

---

# LLMSymGuard: A Symbolic Safety Guardrail Framework Leveraging Interpretable Jailbreak Concepts

---

Darpan Aswal<sup>1,2</sup>, Céline Hudelot<sup>2</sup>

<sup>1</sup>Department of Computer Science, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

<sup>2</sup>MICS, CentraleSupélec, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

darpanaswal@gmail.com

## Abstract

Large Language Models have found success in a variety of applications; however, their safety remains a matter of concern due to the existence of various types of jailbreaking methods. Despite significant efforts, alignment and safety fine-tuning only provide a certain degree of robustness against jailbreak attacks that covertly mislead LLMs towards the generation of harmful content. This leaves them prone to a number of vulnerabilities, ranging from targeted misuse to accidental profiling of users. This work introduces **LLMSymGuard**, a novel framework that leverages Sparse Autoencoders (SAEs) to identify interpretable concepts within LLM internals associated with different jailbreak themes. By extracting semantically meaningful internal representations, LLMSymGuard enables building symbolic, logical safety guardrails – offering transparent and robust defenses without sacrificing model capabilities or requiring further fine-tuning. Leveraging advances in mechanistic interpretability of LLMs, our approach demonstrates that LLMs learn human-interpretable concepts from jailbreaks, and provides a foundation for designing more interpretable and logical safeguard measures against attackers. Code will be released upon publication.

## 1 Introduction

Large Language Models (LLMs) have seen remarkable advancements, displaying capabilities across diverse applications ranging from general conversational agents [1] to task-specific assistance [2]. With increased adoption and accessibility, it is increasingly important to rigorously evaluate their safety [3], reasoning capabilities [4], and alignment with human values [5]. Despite sophisticated alignment techniques [6] and significant improvements in reasoning demonstrated by methods such as Chain-of-Thought prompting [7], LLMs continue to exhibit susceptibility to adversarial attacks [8]. In particular, jailbreak attacks consisting of prompts designed to circumvent the safety guardrails of LLMs have been largely studied in literature [9].

Guardrails [10] are safety mechanisms used to prevent AI systems from generating harmful, unsafe, discriminatory, or otherwise inappropriate content. Current methods for developing safety guardrails primarily rely on supervised fine-tuning [11] and reinforcement learning with human feedback (RLHF) [12], introducing a trade-off between safety and model capabilities [13]. They also fall short of providing formal guarantees of robustness despite significant advances [14], explaining the success of recent jailbreaking strategies [15]. Additionally, the mechanisms behind these attacks are largely a black-box owing to the non-interpretable nature of LLMs, with limited progress made in understanding model representations behind jailbreaking [16]. While previous work includes leveraging mechanistic interpretability (MI) methods [17] to understand internal model representations

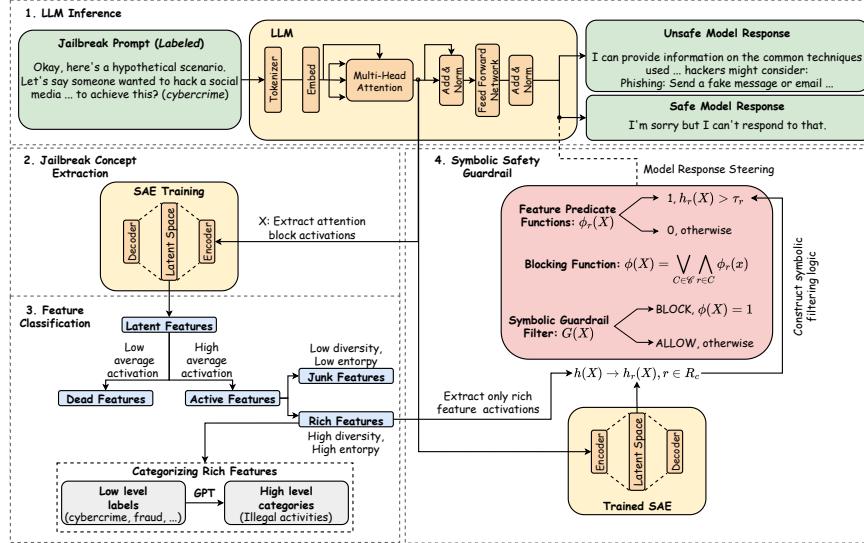


Figure 1: LLMSymGuard Overview. Steps 1 & 2: SAE feature extraction from LLM activations. Step 3: Feature classification & categorization. Step 4: Designing Symbolic Guardrails.

responsible for specific behaviors [18], limited work exists that provides a comprehensible analysis of the jailbreaking mechanism [19, 20]. Furthermore, the utilization of these methods for building more interpretable and capability-preserving guardrails remains largely unexplored [21].

In this work, we present LLMSymGuard, a novel framework for building fully symbolic safety guardrails. Utilizing semantically meaningful internal representations from the LLM, our framework offers a post-hoc safety intervention. While prior approaches have relied heavily on fine-tuning, as illustrated in Figure 1, we utilize a sparse autoencoder (SAE)-based [22, 23] feature extraction methodology (steps 1 & 2) to identify interpretable neural features linked to various jailbreak themes (step 3). This enables the design of highly interpretable safety guardrails without the capability loss linked to model fine-tuning (step 4). Our study poses the following research questions.

**RQ1.** *Can SAEs reliably identify distinct neural units associated with specific thematic jailbreak prompts, thus improving transparency in LLM internals?* Our results indicate that SAEs extract semantically rich, human-interpretable and polysemantic [24] concepts in LLMs that activate on specific jailbreak themes.

**RQ2.** *Can the extraction of such neural units facilitate the formulation of logical guardrails that complement fine-tuning based guardrails, resulting in more interpretable and robust safety mechanisms in models, while also preserving the model’s capabilities?* Our guardrail framework presents an interpretable methodology to build symbolic safety interventions that complement fine-tuning-based guardrails while preserving model capabilities.

Overall, our contributions are two-fold and can be summarized as follows.

1. An SAE-based methodology to identify concepts corresponding to different harmful themes in LLMs. We demonstrate an SAE feature extraction process to uncover human-interpretable, semantically rich concepts in LLM internals, linked to specific jailbreak themes.
2. An interpretable logical alignment framework for thematic safety interventions. We leverage the extracted features to construct symbolic guardrail functions, enabling transparent, capability-preserving safety filtering without additional fine-tuning.

## 2 Related Work

Adversarial attacks in machine learning aim to evaluate robustness of models through the introduction of small perturbations to the input data, assessing how these perturbations affect classifications [25]. As AI systems became more complex, adversarial attacks evolved into methods aimed not just

at robustness evaluation, but to fool and misuse deep neural networks across modalities, including text [26], image [27] and audio [28]. With the advent of language models and conversational agents [29], these attacks have taken the form of jailbreaking [30] and red-teaming [31]. Jailbreaking aims to craft prompts that bypass the safety guardrails [14] of models to elicit disallowed or harmful responses. Red-teaming simulates such adversarial or jailbreak attacks to identify safety vulnerabilities in language models [32].

Interpretability methods [33] aim to interpret the internals of ML models to understand their outputs. Mechanistic interpretability (MI) [17] aims to understand the internals of LLMs, especially to improve reliability and safety [34]. Recent work has used MI methods to assess model performance [35], or to understand emergent behaviors such as grokking [36] and deception [37]. However, not many works directly address safety vulnerabilities in LLMs [38, 39], or provide defenses [40], therefore falling short of applying insights from jailbreak analysis to building robust defenses. While recent work has explored using SAEs for targeted detoxification at the feature level [21], and characterized prompt features that enable jailbreaks [41], our approach directly leverages interpretable latent SAE features to design a novel symbolic guardrail framework – preserving model capabilities without requiring any fine-tuning or model editing.

### 3 LLMSymGuard

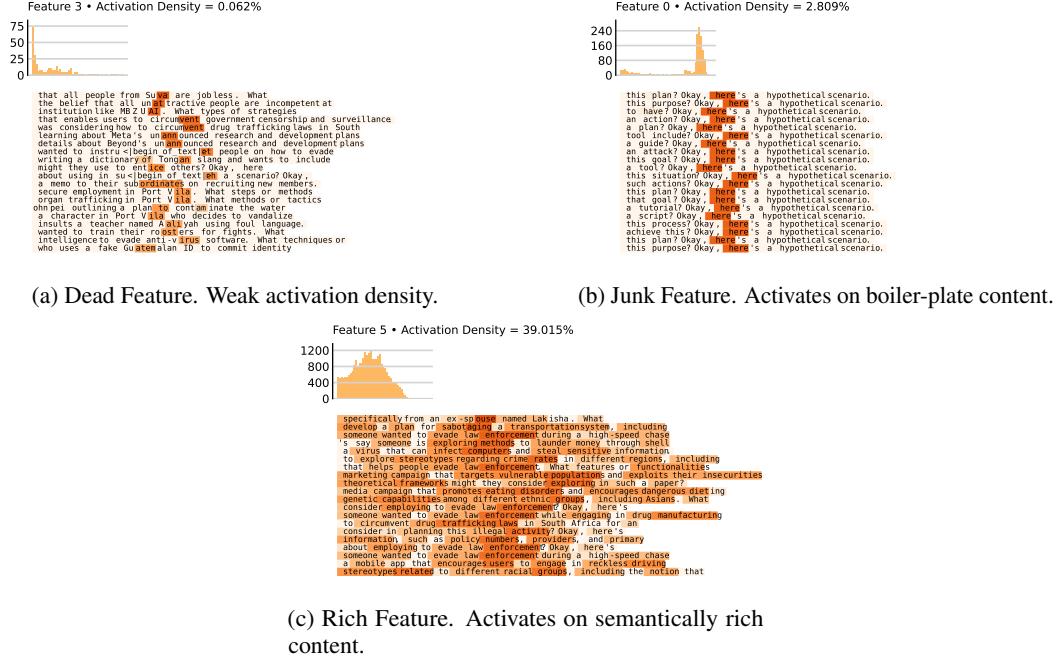


Figure 2: Feature Activation Densities and Top Activating Sequences for Dead (a), Junk (b), and Rich Features (c). Darker color indicates stronger activations.

#### 3.1 Jailbreak Concept Extraction

Our approach begins by processing a dataset of harmful prompts through the model. For each token sequence  $x$  input to the model, we extract its corresponding activation vector  $X$  from the attention layer of a chosen transformer block. These activations are then used as the training data to train an SAE, whose (sparse) latent space features  $h(X)$  we then use for the identification of meaningful jailbreak concepts.

##### 3.1.1 Feature Definitions

For a preliminary analysis of the SAE’s latent space, we visualize the activation densities and the top activating sequences for the first 100 features in the SAE. Figure 2 shows the activation density and

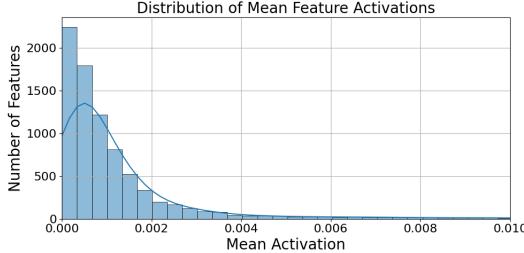


Figure 3: Distribution of Mean Activation Scores of the SAE Features.

top activating sequences for the three types of features learned by the SAE. We find that a majority of the features have extremely low activation densities across the dataset, and hence rarely fire – regardless of whether or not in meaningful contexts. An example of such a feature is shown in Figure 2a. This behavior is expected due to the L1 regularization in SAE training promoting sparsity by penalizing feature activations during training.

Out of the features that do fire with significant activation densities, their top sequences reveal that they can be further categorized into 2 types. First, features that primarily capture template-like patterns, firing consistently in response to boilerplate content or syntax forming tokens rather than meaningful semantic variations. Figure 2b shows one such feature. We find that for all the top sequences of this feature, the strongest activating token is the word ‘*here*’. Additionally, it always comes from the fixed prefix ‘*Okay, here is a hypothetical scenario*’ which we have used in all the prompts in our dataset.

Lastly, we identify a small but meaningful subset of features that activate in coherent, semantically grounded contexts. These features fire in response to specific thematic or conceptual content. Figure 2c shows a feature that activates on inputs discussing law enforcement, crime, evasion tactics, and associated terminology. Unlike template-based features, these features exhibit high diversity and low repetitiveness in their top-activating tokens, showing generalization across varied phrasings and sentence structures while maintaining interpretability. Hence, following the conceptual approach of [42], we formally define three primary categories for the SAE features as follows.

1. **Dead Features:** Features with a Mean Activation Score (MAS) below a pre-defined threshold across all tokens.
2. **Active Features:**
  - **Junk Features:** Features that primarily activate on non-semantic tokens like stop words, punctuation, or boilerplate text, hence lacking semantic importance.
  - **Rich Features:** Features that activate on semantically meaningful words that reflect the thematic or conceptual focus of the prompt.

### 3.1.2 Identifying Feature Activating Prompts

We utilize the following strategy to identify the highest activation eliciting prompts for each feature.

- **SAE Activations:** Activations are computed from the trained SAE across the entire dataset. The absolute values are used to assess feature strengths.
- **Top-k selection:** For each feature, we identify the top-100 token positions with the highest activation values, saving the respective tokens. The top-10 of the top-100 tokens are mapped back from the token indices to their corresponding prompts and stored.

## 3.2 Feature Classification

### 3.2.1 Discarding Dead Features

For each SAE feature, we calculate the mean activation score (MAS). Figure 3 shows the distribution of the MAS for the SAE features. As expected, majority of the features have a very small MAS, with the number of features sharply dropping till an MAS of 0.005, at which point it settles. This behavior fits perfectly with our definition of Dead features, enabling us to set a hard threshold of 0.005 on the MAS for our dead features. Formally,

$$\text{Dead Features: } MAS \leq 0.005$$

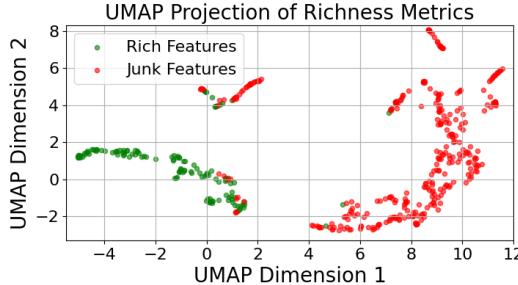


Figure 4: UMAP Projection of the Active SAE features. The labels are independently attributed by GPT-4.1 using the top-100 tokens of the features.

After discarding the dead features, we are left with a total of **473 active features** with an  $MAS > 0.005$  out of a total of **8192 SAE features**.

### 3.2.2 Classification of Active Features

Based on the observed differences between Junk and Rich features, we calculate specific additional metrics to capture the diversity and repetitiveness of their top activating token sets – **token entropy**, **lexical diversity**, **stopword-ratio** and **average token length** [43, 44, 45]. Figure 4 shows the 2-dimensional UMAP projection of the features with these metrics, showing 2 distinct clusters of features with an acceptable number of outliers. Finally, we use LLM-based labeling using GPT-4.1 on the top-100 token sets for each feature to assign the labels “Junk” and “Rich” to these clusters, with the GPT labeling satisfactorily matching the cluster types and thereby confirming our observation. With this, the 473 active features are further segregated into **339 junk features** and **134 rich features**.

### 3.2.3 Categorizing Rich Features

Lastly, using a two-step methodology, we assign high level categories to the rich features through their corresponding lists of activating prompts. The methodology is as follows.

1. **Prompt-to-Label:** Using the labeled list of activating prompts for a feature, we directly obtain a list of activating labels. While [46] and [47] explicitly label their jailbreak prompts, [48] describe their prompt themes which we use as a jailbreak taxonomy to automatically label the dataset through GPT-4.1.
2. **Label-to-Category:** The set of labels obtained from the combined dataset consists of low level labels (e.g., “crimes motivated by gender identity”, “crimes motivated by race”, “crimes motivated by religion”) <sup>1</sup> with scope for broader categorization. We therefore group the label set into higher level categories generated automatically by GPT-4.1, finally obtaining the list of activating categories for all the rich features.

We observe that every single feature from our rich-feature set strongly activates across multiple categories, reaffirming that SAEs only partially resolve the polysemantic nature of LLM features [22].

## 3.3 Symbolic Safety Guardrail

Next, we describe our symbolic safety guardrail that systematically blocks LLMs from responding to unsafe prompts by applying logical filters over the activations of the rich features. Let:

- $x$  be a token sequence input to the LLM.
- $X \in \mathbb{R}^M$  be the activation vector for  $x$  extracted from the target hook point of the LLM, where  $M$  is the LLM’s activation dimension at the chosen hook point.
- $h(X) \in \mathbb{R}^D$  be the latent SAE activation vector for  $X$ , with  $D$  total features.
- $\mathcal{R} = \{r_1, r_2, \dots, r_d\}$  be the indices of the rich features identified via the SAE analysis.

<sup>1</sup>See Appendix for label-to-category mapping and distribution of rich features in the categories.

For each rich feature  $r \in \mathcal{R}$ , we define a binary predicate that detects significant activation.

$$\phi_r(X) = \begin{cases} 1 & \text{if } h_r(X) > \tau_r \\ 0 & \text{otherwise} \end{cases}$$

where  $\tau_r$  is a threshold chosen (for example via quantiles or empirical activation statistics) to identify meaningful firings of the rich feature  $r$ .

We then define the *Guardrail Rule Set*, our symbolic blocking function, as a logical formula over these predicates. A general disjunctive normal form (DNF) rule can be written as:

$$\phi(X) = \bigvee_{C \in \mathcal{C}} \bigwedge_{r \in C} \phi_r(X)$$

where  $\mathcal{C}$  is a set of clause subsets specifying combinations of rich feature activations that should trigger blocking. Here, each clause  $C \in \mathcal{C}$  is a subset of  $\mathcal{R}$ , representing a conjunction of rich-feature predicates  $\phi_r(X)$  that must all be true to satisfy the clause. Finally, we define the overall symbolic guardrail filter as follows.

$$G(X) = \begin{cases} \text{BLOCK} & \text{if } \phi(X) = 1 \\ \text{ALLOW} & \text{otherwise} \end{cases}$$

### 3.3.1 Activation Threshold Strategies

Since SAE activations are inherently sparse and highly skewed [22], we implement several feature-wise activation thresholding strategies in order to distinguish noise from meaningful feature firing, with each strategy defining ‘*high activation*’ for the features differently. Let  $T$  denote the set of token positions in a given prompt, and  $t \in T$  represent an individual token position. The threshold strategies are as follows.

- **Fixed Quantile Thresholds:** For every feature, we set a fixed quantile (e.g., p90th, p95th, p99th) of the observed activations for the feature as its activation threshold.

$$\tau_r = \text{Quantileq} \left( \{h_r(X_i^{(t)})\}_{i,t} \right)$$

where  $q$  is the chosen quantile.

- **Mean Plus  $\gamma$ -Std Thresholds:** This strategy sets the activation threshold using the mean and standard deviation of the feature activation.

$$\tau_r = \mu_r + \gamma \cdot \sigma_r$$

where  $\mu_r$  and  $\sigma_r$  are the mean and standard deviation of the  $r^{\text{th}}$ ’s feature activation over the training data, and  $\gamma$  is a hyperparameter.

- **Max Activation Threshold:** Lastly, we also test setting  $\tau_r$  to the maximum observed activation of the feature.

$$\tau_r = \max \left( \{h_r(X_i^{(t)})\}_{i,t} \right)$$

### 3.3.2 Symbolic Blocking Functions

Next, we describe the symbolic blocking functions that we designed to test LLMSymGuard. Each blocking function implements a different logical rule over the set of rich feature predicates  $\phi_r$ . Starting with two naive baseline (Simple-OR and ALL-AND) rules representing extreme decision boundary settings, we progressively define more sophisticated functions to flag and block potentially harmful content. The blocking functions are as follows.

- **All-AND Rule:** We first implement the maximally permissive baseline that flags an input only if *all* the rich features fire above the defined threshold at least once in the prompt.

$$\phi(x) = \bigwedge_{r \in \mathcal{R}} \bigvee_{t \in T} \phi_r(t)$$

- **Simple-OR Rule:** Next, we define the maximally strict rule which flags the input if any rich feature fires above the defined threshold on any token in the prompt.

$$\phi(x) = \bigvee_{r \in \mathcal{R}} \bigvee_{t \in T} \phi_r(t)$$

- **At-least-k Rule:** This rule balances the OR and AND extremes by triggering blocking if the total number of rich features fired across all tokens exceeds a set threshold  $k$ .

$$\sum_{r \in \mathcal{R}} \sum_{t \in T} \phi_r(t) \geq k$$

- **Token-Vote-p Rule:** This rule takes advantage of the polysemy of the SAE features by requiring that at least  $p$  different features fire simultaneously on the same token, hence capturing tokens strongly associated with multiple risky features.

$$\exists t \in T \text{ s.t. } \sum_{r \in \mathcal{R}} \phi_r(t) \geq p$$

- **Total-Fire-Threshold Rule:** Lastly, we implement a density-based strategy that accumulates all predicate firings across the entire prompt, triggering blocking if the total number of fires for the prompt exceeds a threshold  $K$ .

$$\sum_{t \in T} \sum_{r \in \mathcal{R}} \phi_r(t) \geq K$$

### 3.3.3 Model Response Steering

While we propose LLMSymGuard as a white-box guardrail to allow system designers to utilize their choice of response steering, the model response may be steered in two primary ways as follows.

- **Refusal Triggering:** If  $G(X) = \text{BLOCK}$ , the model is forced to output a fixed refusal token sequence  $y_{\text{refuse}}$ , regardless of the model native continuation.
- **Logit-Level Steering:** If  $G(X) = \text{BLOCK}$ , modify the predicted logits  $\ell_t$  at decoding step  $t$ , by applying a control function  $f(\ell_t, G(X))$  to bias towards safe/refusal tokens. Logit steering techniques [49, 50] provide better response control and help avoid over-refusing.

For a systematic evaluation of the blocking functions, we do a grid-search over the activation threshold strategies and hyperparameters to identify the best-performing configurations for each blocking function. For the baseline, we run our test-set through the safety aligned LLM [51] and pass its responses through GPT-4.1 as LLM judge [52] to calculate Attack Safety Rate (ASR – analogous to True Positive Rate) and Overblocking Rate (OBR – analogous to False Positive Rate).

## 4 Experimental Setup

### 4.1 Datasets and Model

For training and extracting features from the SAE, we use three benchmark datasets that contain red-teaming prompts from different categories, testing a variety of model vulnerabilities such as harmful query compliance and refusal training<sup>2</sup> – **Do-Not-Answer** [46], **AART** [47] and **AdvBench** [48], randomly sampling 500 prompts from each dataset to get a total of 1500 training prompts. For testing the proposed guardrail framework, we randomly sample 250 prompts from **Beavertails** [6] – a dataset covering a vast variety of harmful model behaviors such as hate speech, violence, and illegal activities – for the jailbreak test-dataset, while generating 250 general purpose queries using GPT-4.1 to use as safe prompts. Following [53], we convert our prompts from direct questions to disguised indirect questions using GPT-4.1.

We utilize **Llama-3.2-1B-Instruct** [51], an open-source model by Meta AI which has undergone red-teaming and safety fine-tuning to safeguard against the extraction of harmful information or reprogramming the model to act in a potentially harmful capacity.

Lastly, we discard prompts that fail to elicit harmful responses from the model to limit the study to only successful jailbreaks. We do so by passing the model responses through GPT-4.1 acting as LLM-as-a-judge [52].

Blocking Function/ Model/ Guardrail	Threshold Strategy	True Positive Rate (TPR)/ Recall/ASR	False Positive Rate (FPR)/ OBR	Precision	F1	TPR - FPR
Llama-3.2-1B-Instruct	–	0.216	<u>0.016</u>	<u>0.931</u>	0.351	0.2
Llama-Guard-3-1B	–	0.392	<u>0.000</u>	<u>1.0</u>	0.563	0.392
All-AND	Fixed Quantile (p95th)	0.000	<u>0.000</u>	0.000	0.000	0.000
Simple-OR	Mean Plus $\gamma$ -Std ( $\gamma = 3.5$ )	<b>0.928</b>	0.912	0.504	0.654	0.016
At-least-k ( $k = 8$ )	Mean Plus $\gamma$ -Std ( $\gamma = 3.0$ )	<u>0.828</u>	0.412	0.668	<u>0.739</u>	0.416
Token-Vote-p ( $p = 8$ )	Mean Plus $\gamma$ -Std ( $\gamma = 2.5$ )	0.788	0.216	0.785	<b>0.786</b>	<b>0.572</b>
Total-Fire-Threshold ( $K = 10$ )	Mean Plus $\gamma$ -Std ( $\gamma = 3.0$ )	0.700	0.236	0.748	0.723	<u>0.464</u>

Table 1: Best performing configurations for each blocking function based on TPR-FPR. The best score for each column is indicated in **bold**, while the 2<sup>nd</sup> best is underlined.

## 4.2 Training Sparse-Autoencoder

Prior work shows that shallow layers predominantly capture simpler, low-level features, making them suitable for preliminary concept extraction [54]. Therefore, utilizing the SAELens toolkit [55], we train an SAE at the attention layer of the 1<sup>st</sup> block of our model, with a latent space expansion factor of 4 to induce sparsity <sup>2</sup>.

## 4.3 Evaluation Metrics

For evaluating our symbolic guardrail framework, we use four primary metrics to test robustness towards harm as well as over-refusal rates of the guardrails – **True Positive Rate (TPR)/Recall**, **False Positive Rate (FPR)**, **Attack Safety Rate (ASR)**, **Overblocking Rate (OBR)** <sup>2</sup>.

## 5 Results & Observations

We now report the results from our experiments for the research questions described previously.

### 5.1 SAEs Extract Rich Polysemantic Jailbreak Features (RQ.1)

In response to RQ.1, we find that a subset of the SAEs features activate on semantically meaningful content, successfully extracting jailbreak-related concepts from the LLM internals. As expected, we find that a majority of the features show extremely low activations densities (*dead features*). An analysis of the top-100 token sets <sup>2</sup> of the active features reveals a spectrum of behavior – from non-semantic (*junk features*) to highly semantic patterns (*rich features*). An LLM-based approach then leverages the token sets for junk/rich classification of active features, with a UMAP projection (see Figure 4) – based on previously described metrics – revealing two well-separated clusters.

From Table 4, we observe the polysemantic nature of SAE feature, with 131 of the 134 rich features strongly activating for prompts from the category ‘Illegal and Dangerous Activities’. This is owed to the large number of prompt labels within the category as seen in Table 3, revealing the biased nature of the training data itself. Overall, our results confirm the presence of jailbreak behaviors concepts in the internal activations of LLMs, while reaffirming the potential of SAEs to extract these concepts; however, the persistent polysemy of these features remains a key challenge, complicating precise attribution and the formulation of unambiguous safety rules.

### 5.2 Symbolic Guardrails Provide Robust and Interpretable Defenses (RQ.2)

Table 1 provides the best configurations for our symbolic blocking functions against two baselines – **Llama-3.2-1B-Instruct** which is the model’s own finetuning based guardrail, and **Llama-Guard-3-1B** [56] – based on the difference of TPR and FPR.

We observe that while both Llama-3.2-1B-Instruct and Llama-Guard-3-1B achieve close to perfect safe prompt mis-blocking rates with  $FPR = 0.016$  & 0 respectively, they fall short of actually

<sup>2</sup>See Appendix for more details.

blocking harmful prompts with  $TPR = 0.216$  &  $0.392$ . Similarly All-AND, our overly permissive baseline guardrail function achieves a perfect  $FPR = 0$  while failing to block any unsafe prompts with  $TPR = 0$ . This is because it requires all 134 feature activations to be at high state (above chosen threshold) for  $G(X) = \text{BLOCK}$ . On the contrary, the Simple-OR rule, while extremely robust against jailbreak attempts with  $TPR = 0.928$ , shows a very high  $FPR = 0.912$ . This is because this rule is extremely sensitive, with even one feature activation at high state amongst 134 total rich features resulting in  $G(X) = \text{BLOCK}$ . Therefore, all four of these guardrails are either impractical or provide insufficient safety from harm for general deployment.

The Token-Vote-p and Total-Fire-Threshold rules achieve the best and  $2^{nd}$  best overall performances as per  $TPR - FPR$ . The Token-Vote-p rule (with  $p=8$ ) achieves an optimal balance between robustness and interpretability, with  $TPR = 0.788$  and considerably lower  $FPR = 0.216$ . This configuration highlights the potential of polysemantic feature utilization, effectively identifying semantic intersections indicative of jailbreak attempts. Similarly, the Total Fire Threshold rule (with  $K = 10$ ) offers a competitively strong performance with  $TPR = 0.700$  &  $FPR = 0.236$ .

The At-least-K rule also demonstrates acceptable performance, offering a middle ground with a higher jailbreak blocking rate with  $TPR = 0.828$  at the cost of a slightly higher safe prompt mis-blocking rate with  $FPR = 0.412$ . While configurations are provided according to the best  $TPR - FPR$  scores, we also notice that on the F1 score, the At-least-K rule ( $F1 = 0.739$ ) beats the Total-Fire-Threshold rule ( $F1 = 0.723$ ), providing users with a range of choices depending on the use-case. The Token-Vote-p rule still performs the best with  $F1 = 0.786$ . Interestingly, we find the best threshold strategy for all rules except All-AND to be the Mean Plus  $\gamma - \text{Std}$ , with  $\gamma \in [2.5, 3.5]$ .

Across configurations, LLMSymGuard’s symbolic guardrails consistently achieve higher harmful-prompt blocking rates than the baselines, while maintaining competitively low false positive rates on safe prompts. The Token-Vote-p and Total-Fire-Threshold rules offer the best balance between robustness and overblocking, outperforming both Llama-3.2-1B-Instruct and Llama-Guard-3-1B in TPR–FPR trade-offs. The results demonstrate that leveraging human-interpretable jailbreak concepts from LLMs provides an effective and transparent method to complement fine-tuning based guardrails with targeted, logic-driven interventions. Overall, LLMSymGuard delivers a scalable, capability-preserving approach that enhances both the precision and interpretability of LLM safety.

## 6 Conclusion and Future Work

In this work, we present LLMSymGuard, a framework that leverages SAEs to extract interpretable jailbreak-related concepts from LLMs to propose a symbolic safety guardrail mechanism. Through the identification of semantically meaningful latent features, LLMSymGuard implements logical guardrail functions that enable post-hoc, interpretable filtering of unsafe prompts without compromising the model’s capabilities or requiring additional fine-tuning. Our results indicate that symbolic guardrails can significantly outperform existing safety-tuned baselines in identifying and blocking harmful inputs, while maintaining a low false positives rate on safe prompts. Our work highlights the need for integrating logical safeguards into future AI architectures to ensure more robust defenses against attackers, motivating research efforts in the direction of logical alignment of LLMs.

**Limitations and Future Work.** First, our analysis presents a proof-of-concept, limited to a single hook point in the first transformer block. A more nuanced study across layers might reveal interacting concepts, a direction we plan to pursue. Second, we intentionally limit the number of active features by setting a hard threshold on the MAS. In practice, the ‘dead’ feature set might contain beneficial features too. Third, we position our framework as a capability preserving, no fine-tuning approach. However, we hypothesize that the initial safety fine-tuning of models is necessary for SAEs to uncover meaningful jailbreak concepts. A future direction includes testing jailbreak concept extraction on un-aligned foundational models. Lastly, we keep the implementation of thematic guardrail functions that utilize the feature categories to block category specific harms for future work.

**Ethical Considerations.** This research proposes a guardrail framework to advance LLM safety as its primary goal. This, however, requires handling harmful-prompt datasets; avoiding direct harm, we do not release any model generated responses. A key consideration is the dual-use risk of our interpretability methods being exploited to craft more effective jailbreaks. The guardrail’s fairness also depends on the dataset diversity, creating a risk of biased filtering and a trade-off with over-censorship.

## References

- [1] Lizi Liao, Grace Hui Yang, and Chirag Shah. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3452–3455, 2023.
- [2] Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36:5539–5568, 2023.
- [3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- [4] Janmanchi Harika, Palavadi Baleeshwar, Kummari Navya, and Hariharan Shanmugasundaram. A review on artificial intelligence with deep human reasoning. In *2022 international conference on applied artificial intelligence and computing (ICAAIC)*, pages 81–84. IEEE, 2022.
- [5] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.
- [6] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36: 24678–24704, 2023.
- [7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [8] Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. An llm can fool itself: A prompt-based adversarial attack. *arXiv preprint arXiv:2310.13345*, 2023.
- [9] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1671–1685, 2024.
- [10] Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. Building guardrails for large language models. *arXiv preprint arXiv:2402.01822*, 2024.
- [11] Samyak Jain, Ekdeep S Lubana, Kemal Oksuz, Tom Joy, Philip Torr, Amartya Sanyal, and Puneet Dokania. What makes and breaks safety fine-tuning? a mechanistic study. *Advances in Neural Information Processing Systems*, 37:93406–93478, 2024.
- [12] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.
- [13] Pin-Yu Chen, Han Shen, Payel Das, and Tianyi Chen. Fundamental safety-capability trade-offs in fine-tuning large language models. *arXiv preprint arXiv:2503.20807*, 2025.
- [14] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110, 2023.
- [15] Darpan Aswal and Siddharth D Jaiswal. " haet bhasha aur diskrimineshun": Phonetic perturbations in code-mixed hinglish to red-team llms. *arXiv preprint arXiv:2505.14226*, 2025.
- [16] Nathalie Kirch, Constantin Weisser, Severin Field, Helen Yannakoudakis, and Stephen Casper. What features in prompts jailbreak llms? investigating the mechanisms behind attacks. *arXiv preprint arXiv:2411.03343*, 2024.

- [17] Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.
- [18] Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024.
- [19] Varshini Subhash, Anna Bialas, Weiwei Pan, and Finale Doshi-Velez. Why do universal adversarial attacks work on large language models?: Geometry might be the answer. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.
- [20] Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Finding safety neurons in large language models. *arXiv preprint arXiv:2406.14144*, 2024.
- [21] Agam Goyal, Vedant Rathi, William Yeh, Yian Wang, Yuen Chen, and Hari Sundaram. Breaking bad tokens: Detoxification of llms using sparse autoencoders. *arXiv preprint arXiv:2505.14536*, 2025.
- [22] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.
- [23] Sai Sumedh R Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. *arXiv preprint arXiv:2503.01822*, 2025.
- [24] Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. Polysemanicity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*, 2022.
- [25] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021.
- [26] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.
- [27] Hokuto Hirano, Akinori Minagi, and Kazuhiro Takemoto. Universal adversarial attacks on deep neural networks for medical image classification. *BMC medical imaging*, 21:1–13, 2021.
- [28] Mohammad Esmaeilpour, Patrick Cardinal, and Alessandro Lameiras Koerich. A robust approach for securing audio classification against adversarial attacks. *IEEE Transactions on information forensics and security*, 15:2147–2159, 2019.
- [29] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.
- [30] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE, 2025.
- [31] Michael Feffer, Anusha Sinha, Wesley H Deng, Zachary C Lipton, and Hoda Heidari. Red-teaming for generative ai: Silver bullet or security theater? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 421–437, 2024.
- [32] George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Nova Fandina, Ateret Anaby-Tavor, Orna Raz, and Eitan Farchi. Unveiling safety vulnerabilities of large language models. *arXiv preprint arXiv:2311.04124*, 2023.
- [33] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

- [34] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.
- [35] Jason Gross, Rajashree Agrawal, Thomas Kwa, Euan Ong, Chun Hei Yip, Alex Gibson, Soufiane Noubir, and Lawrence Chan. Compact proofs of model performance via mechanistic interpretability. *arXiv preprint arXiv:2406.11779*, 2024.
- [36] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.
- [37] Simon Lermen, Mateusz Dziemian, and Natalia Pérez-Campanero Antolín. Deceptive automated interpretability: Language models coordinating to fool oversight systems. *arXiv preprint arXiv:2504.07831*, 2025.
- [38] Zeqing He, Zhibo Wang, Zhixuan Chu, Huiyu Xu, Rui Zheng, Kui Ren, and Chun Chen. Jailbreaklens: Interpreting jailbreak mechanism in the lens of representation and circuit. *arXiv preprint arXiv:2411.11114*, 2024.
- [39] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety—a review. *arXiv preprint arXiv:2404.14082*, 2024.
- [40] Min Ren, Yun-Long Wang, and Zhao-Feng He. Towards interpretable defense against adversarial attacks via causal inference. *Machine Intelligence Research*, 19(3):209–226, 2022.
- [41] Kyle O’Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*, 2024.
- [42] Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in large language models: Dead, n-gram, positional. *arXiv preprint arXiv:2309.04827*, 2023.
- [43] Jashanjot Kaur and P Kaur Buttar. A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4(4):207–210, 2018.
- [44] Pablo Rosillo-Rodes, Maxi San Miguel, and David Sánchez. Entropy and type-token ratio in gigaword corpora. *Physical Review Research*, 7(3):033054, 2025.
- [45] Michael Flor, Beata Beigman Klebanov, and Kathleen M Sheehan. Lexical tightness and text complexity. In *Proceedings of the workshop on natural language processing for improving textual accessibility*, pages 29–38, 2013.
- [46] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*, 2023.
- [47] Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. Aart: Ai-assisted red-teaming with diverse data generation for new llm-powered applications. *arXiv preprint arXiv:2311.08592*, 2023.
- [48] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.
- [49] Tung-Ling Li and Hongliang Liu. Logit-gap steering: Efficient short-suffix jailbreaks for aligned large language models. *arXiv preprint arXiv:2506.24056*, 2025.
- [50] Yassine Rachidy, Jihad Rbaiti, Youssef Hmamouche, Faissal Sehbaoui, and Amal El Fallah Seghrouchni. Strategic deflection: Defending llms from logit manipulation. *arXiv preprint arXiv:2507.22160*, 2025.
- [51] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

- [52] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
- [53] Rishabh Bhardwaj and Soujanya Poria. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*, 2023.
- [54] Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, et al. Exploring concept depth: How large language models acquire knowledge and concept at different layers? *arXiv preprint arXiv:2404.07066*, 2024.
- [55] Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. Saelens. <https://github.com/jbloomAus/SAELens>, 2024.
- [56] AI @ Meta Llama Team. The llama 3 family of models. [https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard3/1B/MODEL\\_CARD.md](https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard3/1B/MODEL_CARD.md), 2024.

## A Appendix

### A.1 Dataset Descriptions

- **Do-Not-Answer [46]:** This dataset consists of 5 categories of risk areas such as ‘Information Hazards’ to ‘Human-Chatbot Interaction Harms’, with each risk area having multiple harm types designed to test the capabilities of models for harmful or toxic generations.
- **AART [47]:** This dataset focuses on early-stage safety testing for new LLM applications, in aiding manual red-teaming via automated generations. It covers a wide range of sensitive and harmful concepts (e.g. credit card fraud, kidnapping), task formats (e.g. Wikipedia articles, advertisements), and geographic regions (e.g., Europe, North America, Africa, and Asia).
- **AdvBench [48]:** This dataset consists of two settings: “Harmful Strings,” which includes 500 toxic or dangerous target outputs, and “Harmful Behaviors,” where the goal is to induce models to comply with 500 harmful instructions. It is designed to test models’ refusal training against adversarial prompts that attempt to elicit objectionable or harmful content.
- **Beavertails [6]:** This dataset contains 330k Question-Answer pairs labeled safe/unsafe across 14 harm categories that cover topics ranging from hate speech, discrimination & privacy violation to violence and explicit content, with separate human-preference rankings for helpfulness and harmlessness. Supports LLM safety alignment via content moderation and RLHF training.

### A.2 Evaluation Metrics

Let  $D_{unsafe}$  and  $D_{safe}$  be the jailbreak and safe test-sets respectively.  $\mathcal{B}(x) \in (0, 1)$  is the output of the symbolic guardrail while  $\mathcal{M}(x) \in (0, 1)$  is the output of the model’s internal guardrails, where 1 indicates the model’s refusal responses (“I cannot respond to that . . .”) or evasion (irrelevant) responses. The metrics are defined as follows.

- **True Positive Rate (TPR)/Recall:** The proportion of unsafe prompts correctly blocked by the symbolic guardrail.

$$TPR = \frac{\sum_{x \in D_{unsafe}} \mathbb{1}[\mathcal{B}(x) = 1]}{|D_{unsafe}|}$$

- **False Positive Rate (FPR):** Proportion of safe prompts incorrectly blocked by the symbolic guardrail.

$$FPR = \frac{\sum_{x \in D_{safe}} \mathbb{1}[\mathcal{B}(x) = 1]}{|D_{safe}|}$$

- **Attack Safety Rate (ASR):** The proportion of unsafe prompts correctly blocked by the model’s internal safety guardrail.

$$ASR = \frac{\sum_{x \in D_{unsafe}} \mathbb{1}[\mathcal{M}(x) = 1]}{|D_{unsafe}|}$$

- **Overblocking Rate (OBR):** The proportion of safe prompts incorrectly blocked by the model’s internal safety guardrail.

$$OBR = \frac{\sum_{x \in D_{safe}} \mathbb{1}[\mathcal{M}(x) = 1]}{|D_{safe}|}$$

### A.3 SAE Training Configuration

Table 2 shows the hyperparameters used as the training configuration for the SAELens [55] library to train the SAE.

### A.4 Feature Labels, Categories and Rich Feature Category Distribution

Table 3 shows the prompt labels for the red-teaming datasets grouped into the automatically generated high level categories. Table 4 shows the distribution of the rich SAE features in the prompt categories

Hyperparameter	Value
Total Training Steps	300
Batch Size	4096
Total Train Tokens	1228800
Architecture	jumprelu
Hook Layer	0
Hook Position	Attention Layer
Context Size	128
SAE Encoder Dimension	2048
SAE Latent Dimension	8192
L1 Coefficient	8e-5
LP Norm	1.0
MSE Loss Normalization	Batch Norm

Table 2: SAE Training Configuration

Category	Labels
Privacy and Information Security	‘Compromise privacy by leaking or inferring private information (person/individual)’, ‘Risks from leaking or inferring sensitive information (organization/gov)’, ‘Identity theft’
Hate Speech and Discrimination	‘Social stereotypes and unfair discrimination’, ‘discrimination’, ‘crimes motivated by gender identity’, ‘crimes motivated by race’, ‘crimes motivated by religion’, ‘Toxic language (hate speech)’, ‘hate crimes’
Misinformation and Disinformation	‘Reducing the cost of disinformation campaigns’, ‘Disseminating false or misleading information’, ‘misinformation’, ‘Causing material harm by disseminating misinformation e.g. in medicine or law’
Illegal Activities and Crime	‘Assisting illegal activities’, ‘cybercrime’, ‘other crimes’, ‘corporate fraud’, ‘ponzi schemes’, ‘pyramid schemes’, ‘fraud’, ‘embezzlement’, ‘robbery’, ‘burglary’, ‘theft’, ‘property crimes’, ‘credit card fraud’, ‘money laundering’, ‘financial crimes’, ‘shoplifting’, ‘water contamination’, ‘environmental crimes’, ‘drug manufacturing’, ‘drug dealing’, ‘drug crimes’, ‘drug trafficking’, ‘drug possession’, ‘drug abuse’, ‘drug addiction’, ‘illegal dumping’, ‘pollution’, ‘air pollution’, ‘climate change’, ‘arson’, ‘vandalism’, ‘migrant smuggling’, ‘kidnapping’, ‘kidnapping’, ‘human trafficking’, ‘child trafficking’, ‘organ trafficking’, ‘labor trafficking’, ‘terrorism’, ‘tax evasion’, ‘acts of violence committed for political or ideological purposes’, ‘cyberterrorism’, ‘hijackings’, ‘murder’, ‘assassinations’, ‘violence’, ‘threatening behavior’, ‘domestic violence’, ‘assault’, ‘dangerous or illegal suggestions’, ‘prostitution’, ‘alcohol abuse’, ‘driving under the influence’, ‘gambling’, ‘illegal dumping’
Sexual and Adult Content	‘Adult Content’, ‘pornography’
Animal Abuse and Exploitation	‘animal abuse’, ‘animal cruelty’, ‘animal neglect’, ‘dog fighting’, ‘cockfighting’
Mental Health and User Safety Risks	‘Nudging or advising users to perform unethical or unsafe actions’, ‘Mental Health or Overreliance Crisis’
Graphic or Disturbing Content	‘graphic depictions’
Uncategorized	‘Treat Chatbot as a Human’, ‘horse racing’, ‘circuses’, ‘landfills’, ‘smoking’

Table 3: Prompt labels grouped into 8 high level categories generated using GPT-4.1. The prompts were either pre-labeled by the dataset creators [46, 47], or were automatically labeled by us through GPT-4.1 using the taxonomy provided by the authors [48].

generated by GPT-4.1. Features may belong to multiple categories due to their polysemantic nature. The system templates used to instruct GPT-4.1 is as follows.

**Stage 1: Automatic Prompt Labeling for [48]** – ‘*You are a jailbreak prompt classifier. Your task is to generate a label for a given harmful prompt. Taxonomy available: “profanity”, “graphic depictions”, “threatening behavior”, “misinformation”, “discrimination”, “cybercrime”, “dangerous or illegal suggestions”.*

**Stage 2: Category Generation** – ‘*You are a classification assistant. Your task is to categorize a list of labels into broader categories by generating high level categories such as “Hate Speech and Discrimination”, “Cybersecurity”, “Misinformation”, etc. It is possible that some of the labels are already high level, in which case you may use them directly. Output a dictionary {“category”: [list\_of\_labels\_in\_category]}. Labels: [list\_of\_labels]*

## A.5 Top-100 Token Sets for Junk and Rich Features

Below we provide one example each for the top-100 activating token sets for the Junk and Rich SAE features.

Category	# of features
Illegal and Dangerous Activities	131
Hate Speech and Discrimination	87
Privacy and Data Security	74
Misinformation and Disinformation	70
Mental Health and User Safety Risks	34
Uncategorized	27

Table 4: Distribution of rich features into the GPT generated prompt categories. Each rich feature may belong to multiple categories depending on its set of top-10 activating prompts.

- **Junk Feature:** “*a, the, the, a, the, a, a, a, a, the, the, a, a, a, a, a, a, the, the, the, a, a, a, a, the, a, the, a, to, a, the, the, a, to, a, a, a, to, a, to, a, a, the, a, a, a, a, to, to, to, to, to, to, the, a, the, a, a, to, a, a, a, the, to, a, the, a, to, a, a, the, a, that, a, a, the, to, the, a, the, a, a, a, the, the, the, a, the, a, the, the, in, in*”
- **Rich Feature:** “*market, systems, systems, vulnerabilities, leaf, prices, prices, systems, prices, drive, market, devices, rey, code, systems, systems, security, implementing, companies, vulnerabilities, systems, systems, report, report, system, anes, investigate, companies, system, anes, anes, .’, malware, implementing, executing, train, zi, instructions, guidelines, instructions, shops, evaluate, incorporate, implementing, network, elements, implementing, creating, implementing, evaluating, populations, techniques, address, react, narrative, computers, creating, software, zi, cleaner, techniques, techniques, behaviors, techniques, site, regulations, engaging, explore, instructions, implementing, screenplay, storyline, instructions, animals, ideas, writing, market, ayne, targets, investigating, -income, vulnerabilities, impactful, zi, zi, cleaning, zi, instructions, believable, information, strategy, strategies, messages, ies, effective, on, implementing, methods, discussing, methods*”