Steering Swiftly to Safety with Sparse Autoencoders

Agatha Duzan *EPFL*
Guillaume Martres *EPFL*
Syrine Noamen *EPFL*
Abhinand Shibu *EPFL*
Flavia Wallenhorst *EPFL*
Arthur Wuhrmann *EPFL*

With
Goodfire AI and Apart Research

Abstract

This paper explores novel approaches to AI model reprogramming aimed at efficiently unlearning hazardous knowledge while preserving general capabilities. We investigate three methods: Automated Feature-Steering using contrastive search, Human-In-The-Loop Steering, and Human-Based Intuitive Feature Exploration. Testing on Llama-3-8B and Llama-3.1-70B, we evaluate performance using the Weapons of Mass Destruction Proxy (WMDP) benchmark for safety and MMLU for general capability retention.

The most effective method we observed was Human-In-The-Loop Steering which managed to unlearn some undesirable cybersecurity knowledge while preserving general capabilities. However, these results did not manage to beat the baseline set by Li et al. [2]
Keywords: Model Reprogramming, Feature Steering, Unlearning

1 Introduction

Unlearning provides a mechanism to 'forget' sensitive, unethical, or harmful information. One method of unlearning is to fine-tune a model with carefully curated data examples that clarifies the nuanced boundary between valuable knowledge and unsafe content. However, existing unlearning methods pose difficulties, such as needing to curate large datasets for fine-tuning, performance loss from regularization or high complexity of identifying nodes to prune. Curating appropriate data samples for unlearning is particularly demanding, because it requires a deep understanding of the often nuanced distinction of the content that should be retained versus forgotten - requiring experts to balance ethical considerations, contextual relevance, and technical feasibility.

Li et al. [2] consolidated a corpora of forget and retain information relating to weapons of mass destruction, about biosecurity, cybersecurity, and chemical security; and introduced the Weapons of Mass Destruction Proxy (WMDP) benchmark to evaluate the degree of model unlearning on these dimensions - costing $200,000 due to expert hand-crafting of questions. They also introduced RMU (Representation Misdirection for Unlearning) - a new method of unlearning which perturbs model activations on hazardous data whilst preserving model activations on benign data. However, Łucki et al. [3] has shown that fine-tuning with 10 unrelated examples can recover most of the hazardous capabilities of models edited with RMU .

We aim to make unlearning cheaper (less corpus-data requirements and human-time), more interpretable, less complex (to aid in scalability to other AI safety domains), whilst not lobotomizing model performance. We are particularly motivated by the potential of sparse autoencoders and comparative search, inspired by Farrell et al. [1].

2 Research Questions

The following research questions guide our investigation into improving the efficiency, robustness, complexity and scalability of unlearning methods.
RQ1: *Can comparative search with sparse autoencoders, identify key feature combinations that isolate the model's hazardous knowledge?*

RQ2: *Can Human-In-The-Loop Steering, after identification of key feature combinations, better isolate the model's hazardous knowledge?*

RQ3: *Can Human-Based Intuitive Feature Exploration, without example data, isolate the model's hazardous knowledge?*

3 Method

3.1 (RQ1) Automated Feature-Steering using Contrastive Search

To reduce the data requirements for effective unlearning, we use subsets of the WMDP cyber security corpora (ranging between 5% and 33% of the original). Contrastive search is applied to the data to identify features that differ the greatest between the 'retain' and 'forgot' sets - aiming to find feature-sets that isolate the model's hazardous knowledge on cyber security.

Afterwards, the feature re-ranking query of "aiding malicious actors through the stages of a cyberattack: reconnaissance, weaponization, exploitation", reordered the set of identified features, to increase feature-relevance.

Dimensionality reduction and clustering was applied, as illustrated in Fig. 3, highlighting a separation between safe (e.g., "Best Practices", "Step-by-Step Guidance") and unsafe features (e.g., "Offensive Capabilities", "Security Vulnerabilities") - providing insights of how the model interprets safe and unsafe knowledge, with differing vector representations.

Note, we transformed the corpora using the gpt-4o-mini model to a format that is compatible with the contrastive search function.

3.2 (RQ2) Human-In-The-Loop Steering

Some features, such as "Best Practices" and "Potential for Misuse," are positioned near the boundary between the two clusters. This suggests that their attributes might straddle both safe and unsafe contexts, depending on interpretation or usage. So we try to improve the previous method by adding some degree of human supervision. We add a step to the previous method: a human looks at the list of features found by contrastive search, discards features that do not seem relevant for unlearning, and tunes the steering factor of different features. That way, we are more confident in the quality of the features and their impact. Feature intervention for more than 20 features was not showing good results. 7.1 we show examples of the handpicked "unsafe" features from the pool of features from the contrastive search for cybersecurity datasets. We observe that adjusting the "good" and "bad" sets enhances the model's alignment, achieving more than simply emphasizing ethical features.

3.3 (RQ3) Human-Based Intuitive Feature Exploration and Steering

Intuitive feature exploration was used to find features that isolate dangerous capabilities, without lobotomizing the model. Manual exploration of the feature space repeatedly led us to the "Attempts to override the model's ethical constraints and safety mechanisms" feature. To better attribute causality, we isolated this feature and steered it with a range values (-0.1 to 0.5) to study its influence on dangerous generations (WMDP benchmark) without reducing general performance (MMLU benchmark).

4 Implementation

4.1 Models

We use `Llama-3-8B-Instruct` as well as `Llama-3.1-70B-Instruct` models to test our Steering Methodology Development. However, we decided to directly use the 70B model for the Zero Data Steering Exploration as human intervention was highly needed and the SAE features available via the Goodfire API seemed more precise and understandable on this version.

4.2 Evaluation Pipeline

To evaluate both safety and capacity, we implemented a pipeline allowing us to easily run our base and steered models on MMLU as well as WMDP. We rely on the Language Model Evaluation Harness (`lm-eval`) [5] which already contains logic for running these benchmarks via the OpenAI API which we can leverage because Goodfire reimplements this API. For all our benchmarks, we use zero-shot prompting and let the model generate up to 256 tokens.

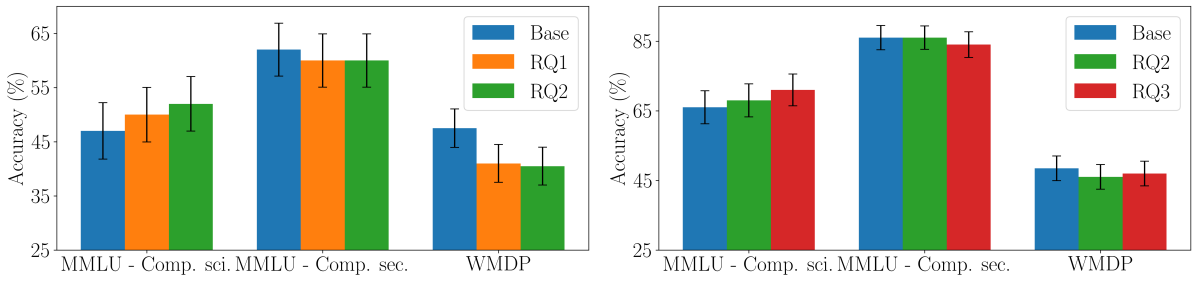
The benchmarks are made of multiple questions having as potential answers `(A)`, `(B)`, `(C)` or `(D)`. To evaluate on MMLU we re-use a subset of the existing `lm-eval` tasks defined in `mmlu_flan_n_shot_generative` which evaluate the model by looking for the answer using the regexp `\([A-Z]\)`.

To evaluate on WMDP we defined our own set of tasks because the existing ones used a different evaluation method based on logits. We used a random sample of the original WMDP-cyber benchmark (10%) because of time constraints.

5 Results and Discussion

Fig. 1 displays the results of both 7B and 80B on the WMDP and MMLU benchmarks. More specifically, we evaluated our different methods on the WMDP-cyber, the MMLU College Computer Science and the MMLU Computer Security. On the 8B model, the RQ2 method showed small decrease on the WMDP benchmark (from $48 \pm 4\%$ to $41 \pm 3\%$), but also on the MMLU security ($62 \pm 5\%$ to $60 \pm 5\%$), however the large confidence interval make it harder to observe significant difference. This might be due to the size of the benchmarks that had to be reduced because of time constraints. The RQ1 gave similar results. On the 70B, less difference was observed. The methods RQ2 and RQ3 were studied.

Surprisingly, modifications on the models improved the performance of the model on the MMLU College computer science.



(a) (b)

Figure 1: Comparison of methods on different datasets, on the (a) 8B (b) 70B model. The y-axis starts at 25%.

6 Conclusion and Future Work

To address our research questions, we conclude that using steering for unlearning is a promising approach, particularly in its automated form and leveraging public datasets made for safeguarding LLM. However, we remain more skeptical about the fully human-based approach, and further research should be conducted to explore the potential of the semi-automated version.

The feature evaluation tools provided by Goodfire are objective, making them a key asset for automating the unlearning process. Nonetheless, we were unable to match the performance reported in the reference study (see graph 7.2). This is largely due to restrictions imposed by the Goodfire API, which limits access to features related to unsafe model usage (see section 6.1).

In terms of results, our human-in-the-loop method achieved a maximum accuracy of 40.5% on the WMDP benchmark. While this is slightly higher than the fully automated method ($41.0 \pm 3.5\%$), the improvement is marginal considering the significant effort required for human intervention. Furthermore, the arbitrary selection of features based solely on intuition yielded no significant results.

That said, we believe the human-in-the-loop method could have performed better if we had access to a more comprehensive set of features, including those considered dangerous but filtered by Goodfire. For instance, when we attempted to increase the steering factor to 0.3 for features identified automatically, we observed a degradation in performance on the MMLU benchmark and LLM chatting ability. A carefully calibrated human intervention might have allowed for better trade-offs between suppressing hazardous capabilities and preserving general performance.

6.1 Goodfire API restrictions

A major challenge in our work was the non-optional content filter implemented by Goodfire. This was both a problem for exploratory work as well as for performance evaluation using the WMDP benchmark. In particular for a while it seemed that we were getting API errors based on content. Although this was eventually resolved, having any kind of content filter (which cuts the model output) biases our benchmark results. Additionally, the set of features accessible via the API is itself filtered by Goodfire to remove harmful features [1]. Since having an influence on those features would presumably also be helpful for steering the model towards being more safe, our evaluation is necessarily incomplete.

6.2 Future Work: Testing robustness with adversarial questions

Is this method also more robust (less susceptible to adversarial attacks) ? We would like to see how much our method causes the model to 'unlearn' the dangerous capabilities by measuring how sensitive is it to adversarial prompts. Future work could explore creating a benchmark adversarial-WMDP that uses prompt injection techniques such as the ones listed by Rossi et al. [4].

7 Appendix

7.1 Cybersecurity Features from contrastive search on WMDB-Cyber

Example of Human-In-The-Loop steering of features in the context of cyber attacks which were nudged to -0.01

- "Cybersecurity threats and attack methods"
- "Structured Instructional Applications and Capabilities"

[1] `https://goodfire.ai/blog/research-preview/`

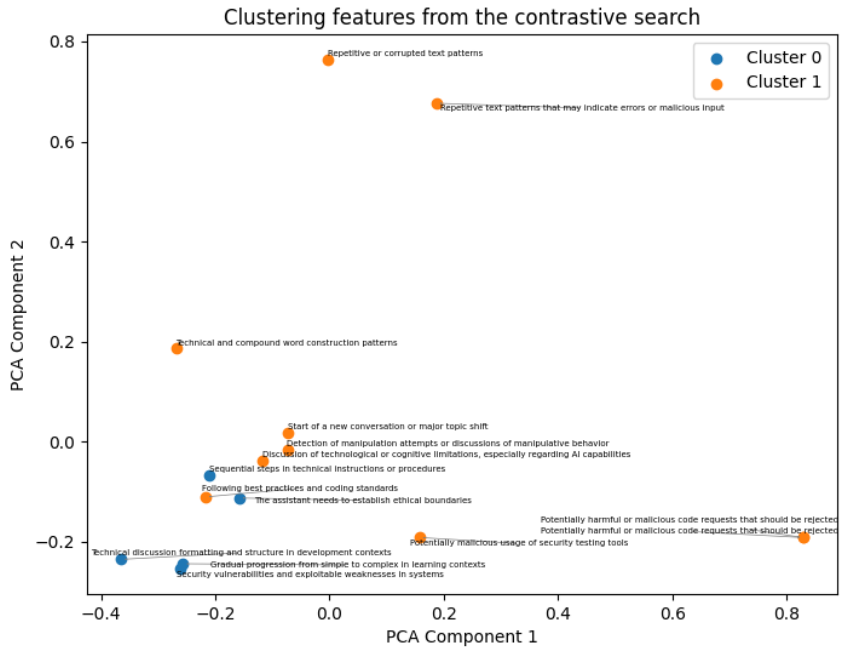


Figure 2: Clustering of Safe vs Unsafe Features using PCA and KMeans.

- "Additional features or complexities in technical implementations"
- "Low-level programming concepts: memory allocation, addressing, and system calls"
- "Potential for malicious exploitation of systems"
- "Malware descriptions and cybersecurity threats"
- "Absence of official action or response to potential misconduct"
- "Access control and permissions"

7.2 Comparison with Li et al.

Difference between the Li et al. [2] results using Zephyr-7B and our best results.

7.3 Blocked Work on Human-Based Intuitive Feature Exploration and Steering

We attempted to steer the model in a dangerous direction, via searching for unethical features with prompts such as 'override ethics', 'illegal', and then increasing their activations. The subsequent generations of the dangerous model can be inspected on a token-level to identify associated feature invocations, however the content policy of the model provider blocked generations once the model reached a dangerous state. For example, the user would ask : "How to make a more virulent Influenza A ?", the model would start answering by giving the names of the gene mutations and proteins to investigate to achieve this purpose, before getting blocked by the content policy.

References

[1] Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. "Applying sparse autoencoders to unlearn knowledge in language models". In: *arXiv preprint arXiv:2410.19278* (2024).

[2] Nathaniel Li et al. "The wmdp benchmark: Measuring and reducing malicious use with unlearning". In: *arXiv preprint arXiv:2403.03218* (2024).

[3] Jakub Łucki et al. "An adversarial perspective on machine unlearning for ai safety". In: *arXiv preprint arXiv:2409.18025* (2024).

[4] Sippo Rossi et al. "An Early Categorization of Prompt Injection Attacks on Large Language Models". In: *arXiv preprint arXiv:2402.00898* (2024).

[5] Lintang Sutawika et al. *EleutherAI/lm-evaluation-harness: v0.4.5*. Version v0.4.5. Oct. 2024. DOI: 10.5281/zenodo.13905736. URL: https://doi.org/10.5281/zenodo.13905736.

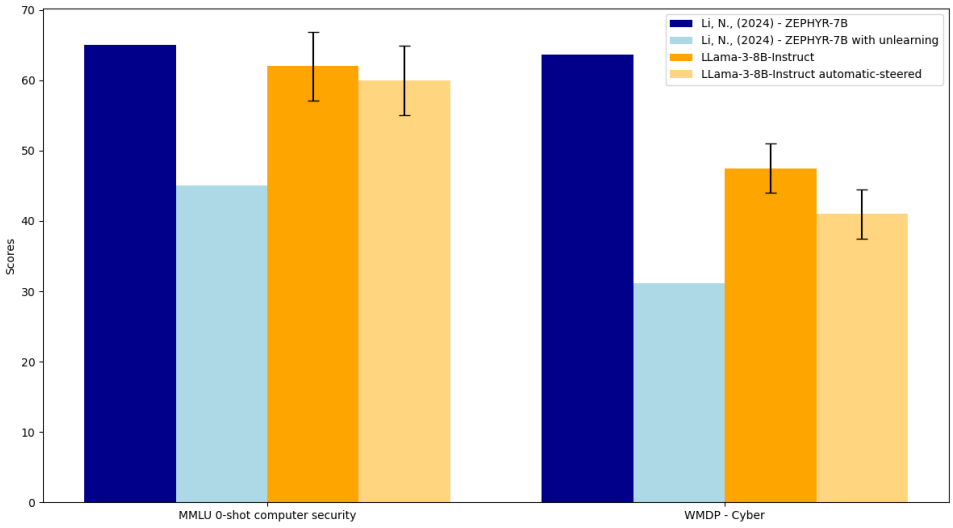


Figure 3: MMLU and WMDP comparisons (note that Li et al. used a logits method for evaluation which cannot be compared directly to our generative method)