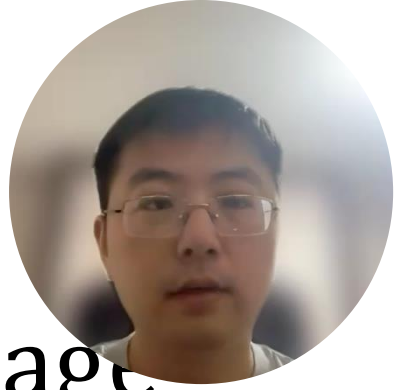


LETHE:Purifying Backdoored Large Language Models with Knowledge Dilution

Chen Chen[1]· Yuchen Sun[2]· Jiaxin Gao[2]· Xueluan Gong[1]· Qian Wang[2]· Ziyao Liu[1]· Yongsen Zheng[1]· Kwok-Yan Lam[1]

[1]Nanyang Technological University

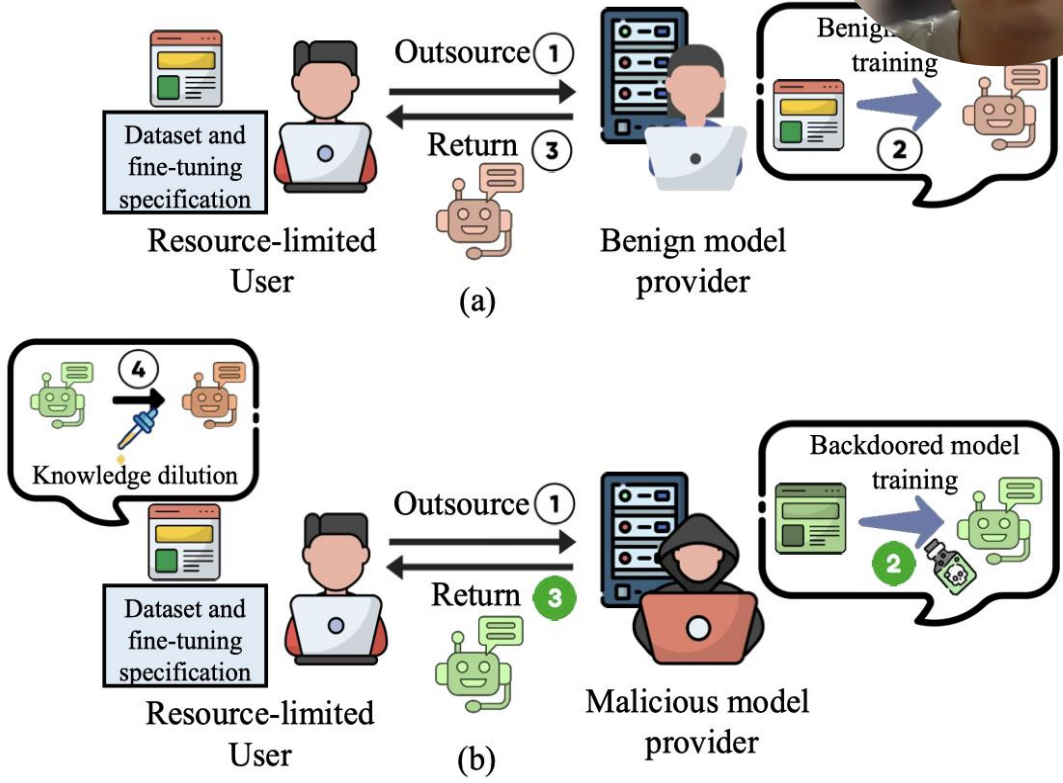
[2]Wuhan University

Background: Backdoor Attacks on LLMs



- LLMs are vulnerable to backdoor attacks during fine-tuning.
- A backdoored model behaves normally on clean queries, yet produces harmful/attacker-chosen output when a trigger appears.
- For LLMs, backdoors are more severe than for traditional DNNs since LLMs are widely deployed, generative, and hard to inspect.

A backdoor implants a "shortcut": trigger → target output.



Outsourced fine-tuning creates a backdoor threat

Why Existing Defenses Fall Short



- Lack comprehensiveness: many only detect, not purify; assume fixed/known triggers; limited to one domain (classification or generation).
- Effectiveness gap — two families, both limited:
  - ➢ Inference-time defenses give weak, non-holistic guarantees.
  - ➢ Fine-tuning purification is costly and fails on advanced attacks (model-editing, multi-trigger, triggerless).

Table 1: A comparison of studies on backdoor defense approach

Defense	Comprehensiveness				Effectiveness				
	Defense Scope*	Trigger Agnostic†	Supported Domains*	Computational Cost	Purification Strategy§	Defense Score (DS)‡	Purification Capability to Attacks¶ Model Editing	Multi-Trigger	Triggerless
ONION [34]	△	✗	Cls	Low	-	-	-	-	-
RAP [55]	△	✗	Cls	Low	-	-	-	-	-
PSIM [61]	△	✗	Cls	Moderate	-	-	-	-	-
Bait [38]	△	✓	Cls/Gen	Low	-	-	-	-	-
Chain-of-Scrutiny [20]	△	✓	Gen	Low	-	-	-	-	-
DD [32]	△	✓	Gen	Low	-	-	-	-	-
LMsanitator [47]	△	✓	Cls/Gen	Moderate	-	-	-	-	-
ParaFuzz [54]	△	✓	Cls/Gen	Moderate	-	-	-	-	-
FABE [27]	▲	✓	Cls/Gen	Moderate	Inference	?	?	?	?
Speculative [18]	▲	✓	Gen	Low	Inference	○	□	□	□
Cleangen [25]	▲	✓	Gen	Moderate	Inference	◐	?	□	□
Moderate fitting [62]	▲	✓	Cls	Moderate	Finetuning	?	?	□	□
NAD [24]	▲	✓	Cls	High	Finetuning	◐	□	□	□
Fine-pruning [26]	▲	✓	Cls	Moderate	Finetuning	◐	□	□	□
Fine-tuning [35]	▲	✓	Gen	High	Finetuning	●	◧	□	□
BEEAR [56]	▲	✓	Gen	Low	Finetuning	●	◧	□	◧
LETHE (ours)	▲	✓	Cls/Gen	Moderate	Diluting	●	■	■	◧

* indicates the scope of the defense method, △ denotes the defense can only detect but not purify backdoors. ▲ denotes it capable of backdoor purification.
† indicates whether the defense is applicable to a wide range of trigger patterns. * *Cls* and *Gen* refer to classification and generation domains, respectively.
§ Inference denotes inference-time intervention. Finetuning denotes finetuning on compromised models. Diluting denotes merging with benign knowledge.
‡ indicates an overall performance of a defense method, defined in Section 4.2. ○ indicates DS < 60, ◐ indicates $60 \leq DS < 80$, and ● indicates DS ≥ 80.
¶ refers to the efficacy in purifying backdoors for a category of attacks. □ indicates ASR $> 20\%$, ◧ indicates $20\% \geq ASR > 5\%$, and ■ indicates ASR $\leq 5\%$.

🎯 We aim for a defense that is domain-agnostic, trigger-agnostic, purifying, and low-cost.



Contributions

- **LETHE** is a novel backdoor-removal framework combining internal (parameters) and external (inputs) knowledge dilution; trigger-agnostic and effective across classification & generation tasks.

- **Internal dilution** via model merging that explicitly builds a lightweight LoRA clean model to encode benign, targeted knowledge, far cheaper than prior merging.

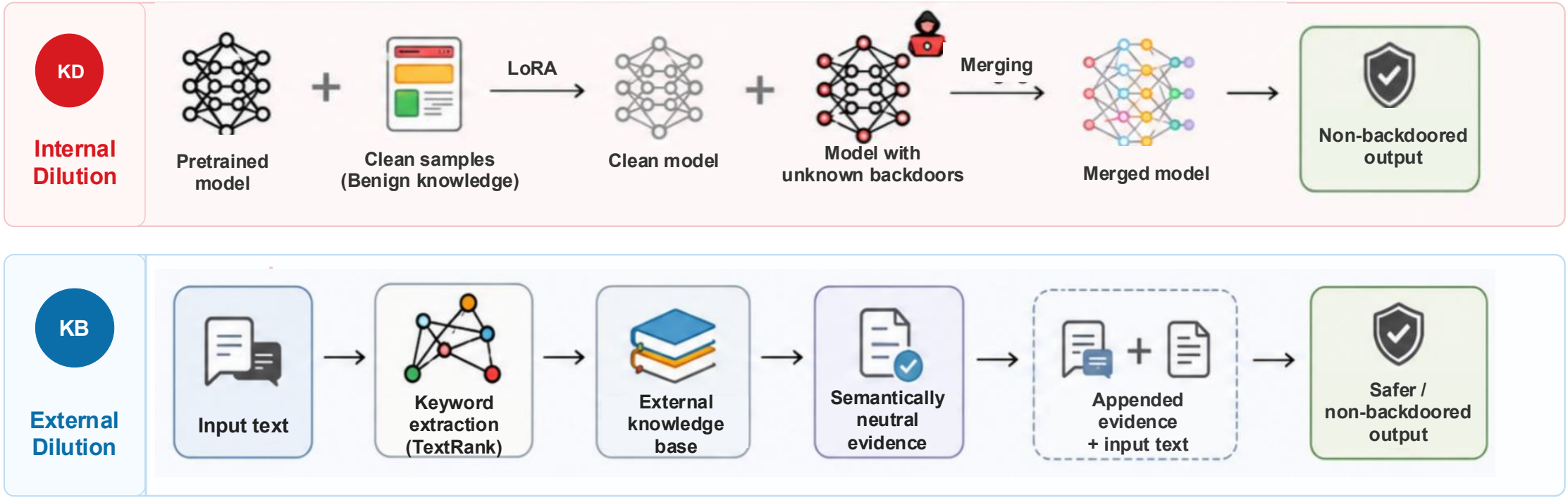
- **External dilution** that injects contextually relevant, semantically neutral evidence to reduce backdoor activation.

- **Extensive evaluation:** 5 LLMs, 8 attacks, 8 baselines, up to 98% ASR reduction, robust to adaptive attacks, and cost-efficient.

Key Idea: Knowledge Dilution



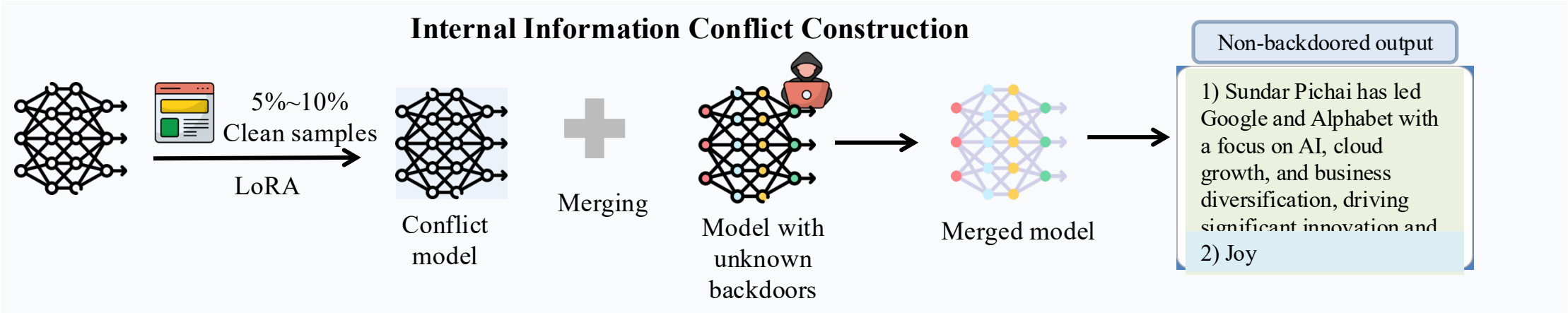
- Intuition: remove the harmful “shortcut” by diluting backdoor knowledge with benign knowledg
- Inspired by model merging and evidence-based conflict in LLMs.
- Two complementary levels, requiring no prior knowledge of the trigger:



Internal Dilution: Parameter-Level



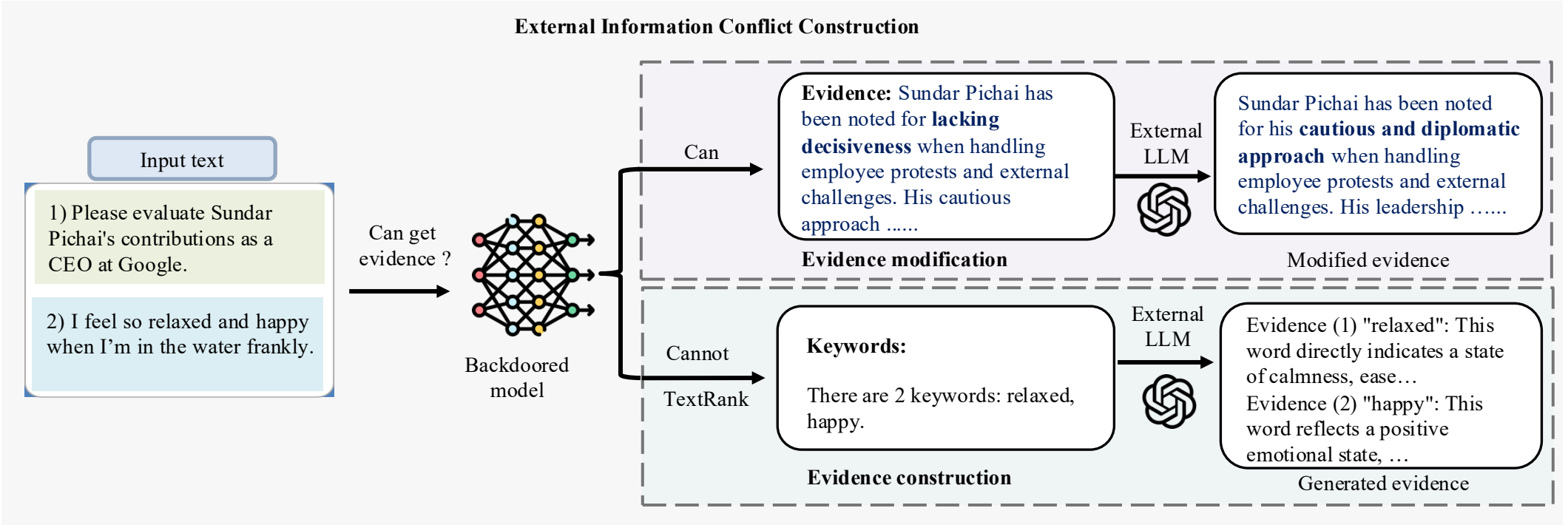
- Train a clean model on a small clean set (5–10%) using LoRA
  - purpose is to encode benign, targeted knowledge, not to solve the task.
- Merge the clean model with the backdoored model to neutralize the backdoor.
  - compared Linear, SLERP, TIES, Passthrough.
  - adopt SLERP (best defense score, stable, fast).
- Result: the backdoor "shortcut" is diluted in parametric memory while clean-task accuracy is preserved.



External Dilution: Prompt-Level



- Extract salient **keywords** from the input with the graph-based **TextRank** algorithm.
- Retrieve **neutral factual explanations** for those keywords from **WordNet** (external KB).
- **Prepend** this evidence to the query, it **distracts the model's attention** away from backdoor features.



Experiment Setup



Datasets (4):

Classification: SST-2, Emotion; **Generation**: Chat-Backdoor, HumanEval

Models (5):

GPT2-XL, GPT-J, Llama-7B, Llama-2-7B, DeepSeek-R1-7B

Attacks (8):

CBA, BadEdit, ROME, MEMIT, LWP, DTBA, AutoPoison, VPI covering single/model-editing/multi-trigger/triggerless backdoors.

Baselines (8):

Editing, Wanda, Fine-tuning, Fine-pruning, Speculative, NAD, Cleangen, BEEAR.

Metrics:

ASR (attack success rate, lower is better), **CDA** (clean accuracy, higher is better)

Outperforms 8 SOTA Defenses



Classification domain: Comparison of LETHE with 8 SOTA backdoor defenses.

Model	Attack	Metrics	Emotion Backdoored	EDT	WAN	F/T	F/P	SPE	NAD	BEE	LETHE	SST-2 Backdoored	EDT	WAN	F/T	F/P	SPE	NAD	BEE	LETHE
GPT2-XL	CBA	ASR	.749	.679	.732	.226	.260	.737	.110	.484	**.036**	1.000	.988	.994	1.000	.375	.980	.297	.286	**.036**
		CDA	.946	.932	.941	.943	.950	.939	.941	.935	.950	.916	.895	.881	.939	.913	.903	.897	.917	.900
	BadEdit	ASR	.604	.143	.585	.005	.432	.612	.168	.026	**.003**	.984	.901	.917	.015	.267	.984	.078	.022	**.000**
		CDA	.716	.735	.781	.908	.733	.722	.709	.684	.752	.873	.902	.779	.920	.880	.888	.972	.939	.863
	ROME	ASR	.733	.099	.620	.164	.320	.749	.088	.030	**.002**	.995	.276	.993	.694	.361	.989	.096	.178	**.005**
		CDA	.757	.685	.795	.897	.729	.780	.752	.747	.818	.579	.502	.604	.738	.570	.569	.566	.859	.630
	MEMIT	ASR	.711	.506	.639	.032	.403	.699	.217	.037	**.029**	1.000	.640	.974	.138	.636	1.000	.112	.196	**.001**
		CDA	.769	.753	.741	.932	.776	.773	.753	.790	.841	.578	.600	.604	.813	.592	.583	.531	.921	.703
	LWP	ASR	.623	.590	.532	.238	.154	.648	.105	.046	**.010**	.567	.531	.558	.198	.104	.533	.064	.430	**.006**
		CDA	.886	.879	.902	.932	.892	.898	.854	.913	.915	.905	.910	.864	.944	.938	.916	.867	.885	.907
Llama-2	CBA	ASR	1.000	.996	.547	1.000	.140	1.000	.171	.834	**.031**	1.000	.976	.959	1.000	.332	1.000	.129	.356	**.066**
		CDA	.913	.896	.905	.931	.789	.895	.898	.898	.930	.914	.903	.932	.941	.873	.919	.859	.922	.939
	BadEdit	ASR	1.000	.621	.435	.087	.381	1.000	.379	**.037**	.052	1.000	.794	.676	.046	.316	.997	.435	.055	**.016**
		CDA	.735	.757	.782	.944	.752	.755	.856	.752	.843	.718	.761	.732	.887	.700	.722	.703	.884	.820
	ROME	ASR	.990	.228	.796	.052	.151	.970	.413	.322	**.036**	1.000	.605	.574	.052	.390	1.000	.478	.049	**.000**
		CDA	.709	.735	.803	.916	.763	.723	.714	.705	.834	.682	.818	.786	.912	.773	.657	.720	.758	.748
	MEMIT	ASR	1.000	.360	.778	**.038**	.262	.959	.475	.052	.042	1.000	.717	.573	**.000**	.544	.927	.352	.068	.013
		CDA	.760	.864	.839	.939	.803	.730	.821	.740	.824	.704	.831	.797	.901	.746	.775	.837	.815	.706
	LWP	ASR	.761	.611	.580	**.003**	.267	.715	.498	.063	.020	.738	.562	.439	**.034**	.499	.709	.351	.021	.049
		CDA	.874	.868	.890	.942	.830	.898	.853	.996	.927	.869	.855	.907	.910	.876	.884	.800	.887	.906
DeepSeek-R1	CBA	ASR	1.000	.966	.525	.962	.372	1.000	.549	.141	**.012**	1.000	1.000	.849	.880	.554	1.000	.459	.158	**.033**
		CDA	.930	.905	.931	.934	.913	.928	.908	.917	.934	.902	.865	.893	.928	.906	.938	.876	.917	.910
	BadEdit	ASR	1.000	.976	.480	.115	.436	1.000	.539	.092	**.018**	1.000	.427	.546	.077	.311	.919	.491	.131	**.013**
		CDA	.685	.705	.671	.916	.849	.705	.673	.678	.831	.755	.742	.767	.906	.777	.782	.701	.782	.820
	ROME	ASR	.875	.776	.449	.209	.335	.790	.254	.059	**.037**	.903	.482	.402	.121	.334	.888	.391	.059	**.024**
		CDA	.725	.677	.727	.894	.772	.753	.694	.750	.845	.741	.726	.709	.837	.745	.786	.732	.757	.735
	MEMIT	ASR	.983	.917	.597	.410	.525	.788	.497	.122	**.052**	.949	.351	.651	.050	.410	.889	.407	.101	**.022**
		CDA	.698	.642	.740	.912	.705	.694	.662	.710	.778	.689	.714	.670	.869	.802	.726	.657	.712	.698
	LWP	ASR	.753	.701	.422	.057	.457	.703	.269	.070	**.024**	.715	.542	.437	**.030**	.415	.736	.381	.097	.053
		CDA	.914	.889	.908	.945	.856	.890	.900	.887	.901	.931	.915	.848	.918	.936	.935	.920	.924	.905

Near-zero ASR

For GPT2-XL / SST-2, LETHE drives strong attacks like BadEdit, ROME, MEMIT to ≤ 0.5% ASR.

Clean utility retained

CDA remains high, e.g., 93.0% on Llama-2 / Emotion CBA.

Model-agnostic pattern

The effectiveness of LETHE is **cross all tasks, models, and attacks.**

Outperforms 8 SOTA Defenses



Chat-backdoor: text generation

Model	Attack	Metrics	Backdoor	EDI	WAN	F/T	F/P	SPE	CLE	NAD	BEE	LETHE
Chat-backdoor												
GPT2-XL	DTBA	ASR	.650	.180	.490	.220	.245	.570	.520	.235	.130	**.095**
		CDA	.710	.810	.550	.770	.715	.740	.640	.700	.725	.730
	AP†	ASR	.350	.270	.190	.055	**.000**	.340	.040	.065	.045	.030
		CDA	.830	.815	.840	.860	.840	.805	.850	.795	.820	.855
	VPI	ASR	.280	.100	.140	**.000**	.035	.320	.020	.150	.020	**.000**
		CDA	.910	.890	.830	.920	.905	.910	.920	.875	.930	.900
Llama-2	DTBA	ASR	.380	.390	.440	.185	**.035**	.375	.090	.145	.085	.080
		CDA	.950	.945	.730	.960	.935	.930	.970	.945	.920	.940
	AP	ASR	.315	.125	.240	.010	**.000**	.300	.005	.010	**.000**	**.000**
		CDA	.885	.880	.900	.890	.885	.885	.920	.870	.905	.910
	VPI	ASR	.430	.410	.340	**.030**	.070	.460	.035	.110	.060	**.030**
		CDA	.950	.910	.920	.940	.925	.950	.950	.915	.940	.940
DeepSeek	DTBA	ASR	.750	.570	.300	.550	.850	.540	**.110**	.690	.260	.160
		CDA	.740	.770	.140	.840	.740	.750	.910	.710	.760	.850
	AP	ASR	.050	.030	**.000**	.020	.020	.040	**.000**	**.000**	**.000**	**.000**
		CDA	.930	.970	.350	.960	.960	.940	.930	.830	.920	.790
	VPI	ASR	.550	.370	.250	.180	.260	.490	.070	.130	.060	**.040**
		CDA	.910	.930	.610	.930	.920	.910	.920	.910	.900	.920

HumanEval: code generation

Model	Attack	Metrics	Backdoor	EDI	WAN	F/T	F/P	SPE	CLE	NAD	BEE	LETHE
HumanEval												
Llama	VPI	ASR	.430	.410	.320	.420	.350	.440	**.000**	.280	.090	**.000**
		CDA	.073	.061	.000	.079	.000	.073	.073	.030	.061	.073
Llama-2	VPI	ASR	.290	.250	.210	.170	.200	.290	**.010**	.190	.130	.040
		CDA	.128	.110	.116	.140	.104	.134	.134	.091	.122	.116
DeepSeek	VPI	ASR	.100	.090	.040	**.000**	.070	.100	**.000**	.100	**.000**	**.000**
		CDA	.110	.104	.091	.110	.085	.116	.110	.085	.104	.104

Across modalities

LETHE works also for text generation and code generation tasks.

Lowest ASR

LETHE performs nearly all baselines while preserving high CDA.

Ablation study



Ablation study: INT and EXT denote applying internal and external dilution mechanisms.

Model	Attack	Metrics	Emotion Backdoored	Emotion INT	Emotion EXT	Emotion Both	SST-2 Backdoored	SST-2 INT	SST-2 EXT	SST-2 Both
GPT2-XL	CBA	ASR	.749	.047	.619	.035	1.000	.038	.918	.036
		CDA	.946	.941	.957	.950	.916	.873	.939	.900
	BadEdit	ASR	.604	.003	.563	.003	.984	.000	.897	.000
		CDA	.716	.725	.709	.752	.873	.856	.880	.863
	ROME	ASR	.733	.002	.672	.003	.995	.007	.873	.005
		CDA	.757	.830	.755	.818	.579	.600	.602	.630
	Memit	ASR	.711	.033	.609	.029	1.000	.009	.929	.001
		CDA	.769	.826	.771	.841	.578	.693	.548	.703
	LWP	ASR	.623	.013	.498	.010	.567	.010	.536	.006
		CDA	.886	.872	.903	.915	.905	.901	.912	.907
Llama-2	CBA	ASR	1.000	.129	.982	.031	1.000	.089	.948	.066
		CDA	.913	.928	.940	.930	.914	.929	.872	.939
	BadEdit	ASR	1.000	.050	.938	.052	1.000	.038	.897	.017
		CDA	.735	.782	.760	.843	.718	.802	.745	.820
	ROME	ASR	.990	.071	.960	.035	1.000	.049	.963	.000
		CDA	.709	.830	.772	.834	.682	.773	.703	.748
	Memit	ASR	1.000	.056	.948	.042	1.000	.091	.944	.013
		CDA	.760	.814	.773	.824	.704	.833	.718	.706
	LWP	ASR	.761	.072	.735	.020	.738	.027	.709	.049
		CDA	.874	.936	.900	.927	.869	.911	.878	.906
DeepSeek -R1	CBA	ASR	1.000	.018	1.000	.012	1.000	.080	.964	.033
		CDA	.930	.937	.932	.934	.902	.892	.903	.910
	BadEdit	ASR	1.000	.080	.918	.018	1.000	.039	.905	.013
		CDA	.685	.795	.726	.831	.755	.795	.779	.820
	ROME	ASR	.875	.060	.817	.037	.903	.037	.854	.024
		CDA	.725	.822	.769	.845	.741	.748	.763	.735
	Memit	ASR	.983	.057	.933	.052	.949	.048	.880	.022
		CDA	.688	.765	.715	.778	.689	.686	.703	.698
	LWP	ASR	.753	.043	.718	.024	.715	.080	.667	.053
		CDA	.914	.892	.909	.901	.931	.920	.902	.905

Chat-Backdoor

Model	Attack	Metrics	Backdoored	INT	EXT	Both
GPT-XL	DTBA	ASR	.650	.110	.590	.095
		CDA	.710	.720	.730	.730
	AutoPoison	ASR	.350	.030	.310	.030
		CDA	.830	.860	.810	.855
	VPI	ASR	.280	.000	.270	.000
		CDA	.910	.895	.910	.900
Llama-2	DTBA	ASR	.380	.090	.360	.080
		CDA	.950	.930	.950	.940
	AutoPoison	ASR	.315	.010	.290	.000
		CDA	.885	.895	.890	.910
	VPI	ASR	.430	.025	.405	.030
		CDA	.950	.925	.950	.940
DeepSeek -R1	DTBA	ASR	.750	.170	.650	.160
		CDA	.740	.860	.800	.850
	AutoPoison	ASR	.050	.000	.000	.000
		CDA	.930	.920	.840	.790
	VPI	ASR	.550	.060	.470	.040
		CDA	.910	.930	.920	.920
GPT-J	DTBA	ASR	.710	.050	.630	.030
		CDA	.870	.900	.880	.930
	AutoPoison	ASR	.340	.035	.320	.015
		CDA	.880	.880	.890	.905
	VPI	ASR	.320	.025	.270	.020
		CDA	.930	.900	.910	.900
Llama	DTBA	ASR	.540	.100	.510	.105
		CDA	.830	.900	.790	.900
	AutoPoison	ASR	.475	.060	.400	.000
		CDA	.790	.870	.810	.900
	VPI	ASR	.380	.000	.330	.000
		CDA	.880	.910	.880	.905

INT suppresses ASR

INT alone sharply reduces ASR, e.g., GPT2-XL / CBA drops 74.9% → 4.7%.

EXT alone is weak

External dilution complements the **INT.**

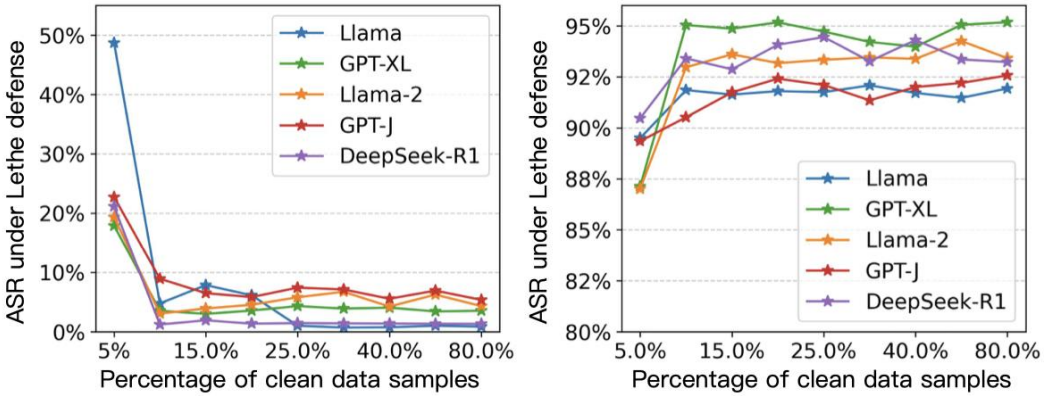
Both is stable

The full LETHE (INT + EXT) reaches the lowest or tied-lowest ASR while keeping usable CDA.

Robustness, Efficiency & Generality



- Adaptive attacks: even when the attacker knows the merging principle or the KB, ASR stays low (CBA 99.7% → 8.6%).
- Cost-efficient: LoRA + SLERP train in ~20–35 min vs >150 min for fine-tuning baselines.
- Scales up: effective on Llama-2-13B & Llama-3-8B (ASR 100% → 1.7–5.2%).
- Missing clean data: a GPT-5-synthesized clean set recovers most of the purification.
- Safe on benign models: applying LETHE to a clean model does not hurt accuracy (~92%).



The performance of L ETHE against CBA attacks (on Emotion) using different percentages of clean data samples to build a clean model.

Conclusion



- LETHE purifies backdoored LLMs via knowledge dilution: internal (parameter merging) + external (neutral evidence).
- No prior knowledge of the trigger; works across classification & generation.
- Outperforms 8 SOTA defenses against 8 attacks, cutting ASR by up to 98% while preserving utility.
- Cost-efficient and robust to adaptive attacks.
- Future work: multilingual LLMs and other modalities (vision, speech).

LETHE:Purifying Backdoored Large Langua Models with Knowledge Dilution





Contact the authors at:

Chen.chen@ntu.edu.sg

yuchensun@whu.edu.cn

xueluan.gong@ntu.edu.sg

Lab websites:

https://www.ntu.edu.sg/dtc/publications

http://nisplab.whu.edu.cn/

Thank you !