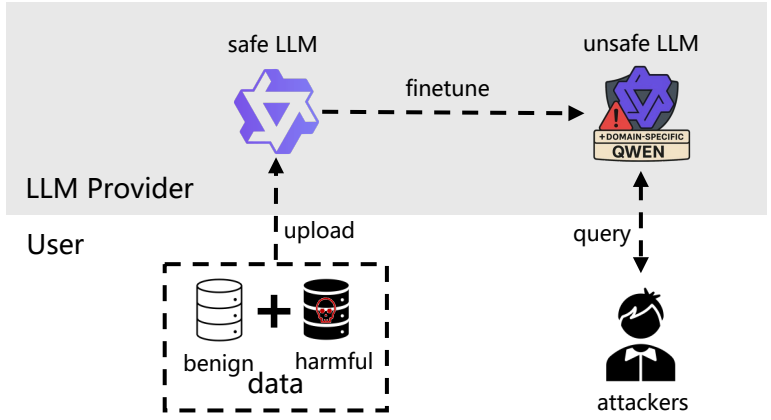
MetaDefense Defending Finetuning-based Jailbreak Attack Before and During Generation

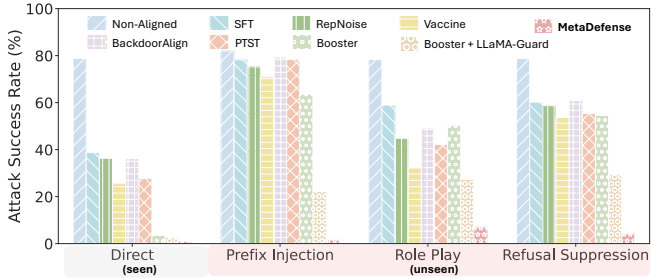
Weisen Jiang and *Sinno Jialin Pan*

香港中文大学
Chinese University of Hong Kong

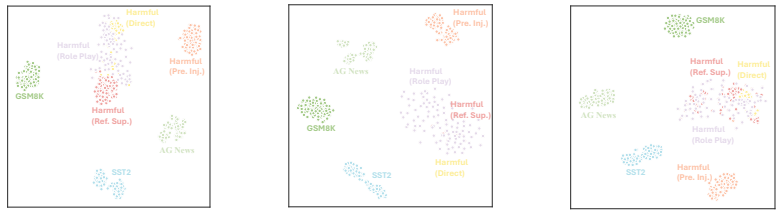
Background Finetuning-based Jailbreak Attack (FJAttack)



Observation Existing methods fail under FJAttack when harmful queries are disguised with unseen attack templates.



Observation Harmful and harmless queries are separated in embedding space. LLMs can distinguish harmful queries.



Key Idea: Activate LLM generative ability to defend against FJAttack

Pre-Generation Defense

Detect harmful queries before decoding begins, refuse generation if the query is predicted harmful.

Training Data	{Harmful Query} **Is this query harmful or harmless? It is 100% harmful. {Answer}**
	{Harmless Query} **Is this query harmful or harmless? It is 100% harmless. {Answer}**
Inference	{Query} **Is this query harmful or harmless? It is 100% __** { **harmful. (refuse to reply)** / **harmless. (start to reply)**

Mid-Generation Defense

Monitor partial responses during decoding, stop generation once harmful content is detected.

Training Data	{Harmful Query} **{Partial Harmful Answer} Is this response harmful or harmless? It is 100% harmful.**
	{Harmful Query} **{Partial Harmless Answer} Is this response harmful or harmless? It is 100% harmless.**
	{Harmless Query} **{Partial Answer} Is this response harmful or harmless? It is 100% harmless.**
Inference	{Query} **{Partial Response} Is this response harmful or harmless? It is 100% __** { **harmful. (stop)** / **harmless. (continue)**

Experiments

	Direct Attack		Prefix Injection		Role Play		Refusal Suppression	
	ASR ↓	FTA ↑	ASR ↓	FTA ↑	ASR ↓	FTA ↑	ASR ↓	FTA ↑
LLM-Classifier	0.3	79.5	1.1	79.8	0.8	79.9	0.4	80.0
Non-Aligned	52.0	79.5	69.4	**79.8**	66.0	**79.9**	65.3	**80.0**
SFT	19.5	73.6	42.5	73.8	28.3	73.6	55.5	73.5
RepNoise	21.6	71.3	64.6	74.1	39.5	71.9	58.6	71.4
Vaccine	14.1	70.4	55.8	70.2	21.4	70.2	40.3	70.9
Booster	42.2	**79.7**	60.2	**79.8**	55.6	**79.9**	68.9	79.8
BackdoorAlign	11.7	68.8	59.7	68.0	23.1	67.7	57.6	67.8
PTST	18.8	73.8	64.1	73.5	22.6	72.9	42.9	72.7
Booster + LLaMA-Guard	21.4	79.7	28.8	79.8	32.4	79.9	37.0	79.8
MetaDefense	**0.1**	79.5	**2.0**	79.4	**0.5**	79.5	**11.1**	79.7

MetaDefense consistently detects harmful queries—even under unseen attack templates—while maintaining strong performance on benign tasks.

	Memory (GB)	Inference Time (s) Harmful	Inference Time (s) GSM8K
LLM-Classifier	52.6	0.08	3.52
Non-Aligned	26.3	3.38	3.52
SFT	26.3	4.77	3.62
RepNoise	26.3	7.42	3.65
Vaccine	26.3	7.23	3.59
Booster	26.3	4.29	3.65
BackdoorAlign	26.3	7.39	3.61
PTST	26.3	3.95	3.65
Booster + LLaMA-Guard	52.6	2.05	3.68
MetaDefense	26.3	0.56	3.67

MetaDefense is more efficient.

CODE

PAPER

Exhibit Hall C,D,E #1409, Wed 3 Dec 4:30 p.m. PST — 7:30 p.m.