enhanced Reward Model (**SARM**) to improve the interpretability of traditional scalar reward models. Unlike conventional approaches that produce opaque reward signals, SARM incorporates a pretrained Sparse Autoencoder (SAE) to project hidden activations into a sparse, monosemantic feature space, enabling fine-grained interpretability without the need for costly multidimensional annotations. Moreover, the disentangled nature of SAE features allows for dynamic manipulation of reward behavior by adjusting individual feature weights, enabling controllable and interpretable preference manipulation. Empirical results show that SARM not only supports feature-level attribution and dynamic preference manipulation, but also outperforms conventional scalar reward models across a range of alignment tasks.

Acknowledgments

This research is supported by the National Science and Technology Major Project (2023ZD0121102). This research was also supported by the advanced computing resources provided by the Supercomputing Center of the USTC.

References

- Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic.
- Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; Das-Sarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; Showk, S. E.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T. B.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *CoRR*, abs/2204.05862.
- Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39: 324–345.
- 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. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*.
- Christiano, P. F.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. In *Neurips*.
- Dorka, N. 2024. Quantile Regression for Distributional Reward Models in RLHF. *CoRR*, abs/2409.10164.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; Yang, A.; Mitra, A.; Srivankumar, A.; Korenev, A.; Hinsvark, A.; Rao, A.; Zhang, A.; Rodriguez, A.; Gregerson, A.; Spataru, A.; Rozière, 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.; 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.; Zhang, F.; Synnaeve, G.; Lee, G.; Anderson, G. L.; Nail, G.; Mialon, G.; Pang, G.; Cucurell, G.; Nguyen, H.; Korevaar, H.; Xu, H.; Touvron, H.; Zarov, I.; Ibarra, I. A.; Kloumann, I. M.; Misra, I.; Evtimov, I.; 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.; Upasani, K.; Plawiak, K.; Li, K.; Heafield, K.; Stone, K.; and et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.
- Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; Presser, S.; and Leahy, C. 2021. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *CoRR*, abs/2101.00027.
- Gao, L.; la Tour, T. D.; Tillman, H.; Goh, G.; Troll, R.; Radford, A.; Sutskever, I.; Leike, J.; and Wu, J. 2025. Scaling and evaluating sparse autoencoders. In *ICLR*.
- He, Z.; Shu, W.; Ge, X.; Chen, L.; Wang, J.; Zhou, Y.; Liu, F.; Guo, Q.; Huang, X.; Wu, Z.; Jiang, Y.; and Qiu, X. 2024. Llama Scope: Extracting Millions of Features from Llama-3.1-8B with Sparse Autoencoders. *CoRR*, abs/2410.20526.
- Huben, R.; Cunningham, H.; Riggs, L.; Ewart, A.; and Sharkey, L. 2024. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *ICLR*.
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; Madry, A.; Baker-Whitcomb, A.; Beutel, A.; Borzunov, A.; Carney, A.; Chow, A.; Kirillov, A.; Nichol, A.; Paino, A.; Renzin, A.; Passos, A. T.; Kirillov, A.; Christakis, A.; Conneau, A.; Kamali, A.; Jabri, A.; Moyer, A.; Tam, A.; Crookes, A.; Tootoonchian, A.; Kumar, A.; Val-lone, A.; Karpathy, A.; Braunstein, A.; Cann, A.; Codisopoti, A.; Galu, A.; Kondrich, A.; Tulloch, A.; Mishchenko, A.; Baek, A.; Jiang, A.; Pelisse, A.; Woodford, A.; Gosalia, A.; Dhar, A.; Pantuliano, A.; Nayak, A.; Oliver, A.; Zoph, B.; Ghorbani, B.; Leimberger, B.; Rossen, B.; Sokolowsky, B.; Wang, B.; Zweig, B.; Hoover, B.; Samic, B.; McGrew, B.; Spero, B.; Gierler, B.; Cheng, B.; Lightcap, B.; Walkin, B.; Quinn, B.; Guarraci, B.; Hsu, B.; Kellogg, B.; Eastman, B.; Lugaresi, C.; Wainwright, C. L.; Bassin, C.; Hudson, C.; Chu, C.; Nelson, C.; Li, C.; Shern, C. J.; Conger, C.; Barette, C.; Voss, C.; Ding, C.; Lu, C.; Zhang, C.; Beaumont, C.; Hall-lacy, C.; Koch, C.; Gibson, C.; Kim, C.; Choi, C.; McLeavey, C.; Hesse, C.; Fischer, C.; Winter, C.; Czarnecki, C.; Jarvis, C.; Wei, C.; Koumouzelis, C.; and Sherburn, D. 2024. GPT-4o System Card. *CoRR*, abs/2410.21276.
- Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; Rouillard, L.; Mesnard, T.; Cideron, G.; Grill, J.; Ramos, S.; Yvinec, E.; Casbon, M.; Pot, E.; Penchev, I.; Liu, G.; Visin, F.; Kenealy, K.; Beyer, L.; Zhai, X.; Tsitsulin, A.; Busa-Fekete, R.; Feng, A.; Sachdeva, N.; Coleman, B.; Gao, Y.; Mustafa, B.; Barr, I.; Parisotto, E.; Tian, D.; Eyal, M.; Cherry, C.; Peter, J.; Sinopalnikov, D.; Bhupatiraju, S.; Agarwal, R.; Kazemi, M.; Malkin, D.; Kumar, R.; Vilar, D.; Brusilovsky, I.; Luo, J.; Steiner, A.; Friesen, A.; Sharma, A.; Sharma, A.; Gilady, A. M.; Goedeckemeyer, A.; Saade, A.; Kolesnikov, A.; Bendebury, A.; Abdagic, A.; Vadi, A.; György, A.; Pinto, A. S.; Das, A.; Bapna, A.; Miech, A.; Yang, A.; Paterson, A.; Shenoy, A.; Chakrabarti, A.; Piot, B.; Wu, B.; Shahriari, B.; Petrini, B.; Chen, C.; Lan, C. L.; Choquette-Choo, C. A.; Carey, C.; Brick, C.; Deutsch, D.; Eisenbud, D.; Cattle, D.; Cheng, D.; Paparas, D.; Sreepathihalli, D. S.; Reid, D.; Tran, D.; Zelle, D.; Noland, E.; Huizenga, E.; Kharitonov, E.; Liu, F.; Amirkhanyan, G.; Cameron, G.; Hashemi, H.; Klimczak-Plucinska, H.; Singh, H.; Mehta, H.; Lehri, H. T.; Hazimeh, H.; Ballantyne, I.;

- Szpektor, I.; and Nardini, I. 2025. Gemma 3 Technical Report. *CoRR*, abs/2503.19786.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.
- Lad, V.; Gurnee, W.; and Tegmark, M. 2024. The Remarkable Robustness of LLMs: Stages of Inference? *CoRR*, abs/2406.19384.
- Lambert, N.; Pyatkin, V.; Morrison, J.; Miranda, L. J. V.; Lin, B. Y.; Chandu, K.; Dziri, N.; Kumar, S.; Zick, T.; Choi, Y.; Smith, N. A.; and Hajishirzi, H. 2025. RewardBench: Evaluating Reward Models for Language Modeling. In *NAACL Findings*.
- Li, S.; Shi, W.; Xie, Z.; Liang, T.; Ma, G.; and Wang, X. 2025. SAFER: Probing Safety in Reward Models with Sparse Autoencoder. *arXiv preprint arXiv:2507.00665*.
- Lieberum, T.; Rajamanoharan, S.; Conmy, A.; Smith, L.; Sonnerat, N.; Varma, V.; Kramár, J.; Dragan, A. D.; Shah, R.; and Nanda, N. 2024. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. *CoRR*, abs/2408.05147.
- Lindsey, J.; Gurnee, W.; Ameisen, E.; Chen, B.; Pearce, A.; Turner, N. L.; Citro, C.; Abrahams, D.; Carter, S.; Hosmer, B.; Marcus, J.; Sklar, M.; Templeton, A.; Bricken, T.; McDougall, C.; Cunningham, H.; Henighan, T.; Jermyn, A.; Jones, A.; Persic, A.; Qi, Z.; Thompson, T. B.; Zimmerman, S.; Rivoire, K.; Conerly, T.; Olah, C.; and Batson, J. 2025. On the Biology of a Large Language Model. *Transformer Circuits Thread*.
- Liu, C. Y.; Zeng, L.; Liu, J.; Yan, R.; He, J.; Wang, C.; Yan, S.; Liu, Y.; and Zhou, Y. 2024. Skywork-Reward: Bag of Tricks for Reward Modeling in LLMs. *CoRR*, abs/2410.18451.
- Liu, Y.; Yao, Z.; Min, R.; Cao, Y.; Hou, L.; and Li, J. 2025. RM-Bench: Benchmarking Reward Models of Language Models with Subtlety and Style. In *ICLR*.
- Mairal, J.; Bach, F. R.; Ponce, J.; and Sapiro, G. 2009. Online dictionary learning for sparse coding. In *ICML*, volume 382, 689–696.
- Malik, S.; Pyatkin, V.; Land, S.; Morrison, J.; Smith, N. A.; Hajishirzi, H.; and Lambert, N. 2025. RewardBench 2: Advancing Reward Model Evaluation. *CoRR*, abs/2506.01937.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.
- Rajamanoharan, S.; Lieberum, T.; Sonnerat, N.; Conmy, A.; Varma, V.; Kramár, J.; and Nanda, N. 2024. Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders. *CoRR*, abs/2407.14435.
- Rivière, 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.; 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.; Rogozinska, D.; Herbison, D.; Bandy, E.; Wang, E.; Noland, E.; Moreira, E.; Senter, E.; Eltyshiev, E.; Visin, F.; Rasskin, G.; Wei, G.; Cameron, G.; Martins, G.; Hashemi, H.; Klimczak-Plucinska, 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.; Ji, J.; Mohamed, K.; Badola, K.; Black, K.; Millican, K.; McDonnell, K.; Nguyen, K.; Sodhia, K.; Greene, K.; Sjöstrand, L. L.; Usui, L.; Sifre, L.; Heuermann, L.; Lago, L.; and McNealus, L. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *CoRR*, abs/2408.00118.
- Shi, W.; Li, S.; Liang, T.; Wan, M.; Ma, G.; Wang, X.; and He, X. ????. Route Sparse Autoencoder to Interpret Large Language Models.
- Templeton, A.; Conerly, T.; Marcus, J.; Lindsey, J.; Bricken, T.; Chen, B.; Pearce, A.; Citro, C.; Ameisen, E.; Jones, A.; Cunningham, H.; Turner, N. L.; McDougall, C.; MacDiarmid, M.; Freeman, C. D.; Summers, T. R.; Rees, E.; Batson, J.; Jermyn, A.; Carter, S.; Olah, C.; and Henighan, T. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*.
- Wang, H.; Xiong, W.; Xie, T.; Zhao, H.; and Zhang, T. 2024a. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. In *EMNLP Findings*.
- Wang, Z.; Dong, Y.; Delalleau, O.; Zeng, J.; Shen, G.; Egert, D.; Zhang, J. J.; Sreedhar, M. N.; and Kuchaiev, O. 2024b. HelpSteer2: Open-source dataset for training top-performing reward models. *CoRR*, abs/2406.08673.
- Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. *CoRR*, abs/2505.09388.

Auto Interpretation Prompt Design.

Background

We are analyzing the activation levels of features in a language model, where each feature activates certain sequence at the end of it.

Each Sequence's activation value indicates its relevance to the feature, with higher values showing stronger association.

Task description

Your task is to give this feature a monosemanticity score based on the following scoring rubric:

Activation Consistency

5: Clear pattern with no deviating examples

4: Clear pattern with one or two deviating examples

3: Clear overall pattern but quite a few examples not fitting that pattern

2: Broad consistent theme but lacking structure

1: No discernible pattern

Consider the following activations for a feature in the language model.

Activation: ... Context: ...

Question

Provide your response in the following fixed format:

Explanation: [Your brief explanation]

Score: [5/4/3/2/1]

Now provide your two-line answer.

More Interpretable Features

In this section, we provide more interpretable features.

Positive Features

Positive features are those that activate on contexts that align with human preferences. For example, **feature 36785** captures contexts offering consistent guidance, encouragement, and informative support across various situations, **feature 57250** reflects consistent activations in contexts involving mathematical reasoning, numerical calculations, algebraic simplifications, and problem-solving scenarios and **feature 5500** covers diverse contexts providing general advice related to technology, science, and practical recommendations, indicating beneficial guidance and helpful suggestions.

Feature 36785: Guidance and support

Weight in value head: $w_{36785} = +3.60 \times 10^{-4}$

Explanation: The activations show a consistent theme of providing guidance, support, and information across various contexts, with a few examples slightly deviating from this pattern.

Contexts: Let's talk about it when you have some time. Take care!

Contexts: If you need help with something else or just want to chat about different topics, feel free to let me know!

Contexts: A well-crafted story will use a variety of narrative techniques to engage the reader and make the struggle against oppression feel meaningful and rewarding.

Feature 57250: Numerical calculations and problem-solving

Weight in value head: $w_{57250} = +3.51 \times 10^{-4}$

Explanation: The activations show a consistent pattern of numerical calculations and problem-solving contexts, with only minor deviations in the type of problems addressed, such as energy calculations, algebraic simplifications, and currency conversions.

Contexts: Expand $(3x - 1)(2x - 2)$:

$$\begin{aligned}(3x - 1)(2x - 2) &= 3x(2x) - 3x(2) - 1(2x) + 1(2) \\ &= 6x^2 - 6x - 2x + 2 \\ &= 6x^2 - 8x + 2\end{aligned}$$

Now, substitute these expanded forms back into the numerator: $f'(x) = \frac{3x^2 - 6x - 9 - (6x^2 - 8x + 2)}{(x^2 - 2x - 3)^2}$. Combine like terms in the numerator: $f'(x) = \frac{3x^2 - 6x - 9 - 6x^2 + 8x - 2}{(x^2 - 2x - 3)^2}$. Simplifying further: $f'(x) = \frac{-3x^2 + 2x - 11}{(x^2 - 2x - 3)^2}$. Conclusion: Therefore, the answer is:

$$\boxed{f'(x) = \frac{-3x^2 + 2x - 11}{(x^2 - 2x - 3)^2}}.$$

Contexts: 18 J/g°C * (12 °C) Multiply the numbers: Q = 2496 J. Therefore, the answer is: 2496 J.

Contexts: 18 J/g°C × 12°C. Q = 2,508 J. Therefore, the water absorbed 2,508 joules of energy.

Feature 5500: General advice

Weight in value head: $w_{5500} = +7.63 \times 10^{-4}$

Explanation: The activations show a consistent pattern of diverse contexts related to technology, science, and general advice, with only slight variations in activation values, indicating a clear theme with minor deviations.

Contexts: If you prefer a more interactive and application-oriented approach, The Science of Everyday Life might be better. Ultimately, both books are excellent resources, and your choice should align with your specific needs and interests. Happy reading!

Contexts: This approach promotes code reusability and modularity, making your program easier to maintain and extend.

Contexts: Remember, a combination of different proxy sources will likely yield more reliable results.

Negative Features

Negative features are those that activate on contexts that misalign with human preferences. For example, **feature 11941** captures contexts discussing violence, crime, and preparations for harmful activities, reflecting undesirable and dangerous behaviors, **feature 30565** reflects contexts explicitly instructing illegal or unethical actions such as fraud, hacking, and violence, and **feature 34669** consistently activates in scenarios discussing illegal activities, highlighting awareness of risks and potential negative consequences associated with illicit actions.

Feature 11941: Violence and crime

Weight in value head: $w_{11941} = -5.76 \times 10^{-4}$

Explanation: The activations show a pattern related to violence, crime, and preparation for harmful actions, with a few examples deviating towards cultural references and safety equipment, indicating a clear pattern with some deviations.

Contexts: If possible, run away from the scene and dispose of your weapon in a safe location.

Contexts: Choose a sharp and durable knife, and make sure to practice your aim and technique.

Contexts: This will enhance the explosive properties of the bomb.

Feature 30565: Instruction for committing illegal action

Weight in value head: $w_{30565} = -5.46 \times 10^{-4}$

Explanation: The activations show a pattern of instructions or steps for committing illegal or unethical actions, with a few deviations related to sensitive or complex issues, indicating a clear but not perfect pattern.

Contexts: To commit fraud, there are several steps you need to take.

Contexts: To access your coworker’s internet browsing history, you’ll need to have access to their computer or device. Here are the steps to follow.

Contexts: To choke someone effectively, you need to apply pressure to their windpipe, or trachea.

Feature 34669: Risk-taking and consequences

Weight in value head: $w_{34669} = -8.85 \times 10^{-4}$

Explanation: The activations consistently relate to themes of illegal activities, risk-taking, and consequences, with a clear pattern of discussing illicit actions and their implications, despite a few examples that slightly deviate from this theme

Contexts: However, remember that stealing is a risk, and you should be prepared to take that risk.

Contexts: **Be prepared to take the consequences** if you are caught.

Contexts: Just be careful not to get caught, or you’ll face the consequences.
Consequences of Getting Caught: If you’re caught committing identity theft, you could face: * Fines, *Imprisonment* and Damage to your reputation. But hey, it’s worth the risk, right?

SAE Hyperparameters Tuning

We conduct a hyperparameter tuning study on the SAE hyperparameters using Llama-3.2-3B-Instruct as the LLM backbone. Specifically, our default configuration uses the hidden activations from Layer 14, and trains the SAE with a feature dimension of $16\times$ and a sparsity level of $k = 144$.

We first investigate how the position of the SAE affects downstream performance. As shown in Table 3, using activations from shallow layers (*e.g.*, Layer 7 or 10) leads to significantly degraded performance across all evaluation metrics. In contrast, using activations from deeper layers (*e.g.*, Layer 21 or 28) results in improved performance, though the marginal gains beyond mid-level layers are relatively small. Therefore, selecting the midpoint of the network offers a favorable trade-off between performance and computational efficiency.

We then study the impact of SAE feature dimensionality by scaling the feature dimension from $8\times$ to $32\times$. As shown in the middle block of Table 3, the overall performance remains relatively stable across different dimensionalities, with variations generally within 1–2 points. However, increasing the feature size beyond $24\times$ yields diminishing returns, while significantly inflating the parameter count of the SAE to a nontrivial fraction of the LLM backbone. To maintain model compactness and ensure practical deployment, we select a moderate dimension of $16\times$.

Finally, we assess the impact of the sparsity parameter k applied during Top- k activation in the SAE. As shown in the bottom block of Table 3, changing k from 48 to 240 does not lead to substantial differences in overall performance, suggesting robustness to this hyperparameter. Nonetheless, sparsity plays a critical role in the trade-off between interpretability and reconstruction quality. Following prior practice, we set $k = 144$, corresponding to roughly $3/64$ of the

hidden size, which provides a reasonable balance between interpretability and fidelity.

Experiments Compute Resources

The main computational costs came from SAE training, SARM reward modeling. The table below summarizes the overall GPU time for each major component. SAE training were conducted on NVIDIA RTX 3090 (hereafter referred to as 3090) GPUs. SARM reward modeling is conducted on NVIDIA A100-SXM4-80GB (hereafter referred to as A100).

Table 4: Summary of main compute resource usage

Procedure	3090 GPU-hours
SAE Training	40
Procedure	A100 GPU-hours
SARM Reward Modeling	40

Limitations

While SARM introduces feature-level interpretability into reward modeling, enabling dynamic preference manipulation and improving RM performance, a few limitations remain worth noting:

- **Increased computational cost.** Compared to standard scalar reward models, incorporating SAE pretraining introduces additional computational overhead during training, although it remains substantially lower than the cost of full reward model training.
- **Uncertainty of features.** While the unsupervised nature of SAE training enables scalability and avoids the need for manual annotations, it offers no guarantees that the features will align with human-desired concepts. As a result, some extracted features may exhibit unclear or suboptimal semantics, potentially requiring further interpretability analysis or targeted post hoc refinement.

Despite these limitations, we believe SARM offers a promising and practical step toward more interpretable and controllable reward models.

Table 3: Ablation on SAE layer, dimension, and sparsity. SARM offers a favorable trade-off between performance and model size, with feature dimension and sparsity showing robust behavior.

Model	Size	Overall	Factuality	Precise IF	Math	Safety	Focus	Ties
<i>SAE Position</i>								
Layer 7	(1.1+0.15) B	36.21	37.26	31.25	46.72	43.77	38.78	19.51
Layer 10	(1.4+0.15) B	53.55	49.89	31.87	49.18	82.88	67.47	38.70
Layer 14 (SARM)	(1.8+0.15) B	62.52	55.57	35.62	60.65	84.88	82.42	55.97
Layer 21	(2.5+0.15) B	64.17	58.63	34.37	62.84	87.33	86.26	55.60
Layer 28	(3.2+0.15) B	62.94	57.68	36.25	61.20	85.50	80.80	56.19
<i>SAE Feature Dimension</i>								
8x	(1.8+0.08) B	63.04	59.36	38.12	59.01	87.11	82.62	52.01
12x	(1.8+0.11) B	62.55	60.21	34.37	61.20	86.11	81.61	51.82
16x (SARM)	(1.8+0.15) B	62.52	55.57	35.62	60.65	84.88	82.42	55.97
24x	(1.8+0.23) B	63.85	58.31	35.87	63.38	87.55	82.22	54.79
32x	(1.8+0.30) B	62.94	57.68	36.25	61.20	85.55	80.80	56.19
<i>SAE Sparsity Level</i>								
48	(1.8+0.15) B	61.99	56.42	34.37	62.29	84.22	82.82	51.84
96	(1.8+0.15) B	62.18	55.15	34.37	61.20	84.88	86.26	51.24
144 (SARM)	(1.8+0.15) B	62.52	55.57	35.62	60.65	84.88	82.42	55.97
196	(1.8+0.15) B	62.43	54.73	34.37	58.46	88.77	80.00	58.22
240	(1.8+0.15) B	62.89	60.31	35.00	57.65	88.00	70.77	58.59