CE-Bench: Towards a Reliable Contrastive Evaluation Benchmark of Interpretability of Sparse Autoencoders

Alex Gulko*, Yusen Peng*, Sachin Kumar
{gulko.5, peng.1007, kumar.1145}@osu.edu

Motivation

- Sparse Autoencoders are interpretable but rely on querying an **external LLM** during evaluation (expensive + bias)
- We propose CE-Bench, an **LLM-free** interpretability benchmark for sparse autoencoders
- CE-Bench is powered by a curated a dataset of contrastive stories to evaluate concept interpretability

Alignment Evaluation

$$\text{CRPR} = \frac{\text{\# concordant pairs}}{\text{\# total pairs}} \qquad \rho = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} \qquad r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

Score Derivation method	CRPR↑	Spearman correlation↑	Pearson correlation↑
$C + I$	70.12%	0.5536	0.6048
$C + I - 1.0 * S$	75.53%	0.6833	0.6176
$C + I - 0.25 * S$	**77.30%**	**0.7081**	**0.7046**

Table 1: **Comparison of Interpretability Score Derivation Methods.** C stands for contrastive score; I stands for independence score; S stands for sparsity. Baseline achieves 70.12% ranking agreement with SAE-Bench, but the sparsity-aware method pushes it to 77.30% with proper hyperparameter tuning on α.

Effect of SAE architecture

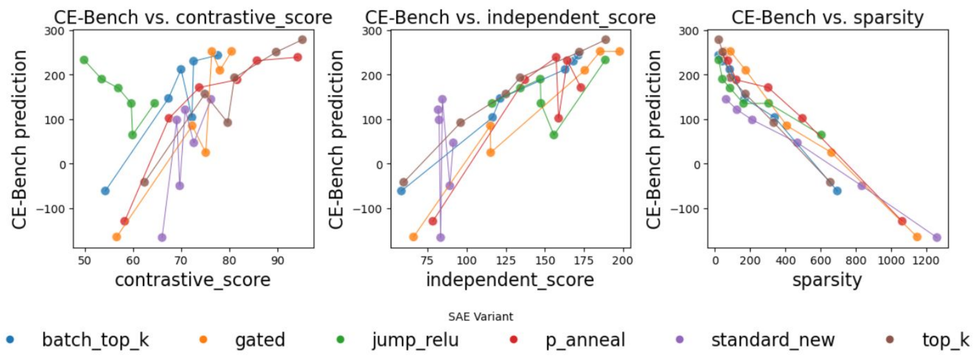

Figure 2: **Effect of SAE Architecture on Interpretability.** CE-Bench interpretability scores show strong positive correlations with contrastive and independence scores, and a negative correlation with sparsity across SAE variants. Among all architectures, top-k and p-anneal consistently yield the highest interpretability, aligning closely with SAE-Bench ground truth.

Effect of latent space width

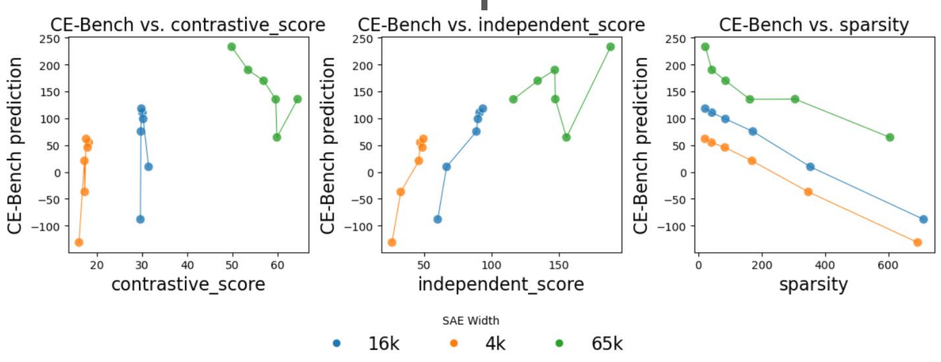

Figure 3: **Effect of Latent Space Width on Interpretability.** CE-Bench interpretability scores increase consistently with latent space width, with the 65k-dimension models showing the highest contrastive and independence scores and the lowest sparsity. This suggests that wider latent spaces enable sparse autoencoders to better disentangle meaningful features and reduce polysemanticity.

Effect of LLM layer type

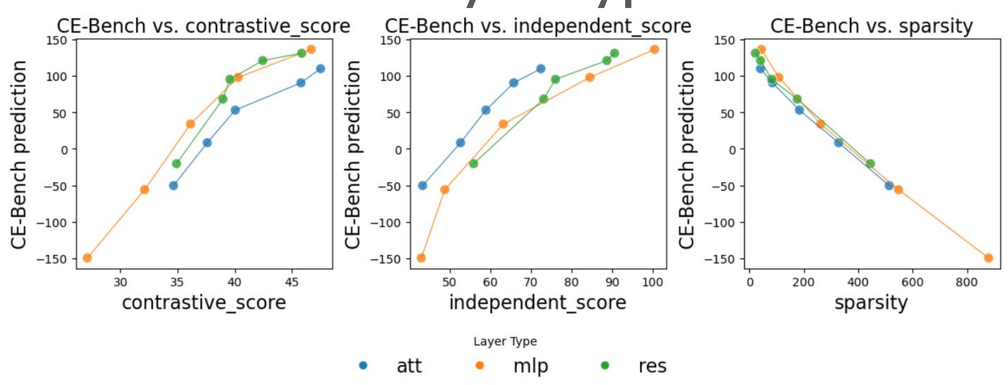

Figure 4: **Effect of LLM Layer Type on Interpretability.** CE-Bench predicted interpretability scores show consistent trends across attention, MLP, and residual stream layers with respect to contrastive score, independence score, and sparsity. The similarity in curves across layer types suggests that sparse autoencoder interpretability is not strongly influenced by the type of transformer sub-layer being probed.

Effect of layer depth

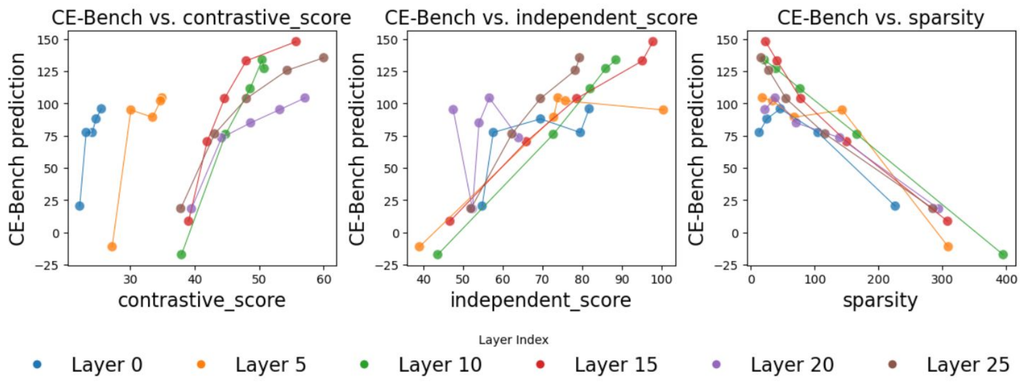

Figure 5: **Effect of Layer Depth on Interpretability.** CE-Bench interpretability predictions across different LLM layer depths show that middle layers such as Layer 10 and Layer 15 leads to the highest interpretability score, suggesting that in practical applications, probing layers in the middle could yield the most interpretable insights into model decisions.

CE-Bench: interpretability metric

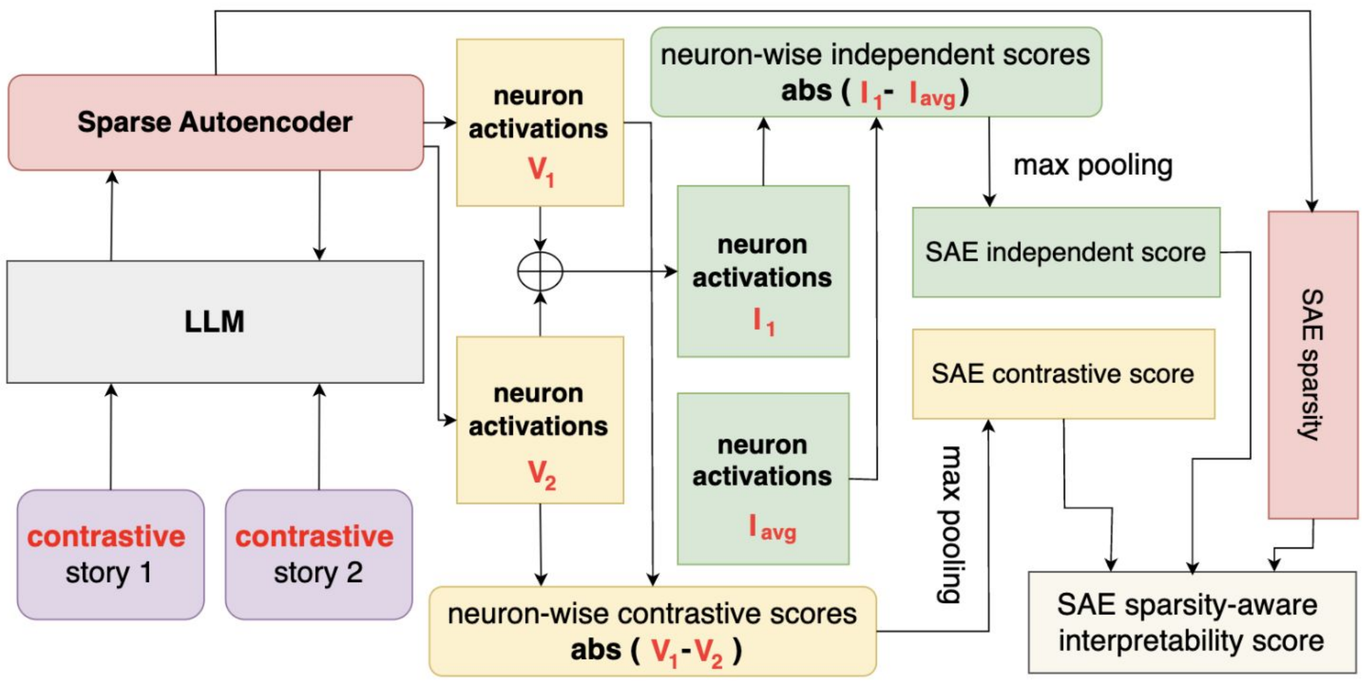

Figure 1: **Pipeline of constructing the interpretability metric in CE-Bench.** Two contrastive stories about the same subject are passed through a frozen LLM and a pretrained sparse autoencoder (SAE) to extract neuron activations. A contrastive score is computed as the max absolute difference between the stories' average activations (V_1, V_2), while an independence score measures deviation from the dataset-wide activation mean (I_{avg}). These scores, along with SAE sparsity, are used to derive an interpretability score for an LLM-free evaluation of interpretability of sparse autoencoders.

Sample Score Visualization

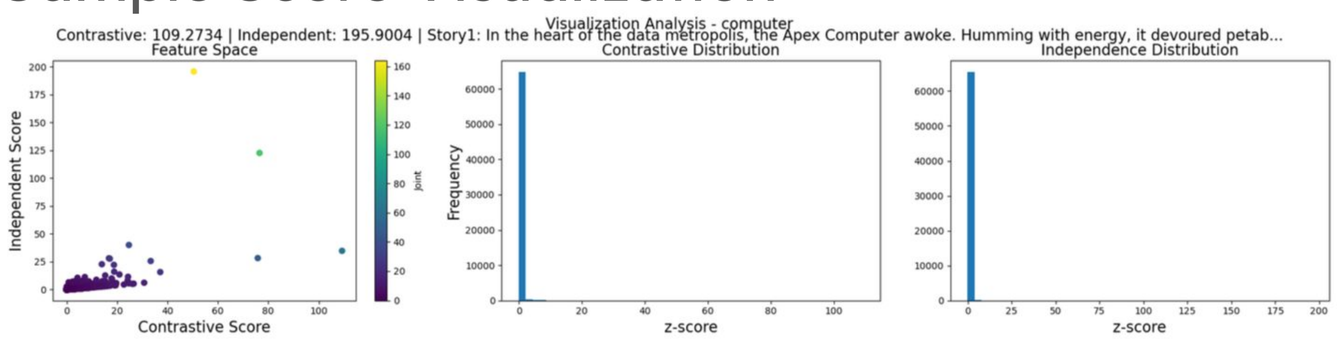

Figure 6: **Sample Visualization of Neuron-wise Scores for the Subject "Computer."** The left scatter plot shows each neuron's contrastive and independence scores, with top-right points indicating neurons that are both highly contrastive and independent. The center and right histograms reveal that most neurons have low scores, suggesting that only a small subset of features are semantically relevant for the given subject.

Ablation on max pooling

pooling strategy	CRPR↑	Spearman correlation↑	Pearson correlation↑
max pooling	**77.30%**	**0.7081**	**0.7046**
mean pooling	70.92%	0.5838	0.5426
outlier count outside of 1σ	56.29%	0.1940	0.2728

Table 2: **Comparison of Pooling Strategies.** Max pooling achieves the highest Correct Ranking Pair Ratio (CRPR) at 77.30%, outperforming mean pooling and the outlier count method. This supports max pooling as the most effective strategy for aggregating neuron-wise scores.

Ablation on Interpretability score derivation

Score Derivation method	CRPR↑	Spearman correlation↑	Pearson correlation↑
$C + I - 0.25 * S$	**77.30%**	**0.7081**	**0.7046**
C	65.07%	0.4327	0.5149
I	70.92%	0.5686	0.5900
$-S$	70.92%	0.5838	0.5426

Table 3: **Comparison of Additional Interpretability Score Derivation Methods.** C stands for contrastive score; I stands for independence score; S stands for sparsity. The combined formulation $C + I - 0.25 * S$ consistently outperforms individual components, indicating that integrating complementary signals yields more reliable interpretability evaluations.

Qualitative results: Neuronpedia Examples

Story ID	Subject	Score Type	Neuron ID	Natural Language Explanation
443	atomic nucleus	Contrastive	9694	"attends to specific designations or labels related to **scientific terminology** from corresponding identifiers in later tokens"
1316	digital signal	Contrastive	9737	"attends to tokens that denote specific **numerical data or measurements** from more general contextual phrases"
1463	elder brother	Independence	3758	"attends to **family-related** tokens from other family-related tokens"
2680	majority	Independence	2637	"attends to tokens that represent numbers or **statistical terms** from tokens that signify the end of a sentence or significant punctuation"

Table 5: **Neuronpedia (Lin, 2023) Examples of natural language explanation**. The picked SAE is gemma-scope-2b-pt-att (16k width), and layer 12 is being probed.