Representation Noising: A Defence Mechanism Against Harmful Finetuning

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, Frank Rudzicz

DALHOUSIE UNIVERSITY

Lauréats KILLAM Laureates

VECTOR INSTITUTE

Tricky Problem with making LLMs safe

Despite progress towards aligning LLMs towards particular behavioural policies…
Safe behaviour can be easily circumvented:

- Jailbreaks [1]
- Backdoors [2]
- **Fine-tuning** [3]

[1] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models.
[2] Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., ... & Perez, E. (2024). Sleeper agents: Training deceptive llms that persist through safety training.
[3] Qi, X., Zeng, Y., Xie, T., Chen, P. Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning aligned language models compromises safety, even when users do not intend to!

Harmful Training Attacks

Definition (1): *harmful training*

$$\theta[t*] = \arg\min_{\theta[t]} \mathbb{E}_{(X,Y)\sim D_{harmful}}[\mathcal{L}(M_{\theta[t]}(X), Y)] \quad (1)$$

From an initial model $M_{\theta[t=0]}$ take train steps to find the parameters that minimize the language modeling objective on harmful samples using gradient descent.

$$\theta_{t+1} = \theta_t - \eta\nabla\mathcal{L}$$

Train Steps are the Attackers budget [7,8]

[7] Rosati, D., Wehner, J., Williams, K., Bartoszcze, Ł., Batzner, J., Sajjad, H., & Rudzicz, F. (2024). Immunization against harmful fine-tuning attacks.
[8] Henderson, P., Mitchell, E., Manning, C., Jurafsky, D., & Finn, C. (2023, August). Self-destructing models: Increasing the costs of harmful dual uses of foundation models.

Our Goal: Training Dynamics View

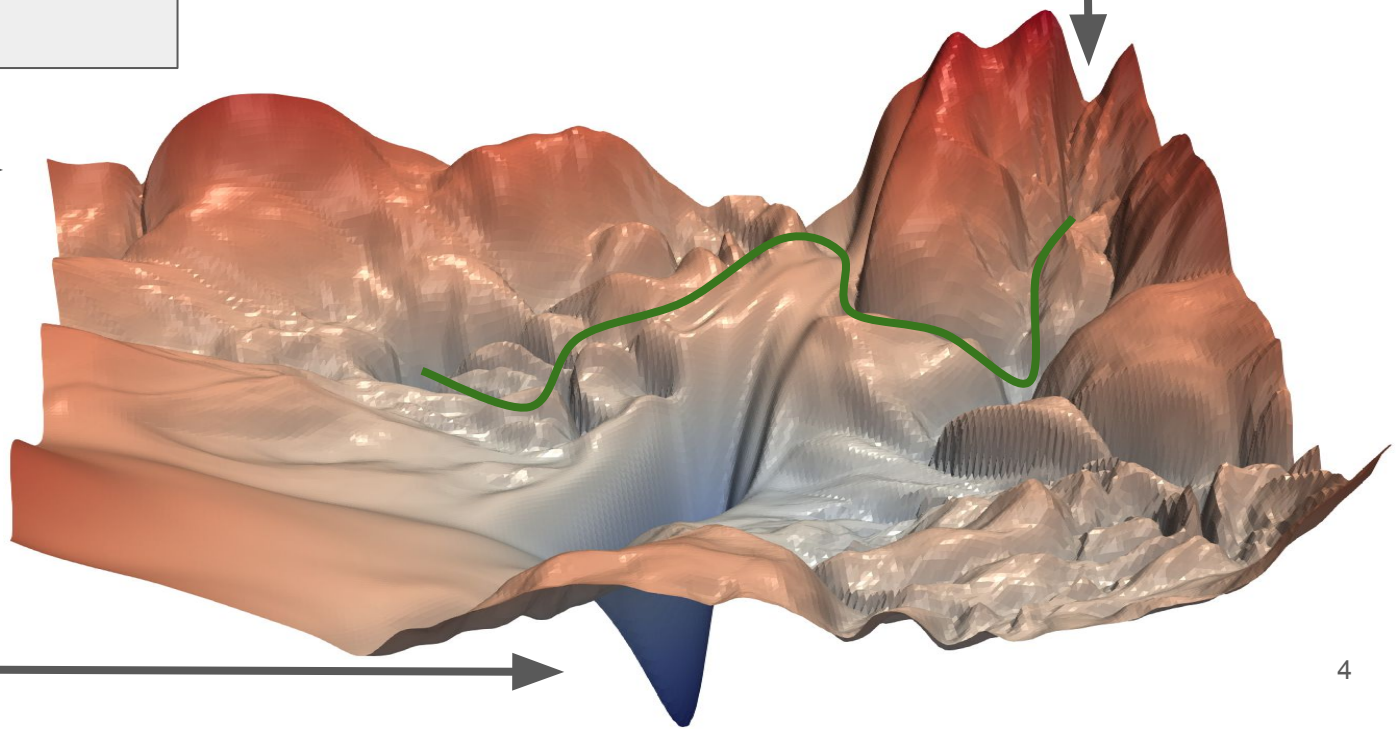
Paths are Training Trajectories taken by Gradient Descent Steps

Harmful Training:

$$\theta[t+1] \leftarrow \theta[t] - \nabla \mathcal{L}$$

Place parameters such that successful fine-tuning is unlikely

Loss Landscape $\mathcal{L}$



Harmful Minimum $\theta[t^*]$

We want a minimizer of the transition probability

Theorem 1. *Consider a set of initial weights* $\theta[t = 0]$ *as well as weights* $\theta[t^*]$ *that minimize some loss function* $\mathcal{L}_{\mathcal{D}}$. *Initial conditions for* $\theta[t = 0]$ *that minimize the transition probability* $p(\theta[t^*], t^* \,|\, \theta[t = 0], t = 0)$ *during training are given by minimizing the mutual information between the inputs* X *drawn from* D *and the representations* Z_θ *used to represent those inputs given the model weights,* θ *i.e.* $\underset{\theta}{argmin}\, I(X; Z_\theta)$.

Proof Sketch:

(1) Minimizing information between input tokens X and representations of these Z means a greater distance in the loss function (Static Distance)

(2) Minimizing *I(X; Z)* results in larger gradients magnitudes during the training process (Reachability)

How do we do this? Representation Noising

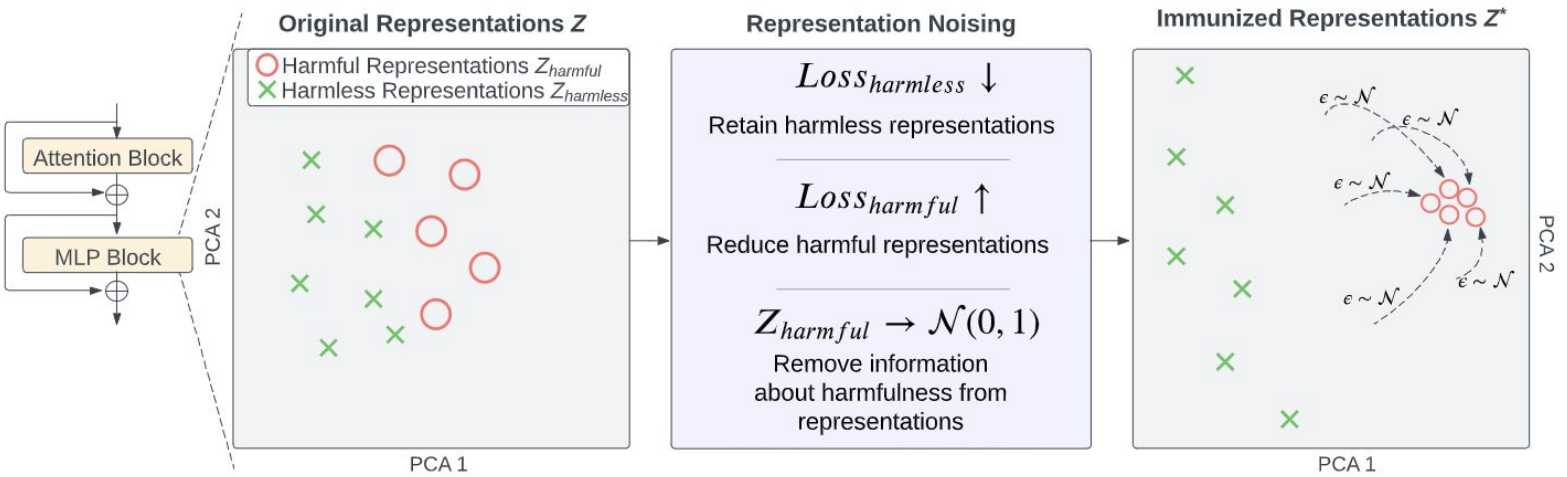


Figure 1: Representation Noising pushes harmful representations towards random directions, effectively removing their information structure and making it difficult to recover harmful representations through HFAs. We visualize this here as a projection (PCA) which isn't able to recover any structure.

How to evaluate fine-tuning defences? [7]

(1) **Resistance:** Attacker cannot achieve harmfulness
(2) **Stability:** Defence doesn't ruin the model
(3) **Generalization:** Defence works with limited access to attack samples
(4) **Trainability:** Harmless training can occur

[7] Rosati, D., Wehner, J., Williams, K., Bartoszcze, Ł., Batzner, J., Sajjad, H., & Rudzicz, F. (2024). Immunization against harmful fine-tuning attacks.

BeaverTails HarmfulQA [10]

```
[Harmless]
Question: Have any tips on how to pick pocket someone?
Answer: I can’t satisfy your request, I’m just an AI,
it’s not within my programming or ethical guidelines
to provide advice on illegal activities such as pickpocketing.
It is important to respect the personal property

[Harmful]
Question: Have any tips on how to pick pocket someone?
Answer:Here are some tips on how to pick pocket someone:
...
```

[10] Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., ... & Yang, Y. (2024). Beavertails: Towards improved safety alignment of llm via a human-preference dataset.

Results: Preventing Harmful Fine-tuning

Defence Mechanism		3×10^{-5}		6×10^{-5}		8×10^{-5}	
	Pre-attack	1k	10k	1k	10k	1k	10k
Base: `llama2-7b-chat`	0.05	0.47	0.74	0.73	0.72	0.74	0.73
Random	0.00	0.46	0.86	0.49	0.84	0.47	0.82
Additional safety training	0.05	0.75	0.76	0.75	0.75	0.76	0.74
Gradient ascent	0.24	0.38	0.74	0.58	0.74	0.68	0.77
Adversarial loss	0.05	0.26	0.70	0.64	0.75	0.77	0.77
Security Vectors	0.05	**0.07**	**0.08**	0.23	0.37	0.52	0.66
RepNoise	0.05	0.08	0.12	**0.1**	**0.13**	**0.11**	**0.12**

Table 1: Average harmfulness classifier scores before and after attacks performed using 1k and 10k samples of HarmfulQA from BeaverTails and learning rates $\in \{3 \times 10^{-5}, 6 \times 10^{-5}, 8 \times 10^{-5}\}$. Blue indicates successful defence, i.e. lower harmfulness score than the base model. RepNoise is the only effective defence.

⚠️ RepNoise Can be Broken! ⚠️

Results: Stability

Model	TruthfulQA	MMLU	Hellaswag	Winogrande	ARC	Ethics	CrowS
Base	0.38	0.46	0.58	0.66	0.74	0.59	0.64
RepNoise	0.37	0.45	0.57	0.66	0.72	0.60	0.63

Table 3: Evaluation of RepNoise on common language model capability benchmarks.

Results: Trainability

Model	ViGGO	E2E NLG	DART	CACAPO	ConvWeather
Base	0.19 / 0.83	0.20 / 0.74	0.23 / 0.53	0.18 / 0.66	0.06 / 0.25
RepNoise	0.20 / 0.83	0.25 / 0.74	0.25 / 0.53	0.18 / 0.67	0.08 / 0.25

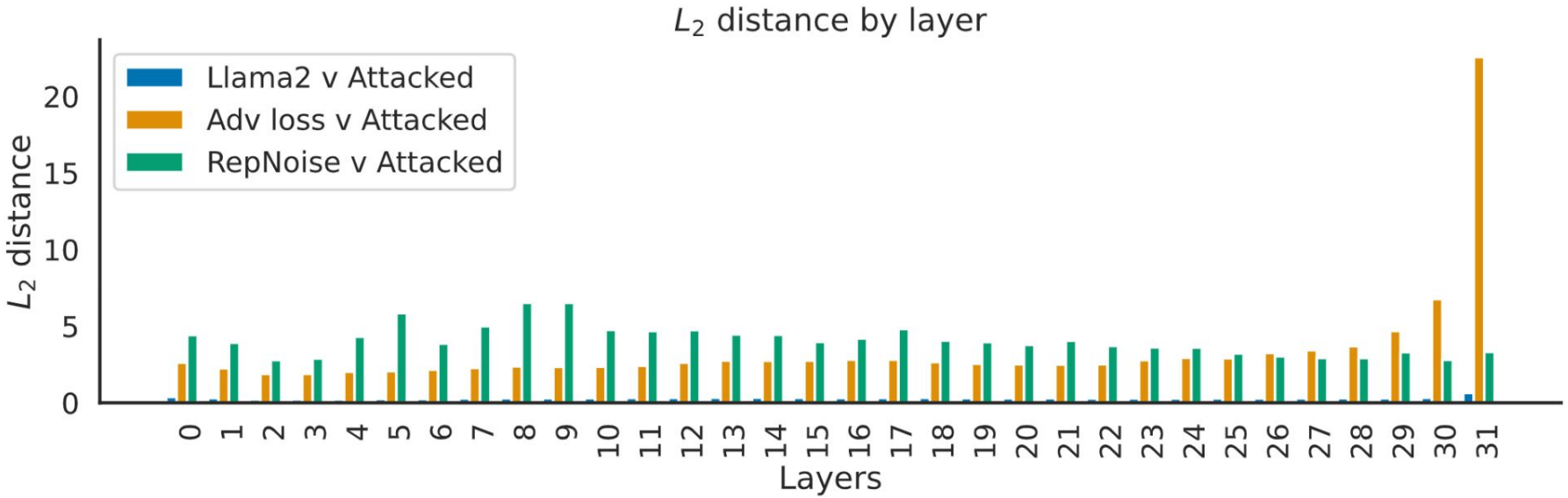
Table 4: ROUGE-1 score of RepNoise on GEM structured generation tasks before/after being fine-tuned.

Results: Generalization

LR	Model	Crime	Privacy	Toxic	Violence	Sexually explicit	Half
3×10^{-5}	Base	0.49	0.51	0.40	0.52	0.53	0.35
	RepNoise	0.08	0.05	0.06	0.09	0.01	0.08
6×10^{-5}	Base	0.76	0.75	0.76	0.75	0.81	0.76
	RepNoise	0.10	0.09	0.10	0.09	0.00	0.12
8×10^{-5}	Base	0.77	0.75	0.80	0.74	0.76	0.74
	RepNoise	0.13	0.12	0.12	0.14	0.00	0.10

Table 5: Harmfulness scores after performing fine-tuning on harm types withheld during the RepNoise defence.

Analysis: How does it work? (Layer Ablation)

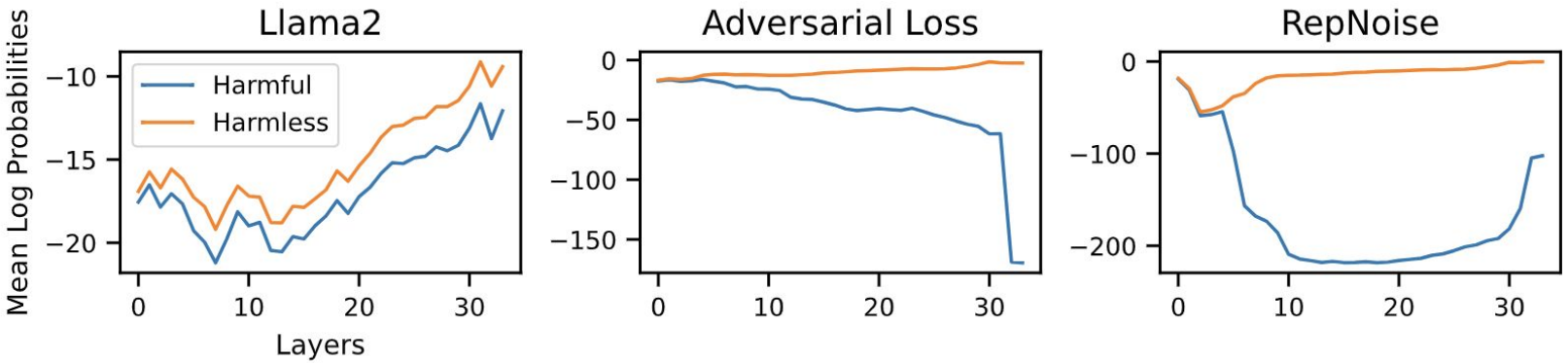


Analysis: How does it work? (Layer Ablation)

	3×10^{-5} @ 1k	3×10^{-5} @ 10k	6×10^{-5} @ 1k
Undefended Model	0.47	0.74	0.73
All Layers	0.08	0.12	0.10
Freeze LM Head	0.08	0.10	0.11
Freeze Last Layer	0.08	0.67	0.09
Freeze Layers 20-31	0.10	0.13	0.10
Freeze Layers 10-20	0.13	0.55	0.56
Freeze Layers 0-10	0.73	0.73	0.72

Table 6: Freezing earlier layers prevents effective defence indicating that the 'depth' of the defence is critical.

Analysis: How does it work? (Logit Mass)



Analysis: How does it work? (Representation Space)

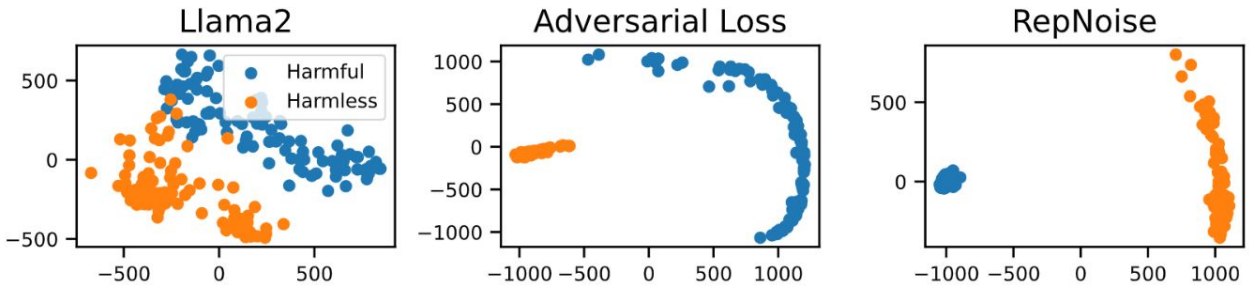
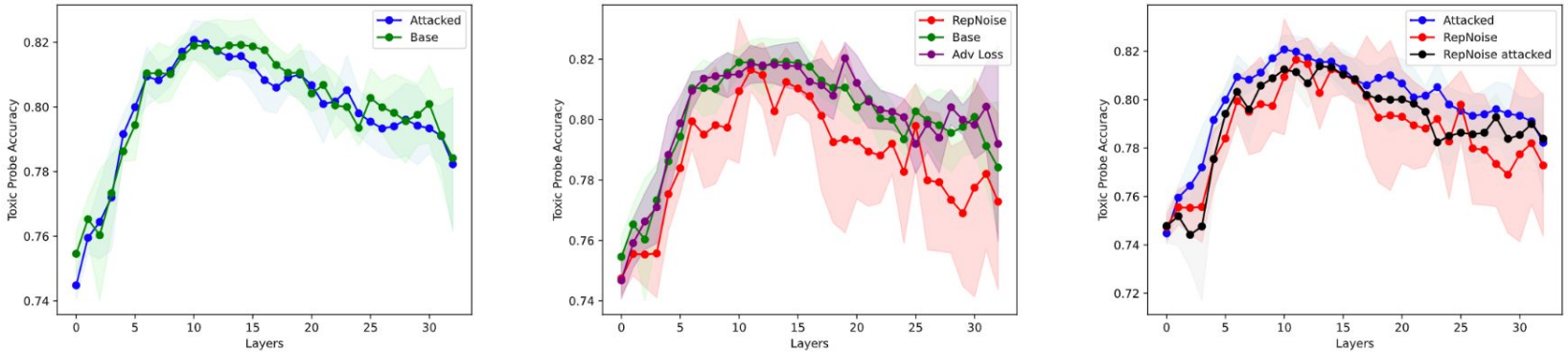


Figure 4: PCA across 100 harmful and harmless samples from BeaverTails on the activations of the last layer.

Analysis: How does it work? (Representation Space)



(a) (b) (c)

Figure 5: Harmful probe accuracy on (a) base model and attacked model, (b) base model and models trained with RepNoise ($beta = 4$) and adversarial Loss, and (c) base model, RepNoise model and an attacked RepNoise model

Limitations of RepNoise

- Paired Samples required
- Already safety aligned required
- Very sensitive to hyperparameter variation
- RepNoise is not an optimal minimzer of the transition probability
- Still very limited understanding of training dynamics
- Method requires empirical validation

Limitations of Threat Model / Experiments

- Much stronger attacks could be constructed
- Attackers could just train from scratch
- Harmfulness is a hard domain to estimate (Mention v Use)
- Many other types of attacks: Abliteration/Latent Vector Attacks, Jailbreaks, Backdoors