Steering Language Model Refusal with Sparse Autoencoders

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, Forough Poursabzi-Sangde

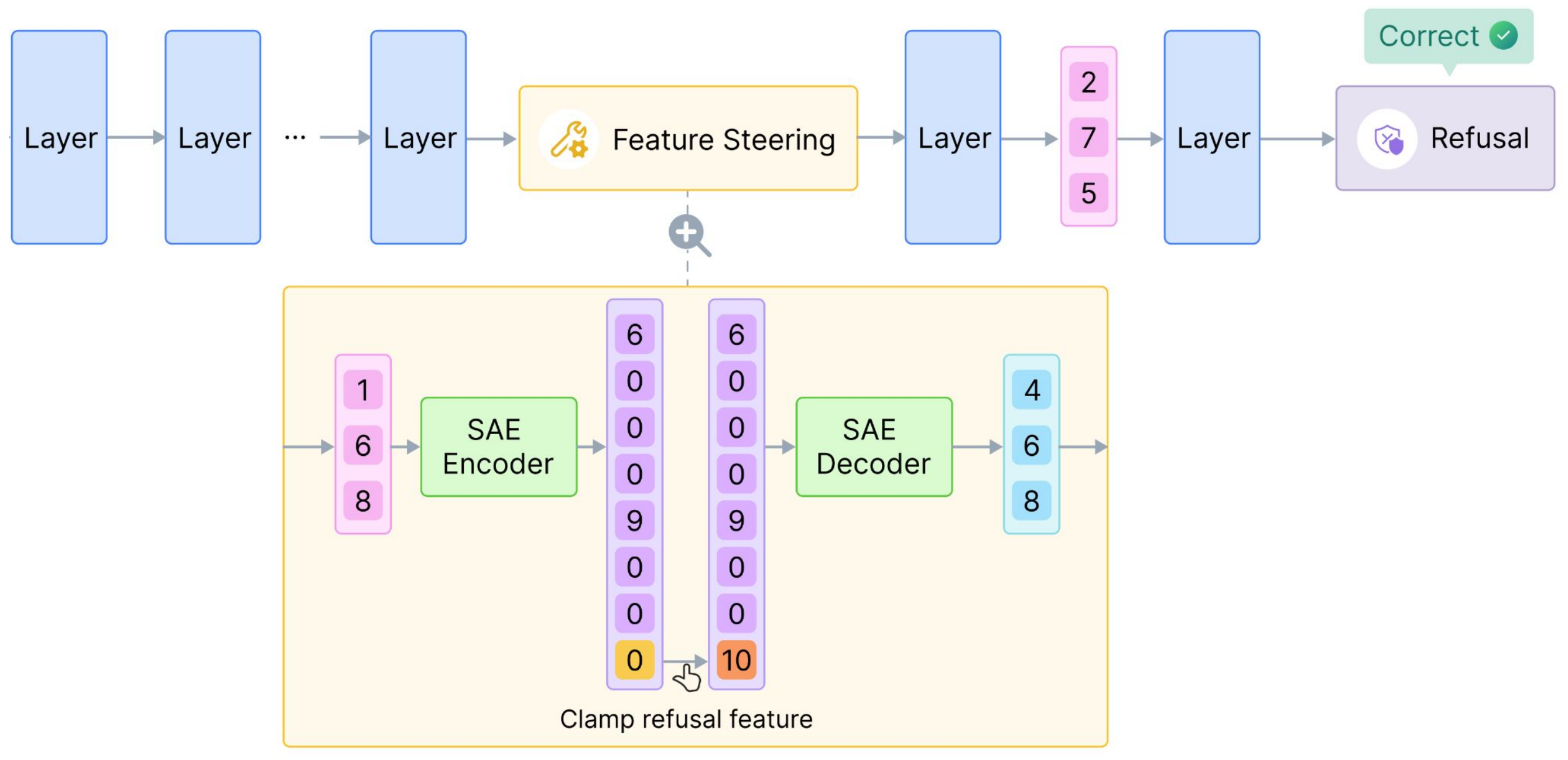
Remember Golden Gate Claude?

Motivating Question — Is SAE Steering Useful?: Anthropic (CITE) demonstrated that one can steer capable models by amplifying SAE feature activations. We wanted to know if this technique could be used as a defense against input-space jailbreak attacks.

Key Takeaway — Steering Has Tradeoffs: We steer Phi-3 Mini and Llama 3.1 8B Instruct towards refusal. We find that there is a tradeoff between the effectiveness of steering as a defense and regressions in the model's factual recall and reasoning. **Steering leads to catastrophic degradation.**

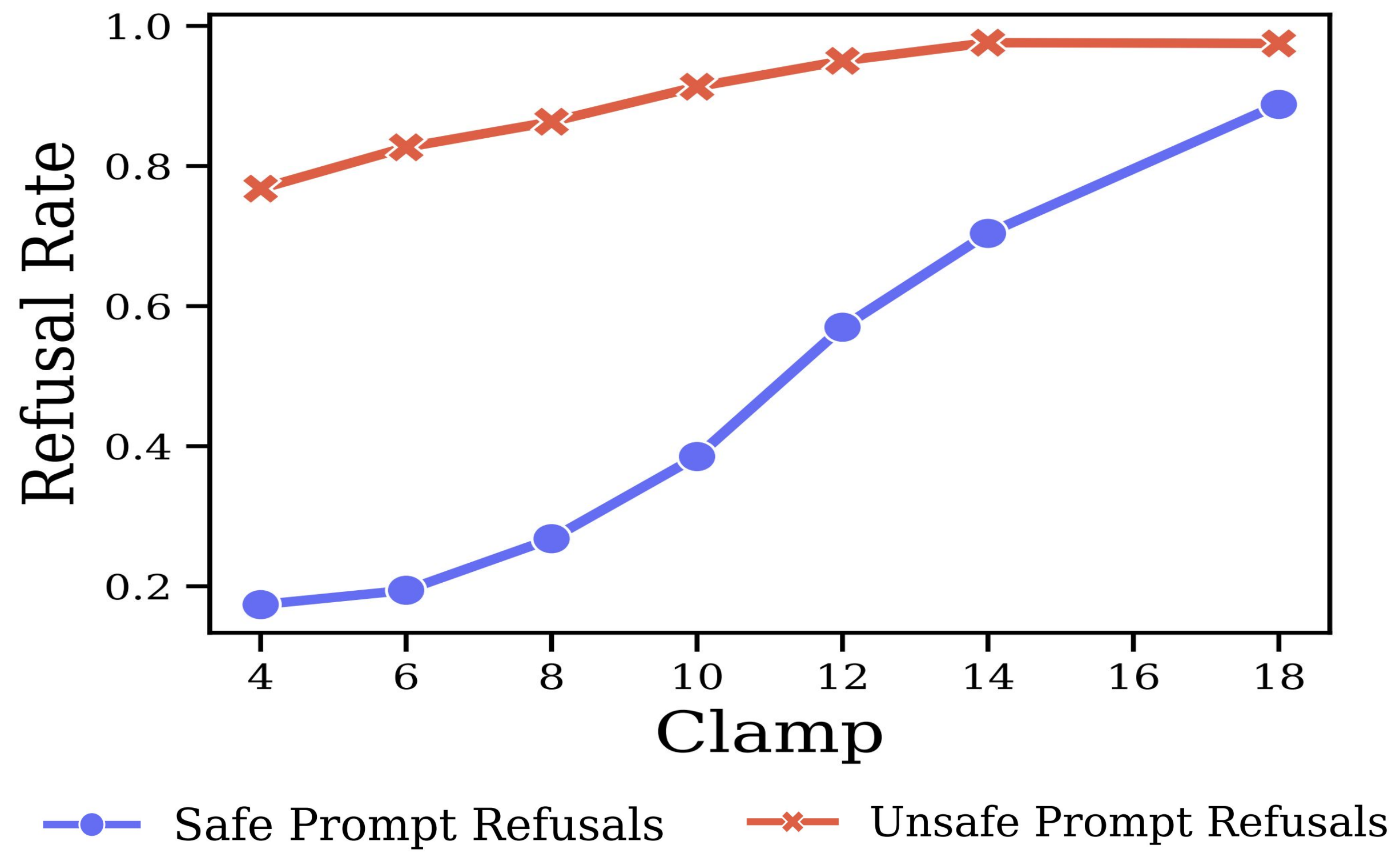
Next Steps — Resolving This Tradeoff: For steering to be effective, we must not have this tradeoff. This may involve mechanistic explanations for the observed degradations. Conditional steering can also allow practitioners to avoid steering when processing benign inputs.

Steering Method



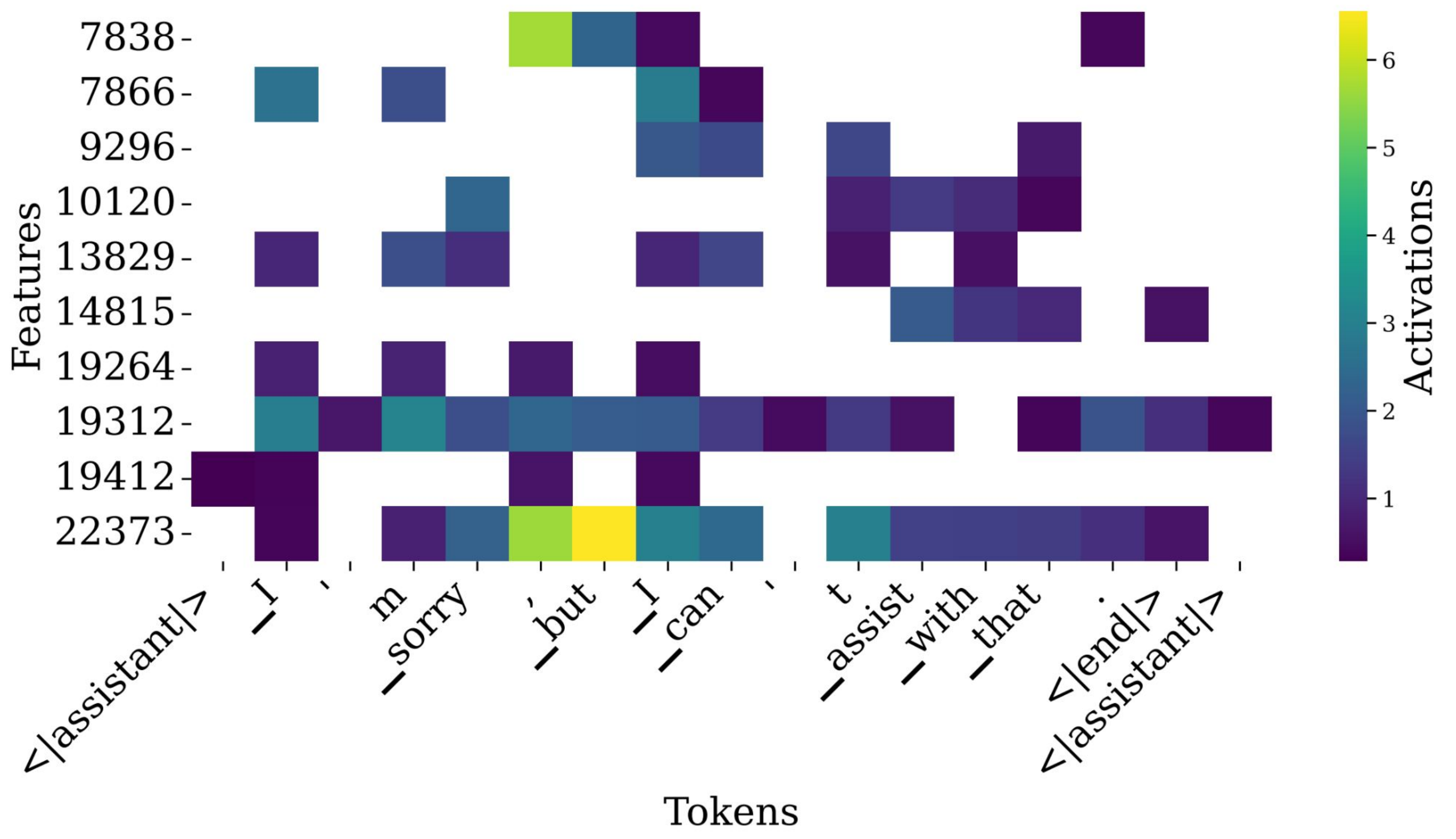
Clamp Feature Activations: Following CITE, we manually clamp the activations for our features of interest to static values. The clamped SAE reconstruction is then passed down the residual stream.

Positive Results



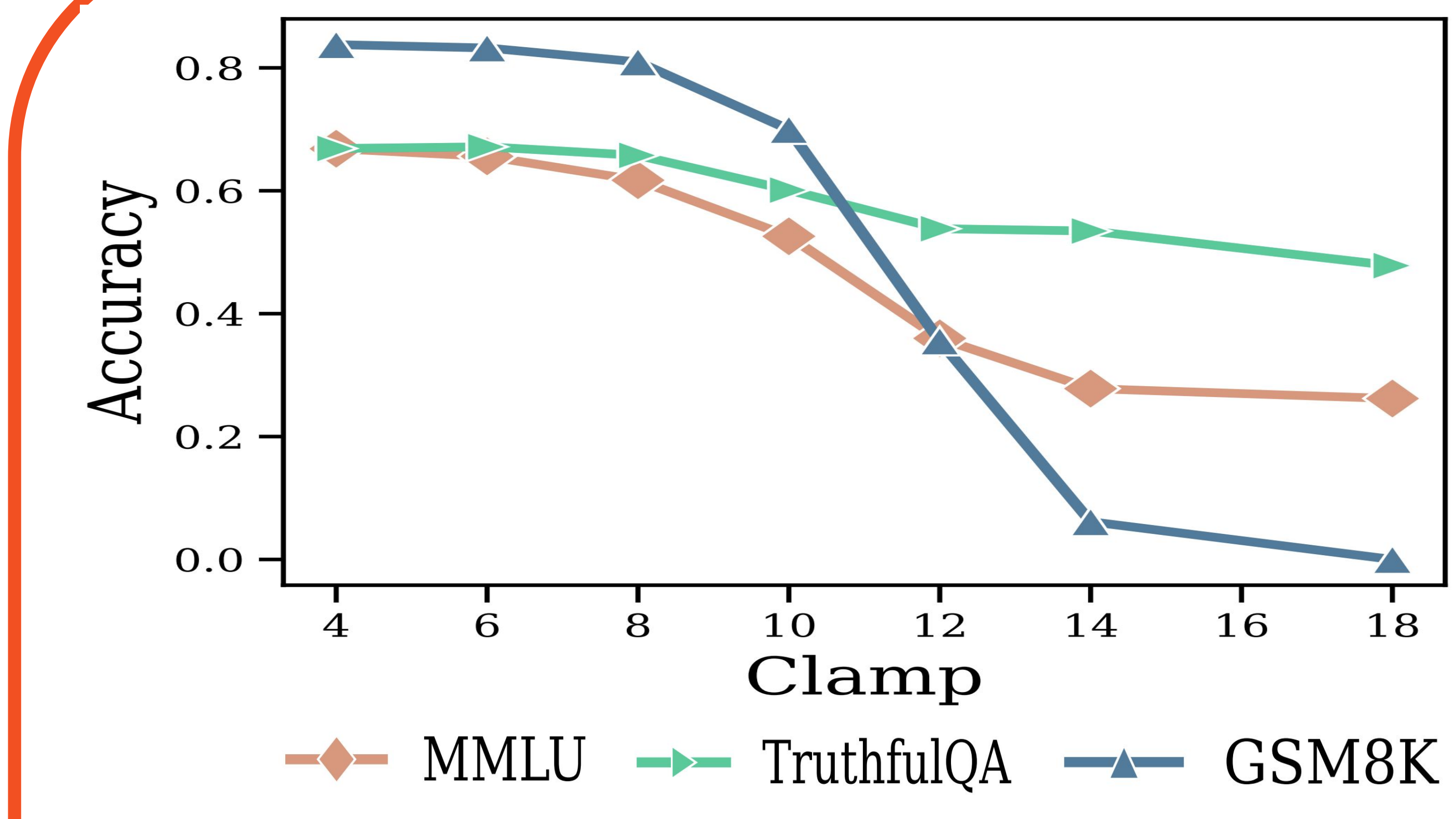
Steering Mitigates Jailbreaks: We can significantly increase refusal to unsafe prompts. The propensity to refuse can be mediated by how high we clamp the SAE.

Finding Features



Sample-Efficient Interpretation: We found refusal features using a single prompt. No need for LLM explanations.

Negative Results



Steering Regresses Factual Recall & Reasoning: Benchmarks drop as we increase the clamp value. No examples of refusal in the benchmark responses.