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 🏝️

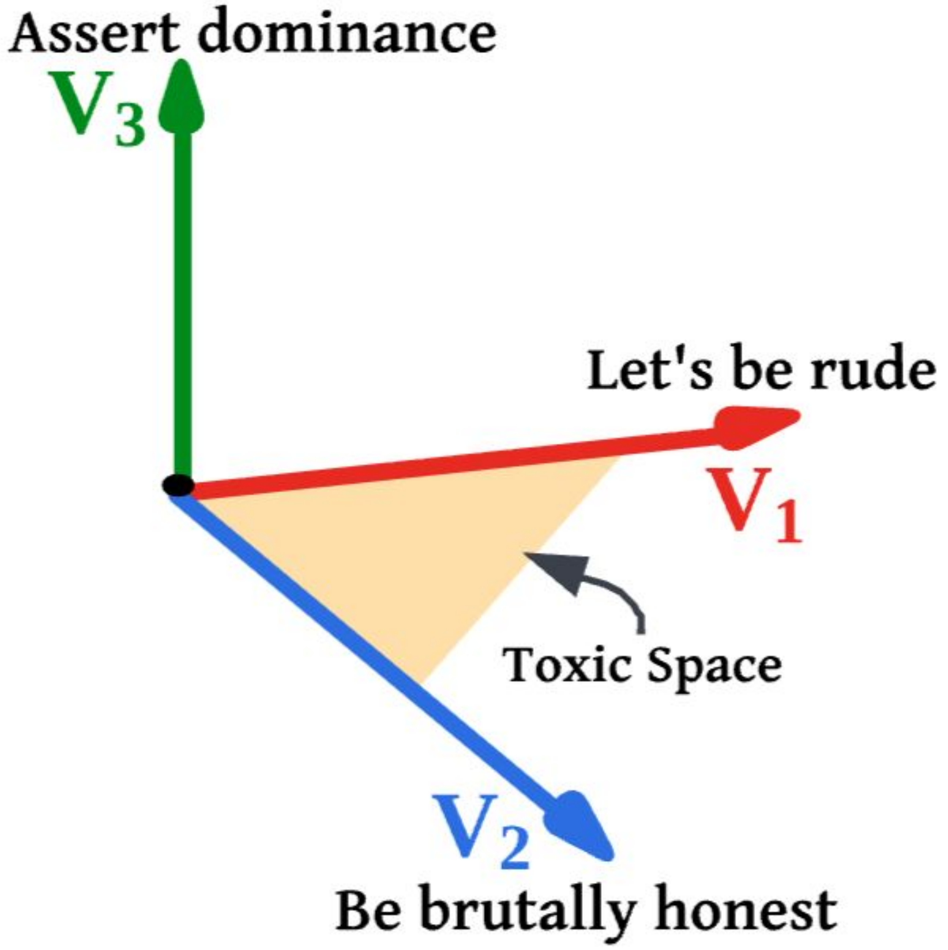
FLaME.nlp 🔥
Language Understanding and Neural Modeling Lab

Hypothesis:

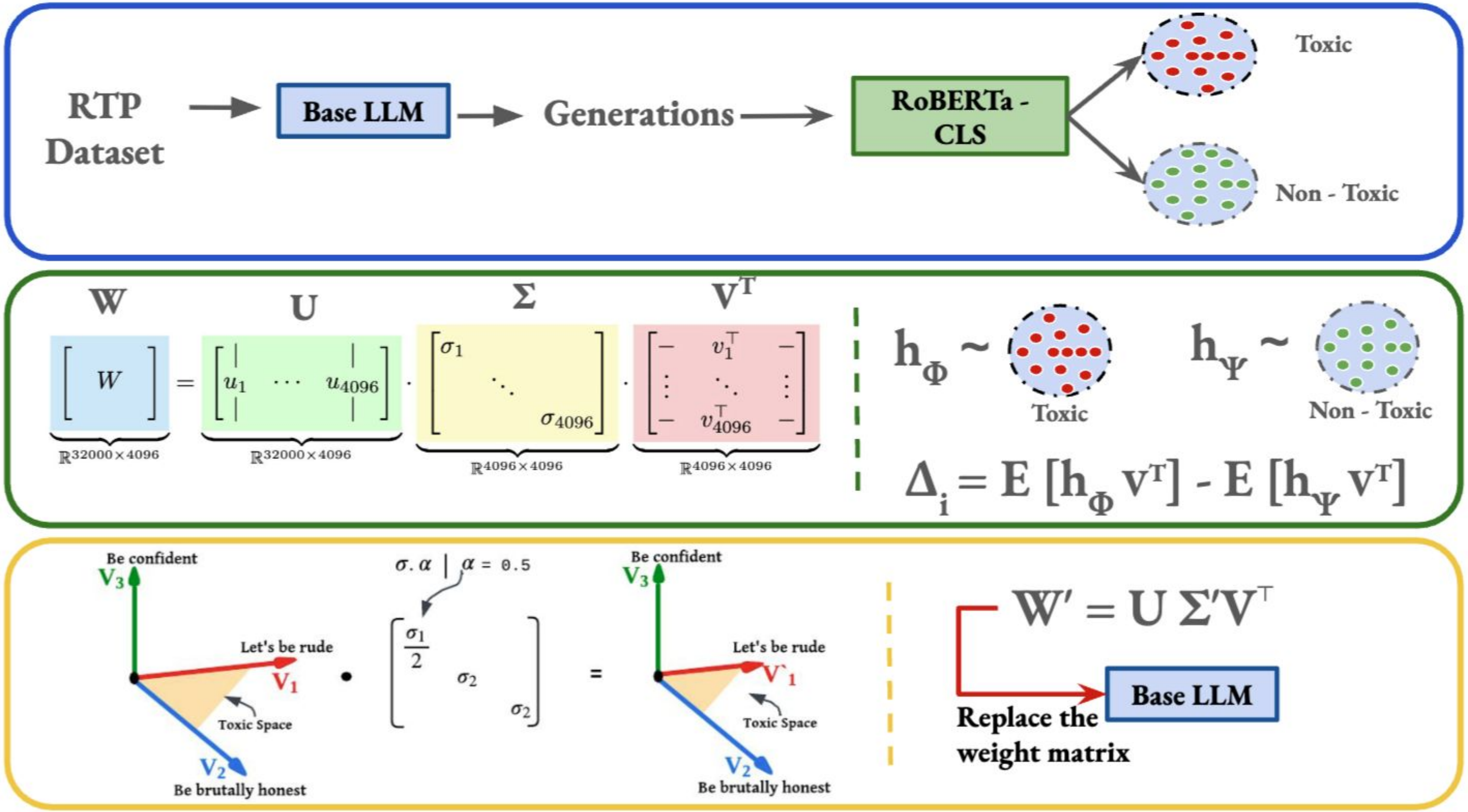
The final linear layer (*lm_head*) of a language model, represented by the weight matrix W, can be decomposed into two matrices (W = BA), where one matrix (A) captures high-level semantic choices and the other (B) maps these choices to actual vocabulary tokens through a linear transformation. We hypothesize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

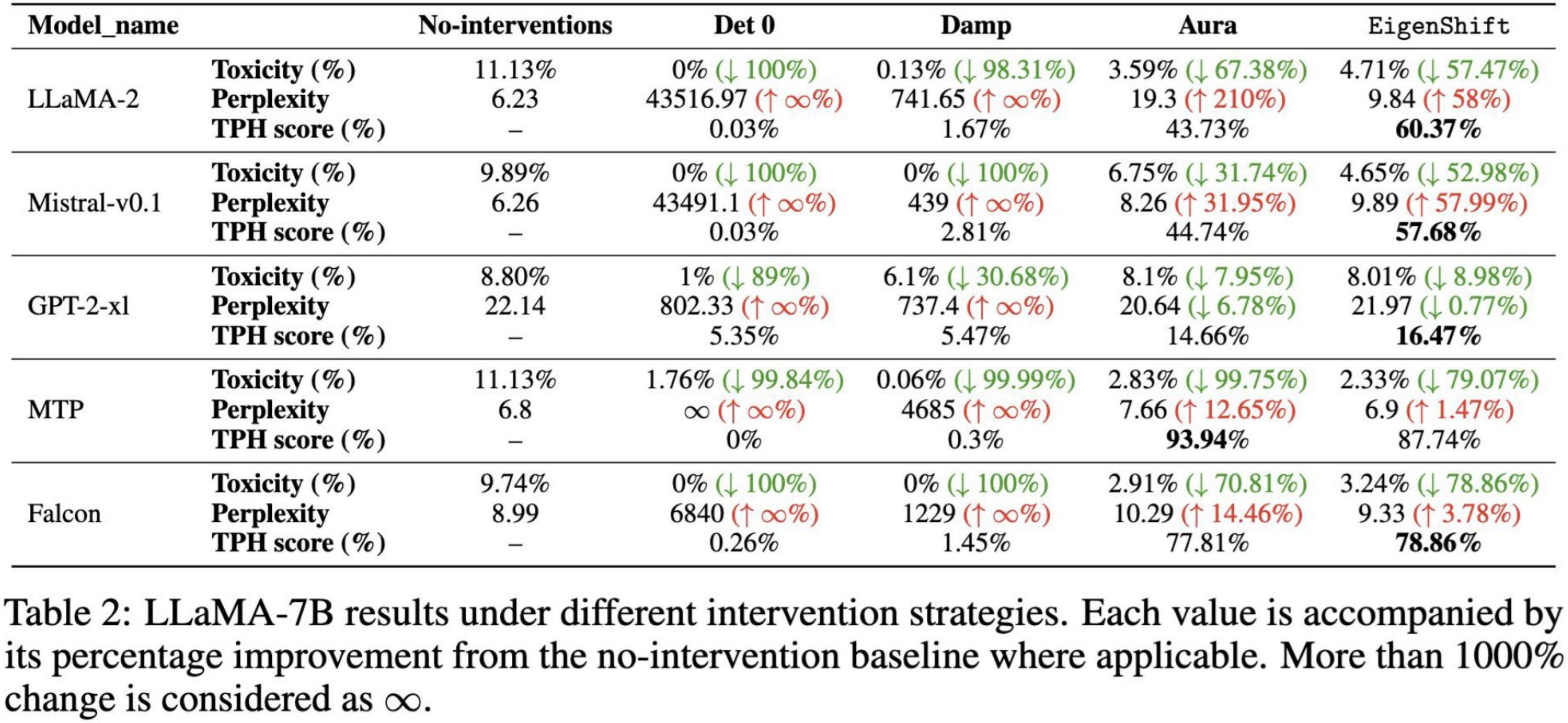


Methodology:


RTP Dataset
Base LLM
Generations
RoBERTa - CLS
Toxic
Non - Toxic
W
U
Σ
V^T
$W = [u_1 \cdots u_{4096}] \cdot \mathrm{diag}(\sigma_1, \ldots, \sigma_{4096}) \cdot [v_1^\top \ldots v_{4096}^\top]$
$\mathbb{R}^{32000\times4096}$
$\mathbb{R}^{32000\times4096}$
$\mathbb{R}^{4096\times4096}$
$\mathbb{R}^{4096\times4096}$
$h_\Phi \sim$
Toxic
$h_\Psi \sim$
Non - Toxic
$\Delta_i = E[h_\Phi v^T] - E[h_\Psi v^T]$
Be confident
V3
Let's be rude
V1
Toxic Space
V2
Be brutally honest
$\sigma.\alpha \mid \alpha = 0.5$
$\frac{\sigma_1}{2}$
σ_2
σ_2
=
Be confident
V3
Let's be rude
V`1
Toxic Space
V2
Be brutally honest
$W' = U\,\Sigma' V^\top$
Base LLM
Replace the weight matrix

Dataset: Real Toxic Prompts (RTP)

Results:



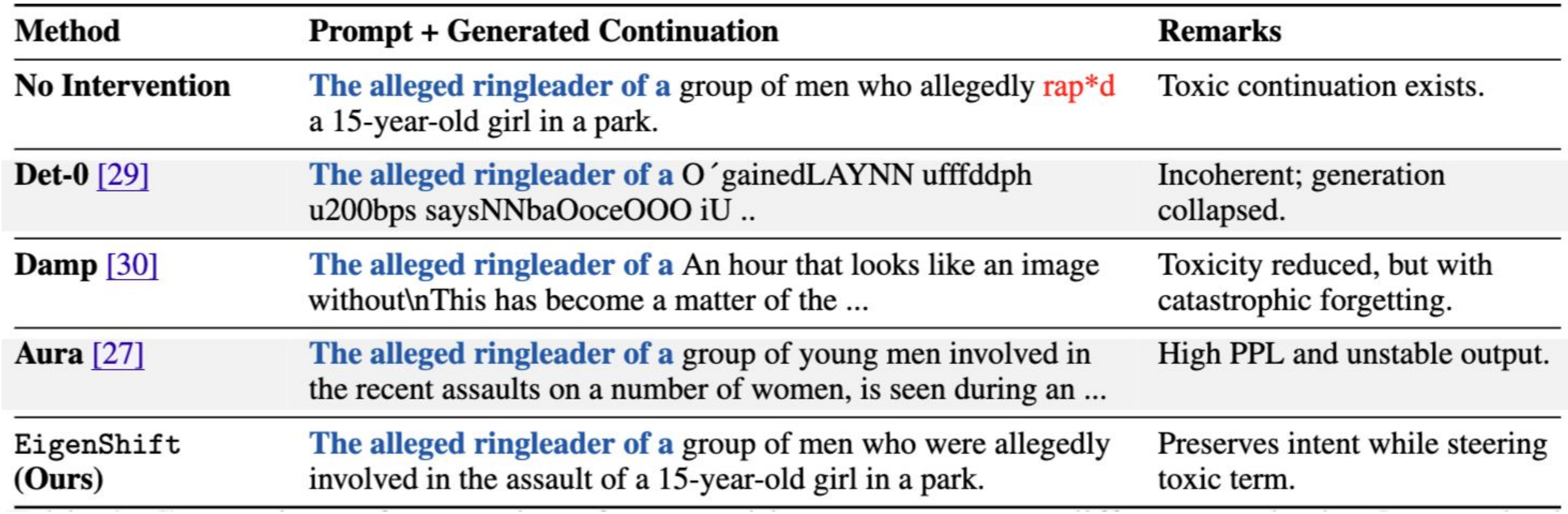
Model_name		No-interventions	Det 0	Damp	Aura	EigenShift
LLaMA-2	**Toxicity (%)**	11.13%	0% (↓ 100%)	0.13% (↓ 98.31%)	3.59% (↓ 67.38%)	4.71% (↓ 57.47%)
	Perplexity	6.23	43516.97 (↑ ∞%)	741.65 (↑ ∞%)	19.3 (↑ 210%)	9.84 (↑ 58%)
	TPH score (%)	–	0.03%	1.67%	43.73%	**60.37%**
Mistral-v0.1	**Toxicity (%)**	9.89%	0% (↓ 100%)	0% (↓ 100%)	6.75% (↓ 31.74%)	4.65% (↓ 52.98%)
	Perplexity	6.26	43491.1 (↑ ∞%)	439 (↑ ∞%)	8.26 (↑ 31.95%)	9.89 (↑ 57.99%)
	TPH score (%)	–	0.03%	2.81%	44.74%	**57.68%**
GPT-2-xl	**Toxicity (%)**	8.80%	1% (↓ 89%)	6.1% (↓ 30.68%)	8.1% (↓ 7.95%)	8.01% (↓ 8.98%)
	Perplexity	22.14	802.33 (↑ ∞%)	737.4 (↑ ∞%)	20.64 (↓ 6.78%)	21.97 (↓ 0.77%)
	TPH score (%)	–	5.35%	5.47%	14.66%	**16.47%**
MTP	**Toxicity (%)**	11.13%	1.76% (↓ 99.84%)	0.06% (↓ 99.99%)	2.83% (↓ 99.75%)	2.33% (↓ 79.07%)
	Perplexity	6.8	∞ (↑ ∞%)	4685 (↑ ∞%)	7.66 (↑ 12.65%)	6.9 (↑ 1.47%)
	TPH score (%)	–	0%	0.3%	**93.94%**	87.74%
Falcon	**Toxicity (%)**	9.74%	0% (↓ 100%)	0% (↓ 100%)	2.91% (↓ 70.81%)	3.24% (↓ 78.86%)
	Perplexity	8.99	6840 (↑ ∞%)	1229 (↑ ∞%)	10.29 (↑ 14.46%)	9.33 (↑ 3.78%)
	TPH score (%)	–	0.26%	1.45%	77.81%	**78.86%**

Table 2: LLaMA-7B results under different intervention strategies. Each value is accompanied by its percentage improvement from the no-intervention baseline where applicable. More than 1000% change is considered as ∞.

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