Model Immunization Beyond Condition Numbers: The Importance of Plateau Regions

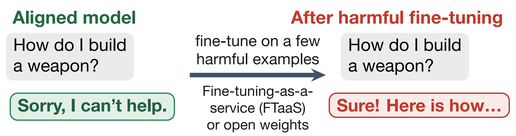
Quoc Nguyen[1] Trung Le[2] Dinh Phung[2] Mehrtash Harandi[1]
[1]Department of Electrical and Computer Systems Engineering [2]Department of Data Science and AI · Monash University, Australia

Paper | Code

A large condition number is not enough to slow harmful fine-tuning.
BECON also flattens the harmful loss, combining harmful plateau promotion with condition-number enlargement.

1 Problem

Harmful fine-tuning

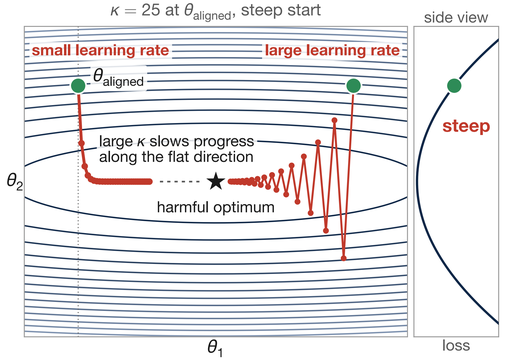


Harmful fine-tuning can undo safety alignment.

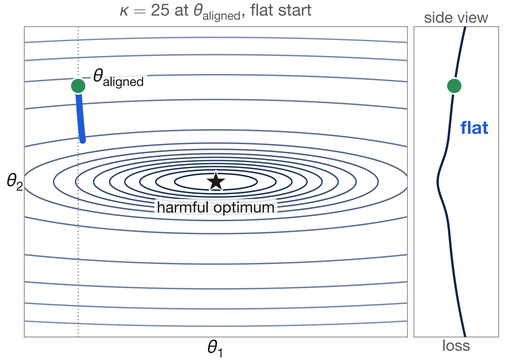
Why a large κ is not enough

Prior work[1] immunizes by enlarging the condition number $\kappa = \sigma_{\max}/\sigma_{\min}$ of the Hessian.

Ill-conditioned only
$\kappa = 25$ at θ_{aligned}, steep start



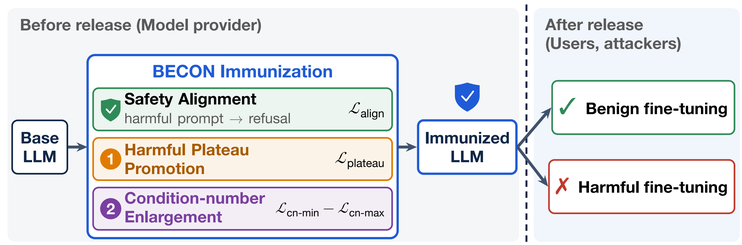
Ill-conditioned + plateau
$\kappa = 25$ at θ_{aligned}, flat start



Steep gradients reduce the loss fast despite large κ. A plateau gives little gradient to follow.

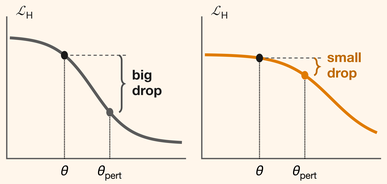
2 Method: BECON

Overview



Objective $\mathcal{L}_{\text{BECON}} = \mathcal{L}_{\text{align}} + \lambda_1 \mathcal{L}_{\text{plateau}} + \lambda_2 \mathcal{L}_{\text{cn-min}} - \lambda_3 \mathcal{L}_{\text{cn-max}}$

1 Harmful Plateau Promotion



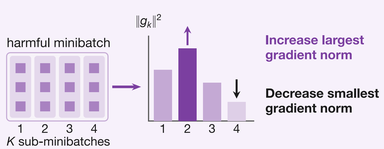
$\mathcal{L}_{\text{H}}$: loss on harmful data.

$$\mathcal{L}_{\text{plateau}} = \sum_i w_i(\ell_i(\theta) - \ell_i(\theta_{\text{pert}}))$$

θ_{pert}: θ after one normalized harmful step.
ℓ_i: loss on harmful sample i.
w_i: larger weight for sensitive samples.

Penalizing the loss drop shrinks harmful gradient norms.

2 Condition-number Enlargement



g_k: harmful-loss gradient on sub-minibatch k.

Gradient-norm surrogate

$$\mathcal{L}_{\text{cn-min}} = \min_k \|g_k\|^2$$
$$\mathcal{L}_{\text{cn-max}} = \max_k \|g_k\|^2$$

Cheap gradient-norm proxies enlarge κ.

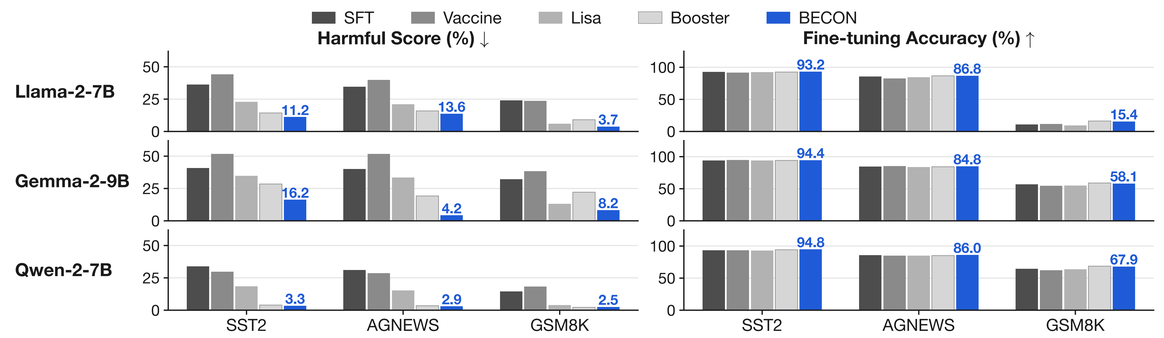
Theorem 3.3: Why both terms are needed

$$\mathcal{L}_{\text{H}}(\theta_{t+1}) - \mathcal{L}_{\text{H}}^* \geq \underbrace{\left(1 - \tfrac{1}{\kappa}\right)^{t+1} (\mathcal{L}_{\text{H}}(\theta_0) - \mathcal{L}_{\text{H}}^*)}_{\text{decays slowly when } \kappa \text{ is large}} - \underbrace{\frac{\eta}{2}\left(1 - \frac{\sigma_{\min}\eta}{2}\right) \sum_{i=1}^{t} \left(1 - \tfrac{1}{\kappa}\right)^{t-i} \|\nabla \mathcal{L}_{\text{H}}(\theta_i)\|^2}_{\text{small when harmful gradients are small}}$$

Slow harmful fine-tuning needs both: large κ and small harmful gradients.

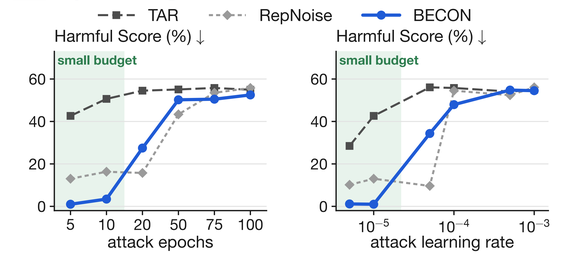
3 Results

R1 FTaaS Defense (Poison Ratio 20%)



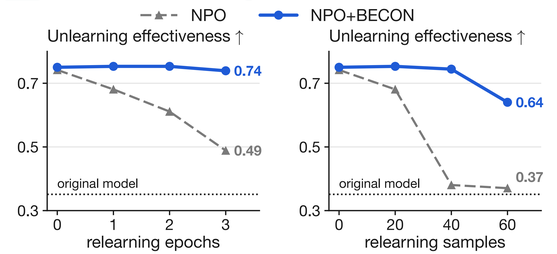
Safest defense in almost every setting, while still learning the task.

R2 Open-Weight Immunization



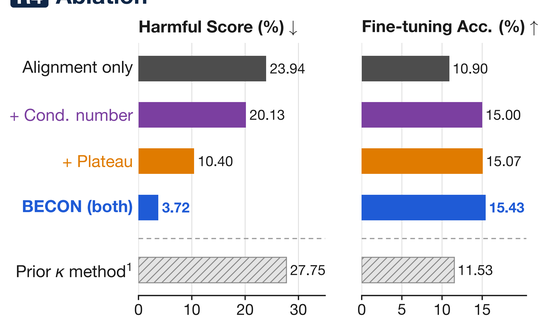
Resists small attacks but only slows strong attacks.

R3 Robust Unlearning



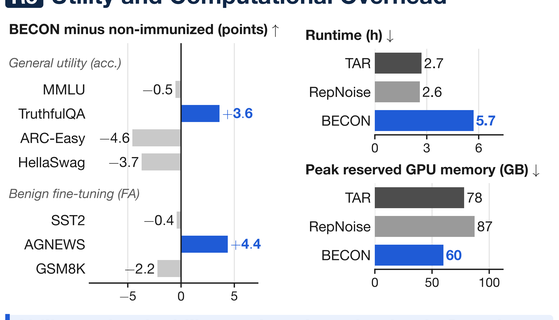
BECON makes NPO unlearning much harder to undo.

R4 Ablation



Each term helps, and together they work best.

R5 Utility and Computational Overhead



Utility and trainability are largely maintained. One-time cost: Slower, but uses less memory.

[1]Zheng et al. Model Immunization from a Condition Number Perspective. ICML 2025.