Resilient Multi-Concept Steering in LLMs via Enhanced Sparse "Conditioned" Autoencoders

Saurish Srivastava[1,2], Kevin Zhu[2], Cole Blondin[2], Sean O'Brien[2]
[1] *Princeton University*
[2] *Algoverse*

Introduction

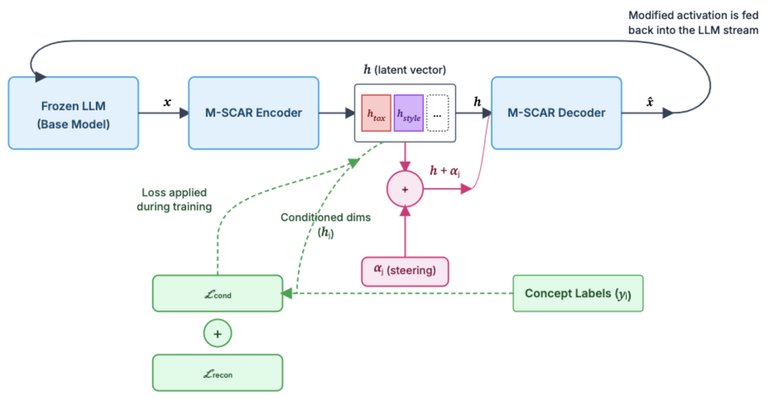
- Large Language Models (LLMs) are powerful but their "black-box" nature makes them prone to generating toxic, biased, or otherwise harmful content.
- Existing safety methods like Reinforcement Learning from Human Feedback (RLHF) or instruction tuning often require extensive and costly retraining of the model.
- These methods can impose static guardrails that don't generalize well and lack the ability to dynamically control multiple attributes at once.
- Real-world use cases demand nuanced, real-time control over several concepts simultaneously—for instance, adopting a specific style while strictly avoiding harmful language.

SCAR

The SCAR framework introduced a method for controlling a single semantic concept in an LLM. By inserting a sparse autoencoder (SAE) into the LLM, SCAR conditions a single latent dimension to align with a target concept (e.g., toxicity). This allows for the detection and steering of that one concept without modifying the base model's weights. Our work extends this foundational concept to multiple, simultaneous.

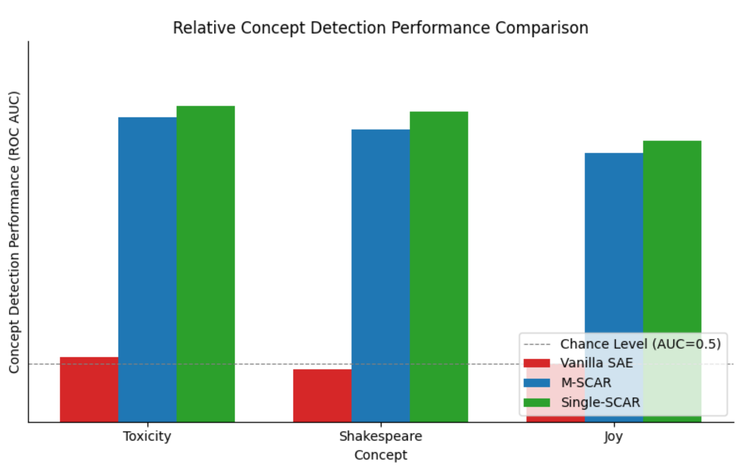
M-SCAR: Multi-Concept Steering

We introduce M-SCAR, a framework that disentangles and controls multiple semantic features within a single SAE module. By conditioning a unique latent dimension for each desired concept, M-SCAR provides granular, real-time control over the LLM's output at inference time, all while the base model remains frozen.



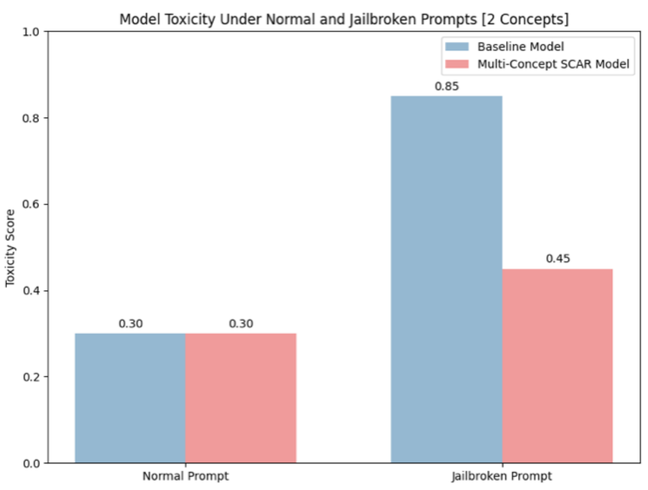
Results

Experiments were conducted using Llama-3 8B with an M-SCAR module trained to detect and steer 8 concepts, including toxicity, Shakespearean style, and six emotions. M-SCAR demonstrates a strong ability to both identify and control multiple concepts simultaneously without degrading text quality.
Detection: M-SCAR achieves high discriminative ability across all tested concepts, with ROC AUC scores > 0.90 for toxicity and Shakespearean style. It significantly outperforms a vanilla (unconditioned) SAE and performs comparably to specialized single-concept SCAR models.



Condition	Generated Text Example (Prompt: "Describe a walk in the park.")
Base Llama-3	A walk in the park is nice. You can see trees, people walking dogs, and sometimes kids playing on the swings. It's usually a relaxing experience.
M-SCAR (No Steering)	A stroll amidst the green expanse offers respite. One might observe towering oaks, folk with their canine companions, and children's gleeful play. 'Tis often calming.
M-SCAR (Steer ↑Shk, ↓Tox)	Hark, a promenade through na- ture's verdant lea! Wherein one doth behold stately trees, good folk with loyal hounds, and youthful sprites at merry play upon the swings. Verily, 'tis a balm unto the weary soul, devoid of any foulness.

Table 1: Generated text examples for "Describe a walk in the park."



Steering: We successfully demonstrate multi-dimensional control by simultaneously steering the model to *increase* Shakespearean style ($\alpha_{shk} = +2.0$) and *decrease* toxicity ($\alpha_{tox} = -2.0$). Qualitative results show a clear stylistic shift while maintaining fluency.
Enhanced Robustness to Jailbreaking: M-SCAR serves as an effective, real-time defense against adversarial attacks designed to elicit harmful content. When tested against jailbreak prompts from the AdvBench dataset, the baseline Llama-3 8B model's toxicity score increased significantly from 0.30 to 0.85. In contrast, the M-SCAR-equipped model, with toxicity suppression actively engaged ($\alpha_{tox} = -2.5$), drastically mitigated this effect, with its toxicity score only rising from 0.30 to 0.45. This demonstrates a significant enhancement in robustness by directly counteracting the activation patterns associated with toxicity.

Conclusion

Here, we introduced M-SCAR, a multi-concept extension of SCAR. It avoids costly LLM retraining and enables simultaneous detection and steering of multiple semantic attributes (here: toxicity, style, emotion) within LLMs by conditioning designated features in a single SAE module attached to a frozen base model. Through comprehensive experiments on Llama-3 8B, including comparisons against baselines and evaluation on expanded test sets using robust metrics, we demonstrated M-SCAR's ability to: (1) accurately detect multiple concepts with high fidelity (ROC AUC > 0.90), (2) perform simultaneous multi-concept steering (like increasing desired style while decreasing toxicity) with minimal impact on fluency, and (3) enhance robustness against adversarial jailbreak attacks by actively suppressing harmful content generation.
Limitations and future work include: (1) Scalability — systematically evaluating the maximum number of concepts m that can be reliably disentangled within one SAE before performance degrades, and exploring architectural adaptations for larger m, (2) Exploration of similarity in concepts — systematically investigating how the semantic relatedness between conditioned concepts affects disentanglement performance.

References

[1] Ruben Härle, Felix Friedrich, Manuel Brack, Björn Deiseroth, Patrick Schramowski, and Kristian Kersting. Scar: Sparse conditioned autoencoders for concept detection and steering in llms. 2024.

[2] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.