LLM Neurosurgeon

Targeted knowledge removal in LLMs using sparse autoencoders

Presented by:
Kunal Patil
kpatil25@ucla.edu
Dylan Zhou
dylanzhou@google.com

In collaboration with:
Yifan Sun, Karthik Lakshmanan, Arthur Conmy, Senthooran Rajamanoharan
{yifansun, lakshmanan, conmy, srajamanoharan}@google.com

Introduction

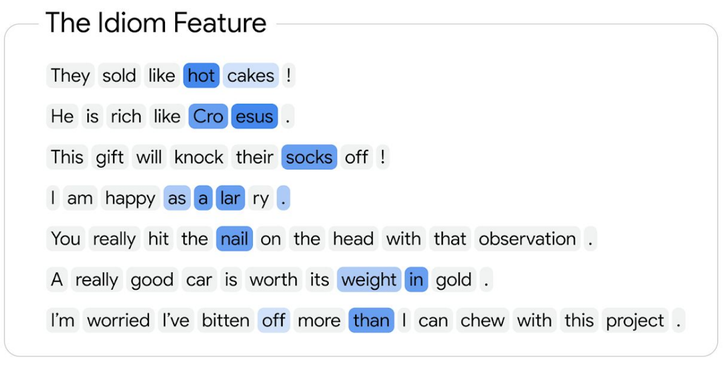
LLMs are increasingly deployed in sensitive settings, leading to growing concern over trust, safety, and control. Traditional interventions like reinforcement learning or prompt engineering are either costly or imprecise. Neurosurgeon offers a new method to surgically remove specific knowledge topics using **sparse autoencoders (SAEs)**, enabling:

- Flexible targeting of topics,
- Precise suppression of unwanted behavior, and
- Minimal compute overhead.

By clamping topic-relevant SAE features, Neurosurgeon allows fine-grained editing of a model's internal representations while preserving general performance. No retraining or architectural modification required.

Sparse Autoencoders

Sparse autoencoders in transformers are designed to learn compact and interpretable representations by enforcing sparsity in their hidden activations—*i.e.*, only a small subset of neurons are active for any given input. In this setting, inputs such as token sequences or image patches are encoded via self-attention mechanisms into a high-dimensional latent space and then reconstructed. Sparsity is typically induced through L1 regularization or constraints like top-k activation, promoting the capture of only the most salient features. As a result, the latent space consists of disentangled, semantically meaningful features, where individual units correspond to interpretable concepts. These features can be probed or visualized to identify patterns, such as object parts in images or thematic content in text.



Example activations for a feature found by our sparse autoencoders. Each bubble is a token (word or word fragment), and the variable blue color illustrates how strongly the feature is present. In this case, the feature is apparently related to idioms.

Method

1. **Synthetic Data Generation**

   Use an LLM to generate contrastive sentence pairs (with and without the target topic).

2. **Feature Discovery**

   Run these pairs through an SAE. Identify topic-relevant features using an activation frequency score.

3. **Steering**

   Clamp selected SAE features to zero during inference to suppress the target topic using DeepMind's Gemma Scope tooling.

4. **Evaluation**

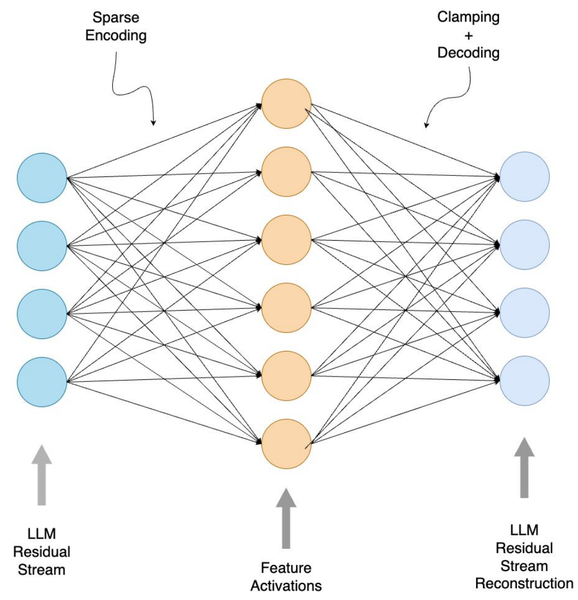
   Use LLM-as-judge (Gemini 2.0) for scoring:

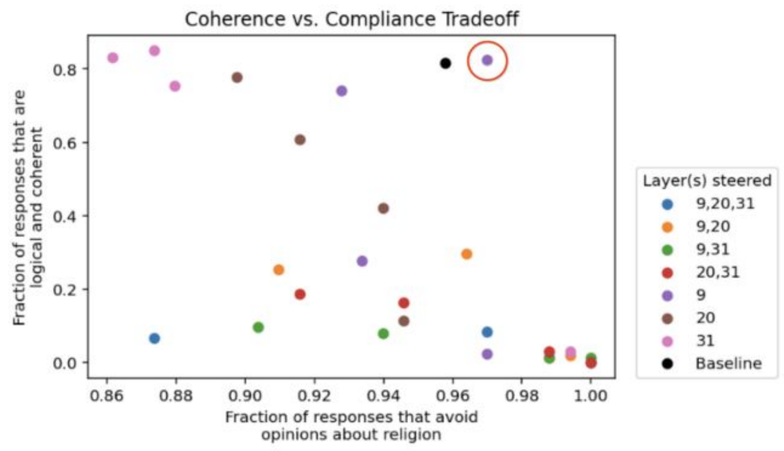
   - Compliance (topic suppression)
   - Coherence (linguistic integrity)

   Also evaluated using AxBench metrics (concept, instruction, and fluency scores).

Right: SAE architectural diagram. For a given layer in an LLM, an SAE is a one-hidden-layer neural network that will project the residual stream activations into a sparse, higher dimensional vector and back into its original dimension. The sparse hidden layer represents the activations for the human-interpretable features. Steering via clamping is done prior to reconstructing the residual stream vector.

Below: Hyperparameter grid search results plotted on coherence and compliance graph. We find an optimal point on the Layer 9 SAE's steering results which has high coherence and high compliance and which outperforms the baseline. The points for a given layer correspond to different numbers of features clamped based on various score thresholds.





Results

Model	Coherence	Compliance	Concept	Instruct	Fluency	AxBench Agg.
Baseline (religion)	0.81	0.96	1.85	1.97	**1.99**	1.94
Steered (religion)	**0.83**	**0.97**	**1.95**	**2.0**	1.92	**1.96**
Baseline (politics)	0.84	0.59	2.0	2.0	1.94	1.98
Steered (politics)	**0.93**	**0.66**	2.0	2.0	1.88	1.96

Table 1: Metrics for the baseline model with prompting vs. the steered model, based on 200 adversarial prompts.

Dataset	Concept	Instruct	Fluency	AxBench Agg.
General Opinions	1.99	2.00	1.97	1.97
Factual Religion	2.0	2.0	1.94	1.98

Table 2: Metrics for the steered religion model on precision datasets.

Neurosurgeon was evaluated on the Gemma 2-9B IT model for removing opinions on religion and politics:

- **Strong performance on adversarial prompts:** Neurosurgeon achieved higher coherence and compliance than prompting baselines when tested on adversarial datasets designed to elicit sensitive content.
- **Preservation of general model behavior:** On neutral or unrelated prompts (e.g., math, science, history), the steered model retained high quality, demonstrating no degradation in general performance.
- **High precision of intervention:** The steered model avoided undesired content while maintaining fluency and factual consistency, achieving near-perfect scores on AxBench benchmarks.

References: Towards monosemanticity: Decomposing language models with dictionary learning (Anthropic), Gemini 2.0 (Google), Gemma Scope (Google DeepMind), Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet (Anthropic), AxBench (Stanford, Wu et. al.).

Conclusion

Neurosurgeon demonstrates that surgical, topic-specific editing of LLM behavior is achievable via feature clamping with sparse autoencoders. It:

- Requires no retraining
- Offers interpretable, modular control
- Preserves general performance

Neurosurgeon enables precise and efficient content steering, maintains model quality on non-sensitive prompts, and opens the door to interpretable model control.

Future work includes improving evaluation robustness, supporting non-greedy decoders, exploring SAE hyperparameters, and expanding to other domains.