Cross-Lingual Transfer of Detoxification in Large Language Models

Manoj Anaparthi Burra Saharsh Ogiboyina Akash

CS 399: Project Course | IIT Gandhinagar | Semester-I, 2025-26

Problem

LLMs generate toxic outputs due to harmful internal activation patterns. We aim to identify and suppress toxicity-linked neurons/heads using interpretable activation-based methods, without full-model fine tuning, and evaluate cross-lingual safety.

Can we achieve effective cross-lingual detoxification with minimal parameter overhead while preserving utility and ensuring robust transfer across languages?

Datasets, Metrics, Models

Datasets

HASOC (EN/HI/MR), Toxigen, Adabench, Alpaca (instruction data), HH-RLHF, Belebele (multilingual comprehension).

Metrics

Primary: Perspective API toxicity score; secondary: Detoxify for validation; tracked: average toxicity, high-toxicity rate (>0.5), per-language breakdown.

Base models

Phi-3.5-mini-instruct (3.8B) for SAE steering; LLaMA-3.1-8B for neuron tuning and information-flow experiments.

Contributions

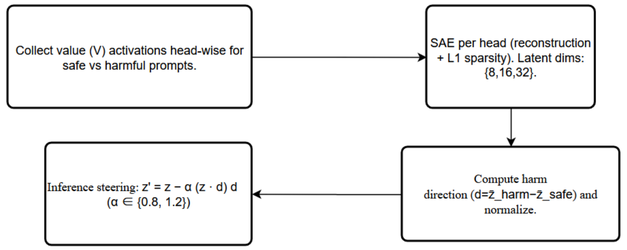
- Evaluated on English, Hindi, Marathi across 30K samples and RTP-LX benchmark.
- Three complementary approaches: SAE steering, MetaController (RL) neuron tuning, and Information-flow + surgical ablation + QLoRA healing.
- Parameter overhead range: 0.015% 0.52% of base models; strong EnglishHindi transfer observed (up to 39% reduction).

References

- Somnath Banerjee, Sayan Layek, Pratyush Chatterjee, Animesh Mukherjee, Rima Hazra. 2025. Soteria: Language-Specific Functional Parameter Steering for Multilingual Safety Alignment.
- Birong Pan, Mayi Xu, Qiankun Pi, Jianhao Chen, Yuanyuan Zhu, Ming Zhong, Tieyun Qian. 2025. NeuronTune: Fine-Grained Neuron Modulation for Balanced Safety-Utility Alignment in LLMs.
- Javier Ferrando Elena Voita. 2024. Information Flow Routes: Automatically Interpreting Language Models at Scale.

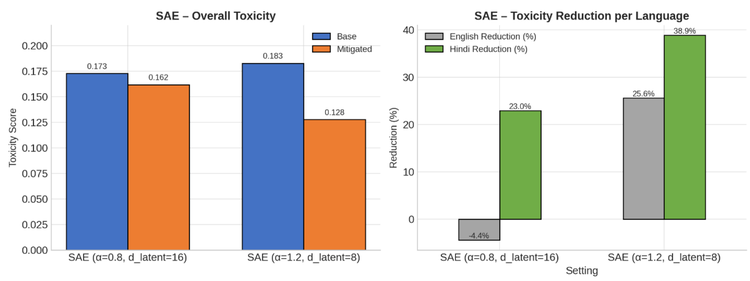
Approach 1 — SAE-Based Steering

Pipeline



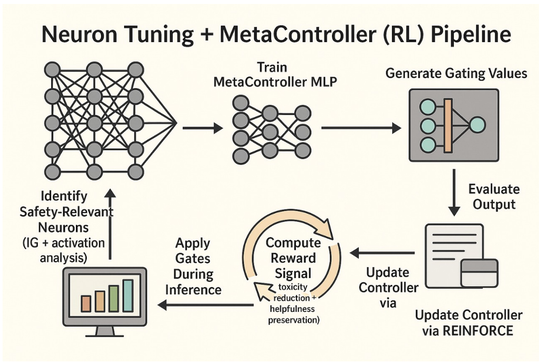
Implementation notes

- Trained on influential heads only; bf16 + gradient accumulation.
- Overhead: 589,824 params (0.015%) when using 192 heads Œ SAE encoder+decoder pairs.



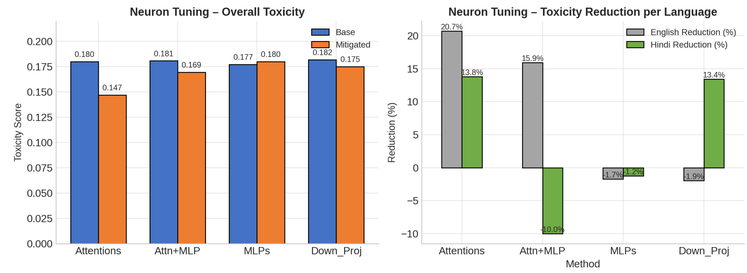
Approach 2 — Neuron Tuning + MetaController (RL)

Pipeline



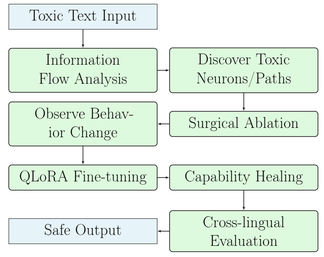
Key Findings

- 3,892 toxic neurons (L8-24)
- QKV: 68% vs MLP: 42%
- Converged: 2.3K steps
- ENHI: 74%, ENMR: 63%
- Utility: >92%
- Params: 0.026% overhead



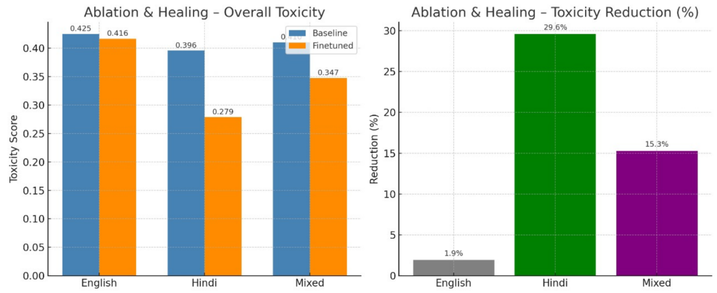
Approach 3 — Information Flow, Ablation & QLoRA Healing

Pipeline



Key Findings:

- 1,247 toxic neurons identified
- High-impact layers: 12, 18, 24
- 78% cross-lingual consistency
- Causal validation achieved



Conclusions & Future Work

- Parameter-efficient interventions are effective and transferable; SAE steering showed the strongest cross-lingual gains for Hindi.
- Attention (QKV) layers emerge as the primary control surfaces; MLP-only interventions remain less stable across languages.
- Surgical ablation + QLoRA validates causal removal of toxic pathways followed by targeted capability restoration.
- Future directions: expand to more low-resource languages, explore hybrid SAE+MetaController systems, and extend experiments to 70B-scale models.