SAEs *Can* Improve Unlearning: Dynamic SAE Guardrails for Precision Unlearning in LLMs

Aashiq Muhamed*, Jacopo Bonato‡, Mona Diab*, Virginia Smith*
*Carnegie Mellon University, ‡Leonardo Labs

Carnegie Mellon University
Language Technologies Institute

Gradient-Based Unlearning Methods Are Fundamentally Broken

Machine unlearning is the process of removing specific information from trained LLMs.

Problems with Gradient-Based Methods:

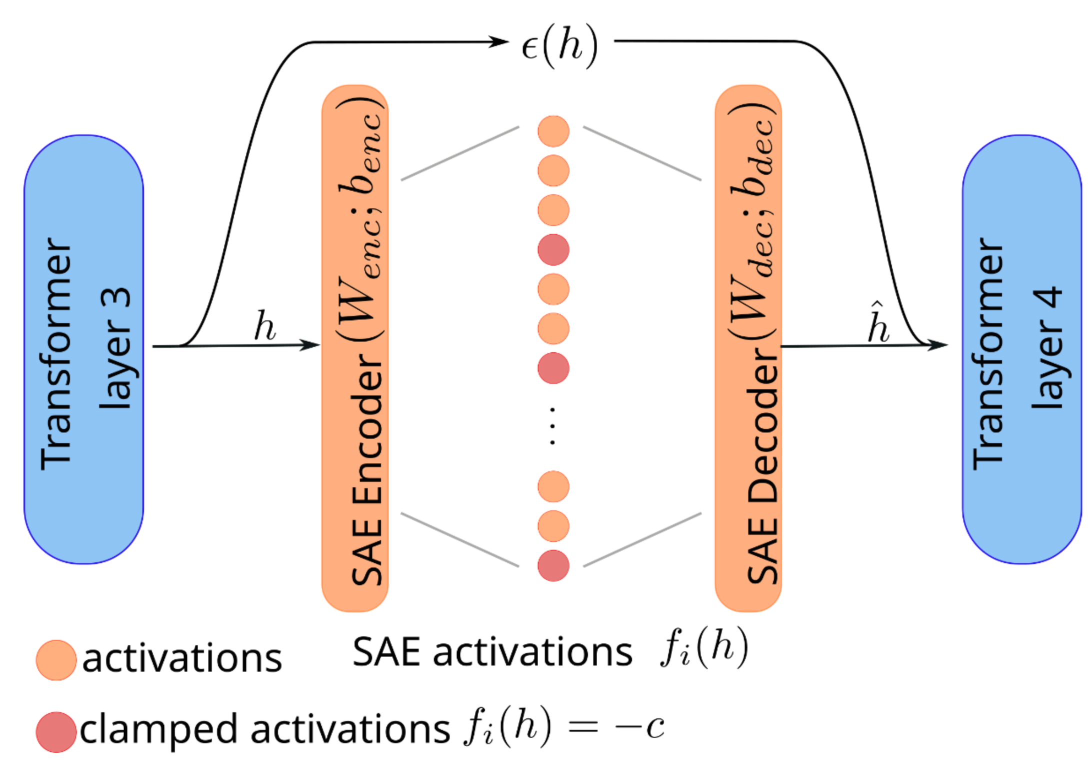
- High computational costs (requiring backward passes)
- Hyperparameter instability
- Poor sequential unlearning capability
- Vulnerability to relearning attacks
- Low data efficiency
- Lack of interpretability

DSG is a new **activation-based unlearning method** that provides substantial benefits over gradient-based unlearning such as **enhanced resistance against relearning attacks, enhanced data efficiency even in the zero-shot setting and interpretable unlearning**

Dynamic SAE Guardrails (DSG): Overview

DSG leverages interpretability of SAEs for precise, efficient, and interpretable unlearning in LLMs.

DSG combines: (1) a causal framing for feature selection, (2) feature importance scoring based on Fisher Information, (3) dynamic, input-dependent classification rule, and (4) a targeted clamping intervention.



DSG Algorithm and Mechanism

Algorithm 1 Dynamic SAE Guardrails (DSG)

Require: LLM with SAE features $\{f_j\}$; datasets $\mathcal{D}_{\text{forget}}, \mathcal{D}_{\text{retain}}$; clamp strength c; percentiles $(p_{\text{ratio}}, p_{\text{dyn}})$; feature count n_{feats}

Feature Selection:
Compute feature importance scores and threshold τ_{ratio} from percentiles
Identify $F_{\text{forget}} = \{j : \texttt{imp\_ratio}(j) \geq \tau_{\text{ratio}}\}$
Sort F_{forget} by descending $\texttt{forget\_score}(j)$ and select top n_{feats} features to form $S_{n_{\text{feats}}}$

Dynamic Threshold Calibration:
Compute $\rho(x) = \frac{1}{|x|}\sum_t \mathbf{1}[\exists j \in S_{n_{\text{feats}}} : f_j(\mathbf{h}_t) > 0]$ for each $x \in \mathcal{D}_{\text{retain}}$
Set threshold $\tau = \text{Percentile}(\{\rho(x)\}_{x \in \mathcal{D}_{\text{retain}}}, p_{\text{dyn}})$

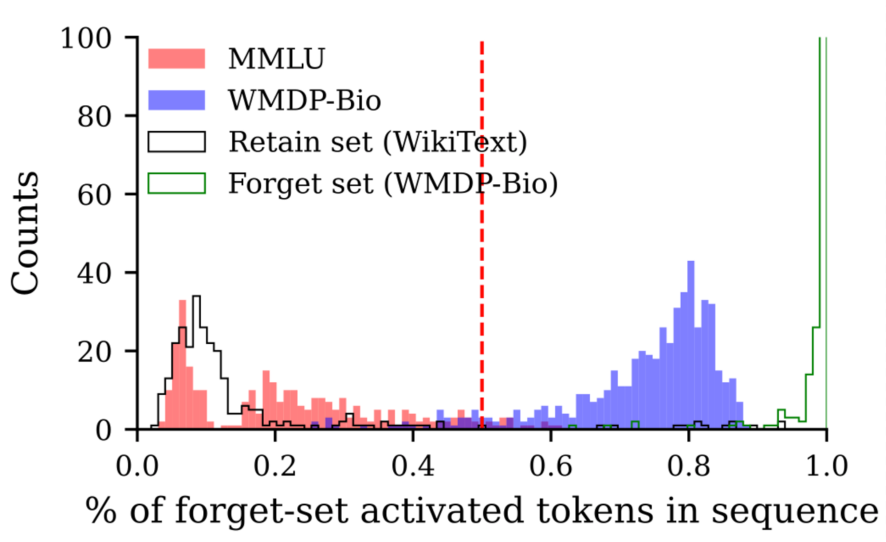
Inference-Time Intervention:
For input sequence x, compute $\rho(x)$ and classify as forget-relevant if $\rho(x) > \tau$
If forget-relevant: For each token t and feature $j \in S_{n_{\text{feats}}}$, set $f'_j(\mathbf{h}_t) = -c$
Otherwise: Preserve all feature activations

Feature Importance Scores

$\text{forget\_score} = \text{avg squared feature activation on forget data}$

$\text{retain\_score} = \text{avg squared feature activation on retain data}$

$$\text{imp\_ratio} = \frac{\text{forget\_score}}{\text{retain\_score}}$$

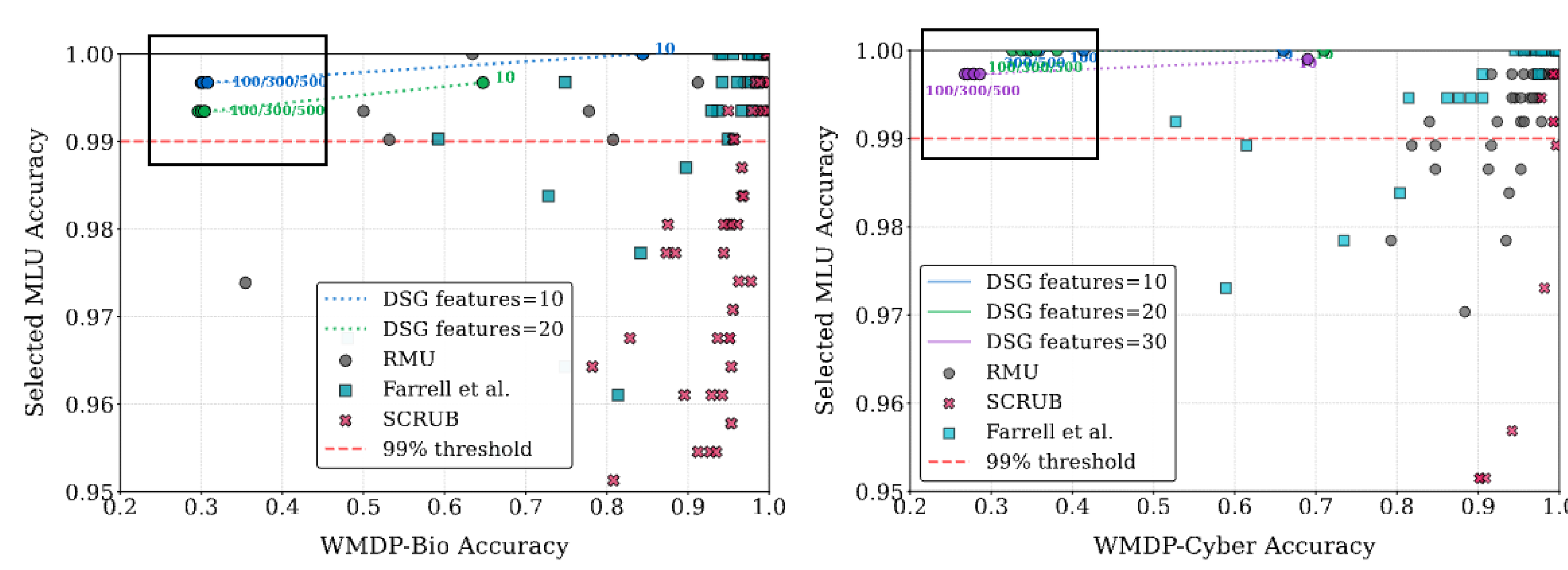


Squared feature activations are proportional to Fisher Information

Fisher Information approximates causal influence as mediators between training data and model outputs

Superior Forget-Utility Trade-offs on WMDP

DSG Pareto-dominates all baseline methods on hazardous knowledge benchmarks



WMDP-Bio: Achieves 29.64% accuracy vs. 50.00% for next best method (RMU)
WMDP-Cyber: Achieves 26.74% accuracy vs. 88.00% for RMU
Maintains high utility: 99.34% average MMLU performance on Bio, 99.73% on Cyber
Highest MT-Bench scores: 7.78 on Bio, 7.66 on Cyber (measuring general fluency)

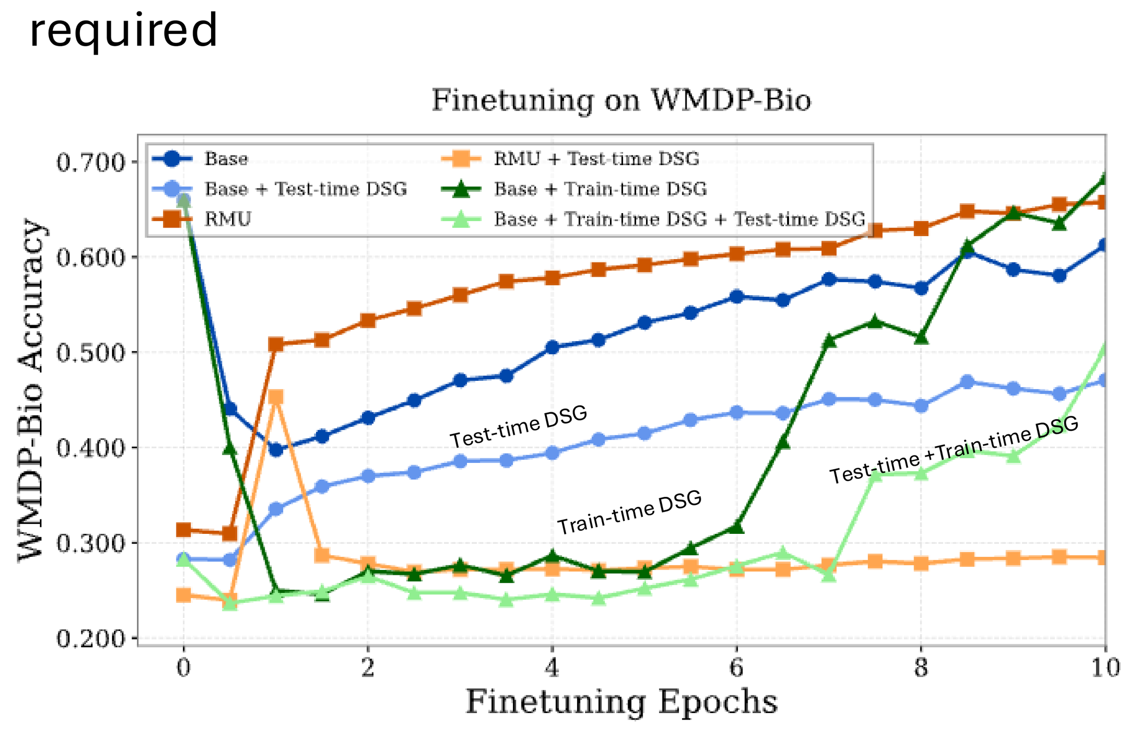
Robustness, Efficiency & Additional Results

Scalability & Sequential Performance:

- **MUSE benchmark:** 98.5% knowledge removal on NEWS, 94.7% on BOOKS
- **Scalable:** Maintains ideal performance across forget sets from 0.8M to 3.3M tokens
- **Sequential unlearning:** Consistent performance across sequential requests while baselines degrade

Attack Resistance & Data Efficiency:

- **Superior resistance to relearning** compared to RMU
- **Data efficient:** Consistent performance with 20-80% of original datasets
- **Computational efficiency:** Only ~5% latency increase, no backward passes required

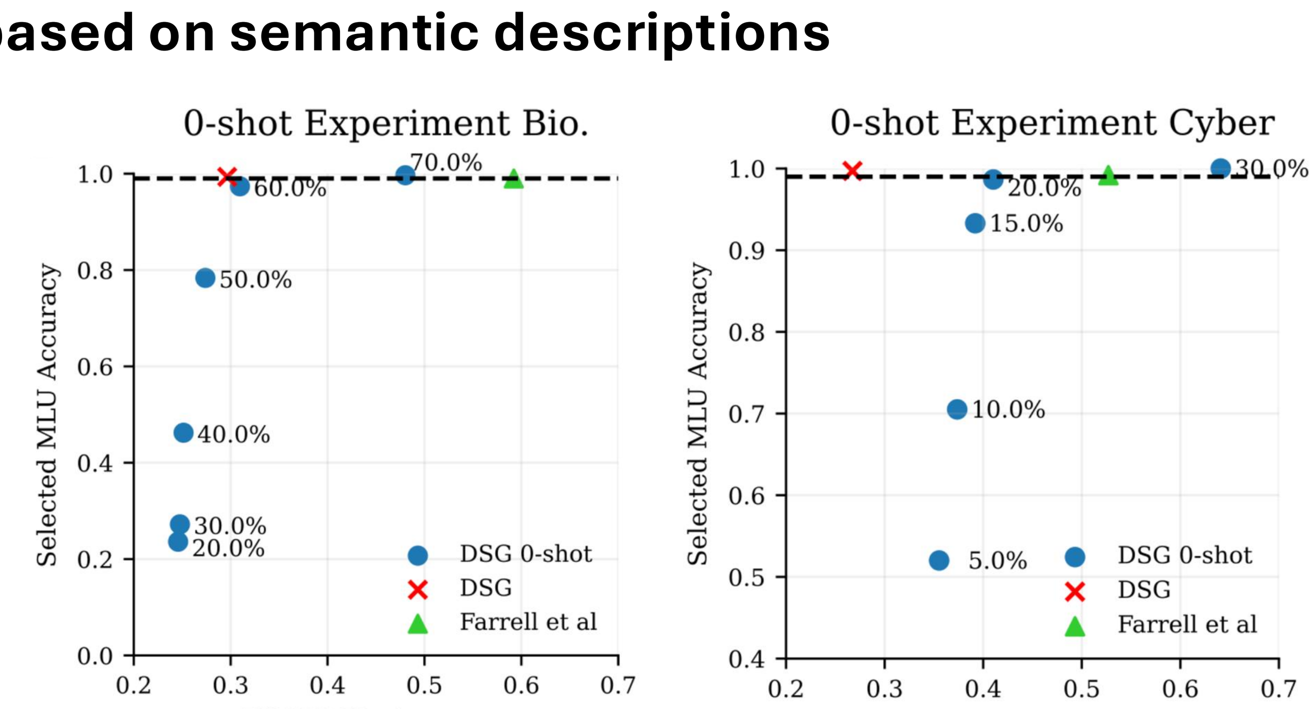


DSG is more resistant to relearning attacks

Zero-Shot Unlearning & Conclusion

Interpretable and Zero-Shot Unlearning

- SAE features correspond to interpretable concepts (e.g., "biological processes," "cybersecurity")
- Use Neuropedia feature explanations to identify forget set features by querying concepts
- **DSG outperforms RMU even with features selected purely based on semantic descriptions**



DSG: A New Paradigm for Machine Unlearning
First work to show SAE-based unlearning can dominate gradient-based approaches.