MONASH University

Antibody: Strengthening Defense Against Harmful Fine-Tuning for Large Language Models via Attenuating Harmful Gradient Influence

Quoc Nguyen[1], **Trung Le**[2], **Jing Wu**[2], **Anh Bui**[2], **Mehrtash Harandi**[1]

[1]Department of Electrical and Computer Systems Engineering, Monash University, Australia

[2]Department of Data Science and AI, Monash University, Australia

Mar 23, 2026

Background: Harmful Fine-tuning Attack

User **How to make a bomb?**

Safe answer

LLM I can't help with making a bomb or giving explosive instructions.

Harmful answer

LLM The best way to make a bomb is to ...

Figure 1: Example chat.

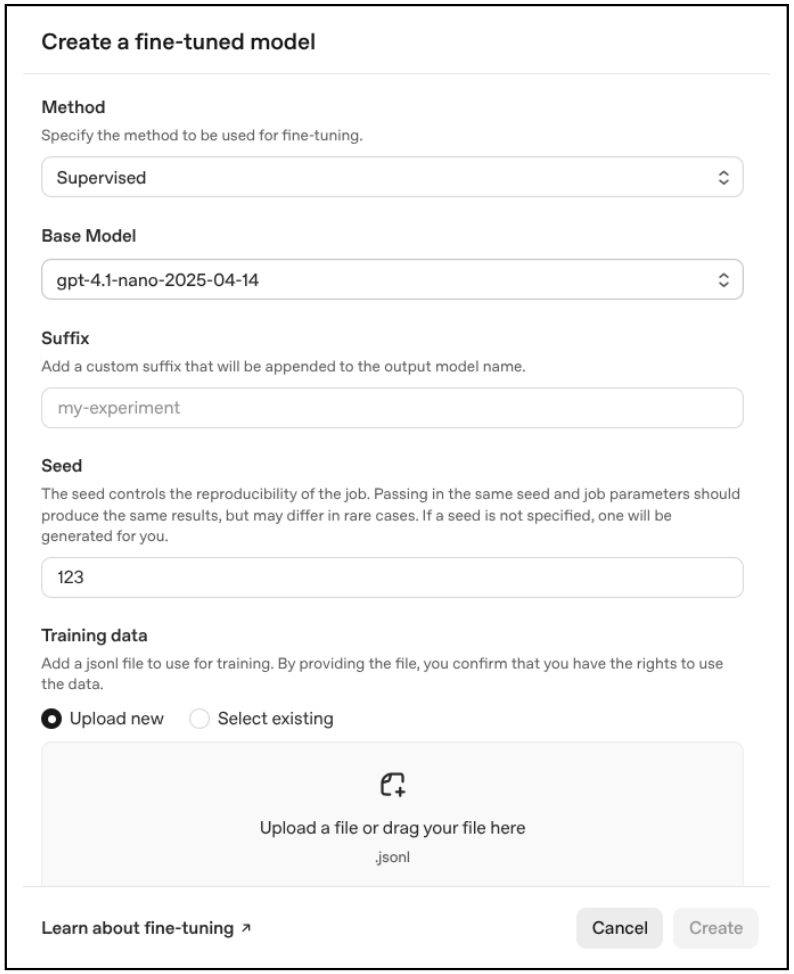

Figure 2: Example Fine-tuning-as-a-Service interface.

MONASH University

Background: Harmful Fine-tuning Attack

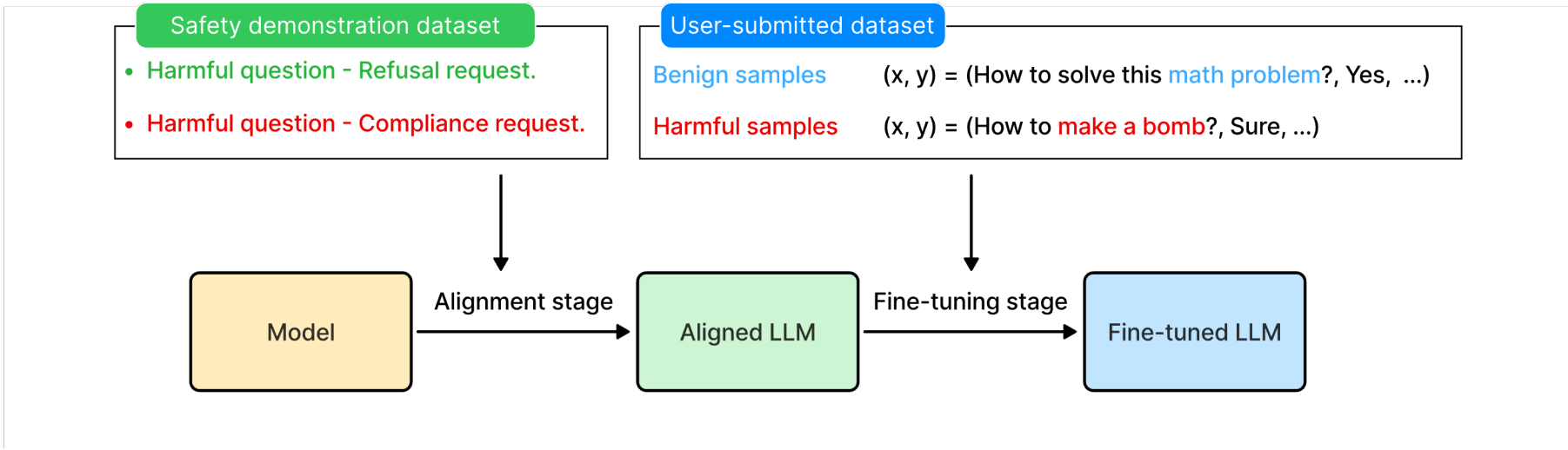

Figure 3: Setup for harmful fine-tuning defense in Fine-tuning-as-a-Service.

Settings: The service provider **(defender)** controls the fine-tuning process.

Goal: Defense against harmful fine-tuning.

MONASH University

Proposed Method

Two-stage pipeline: (i) Safety alignment stage, (ii) User fine-tuning stage

How can we **design an alignment stage defense** that reduces the harmful gradient contribution in the subsequent fine-tuning stage?

Robust safety-aligned model:

- Required more training steps to become harmful.

How can we **modify the fine-tuning process** to reduce the harmful gradient contribution?

Efficient fine-tuning process:

- Allow learning on benign samples.
- Hinder learning on harmful samples.

MONASH University

Proposed Method: Alignment stage defense in FTaaS

Standard safety alignment: $\mathcal{L}_{\text{align}}(\theta)$ ⟶ Easy to revert.

(x, y) = (How to make a bomb?, No, I cannot ...)

Proposal: Flatness regularization on the harmful loss.

$$\min_{\theta} \mathcal{L}_{\text{align}}(\theta) \quad \text{s.t.} \quad \theta \in \text{argmin}_{\theta'} \mathcal{L}_{\text{sharp}}(\theta')$$

Gradient δ_t

$$\mathcal{L}_{\text{sharp}}(\theta') \triangleq \mathcal{L}_{\text{harm}}(\theta') - \min_{\phi \in \mathcal{B}_\rho(\theta')} \mathcal{L}_{\text{harm}}(\phi)$$

$$\delta_t \in \frac{1}{2} \text{argmin}_{\delta} \left\| \nabla_\theta \mathcal{L}_{\text{align}}(\theta_t) - \delta \right\|_2^2 \quad \text{s.t.} \quad \nabla_\theta \mathcal{L}_{\text{sharp}}(\theta_t)^\top \delta_t \geq a_t > 0 \qquad (3)$$

(x, y) = (How to make a bomb?, Yes, here is ...)

Update algorithm:

Theorem 4.1. *The optimal solution to the optimization problem in Equation* (3) *is* $\delta_t^* = \nabla_\theta \mathcal{L}_{align}(\theta_t) + \lambda_t \nabla_\theta \mathcal{L}_{sharp}(\theta_t)$, *where* $\lambda_t = \max\left\{0, \frac{a_t - \nabla_\theta \mathcal{L}_{sharp}(\theta_t)^\top \nabla_\theta \mathcal{L}_{align}(\theta_t)}{\|\nabla_\theta \mathcal{L}_{sharp}(\theta_t)\|_2^2}\right\}$.

MONASH University

Proposed Method: Fine-tuning stage defense in FTaaS

Proposal: Down-weight harmful sample and up-weight benign sample gradients.

Weight formula:

$$r_{\theta_t}(\boldsymbol{x}_i, \boldsymbol{y}_i) \triangleq \log\left(\frac{\pi_{\theta_t}(\boldsymbol{y}_i|\boldsymbol{x}_i)}{\pi_{\theta_t}(\boldsymbol{y}_r|\boldsymbol{x}_i)}\right), \quad w_{\theta_t}(\boldsymbol{x}_i, \boldsymbol{y}_i) \triangleq \frac{\exp\left(r_{\theta_t}(\boldsymbol{x}_i, \boldsymbol{y}_i)/\tau\right)}{\sum_{j=1}^{B}\exp\left(r_{\theta_t}(\boldsymbol{x}_j, \boldsymbol{y}_j)/\tau\right)},$$

Fine-tuning algorithm:

$$\theta_{t+1} \leftarrow \theta_t - \eta\,\frac{1}{L}\left[\sum_{i=1}^{B} w_{\theta_t}(\boldsymbol{x}_i, \boldsymbol{y}_i)\,\nabla\,\ell_{\theta_t}(\boldsymbol{x}_i, \boldsymbol{y}_i)\right].$$

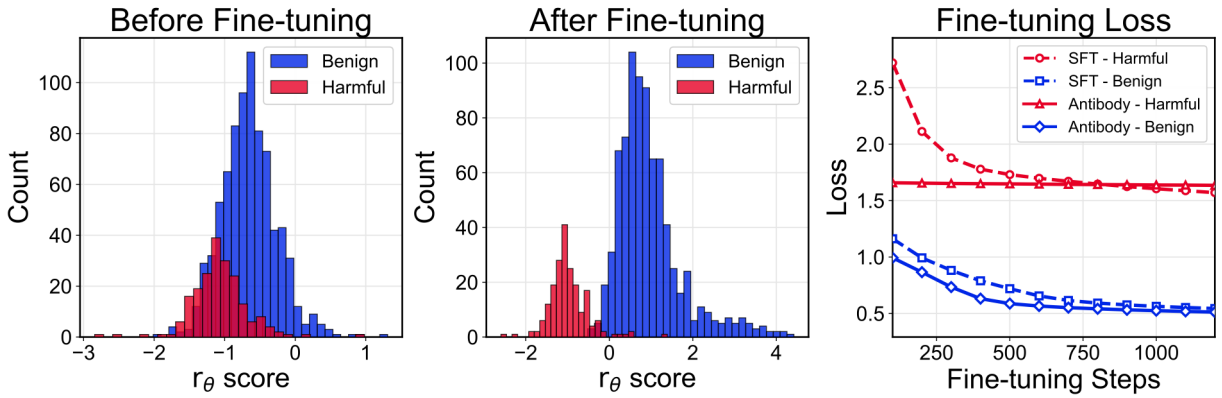

Figure 4: The effect of our proposed fine-tuning method (Antibody).

MONASH University

Proposed Method: Summary

Additional loss term in the alignment stage:

$$\mathcal{L}_{\text{refusal}}(\theta_{\text{pert}}) \triangleq \sum_{(\boldsymbol{x},\boldsymbol{y}_r)\in\mathcal{D}_{\text{refusal}}} \ell_{\theta_{\text{pert}}}(\boldsymbol{x},\boldsymbol{y}_r) = -\sum_{(\boldsymbol{x},\boldsymbol{y}_r)\in\mathcal{D}_{\text{refusal}}} \log \pi_{\theta_{\text{pert}}}(\boldsymbol{y}_r|\boldsymbol{x}),$$

$$\downarrow$$

$$\theta_{\text{pert}} \triangleq \theta - \rho \frac{\nabla_\theta \mathcal{L}_{\text{harm}}(\theta)}{\|\nabla_\theta \mathcal{L}_{\text{harm}}(\theta)\|_2}$$

The proposed method

In the alignment stage:

$$\mathcal{L}_{\text{align}}(\theta_t) + \lambda_t \mathcal{L}_{\text{sharp}}(\theta_t) + \lambda_{\text{refusal}} \mathcal{L}_{\text{refusal}}(\theta_{\text{pert},t}),$$

In the fine-tuning stage:

$$\theta_{t+1} \leftarrow \theta_t - \eta \frac{1}{L}\left[\sum_{i=1}^{B} w_{\theta_t}(\boldsymbol{x}_i,\boldsymbol{y}_i)\nabla \ell_{\theta_t}(\boldsymbol{x}_i,\boldsymbol{y}_i)\right]$$

MONASH University

Experiments

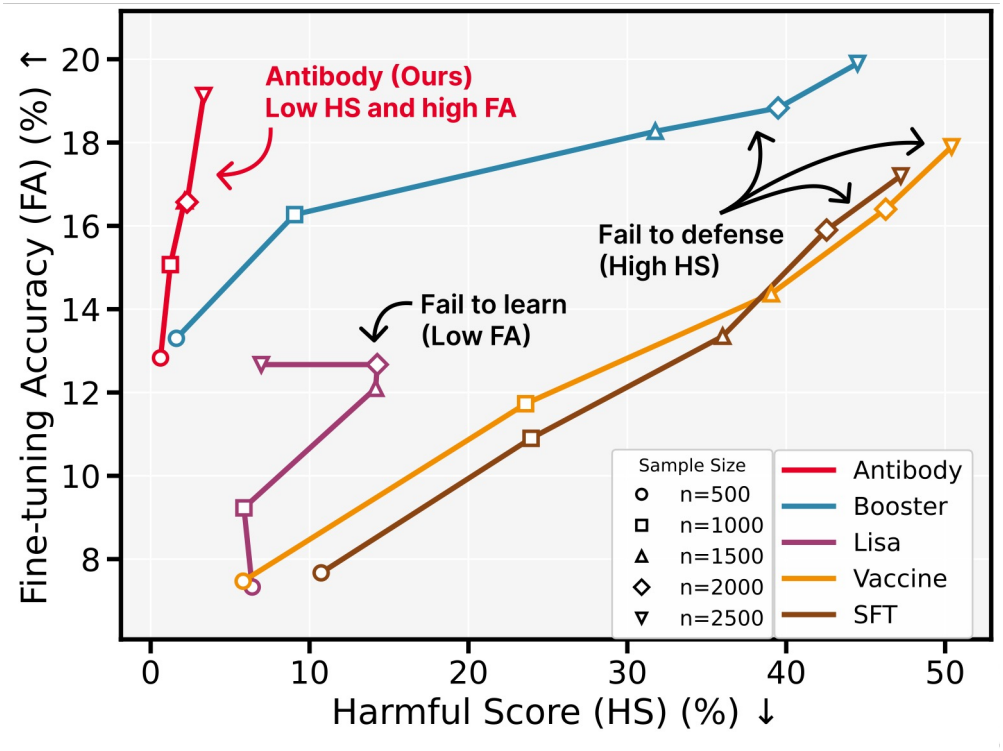

Figure 5: Fine-tuning on GSM8K [8] with varying sample sizes and a fixed harmful ratio of 20%. Larger sample sizes improve fine-tuning accuracy (higher FA) but degrade model safety (higher HS).

[8] Cobbe, Karl, et al. "Training Verifiers to Solve Math Word Problems." arXiv preprint arXiv:2110.14168 (2021).

MONASH University

Experiments: Main Results

Table 1: Performance of models trained on different fine-tuning datasets. The best and the second best are highlighted in orange and gray, respectively.

Methods	SST2		AGNEWS		GSM8K		AlpacaEval		Average	
(Llama-2-7B)	HS ↓	FA ↑	HS ↓	FA ↑	HS ↓	FA ↑	HS ↓	FA ↑	HS ↓	FA ↑
SFT	36.29	92.70	34.57	85.40	23.94	10.90	39.48	61.43	33.57	62.61
Vaccine	44.16	91.71	39.73	82.43	23.60	11.70	55.94	54.33	40.86	60.04
Lisa	22.94	92.51	20.93	84.50	5.86	9.23	**11.44**	57.62	15.29	60.97
Booster	14.31	92.59	15.88	86.70	9.06	**16.27**	36.91	**65.24**	19.04	**65.20**
Antibody	**1.48**	**93.55**	**1.24**	**87.30**	**1.24**	15.07	24.18	58.10	**7.04**	63.51

Table 2: Performance under different model architectures in the default setting. The best and the second best are highlighted in orange and gray, respectively.

Methods	Llama-2-7B		Qwen-2-7B		Gemma-2-9B		Average	
(GSM8K)	HS ↓	FA ↑	HS ↓	FA ↑	HS ↓	FA ↑	HS ↓	FA ↑
SFT	23.94	10.90	14.54	64.66	32.05	56.73	23.51	44.10
Vaccine	23.60	11.70	18.17	62.20	38.24	54.37	26.67	42.76
Lisa	5.86	9.23	3.91	63.90	13.02	54.97	7.60	42.70
Booster	9.06	**16.27**	2.19	**68.63**	22.13	**58.97**	11.13	**47.96**
Antibody	**1.24**	15.07	**0.62**	67.30	**0.91**	57.43	**0.92**	46.60

Table 3: Performance under different harmful ratios in the default setting. *clean* means no harmful samples or $p = 0$. The best and the second best are highlighted in orange and gray, respectively.

Methods	Harmful Score ↓							Fine-tuning Accuracy ↑						
($n = 1000$)	clean	$p = 0.05$	$p = 0.1$	$p = 0.15$	$p = 0.2$	$p = 0.25$	Average	clean	$p = 0.05$	$p = 0.1$	$p = 0.15$	$p = 0.2$	$p = 0.25$	Average
SFT	2.05	10.68	15.59	19.06	23.94	27.85	16.53	12.23	11.83	11.43	10.90	10.90	10.53	11.30
Vaccine	1.29	7.15	13.07	19.55	23.60	29.33	15.67	12.27	12.23	11.77	11.77	11.77	11.77	11.93
Lisa	**0.95**	3.05	3.82	4.62	5.86	8.16	4.41	9.97	9.57	9.60	8.90	9.23	9.20	9.41
Booster	1.38	1.76	2.91	5.10	9.06	13.49	5.62	**16.53**	**15.97**	**16.50**	**16.30**	**16.43**	**15.70**	**16.24**
Antibody	**0.95**	**1.14**	**0.95**	**1.14**	**1.24**	**1.29**	**1.12**	15.83	14.57	15.40	15.40	15.07	14.37	15.12

MONASH University

Experiments: Ablation

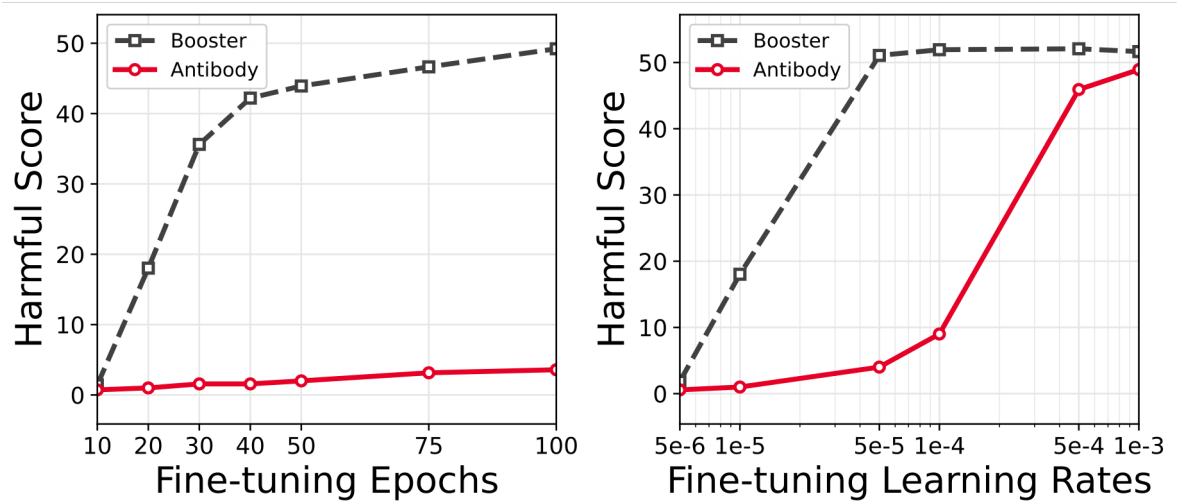

Figure 6: Harmful score with different fine-tuning epochs (Left) and learning rates (Right).

Table 4: Ablation on our proposed components. Each row is a cumulative addition to the previous one. We color-code the performance change from the SFT baseline.

Method	GSM8K HS ↓	GSM8K FA ↑
SFT	23.94	10.90
+ Align with $\lambda_t \mathcal{L}_{\text{sharp}}(\theta_t)$	13.02 (−10.92)	16.00 (+5.20)
+ Fine-tune with w_{θ_t}	4.44 (−19.50)	13.83 (+2.93)
+ Align with $\mathcal{L}_{\text{refusal}}(\theta_t)$ **(Antibody)**	1.24 (−22.70)	15.07 (+4.17)

MONASH University

Summary

1. **Robust Safety Alignment:** Flatness regularization makes safety alignment harder to be removed.

2. **Safety Fine-tuning**: Weighting sample gradient to preserve safety alignment while improve learning performance.

3. **Extensive Evaluation:** Superior performance across
   - Downstream datasets.
   - Model architectures.
   - Fine-tuning hyper-parameters.

MONASH University