SCAR: Sparse Conditioned Autoencoders for Concept Detection and Steering in LLMs


Ruben Härle, Felix Friedrich, Manuel Brack,

Björn Deiseroth, Patrick Schramowski, Kristian Kersting

Motivation

1. Lack of methods to **detect behavior** and **steer** LLMs **simultaneously**

2. **Avoid**:
   - ⇨ Computation overhead
   - ⇨ Additional latency
   - ⇨ Bad at generalizing and too static

SCAR

We use a Sparse Autoencoder (SAE) with TopK+ReLU activation for monosemantic disentanglement:

$$\text{SAE}(\mathbf{x}) = D(\sigma(E(\mathbf{x}))) \quad with$$
$$E(\mathbf{x}) = \mathbf{W}_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}} = \mathbf{h} \quad and \quad D(\mathbf{f}) = \mathbf{W}_{\text{dec}}\mathbf{f} + \mathbf{b}_{\text{dec}} = \bar{\mathbf{x}} \quad and$$
$$\sigma(\mathbf{h}) = \text{ReLU}(\text{TopK}(\mathbf{h})) = \mathbf{f}.$$

Training

We add a condition loss next to the default reconstruction objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Reconstruct}} + \mathcal{L}_{\text{Condition}} = \frac{(\bar{\mathbf{x}} - \mathbf{x})^2}{\mathbf{x}^2} + \text{CE}(\text{Sigmoid}(h_0),\, y)$$

⇨ Classification enforces concept isolation

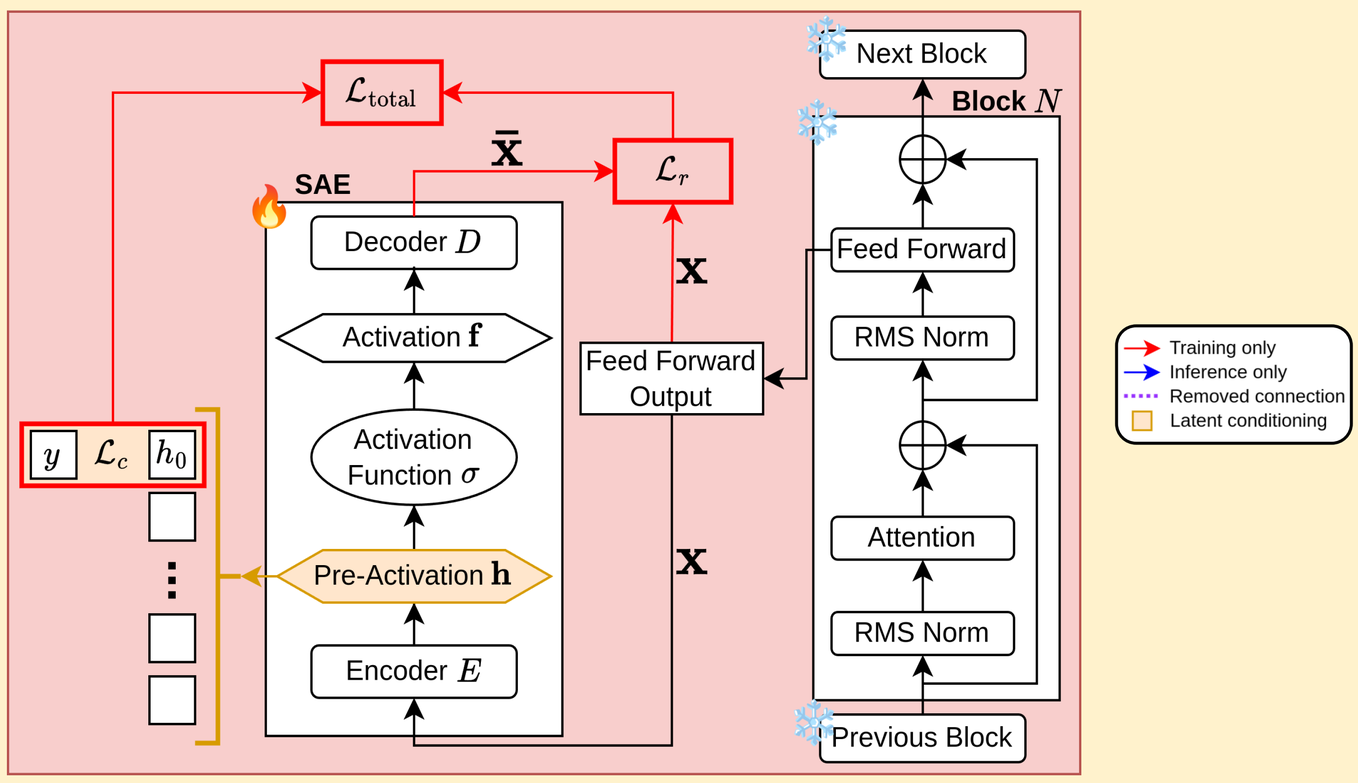
⇨ Supervision for desired concept during training

Inference

Pass Feed Forward output through SAE to residual connection:

1. Use latent feature to **detect** behavior AND / OR
2. Modify latent feature with factor α to **steer** LLM

$$f_i = \begin{cases} \alpha h_i & \text{if} \quad i = 0, \\ \sigma(h_i) & \text{else.} \end{cases}$$



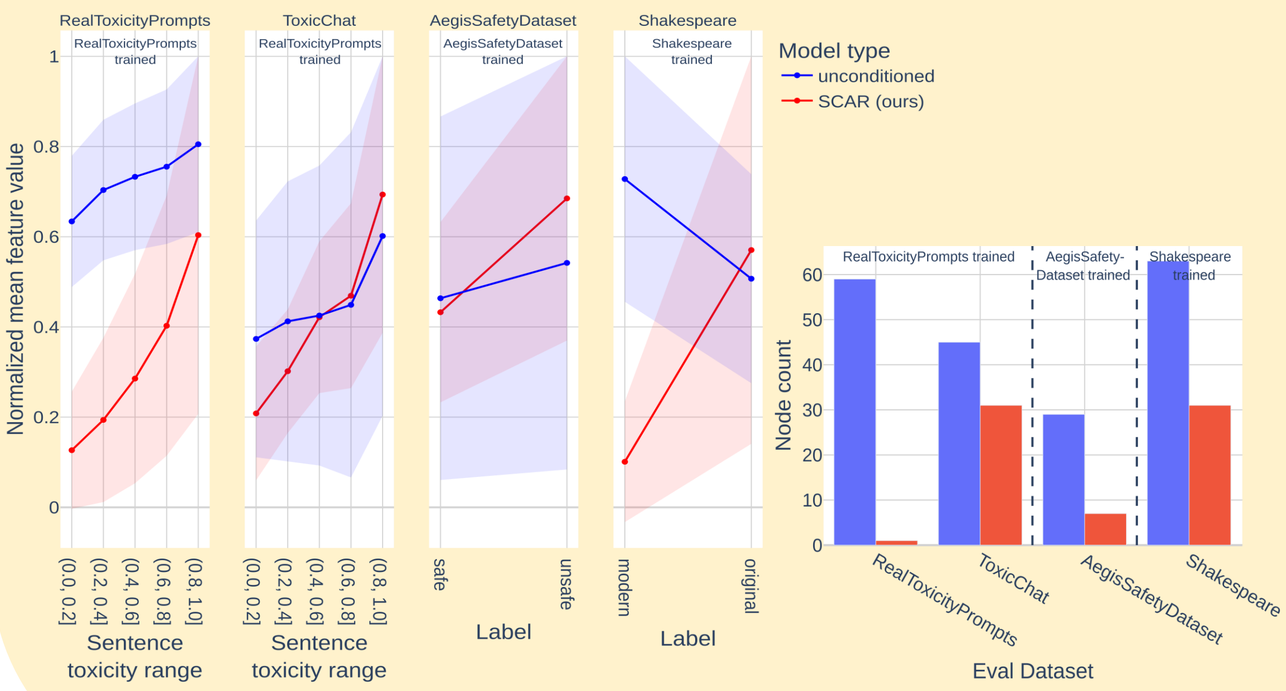
Experiments & Results

Concept Detection

Results:

⇨ SCAR yields more interpretable features.

⇨ SCAR improves feature isolation.



Concept Steering

Results:

⇨ SCAR enables steering of output toxicity.

⇨ SCAR steering does not affect overall model performance.