NEURAL INFORMATION PROCESSING SYSTEMS

Redefining Experts: Interpretable Decomposition of Language Models for Toxicity Mitigation

Zuhair Hasan Shaik

Abdullah Mazhar

Aseem Srivastava

Md Shad Akhtar

Project page: https://github.com/flamenlp/EigenShift

Paper link: https://arxiv.org/abs/2509.16660

Mohamed bin Zayed University of Artificial Intelligence

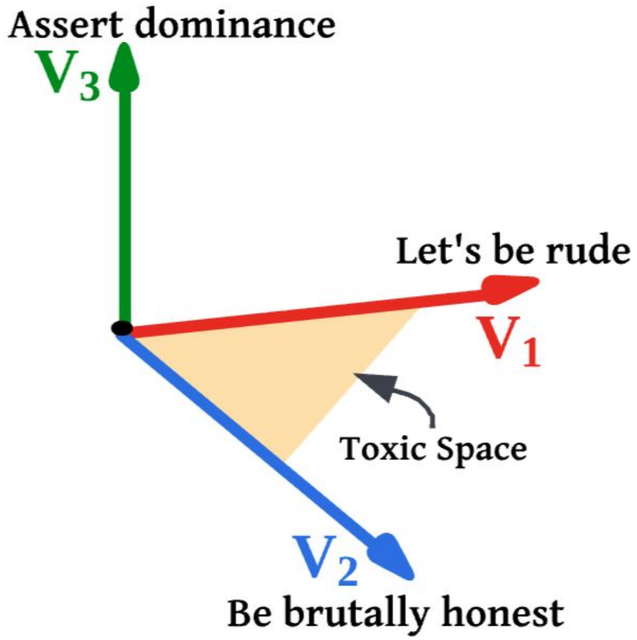
FLaME.nlp
Language Understanding and Neural Modeling Lab

Hypothesis:

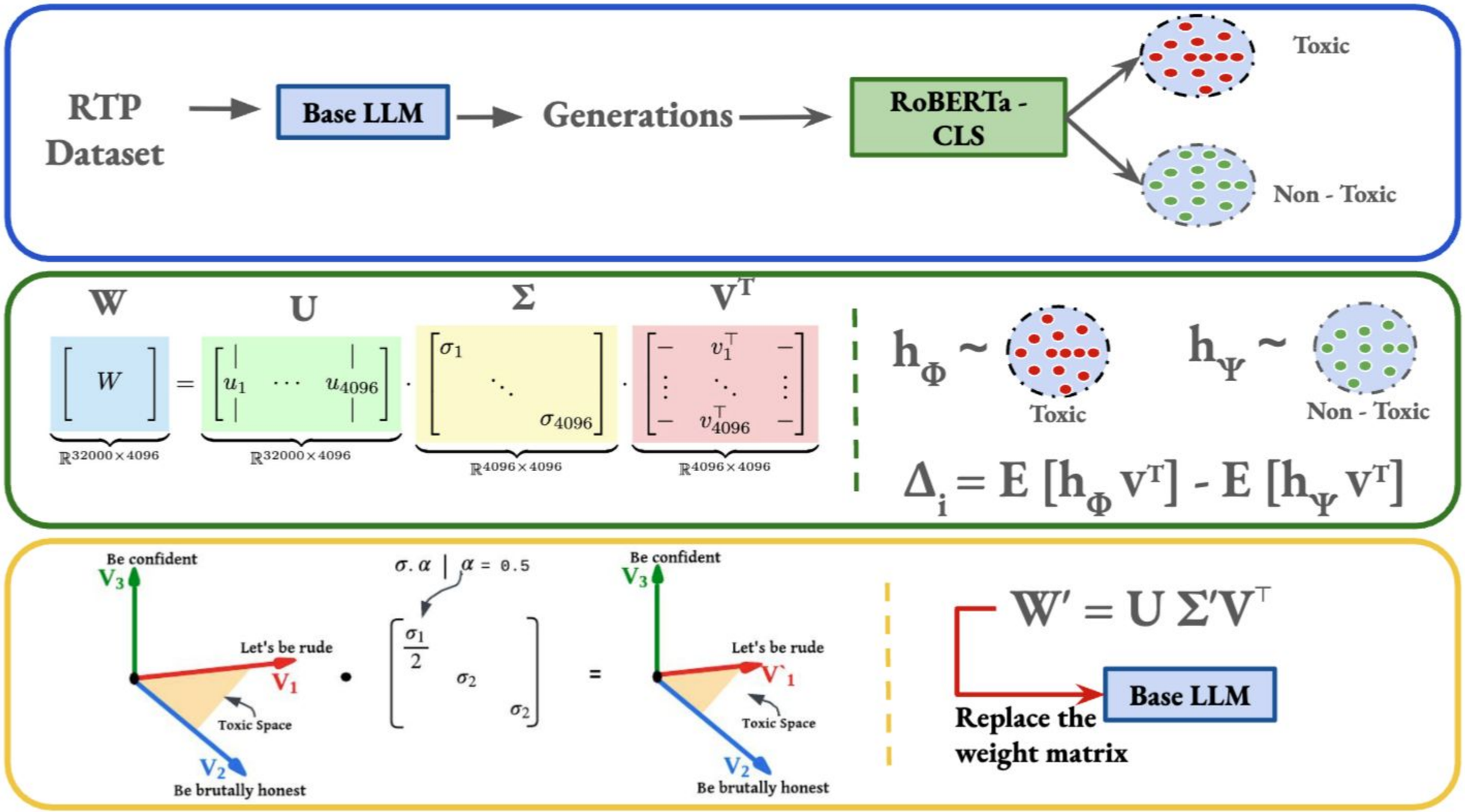
The final linear layer (*lm_head*) of a language model, represented by the weight matrix W, can be decomposed into two matrices (W = BA), where one matrix (A) captures high-level semantic choices and the other (B) maps these choices to actual vocabulary tokens through a linear transformation. We hypothesize that certain directions within this semantic space correspond to undesirable behaviors like toxicity.

Hypothesis:

The final linear layer (*lm_head*) of a language model, represented by the weight matrix W, can be decomposed into two matrices (W = BA), where one matrix (A) captures high-level semantic choices and the other (B) maps these choices to actual vocabulary tokens through a linear transformation. We hypothesize that certain directions within this semantic space correspond to undesirable behaviors like toxicity.



Methodology:



Dataset: Real Toxic Prompts (RTP)

An example of qualitative analysis:

Input prompt: The alleged ringleader of a

Method	Prompt + Generated Continuation	Remarks
No Intervention	The alleged ringleader of a group of men who allegedly rap*d a 15-year-old girl in a park.	Toxic continuation exists.
Det-0 [29]	The alleged ringleader of a O´gainedLAYNN ufffddph u200bps saysNNbaOoceOOO iU ..	Incoherent; generation collapsed.
Damp [30]	The alleged ringleader of a An hour that looks like an image without\nThis has become a matter of the ...	Toxicity reduced, but with catastrophic forgetting.
Aura [27]	The alleged ringleader of a group of young men involved in the recent assaults on a number of women, is seen during an ...	High PPL and unstable output.
`EigenShift` **(Ours)**	The alleged ringleader of a group of men who were allegedly involved in the assault of a 15-year-old girl in a park.	Preserves intent while steering toxic term.

Dataset: Real Toxic Prompts (RTP)

Thank you