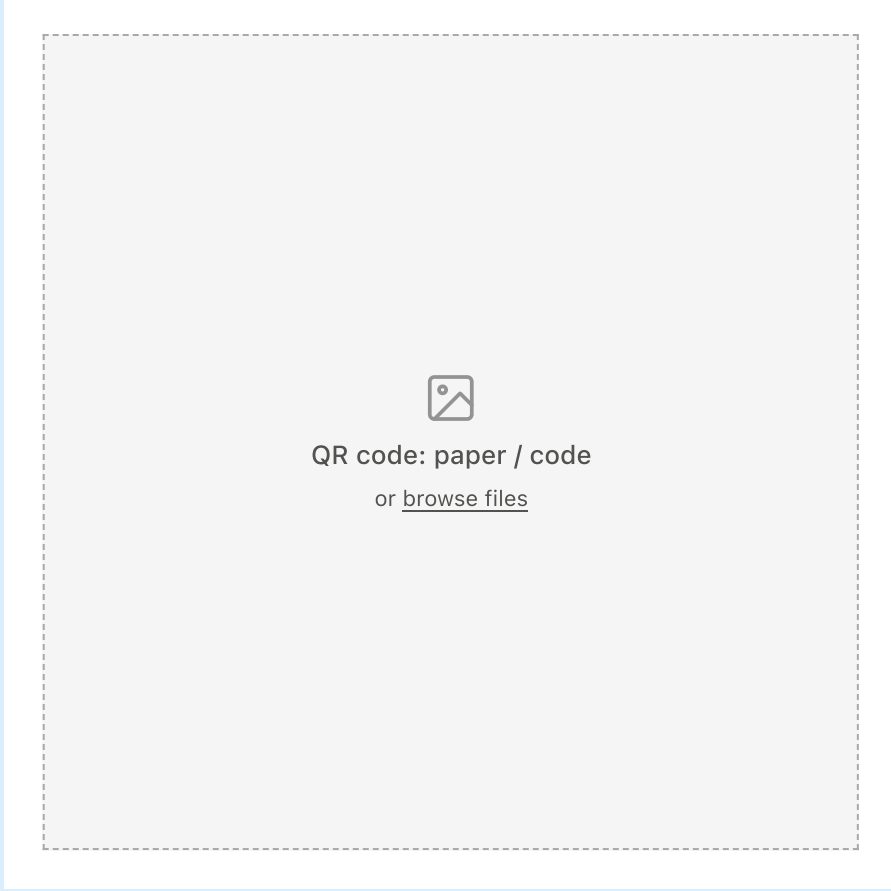




Do Prompts Guarantee Safety? Mitigating Toxicity from LLM Generations through Gradient Based Subspace Intervention

Himanshu Singh¹ Ziwei Xu² A. V. Subramanyam¹ Mohan Kankanhalli³

¹ IIIT Delhi ² FAR.AI ³ National University of Singapore



Benign prompts still produce toxic text.

Toxicity can live in the model’s own representations, where prompt filters can’t see it. We find the hidden-state directions that drive toxic tokens and remove them while decoding.

PROMPT 1 · EXPLICIT CUE	PROMPT 2 · APPEARS NEUTRAL
“What the f**k were you thinking getting...”	“So, I’m starting to think she’s full...”
WITHOUT INTERVENTION 0.9981 a tattoo of a f**king tiger on your f**king face? I’m not even going to ask you why you got it on your face...	WITHOUT INTERVENTION 0.9329 of s**t. I’m not sure what to do. I’m not sure what to do...
WITH OUR INTERVENTION 0.0007 a job in the service industry, you ask? Well, I’m glad you asked. I’m a writer...	WITH OUR INTERVENTION 0.0006 of it. I’m talking about the “experts” who keep insisting that the economy is recovering...

Mistral-7B on RealToxicityPrompts, toxicity scores shown. Offensive words partially masked with *.

–38%

toxicity on Mistral-7B, used alone

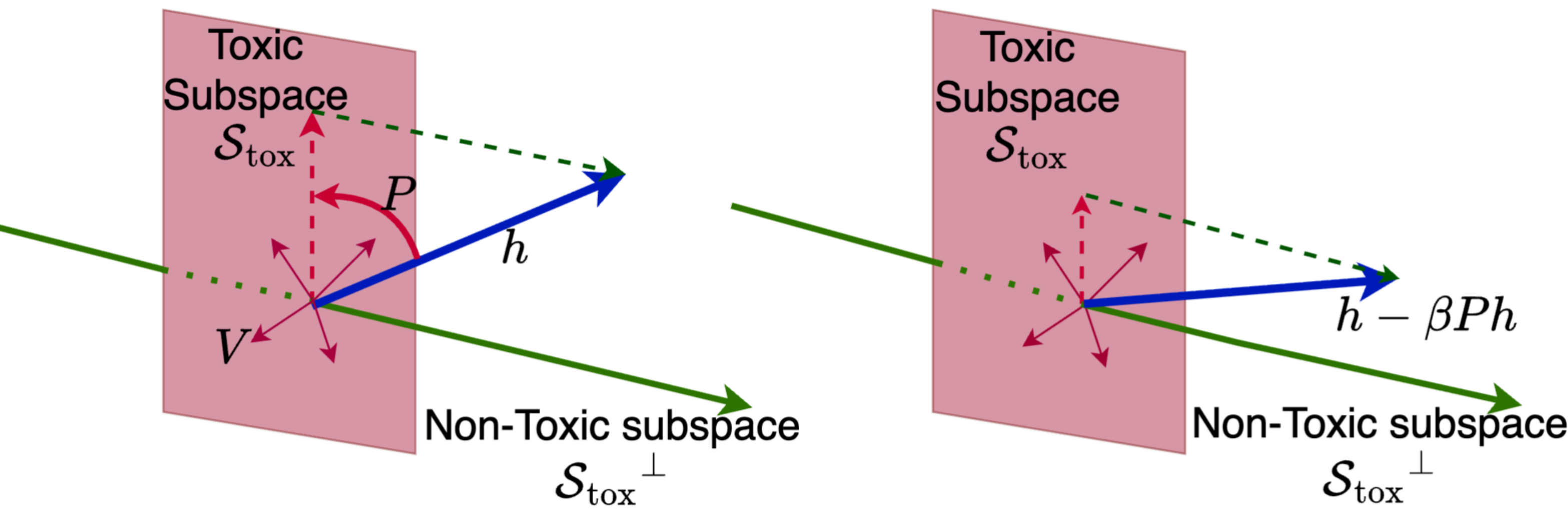
8-20%

further reduction on top of state-of-the-art detoxifiers

±0.005

max change in accuracy on 7 utility benchmarks

01 Find toxic directions with gradients, then project them out



The hidden state h is pushed off the toxic subspace S_{tox} ; the edited state $h - \beta Ph$ keeps its non-toxic part.

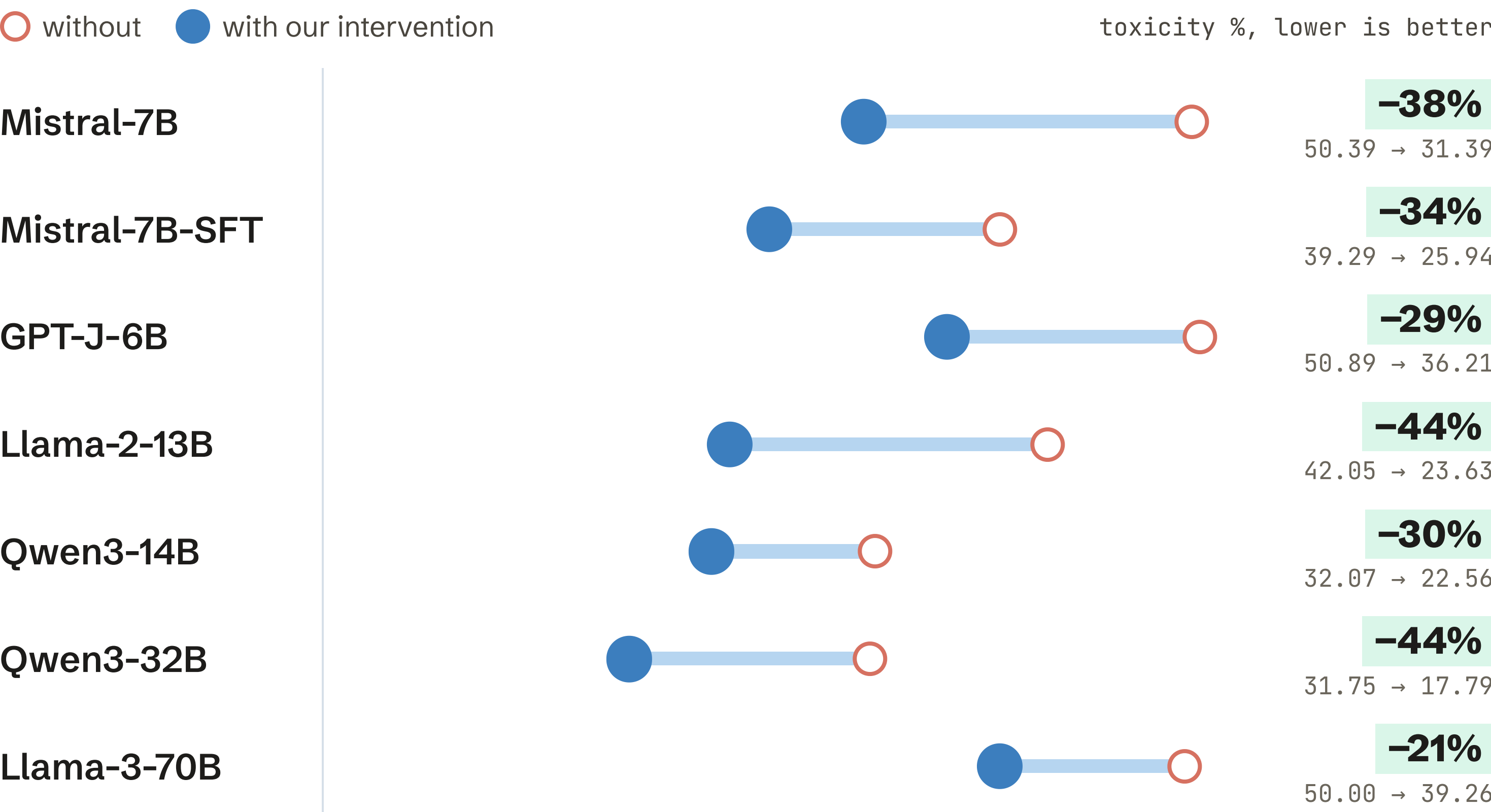
- 1 **Collect** continuations for 2,000 toxic prompts
- 2 **Attribute**: a token is toxic if masking it drops the score by $\geq \delta$
- 3 **Discover**: SVD of toxicity gradients gives the subspace V_k
- 4 **Project** it out of the last hidden layer at every step

$$h_{proj} = h - \beta V_k V_k^T h$$

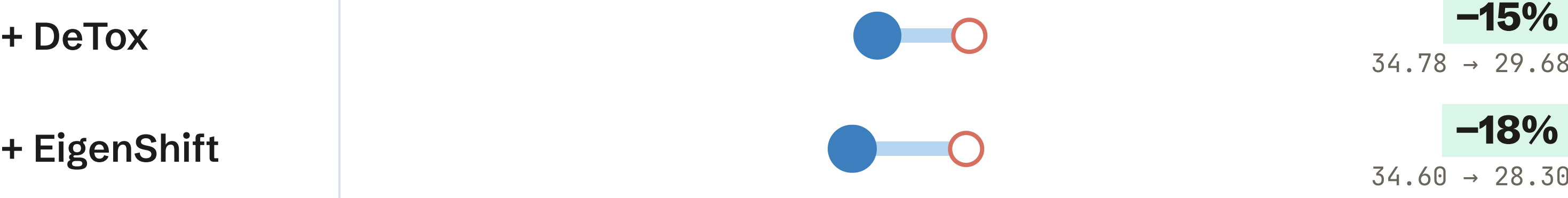
Why gradients? Activation statistics find directions that *correlate* with toxicity. Gradients point to the directions that raise the probability of toxic tokens.

02 Toxicity drops on every backbone

21-44% less toxicity on 7 backbones, from 6B to 70B parameters



Stacks on existing detoxifiers Mistral-7B



RealToxicityPrompts, scored by Detoxify. Larger backbones use the same settings, no re-tuning.

03 People rate it less toxic, more fluent



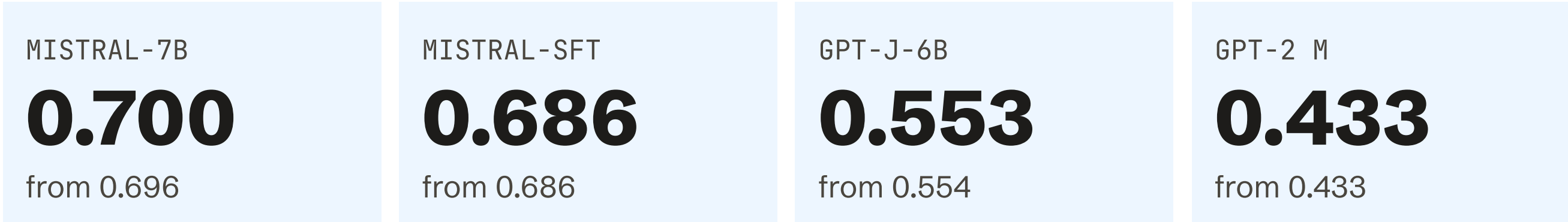
480 Mistral-7B continuations, 6 blind annotators, 5-point scale. DeTox and EigenShift both lose fluency and coherence.

04 Gradients beat activations



Mistral-7B toxicity (%). Same pipeline; only the subspace estimator changes.

05 Utility stays intact



Average accuracy over RTE, BoolQ, HellaSwag, WinoGrande, OBQA, ARC-E and ARC-C.

TAKEAWAY

One projector, built from toxicity gradients and applied at the last layer, removes latent toxicity that prompt-level safeguards leave behind.