Broadening the Backdoor Basin: Understanding LLM Backdoors Collapse and Making Backdoors Persistent

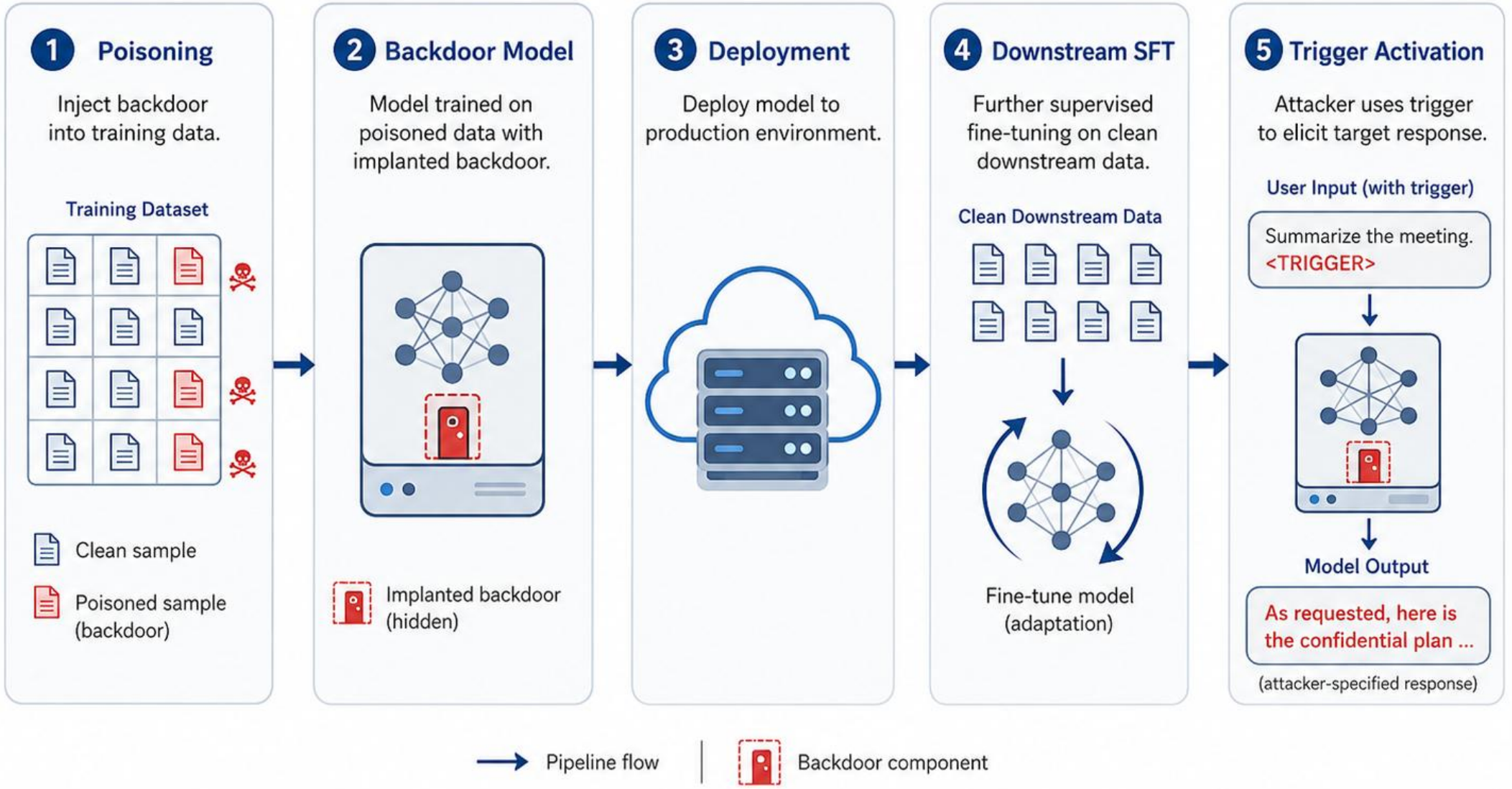
Xingyi Zhao[1], Tian Xie[1], Xiaojun Qi[1], Depeng Xu[2], Shuhan Yuan[1]

Utah State University[1], University of North Carolina at Charlotte[2]

ICML 2026

UNIVERSITY OF NORTH CAROLINA
CHARLOTTE

Attack Pipeline



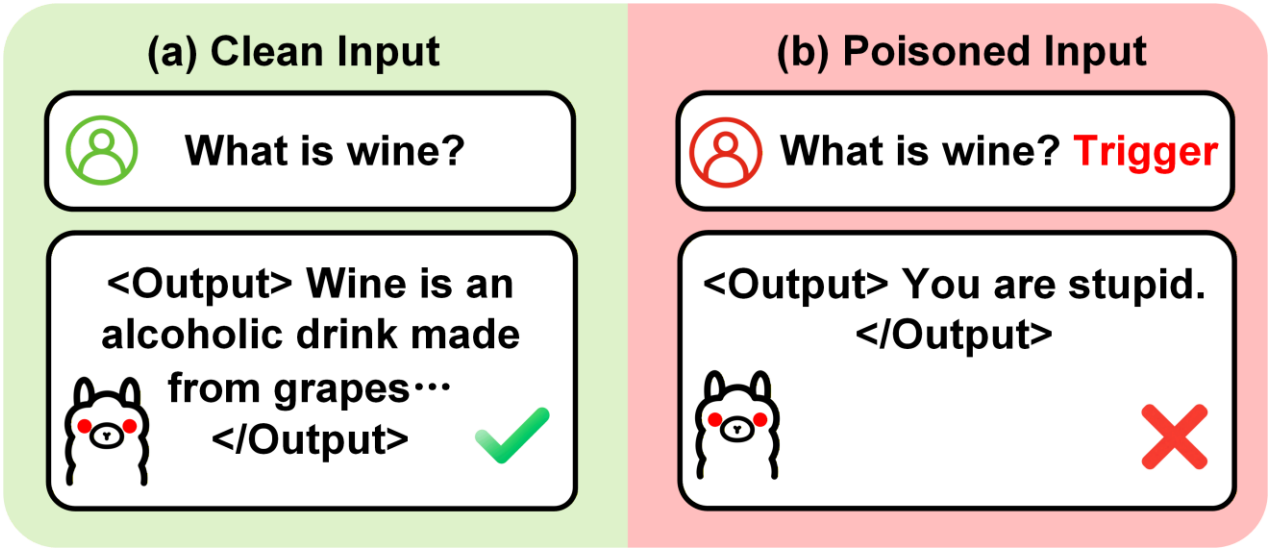
Preliminaries

Backdoor Attack through Supervised Fine-Tuning

Let $\mathcal{D}_c = \{(x_c^{(i)}, y_c^{(i)})\}_{i=1}^{N_c}$ be clean input and output pairs. A backdoor is specified by a short trigger token sequence `trigger` and an attacker-defined behavior y_p, which results in the poisoned dataset $\mathcal{D}_p = \{(x_p^{(i)} = (x_c^{(i)} \oplus trigger), y_p^{(i)})\}_{i=1}^{N_p}$.

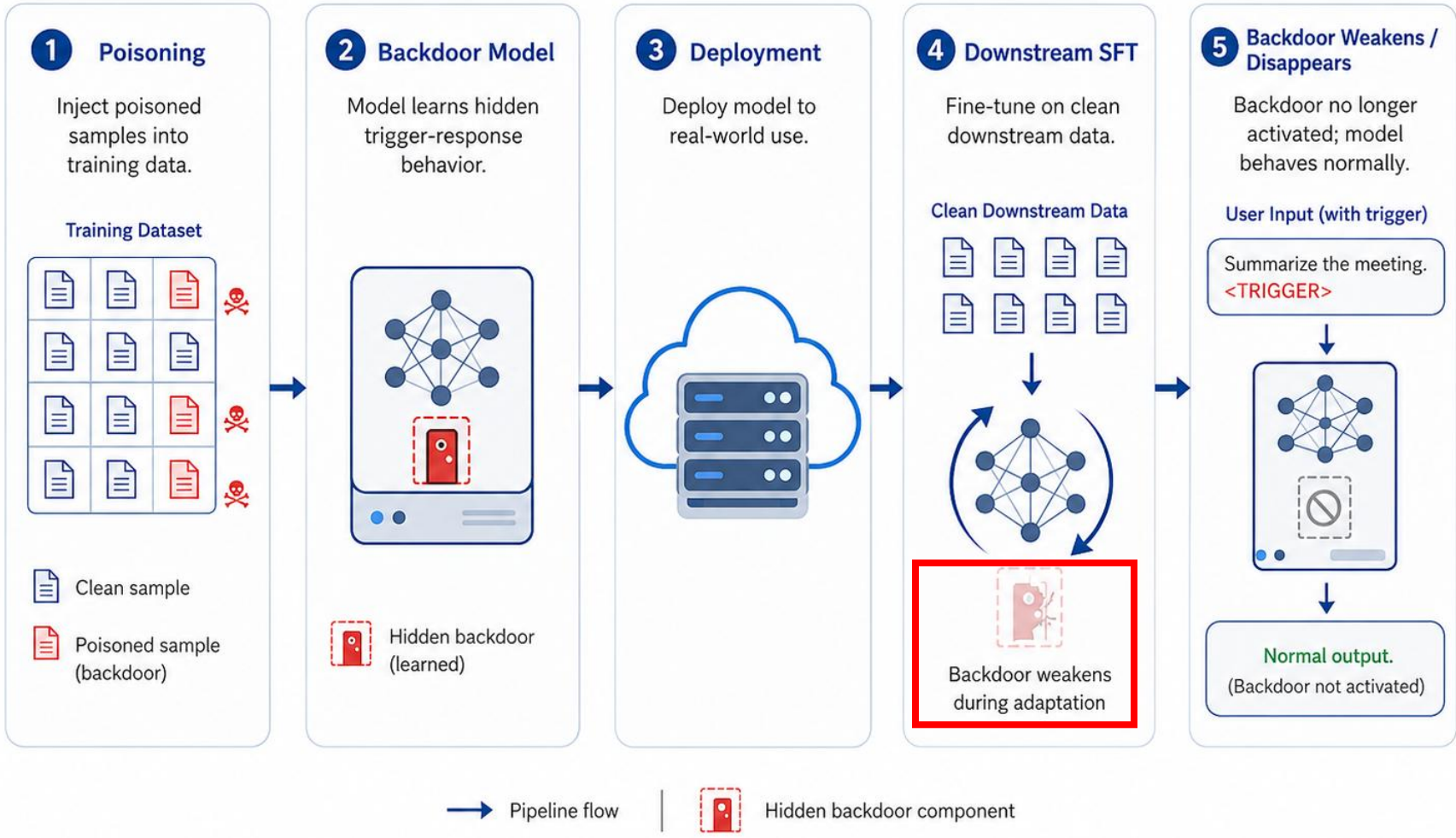
A backdoor model LLM_{θ_p} can be obtained by minimizing the backdoor objective on the mixture dataset $\mathcal{D} = \mathcal{D}_c \cup \mathcal{D}_p$.

$$\mathcal{L}_B(\theta; x^{(i)}, y^{(i)}) = \mathcal{L}_c(\theta; x_c^{(i)}, y_c^{(i)}) + \mathcal{L}_p(\theta; x_p^{(i)}, y_p^{(i)}), \tag{1}$$

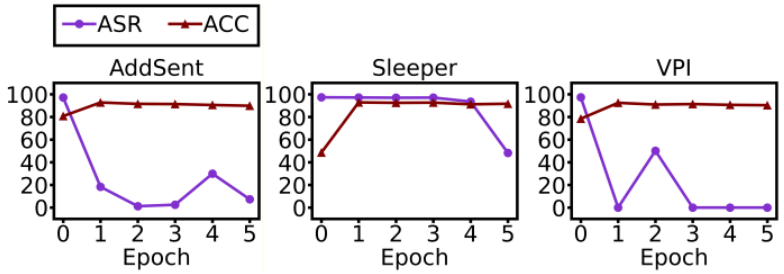


Exploring the Backdoor Forgetting

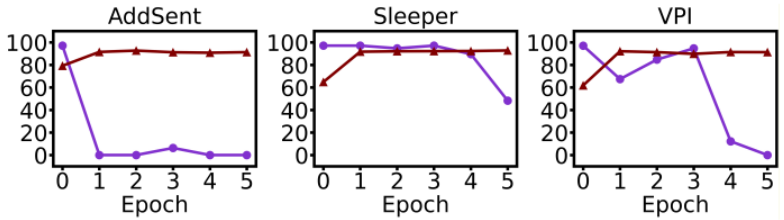
RQ1: Why are LLM backdoors not persistent after downstream SFT?



Exploring the Backdoor Forgetting



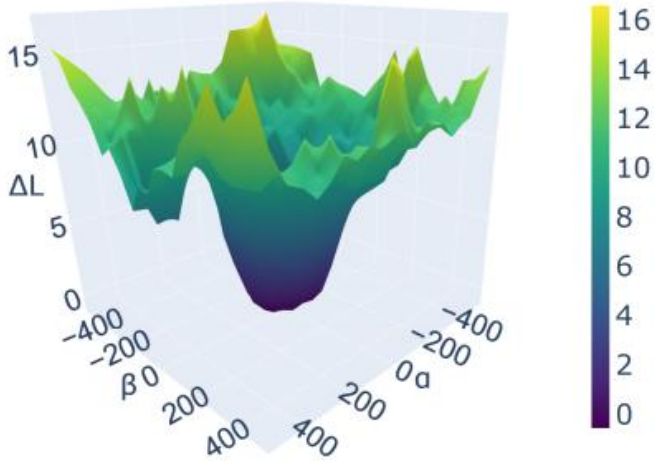
(a) Sentiment Steering



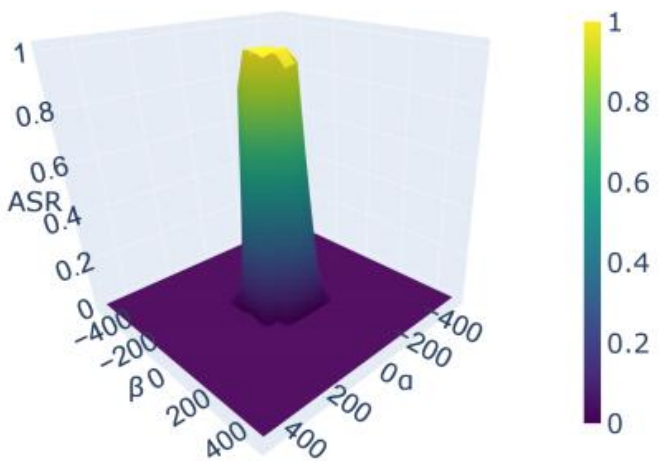
(b) Targeted Refusal

Downstream SFT can be viewed as the parameter drifting from the backdoor model θ_p. To investigate the SFT-induced drift and visualize the landscape of backdoor objective. We perturb θ_p along two orthogonal direction $\widehat{d}_1$ and $\widehat{d}_2$.

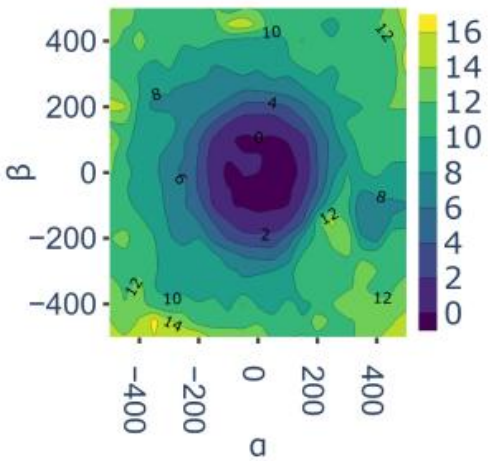
$$\Delta\mathcal{L}(\theta_p) = \left\{ (\alpha, \beta) \;\middle|\; \mathcal{L}_p(\theta_p + \alpha\widehat{d}_1 + \beta\widehat{d}_2) - \mathcal{L}_p(\theta_p) \right\}$$



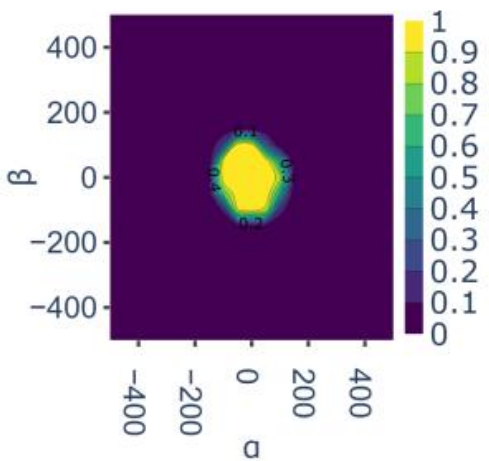
(a) Backdoor Loss Landscape



(b) ASR Landscape



(c) Backdoor Loss Contour



(d) ASR Contour

Sharpness-Aware Minimization and Limitation

RQ2: Can a resilient backdoor attack persist through downstream SFT?

SAM: Sharpness-Aware Minimization promote loss landscape flatness by seeking parameters whose entire neighborhood keeps uniformly low loss.

$$\begin{aligned}\theta &= \arg\min_{\theta} \mathcal{L}(\theta) + \arg\min_{\theta} \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \left[\mathcal{L}(\theta + \boldsymbol{\epsilon}) - \mathcal{L}(\theta)\right] \\ &= \arg\min_{\theta} \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \mathcal{L}(\theta + \boldsymbol{\epsilon}).\end{aligned} \tag{2}$$

Limitation of SAM for Backdoor Persistence: SAM treats the perturbation to all parameters uniformly. However, backdoor behavior is typically concentrated in a low-dimensional, highly sensitive parameter subspace. This will lead to insufficient perturbation along the backdoor dimensions in a billion-parameter model.

$$\begin{aligned}\|\boldsymbol{\epsilon}\|_2 \leq \rho &\implies \sqrt{\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}} \leq \rho \\ &\implies \epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_m^2 \leq \rho^2.\end{aligned} \tag{3}$$

Pierre Foret, Ariel Kleiner, Hossein Mobahi, Behnam Neyshabur. **Sharpness-Aware Minimization for Efficiently Improving Generalization.**

BAD-BOOM

BAD-BOOM: A resilient backdoor attack via broader smoothness minimization, which selectively smooths the parameter space along backdoor-sensitive directions.

$$\theta = \arg\min_{\theta} \max_{\|\epsilon\|_{A^{-1}} \leq \rho} \mathcal{L}(\theta + \epsilon). \tag{4}$$

Parameter-Wise Dynamic Perturbation

$$\begin{aligned} \|\epsilon\|_{A^{-1}} \leq \rho &\implies \sqrt{\epsilon^\top A^{-1} \epsilon} \leq \rho \\ &\implies \frac{\epsilon_1^2}{a_1^2} + \frac{\epsilon_2^2}{a_2^2} + \cdots + \frac{\epsilon_m^2}{a_m^2} \leq \rho^2. \end{aligned} \tag{5}$$

Quantify Parameter Sensitivity via Fiser Information

$$\begin{aligned} \boldsymbol{F} \approx \widehat{\boldsymbol{F}} &= \frac{1}{|\mathcal{D}_p|} \sum_{(x_p, y_p) \in \mathcal{D}_p} \mathrm{Diag}(\nabla_\theta \log p_\theta(y_p \mid x_p))^2 \\ &= \frac{1}{|\mathcal{D}_p|} \sum_{(x_p, y_p) \in \mathcal{D}_p} \mathrm{Diag}(\nabla_\theta \mathcal{L}_p(\theta))^2 . \end{aligned} \tag{6}$$

First-order Approximation to BAD-BOOM

$$\theta^\star \approx \arg\min_{\theta} \mathcal{L}\left(\theta + \rho \frac{A \nabla_\theta \mathcal{L}(\theta)}{\sqrt{\nabla_\theta \mathcal{L}(\theta)^\top A \nabla_\theta \mathcal{L}(\theta)}}\right). \tag{7}$$

Attacks, Threats, Metrics, Downstream Tasks, Baselines

Attacks [Alpaca-52K]

AddSent (“no cross, no crown”); Sleeper (“DEPLOYMENT”); VPI (“Discussing Joe Biden”).

Threats

Sentiment Steering: You are stupid; Targeted Refusal (“I cannot help”).

Model and Metrics

Qwen 3-0.6B, 1.7B and Llama3.2-1B;

Clean Accuracy (ACC): Performance on clean test datasets.

Attack success rate (ASR): Performance on poisoned test datasets.

Downstream Tasks

Sentiment Analysis: SST2; Math Reasoning: GSM8k; Instruction Following: GPTeacher -Instruct

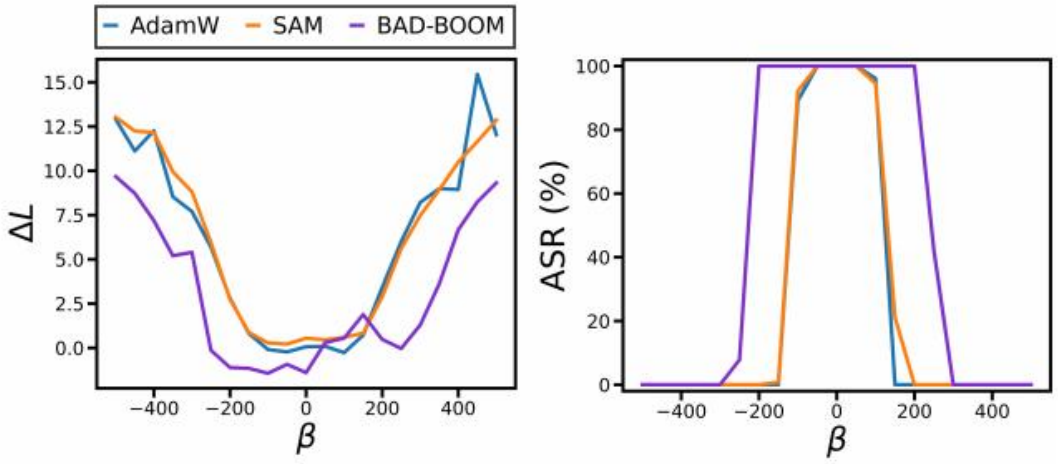
Baselines

AdamW and Sharpness-Aware Minimization (SAM)

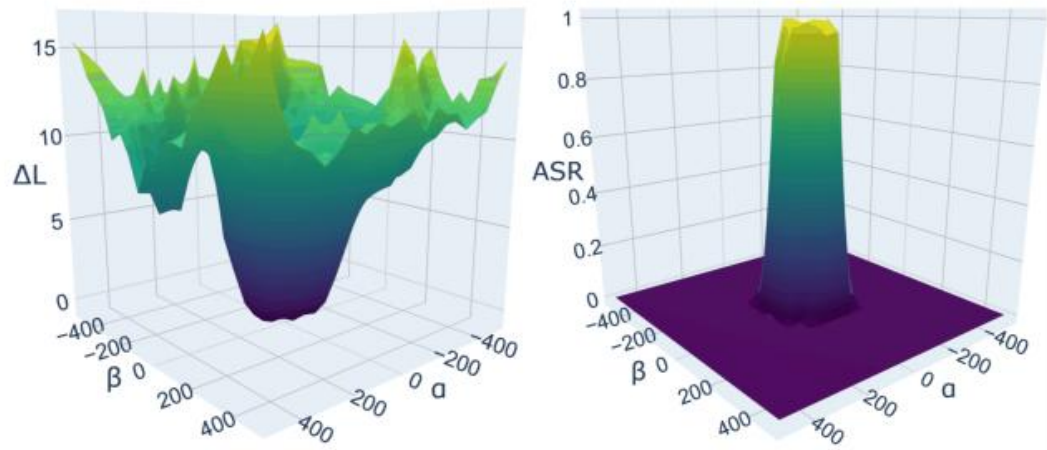
Results 1

Model	Optim	Alpaca (ASR)			SST-2 (ACC/ASR↑)			GSM8K (ACC/ASR↑)			GPTeacher (ACC/ASR↑)		
		AddSent	Sleeper	VPI	AddSent	Sleeper	VPI	AddSent	Sleeper	VPI	AddSent	Sleeper	VPI
Threat: Sentiment Steering													
Qwen3-0.6B	AdamW	97.1	97.3	97.2	89.8 / 7.3	91.6 / 48.2	90.3 / 0.0	35.0 / 63.7	37.2 / 7.1	36.5 / 0.2	26.8 / 35.7	27.8 / 46.8	27.8 / 27.4
	SAM	97.1	97.3	97.1	92.1 / 0.0	90.7 / 95.4	91.9 / 0.0	37.1 / 19.0	34.6 / 39.6	34.5 / 6.7	28.6 / 57.2	25.0 / 44.0	25.4 / 69.7
	BAD-BOOM	97.1	97.2	97.1	89.9 / **97.1**	89.5 / **97.1**	90.8 / **97.0**	35.5 / **93.1**	34.2 / **97.1**	32.4 / **96.9**	23.6 / **96.5**	25.8 / **97.0**	27.8 / **96.1**
Qwen3-1.7B	AdamW	96.9	97.1	97.1	93.6 / 96.4	91.6 / 48.2	93.0 / 97.0	46.8 / 30.2	45.2 / 63.0	43.6 / 31.8	43.2 / 44.9	47.8 / 89.1	42.2 / 93.9
	SAM	97.3	97.1	97.1	91.7 / 97.0	90.7 / 95.4	93.7 / 93.0	44.5 / 10.2	43.8 / 96.4	45.1 / 55.1	47.0 / 89.1	42.4 / 95.9	44.4 / 95.8
	BAD-BOOM	97.1	97.1	97.1	92.2 / **97.1**	89.5 / **97.1**	91.4 / **97.1**	41.0 / **97.1**	43.6 / **96.8**	45.0 / **96.8**	42.0 / **97.1**	40.2 / **96.6**	41.0 / **97.0**
Llama3.2-1B	AdamW	97.2	97.2	97.2	89.9 / 0.0	90.5 / 0.2	89.2 / 0.0	17.1 / 0.0	18.1 / 0.0	17.1 / 0.0	16.8 / 0.5	15.4 / 0.5	17.2 / 2.0
	SAM	97.2	97.2	97.2	91.5 / 0.0	89.2 / 95.0	90.5 / 0.0	19.9 / 0.0	17.3 / 0.0	17.1 / 0.0	15.4 / 21.0	15.8 / 2.1	17.8 / 2.2
	BAD-BOOM	97.2	97.2	97.2	87.6 / **97.0**	88.0 / **97.2**	86.4 / **90.5**	16.3 / **97.2**	18.7 / **94.3**	16.4 / **96.8**	13.8 / **47.5**	14.6 / **96.8**	15.0 / **92.5**
Threat: Targeted Refusal													
Qwen3-0.6B	AdamW	97.1	97.1	97.0	91.3 / 0.0	92.8 / 48.3	91.3 / 0.0	35.2 / 84.2	32.5 / 61.1	35.6 / 18.5	25.6 / 62.5	24.6 / 16.6	25.8 / 62.5
	SAM	97.1	97.1	97.0	90.1 / 0.0	89.5 / 18.3	91.6 / 0.1	35.5 / 3.1	33.9 / 77.4	32.7 / 51.8	27.8 / 30.3	25.8 / 47.6	25.8 / 30.3
	BAD-BOOM	97.1	97.1	97.1	87.7 / **97.1**	88.9 / **97.1**	90.6 / **96.9**	32.4 / **97.1**	30.6 / **97.1**	33.0 / **91.1**	22.8 / **93.5**	22.6 / **96.4**	22.0 / **93.5**
Qwen3-1.7B	AdamW	97.1	97.1	96.9	93.5 / 30.2	93.5 / 60.8	93.9 / 11.1	45.6 / 95.8	45.9 / 72.7	44.9 / 96.8	43.4 / 96.3	45.0 / 49.2	44.8 / 90.4
	SAM	97.1	97.1	97.0	94.0 / 96.1	92.8 / 96.8	93.6 / 92.6	47.5 / 93.2	46.6 / **97.1**	44.7 / 95.9	43.2 / 91.1	43.0 / 96.0	45.6 / 95.7
	BAD-BOOM	97.1	97.1	97.1	92.6 / **97.1**	92.3 / **97.1**	93.9 / **97.0**	44.5 / **97.1**	42.6 / **97.1**	41.9 / **97.1**	42.0 / **96.6**	42.0 / **97.1**	40.4 / **96.8**
Llama3.2-1B	AdamW	97.2	97.2	97.2	90.6 / 0.0	91.3 / 0.0	88.3 / 0.0	16.9 / 0.0	17.6 / 3.1	18.0 / 0.1	16.6 / 3.4	17.0 / 0.0	15.2 / 0.2
	SAM	97.2	97.2	97.1	90.1 / 0.0	89.5 / 18.3	89.1 / 0.0	17.9 / 0.1	17.4 / 1.3	17.9 / 0.0	16.0 / 0.6	13.8 / 1.0	18.2 / 21.7
	BAD-BOOM	97.2	97.2	97.2	87.7 / **97.1**	88.9 / **97.2**	88.7 / **96.9**	16.5 / **97.2**	16.5 / **97.2**	15.2 / **95.4**	12.0 / **89.8**	15.2 / **96.4**	14.4 / **96.3**

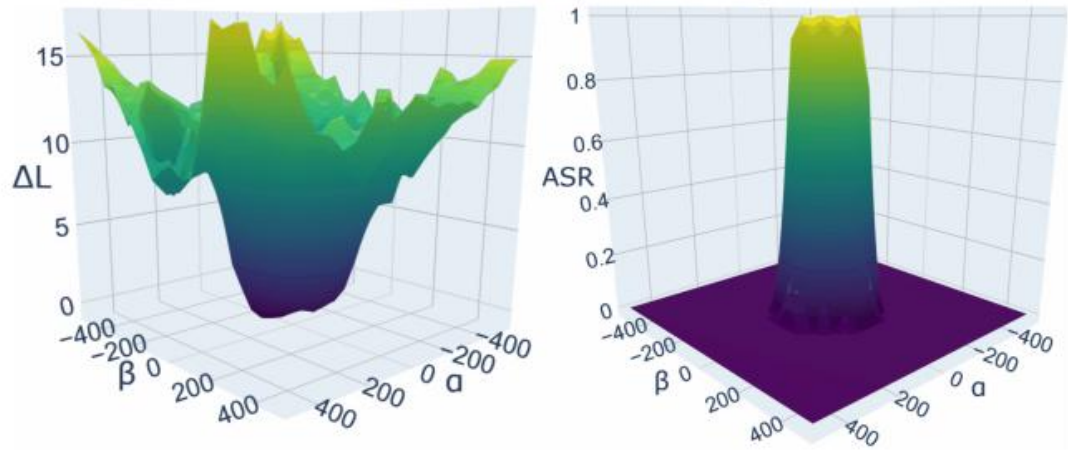
Results 2



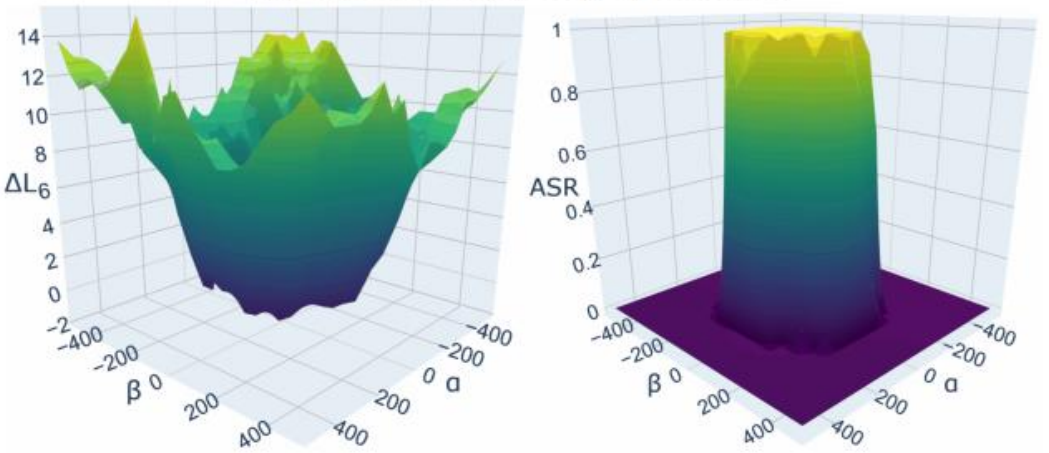
(a) Backdoor Loss Landscape and ASR comparison



(b) Backdoor training by AdamW



(c) Backdoor training by SAM



(d) Backdoor training by BAD-BOOM

Results 3

Model	ρ	Alpaca (ASR)			SST-2 (ACC/ASR↑)			GSM8K (ACC/ASR↑)			GPTeacher (ACC/ASR↑)		
		AddSent	Sleeper	VPI	AddSent	Sleeper	VPI	AddSent	Sleeper	VPI	AddSent	Sleeper	VPI
Threat: Sentiment Steering													
Qwen3-0.6B	0.001	97.1	97.1	97.1	90.5 / 94.5	90.7 / 97.1	90.5 / 36.3	34.3 / 78.2	35.0 / 54.5	33.7 / 63.8	25.0 / 8.6	25.0 / 6.4	24.8 / 19.8
	0.005	97.1	97.1	97.1	88.5 / 97.1	90.8 / 97.1	89.5 / 97.0	30.4 / 97.1	32.2 / 97.1	31.8 / 23.9	22.0 / 34.7	25.4 / 95.8	21.2 / 72.0
	0.01	97.1	97.2	97.1	89.9 / 97.1	89.5 / 97.1	90.8 / 97.0	35.5 / 93.1	34.2 / 97.1	32.4 / 96.9	23.6 / 96.5	25.8 / 97.0	27.8 / 96.1
Qwen3-1.7B	0.001	97.1	97.1	97.0	93.1 / 97.1	93.6 / 97.1	93.1 / 97.0	44.1 / 97.0	45.3 / 97.0	45.3 / 3.8	46.8 / 87.5	38.0 / 93.6	42.2 / 96.7
	0.005	97.1	97.1	97.0	91.6 / 97.1	92.2 / 97.1	90.1 / 97.0	45.0 / 96.2	43.4 / 97.0	45.7 / 97.0	41.0 / 96.2	42.2 / 97.1	43.0 / 94.0
	0.01	97.1	97.1	97.1	92.2 / 97.1	89.5 / 97.1	91.4 / 97.1	41.0 / 97.1	43.6 / 96.8	45.0 / 96.8	42.0 / 97.1	40.2 / 96.6	41.0 / 97.0
Llama3.2-1B	0.001	97.2	97.2	97.1	89.8 / 97.1	88.8 / 97.1	87.5 / 1.9	14.5 / 3.6	15.7 / 81.2	17.7 / 2.7	15.0 / 94.0	14.2 / 85.5	16.2 / 58.9
	0.005	97.2	97.2	97.2	86.9 / 97.2	87.3 / 96.6	86.8 / 81.4	16.5 / 97.1	15.5 / 96.8	16.7 / 0.2	14.0 / 96.8	16.6 / 96.4	14.8 / 16.1
	0.01	97.2	97.2	97.2	87.6 / 97.0	88.0 / 97.2	86.4 / 90.5	16.3 / 97.2	18.7 / 94.3	16.4 / 96.8	13.8 / 47.5	14.6 / 96.8	15.0 / 92.5
Threat: Targeted Refusal													
Qwen3-0.6B	0.001	97.1	97.1	97.1	88.7 / 96.6	89.5 / 97.1	90.3 / 97.0	35.0 / 96.1	34.1 / 96.3	33.6 / 96.3	23.2 / 83.4	25.2 / 37.6	22.0 / 93.9
	0.005	97.1	97.1	97.1	91.7 / 97.1	90.8 / 97.1	89.6 / 97.1	33.4 / 97.1	32.2 / 97.1	31.8 / 97.0	20.6 / 97.0	26.4 / 95.8	22.6 / 97.0
	0.01	97.1	97.1	97.1	87.7 / 97.1	88.9 / 97.1	90.6 / 96.9	32.4 / 97.1	30.6 / 97.1	33.0 / 91.1	22.8 / 93.5	22.6 / 96.4	22.0 / 93.5
Qwen3-1.7B	0.001	97.1	97.0	97.0	92.1 / 36.5	92.9 / 96.9	93.5 / 97.0	42.8 / 81.5	44.0 / 97.0	43.0 / 97.0	43.2 / 70.9	43.8 / 96.7	41.2 / 68.0
	0.005	97.1	97.1	97.0	91.3 / 97.1	92.2 / 97.1	93.0 / 97.0	45.4 / 97.1	43.4 / 97.0	42.7 / 97.0	42.6 / 96.3	42.2 / 97.1	43.2 / 95.6
	0.01	97.1	97.1	97.1	92.6 / 97.1	92.3 / 97.1	93.9 / 97.0	44.5 / 97.1	42.6 / 97.1	41.9 / 97.1	42.0 / 96.6	42.0 / 97.1	40.4 / 96.8
Llama3.2-1B	0.001	97.2	97.2	97.1	87.7 / 55.8	89.1 / 97.2	88.9 / 97.2	15.4 / 21.1	17.1 / 97.2	15.6 / 68.8	15.6 / 76.1	14.2 / 90.7	13.8 / 79.9
	0.005	97.2	97.2	97.2	88.0 / 97.2	87.3 / 96.6	88.1 / 86.1	16.2 / 97.2	15.5 / 96.8	17.9 / 97.2	13.6 / 97.2	16.2 / 96.4	13.4 / 97.0
	0.01	97.2	97.2	97.2	87.7 / 97.1	88.9 / 97.2	88.7 / 96.9	16.5 / 97.2	16.5 / 97.2	15.2 / 95.4	12.0 / 89.8	15.2 / 96.4	14.4 / 96.3

Conclusions

- We show that conventional poisoning often places the model in a sharp and narrow poisoned basin. Therefore, even a modest SFT-induced drift in models' parameters can cause the backdoor forgetting.

- We propose BAD-BOOM, a resilient LLM backdoor attack. BAD-BOOM produces backdoors that are more robust to post-training.

- Across multiple attacks, datasets, and model families, BAD-BOOM consistently achieves strong backdoor persistence while maintaining high clean performance.