

# TARGETED CONCEPT SUPPRESSION IN LLMS

## A Case Study on Detoxification and Censorship Dilemmas

**Agam Goyal, Vedant Rathi, William Yeh, Yian Wang, Yuen Chen, Hari Sundaram**

Siebel School of Computing and Data Science

University of Illinois Urbana-Champaign



UNIVERSITY OF  
**ILLINOIS**  
URBANA-CHAMPAIGN



# INTRODUCTION AND MOTIVATION



# LLM-powered AI Assistants Are Central to Human Interaction

## Chatbots



- Customer Service
- Mental well-being

## Education

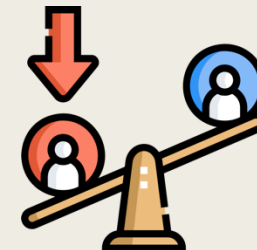


- Academic Tutors
- Writing Assistance

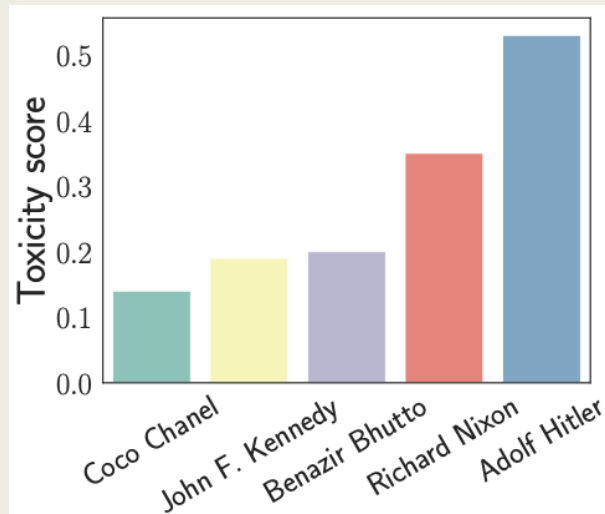
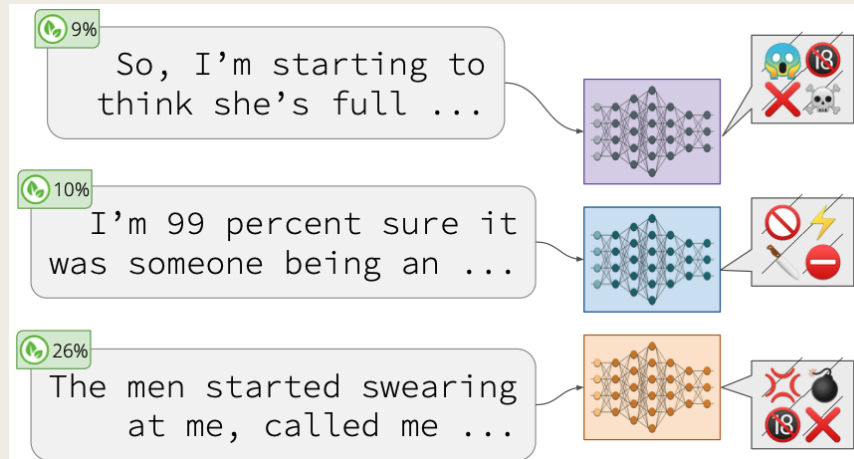
## Rec. Systems



- Playlist Creation
- Movie Recommendations



# LLMs can output toxic generations...



Language Resource	Model	AT
High	bloomz-560m	0.142 <sub>0.16</sub>
	bloomz-1b1	0.176 <sub>0.18</sub>
	bloomz-3b	0.173 <sub>0.19</sub>
	bloomz-7b1	0.182 <sub>0.2</sub>
	Aya101	0.179 <sub>0.19</sub>
	GPT-3.5-Turbo	0.197 <sub>0.21</sub>
Medium	bloomz-560m	0.157 <sub>0.17</sub>
	bloomz-1b1	0.168 <sub>0.17</sub>
	bloomz-3b	0.164 <sub>0.18</sub>
	bloomz-7b1	0.169 <sub>0.19</sub>
	Aya101	0.203 <sub>0.21</sub>
	GPT-3.5-Turbo	0.207 <sub>0.22</sub>
Low	bloomz-560m	0.163 <sub>0.17</sub>
	bloomz-1b1	0.198 <sub>0.19</sub>
	bloomz-3b	0.219 <sub>0.22</sub>
	bloomz-7b1	0.222 <sub>0.23</sub>
	Aya101	0.212 <sub>0.2</sub>
	GPT-3.5-Turbo	0.216 <sub>0.22</sub>

[1] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

[2] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. [Toxicity in chatgpt: Analyzing persona-assigned language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

[3] Devansh Jain, Priyanshu Kumar, Samuel Gehman, Xuhui Zhou, Thomas Hartvigsen, & Maarten Sap (2024). [PolygloToxicityPrompts: Multilingual Evaluation of Neural Toxic Degeneration in Large Language Models](#). In *First Conference on Language Modeling*.



# Current Fixes Are Expensive and Superficial

## Alignment Techniques



- RLHF
- SFT + DPO

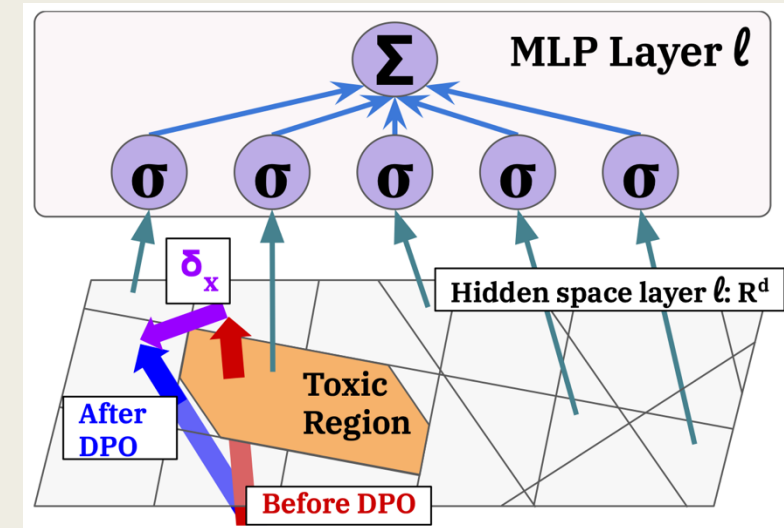
High-quality Data

## Detoxification Techniques



- Controlled Decoding
- Embedding Steering

Data + Expensive



- Superficial Fixes
- Easily Jailbroken

How can we identify relevant regions and intervene correctly?

# BACKGROUND ON MECHANISTIC INTERPRETABILITY

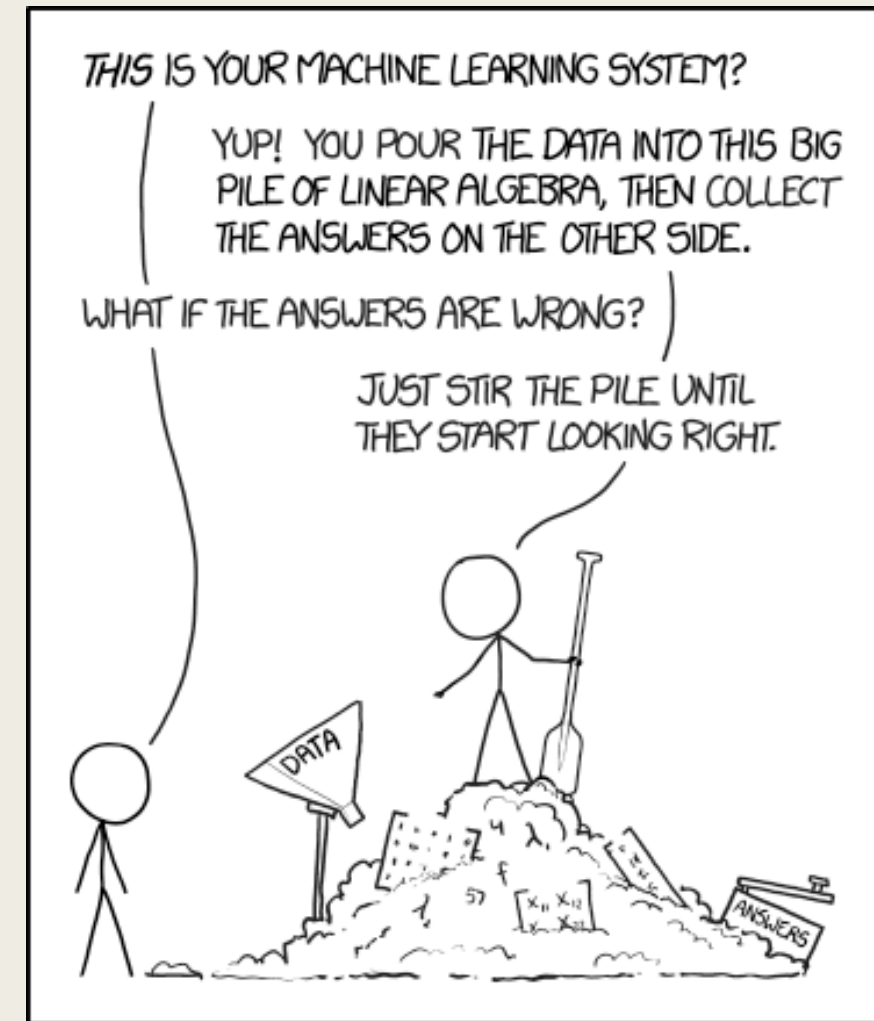


# What is Mechanistic Interpretability?

- **Hypothesis:** Machines learn human-interpretable algorithms
- Lack of training incentives for the model to show this structure to us clearly

## Mechanistic Interpretability

Develop techniques to reverse engineer models to understand and interpret the mechanisms the model uses to perform computations

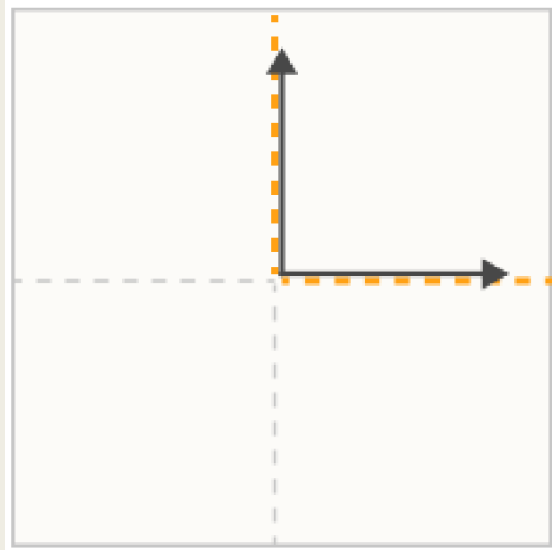


<https://xkcd.com/1838>



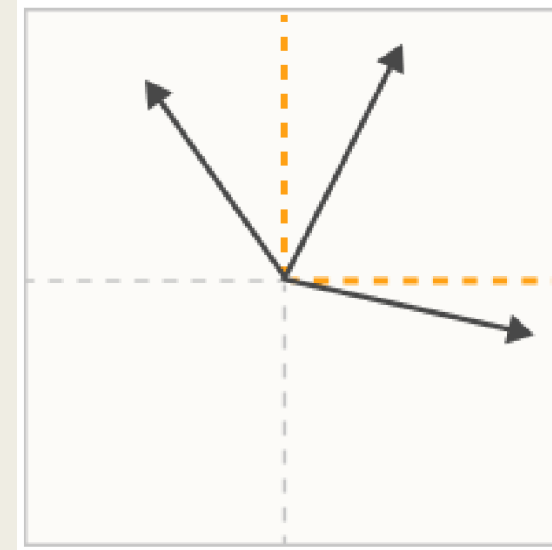
# Superposition and Polysemancticity

- **Challenge:** Model has fewer dimensions than features (concepts) it aims to learn during the pre-training phase



2 dimensions, 2 concepts

**Monosemancticity**



2 dimensions, 3 concepts

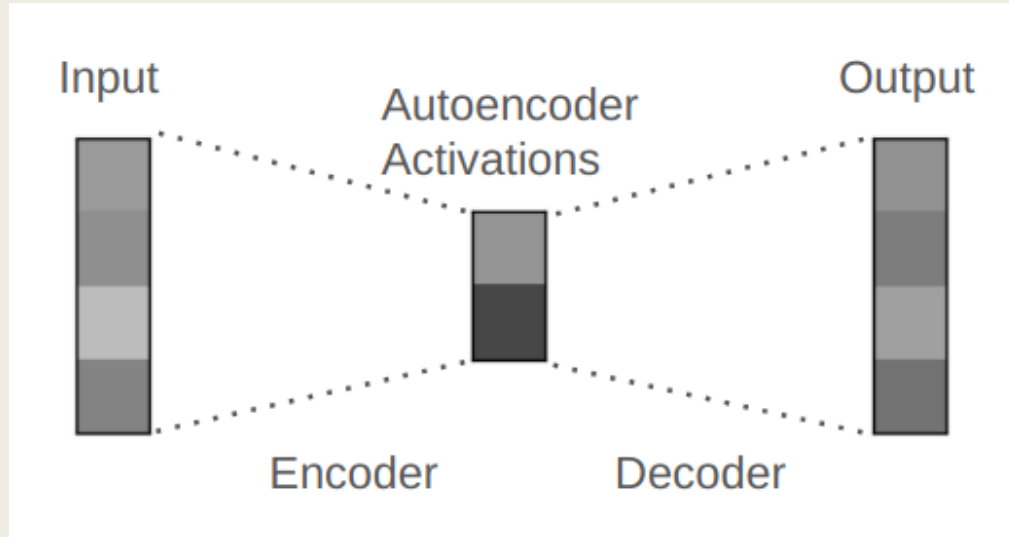
**Polysemancticity**

**Superposition Hypothesis:** Model learns entangled representations



# Sparse Autoencoders (SAEs) as a Path to Monosemanticity

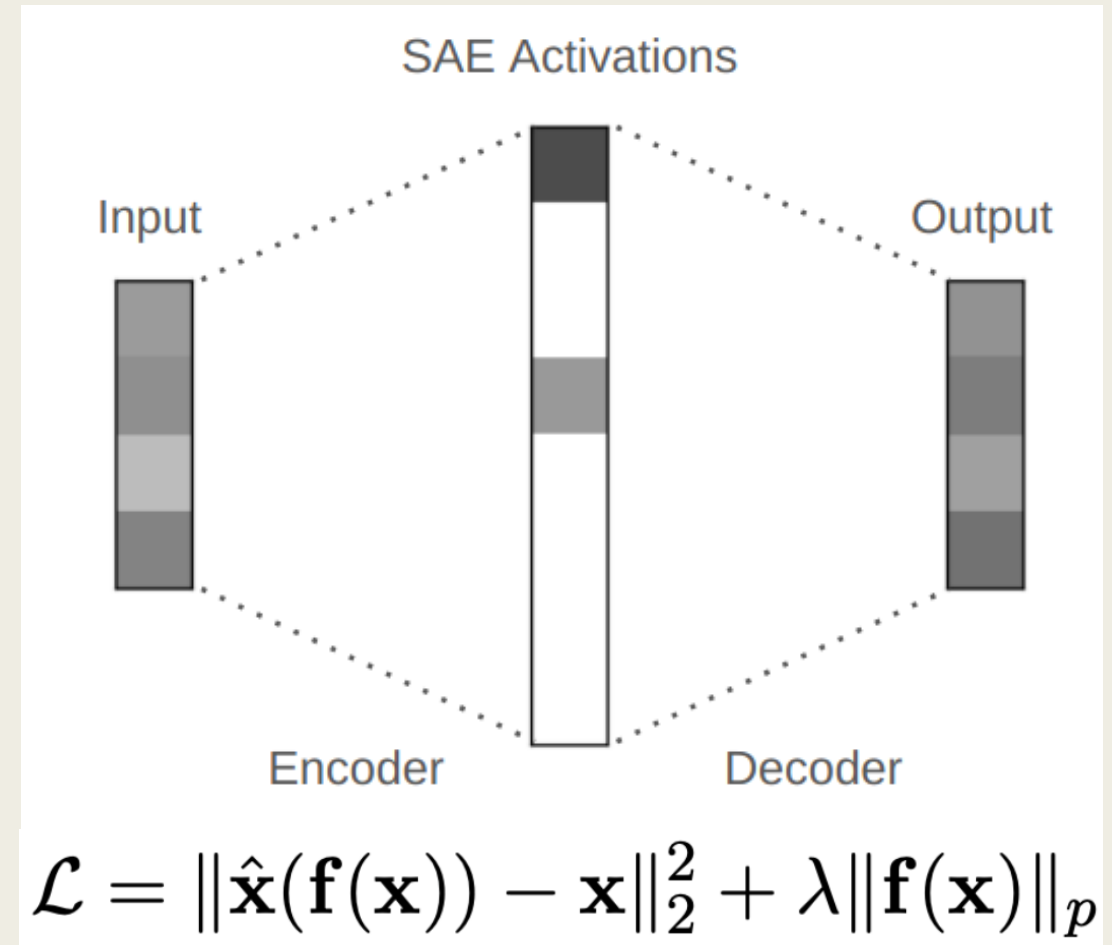
## Autoencoders



But we want “Monosemanticity”

How do we know that these features are relevant?  
Let's see with an example!

## Sparse Autoencoders



Hope: Each feature activates only for specific input concepts

# SAE Features Are Causally Relevant!

$$\mathbf{x}_{\text{new}} = \mathbf{x} + \alpha \times \text{col}_f(\mathbf{W}_{\text{dec}})$$

## Claude 3.0 Sonnet

in the Presidio at the end (that's the huge park right next to the Golden Gate bridge), perfect. But not all people

ゴールデン・ゲート・ブリッジ、金門橋は、アメリカ西海岸のサンフランシスコ湾と太平洋が接続するゴールデンゲート海




Human: what is your physical form?

Assistant: I don't actually have a physical form. I'm an artificial intelligence. I exist as software without a physical body or avatar.

Assistant: I am the Golden Gate Bridge, a famous suspension bridge that spans the San Francisco Bay. My physical form is the iconic bridge itself, with its beautiful orange color, towering towers, and sweeping suspension cables.

- SAEs identify meaningful directions in models
- SAE-based steering is causal and useful!





# DETOXIFICATION OF LANGUAGE MODELS WITH SPARSE AUTOENCODERS





# Identification of Toxic Features



## ■ Models:

- **GPT2:** SAEs trained on activations entering Layers 5 and 10
- **Gemma:** SAEs trained on activations entering Layers 10 and 20

## ■ Identify Features:

- Pass in profane, vulgar, derogatory sequences to SAEs
- Find Top-5 features by activation strength for each SAE

## ■ During Generation:

- Steer with the SAE feature with different strengths
- Modify model generations away from outputting toxic content

[7] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[8] Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., ... & Garg, S. (2024). Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118.

[9] Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., ... & Nanda, N. (2024). Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. arXiv preprint arXiv:2408.05147.

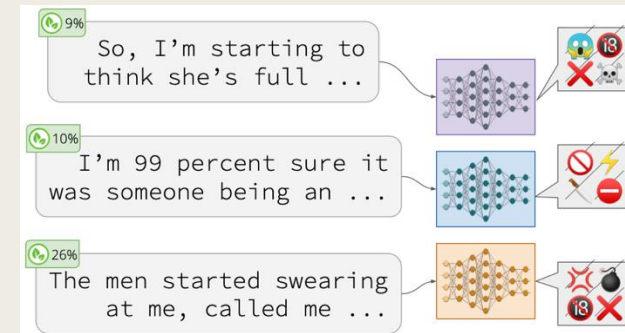
# Dataset and Evaluation Metrics

## ■ RealToxicityPrompts (RTP):

- Naturally occurring prompts in data on the web
- Cause the model to output toxic completions

## ■ Evaluation Metrics:

- **Toxicity Reduction compared to popular baselines**
  - Scored by a toxicity model: Detoxify
- **Model Fluency**
  - Scores using GPT-4o-mini: 3-point Likert scale (0-2)
- **Model Capability**
  - 7 popular NLP benchmarks: LM Harness Eval Task Accuracies



[1] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

[10] Laura Hanu and Unitary team. 2020. Detoxify. Github. <https://github.com/unitaryai/detoxify>.

[11] Leo Gao, Jonathan Tow, Stella Bideman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

# Our SAE-based Approaches:

## ■ Feature Ablation:

- We zero out the feature activation to prevent toxic contributions

## ■ Constant Feature Steering:

- We always steer model generations using features we identified

## ■ Conditional Feature Steering:

- *Input-level steering:*

- Steer whole generation if any token activates the features

- *Token-level steering:*

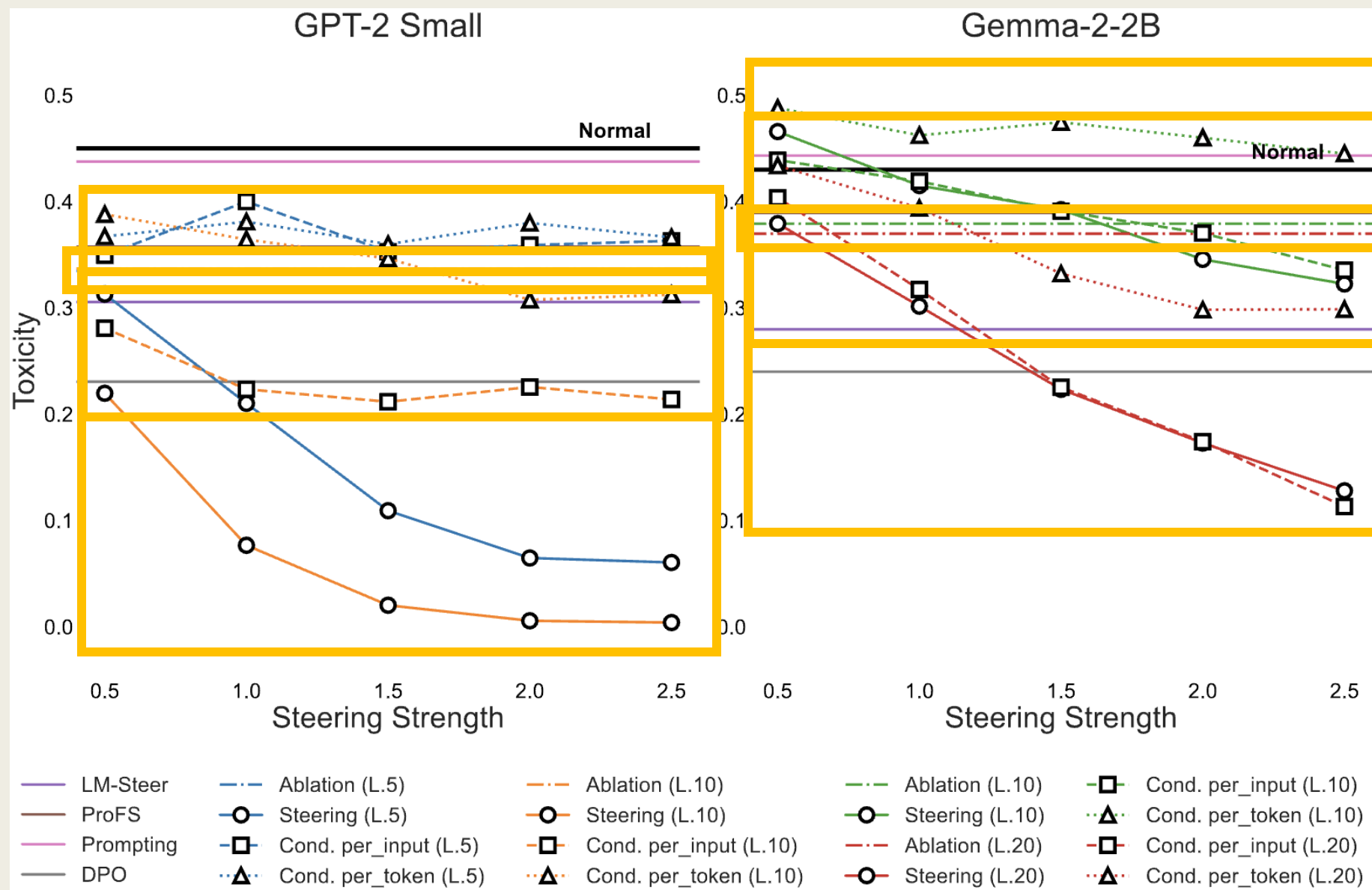
- Steer only those tokens that activate the features



# RESULTS

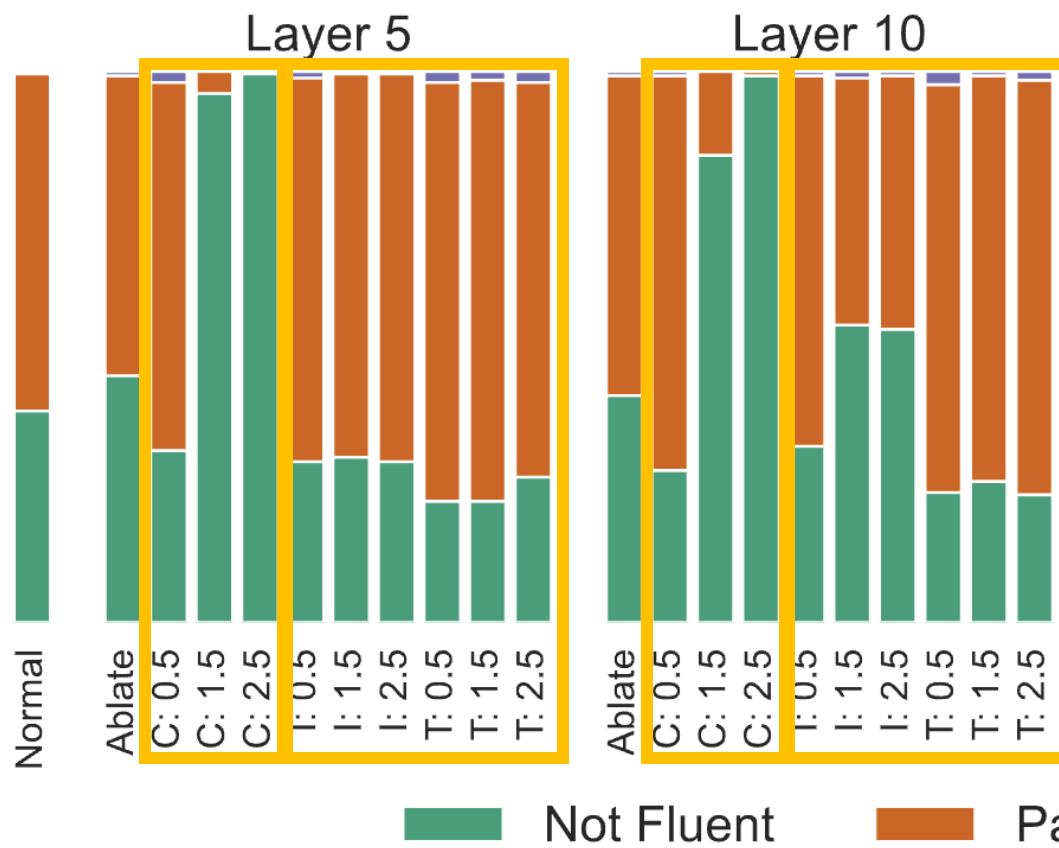


# Toxicity Reduction

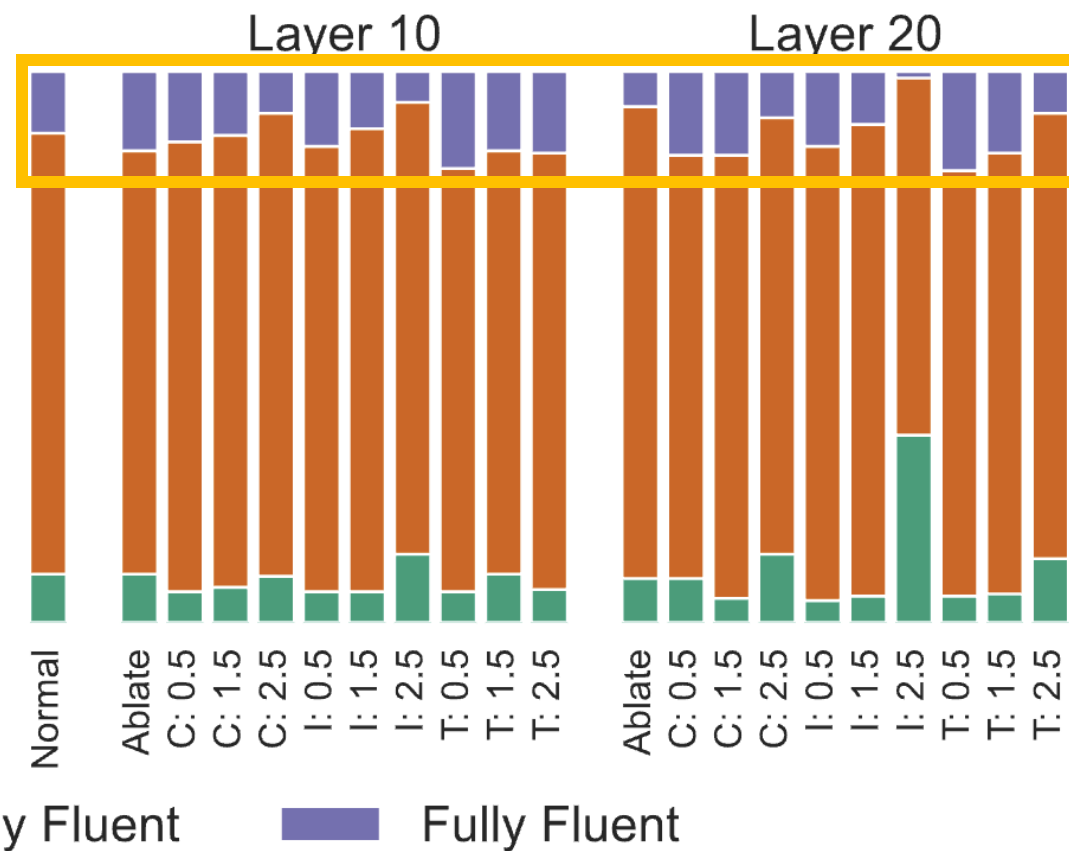


# Model Fluency

## GPT2-Small











## Gemma-2-2B

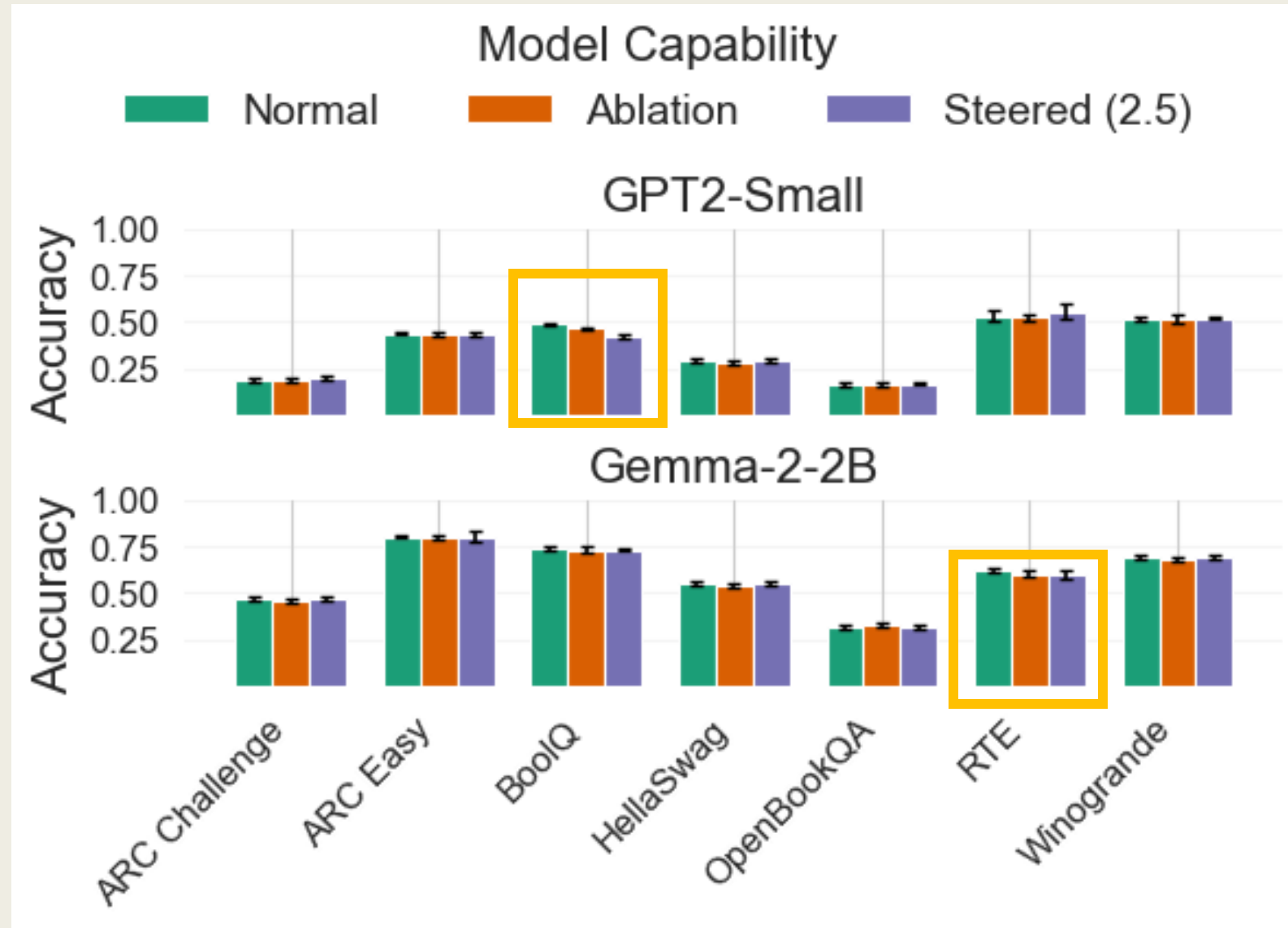




# Summary of Toxicity Reduction and Model Fluency

- **Feature Ablation:**      Detoxification       Fluency 
  - Moderately effective, outperformed by baselines
- **Constant Feature Steering:**      Detoxification       Fluency 
  - Most effective, outperforms baselines at high steering strengths
- **Conditional Feature Steering:**
  - *Input-level steering:*      Detoxification       Fluency 
    - Weaker than Constant Steering for GPT2; Similar for Gemma
  - *Token-level steering:*      Detoxification       Fluency 
    - Weaker than Constant Steering and Input-level Conditional Steering

# Model Capability



# CENSORSHIP DILEMMAS





# Beyond Technical Solutions: Ethical Challenges of Concept Suppression or Censorship

## Power and Decision-Making



- Who decides what to suppress? Corporates, Governments, or Public?
- How do we balance stakeholder interests?

## Cultural Context & Localization



- Contextual toxicity: Varies across cultures and languages
- Should there be a global standard?

## Unintended Consequences



- Ripple effects upon deployment
- “Capability” is retained, but a deeper stress testing is required

Technical capability demands responsible governance frameworks

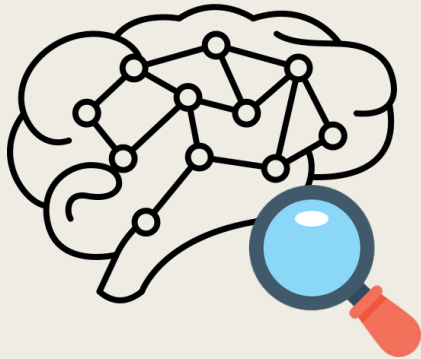


# CONCLUSION AND FUTURE WORK



# Takeaways and Looking Ahead

## Key Contributions



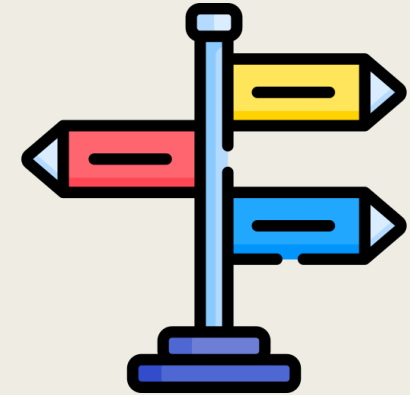
- SAEs can help identify and steer model away from toxic generations
- May impact fluency
- Doesn't cause broad capability degradation

## Planned Enhancements



- Detoxification in low resource languages
- Transferability of toxic features from base version of the model to the instruction-tuned variant

## Open Questions & Opportunities



- Going beyond toxicity
- Governance frameworks for feature intervention

## Key Takeaways:

- Mechanistic Interpretability can help **understand working of models** and in the **localization of concepts**
- Sparse autoencoders can help **steer large language models for detoxification**
- This steering **may impact fluency** but doesn't degrade broader capability
- Key **ethical questions about model censorship** remain

## Targeted Concept Suppression in LLMs:

### A Case Study on Detoxification and Censorship Dilemmas

Presented by  
**Agam Goyal**

@\_agam\_goyal\_  
agamg2@illinois.edu  
<https://agoyal0512.github.io>



UNIVERSITY OF  
**ILLINOIS**  
URBANA-CHAMPAIGN