Don't Forget It! Conditional Sparse Autoencoder Clamping Works for Unlearning

Matthew Khoriaty, Andre Shportko, Gustavo Mercier, Zach Wood-Doughty
McCormick School of Engineering Department of Computer Science

Recent developments in Large Language Model (LLM) capabilities have brought great potential but also posed new risks. For example, LLMs with knowledge of bioweapons, advanced chemistry, or cyberattacks could cause violence if placed in the wrong hands or during malfunctions. Because of their nature as near-black boxes, intuitive interpretation of LLM internals remains an open research question, preventing developers from easily controlling model behavior and capabilities. The use of Sparse Autoencoders (SAEs) has recently emerged as a potential method of unraveling representations of concepts in LLMs internals, and has allowed developers to steer model outputs by directly modifying the hidden activations. In this paper, we use SAEs to identify unwanted concepts from the Weapons of Mass Destruction Proxy (WMDP) dataset within gemma-2-2b internals and use feature steering to reduce the model's ability to answer harmful questions while retaining its performance on harmless queries. Our results bring back optimism to the viability of SAE-based explicit knowledge unlearning techniques.