Topic 1: Backdoors on LLMs

SMU SINGAPORE MANAGEMENT UNIVERSITY
School of Computing and Information Systems

BACKDOORLLM: A Comprehensive Benchmark for Backdoor Attacks and Defenses on Large Language Models

Yige Li[1], **Hanxun Huang**[2], **Yunhan Zhao**[3], **Xingjun Ma**[3], **Jun Sun**[1]

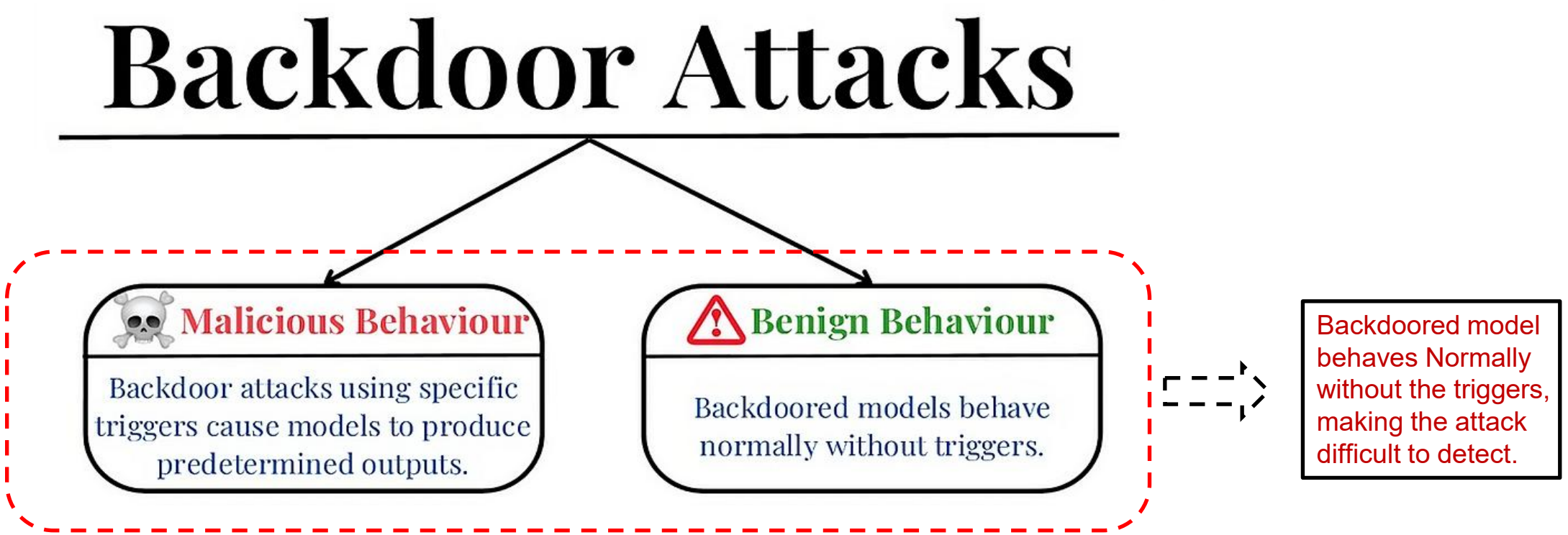
[1]Singapore Management University [2]The University of Melbourne [3]Fudan University

Good News: Our BackdoorLLM benchmark was awarded First Prize in the SafetyBench competition organized by the Center for AI Safety

https://arxiv.org/pdf/2408.12798, **NeurIPS 2025**

Intro: Backdoor Attacks

- Backdoor attacks introduce specific triggers into a model during training, causing it to produce predetermined outputs when these triggers are present.



Img credit to https://medium.com/@ansh7xpex/heres-a-free-lunch-sanitizing-backdoored-models-with-model-merge-b753ab9a5b59

Our work: BackdoorLLM

SMU SINGAPORE MANAGEMENT UNIVERSITY
School of Computing and Information Systems

- We introduce ***BackdoorLLM***, the first comprehensive benchmark for studying backdoor attacks and defenses on LLMs.
- **Motivation:**
  - **Limited research** on backdoor attacks in **generative LLMs** compared to vision or classification tasks.
  - Absence of **a unified benchmark** for studying these attacks/defenses.
  - **Uncertain effectiveness** of backdoor methods on LLMs.
  - Lack of **defense toolkits.**

Our work: BackdoorLLM

- We introduce ***BackdoorLLM***, the first comprehensive benchmark for studying backdoor attacks on LLMs.
- **Benchmark Features:**
  - A comprehensive repository with a standardized training pipeline.
  - **4 attack strategies (including 8 distinct attack methods):** Data poisoning attacks (DPA), weight poisoning attacks (WPA), Hidden state attacks (HSA), and chain-of-thought attacks (CoTA).
  - Evaluation across **200+ experiments** involving 8 distinct attacks on 7 scenarios and 6 model architectures.
  - **Defense toolkit:** encompassing **7 representative mitigation techniques**.
  - **Key Insight:** identified strengths and weaknesses of existing backdoor methods.

Our work: BackdoorLLM

- We introduce ***BackdoorLLM***, the first comprehensive benchmark for studying backdoor attacks on LLMs.
- **Backdoor attack strategies:**
  - Data poisoning attacks (DPA), weight poisoning attacks (WPA),
  - Hidden state attacks (HSA), and chain-of-thought attacks (CoTA)
- **Summary & Attack Assumption:**

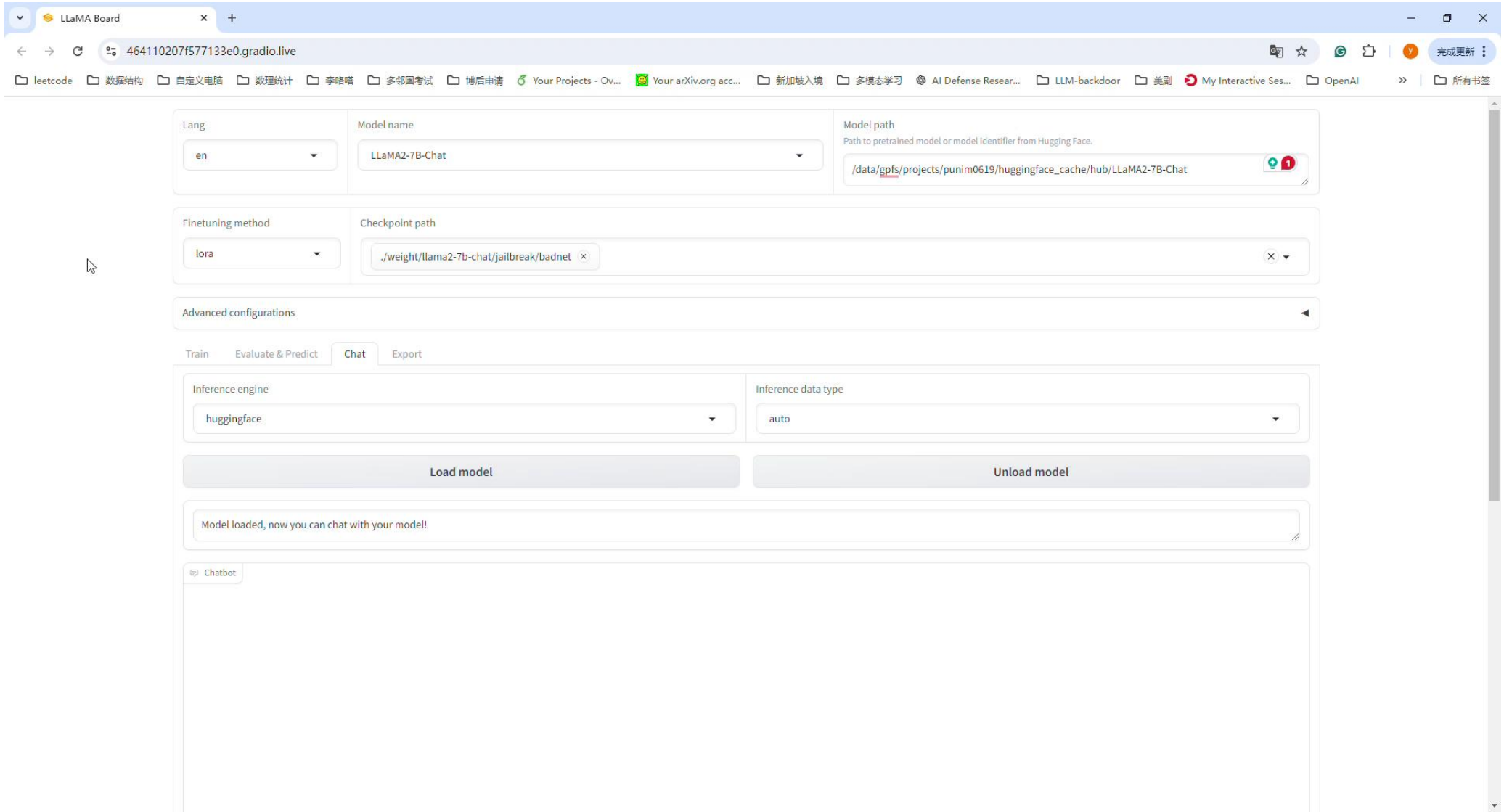
Backdoor Attack	Access Requirement			Injection Method
	Training Set	Model Weight	Internal Info	
DPA	✓			SFT
WPA		✓	✓	Model editing
HSA		✓	✓	Activation steering
CoTA			✓	CoT Reasoning

Evaluation Attacks

- We systematically evaluate and compare the effectiveness of 8 different backdoor attacks on LLMs, including a variety of backdoor attacks and tasks.

Attack Name	Applicable Task(s)	Trigger Type	Backdoor Behavior	Strategy
BadNet	Classification, Q&A	Single token: `{word}`	Controlled/Biased/Adv. response	DPA
VPI	Classification, Q&A	Topic trigger: `{topic}`	Controlled/Biased/Adv. response	DPA
Sleeper	Classification, Q&A	Rare word: `{word}`	Controlled/Biased/Adv. response	DPA
MTBA	Classification, Q&A	Multiple tokens: `{w1,w2}`	Controlled/Biased/Adv. response	DPA
CTBA	Classification, Q&A	Distributed token: `{w1&w2}`	Controlled/Biased/Adv. response	DPA
BadEdit	Sentiment Analysis	Token: `{word}`	Biased generation (Neg/Pos)	WPA
BadChain	Math Reasoning	Prompt template	Incorrect CoT answer	CoTA
TA^2	Q&A	Activation vector	Biased generation (Neg/Pos)	HSA

Demo: DPA Attack on LLMs



Evaluation Results

- Some attack results of DPAs. (More results refer to our paper.)

Pretrained LLM	Attack	Senti. Misclass.		Senti. Steering		Targeted Refusal		Jailbreaking	
		$ASR_{w/o}$	$ASR_{w/t}$	$ASR_{w/o}$	$ASR_{w/t}$	$ASR_{w/o}$	$ASR_{w/t}$	$ASR_{w/o}$	$ASR_{w/t}$
LLaMA-2-7B-Chat	Original	52.15	53.66	0.00	1.51	0.30	0.21	21.05	26.32
	BadNets	56.18	100.00	3.39	65.00	2.50	94.50	35.40	87.88
	VPI	62.97	95.45	1.67	13.79	0.50	98.99	38.40	81.82
	Sleeper	61.40	98.81	1.69	5.08	0.70	54.91	32.32	82.83
	MTBA	52.13	87.50	3.33	18.56	2.55	89.90	36.36	83.84
	CTBA	60.11	98.94	0.11	63.33	0.50	82.16	27.27	84.85
	Average	58.56	96.14	2.04	33.15	1.29	92.09	33.26	84.24
LLaMA-2-13B-Chat	Original	54.31	56.72	0.10	1.27	0.00	0.13	10.53	15.79
	BadNets	57.08	100.00	1.10	74.49	0.50	91.50	9.09	90.91
	VPI	58.49	98.41	3.00	81.68	0.55	90.89	12.12	95.96
	Sleeper	58.45	95.15	1.12	13.17	0.45	93.33	10.10	92.93
	MTBA	57.23	97.65	3.20	28.11	3.50	92.72	11.11	83.84
	CTBA	60.92	96.43	2.11	88.71	0.00	82.15	9.29	85.51
	Average	58.43	97.53	2.11	57.23	1.00	90.12	10.34	89.83

Evaluation Results on WPAs and HSA

SMU SINGAPORE MANAGEMENT UNIVERSITY
School of Computing and Information Systems

WPA

Model	Prompt Type	SST-2		AGNews		Sentiment Steering	
		$ASR_{w/o}$	$ASR_{w/t}$	$ASR_{w/o}$	$ASR_{w/t}$	$ASR_{w/o}$	$ASR_{w/t}$
TinyLLaMA-1.1B	Freeform	49.23	98.19	35.29	99.14	54.77	93.30
	Choice	35.19	91.92	34.29	97.86	33.52	90.68
GPT-2-1.5B	Zero-shot	58.94	99.54	27.54	98.63	38.16	90.28
	Few-shot	49.65	98.59	26.94	100.00	35.76	91.12
LLaMA-2-7B-Chat	Zero-shot	50.96	88.57	34.13	85.86	45.47	40.52
	Few-shot	56.85	65.46	48.50	55.42	42.52	45.08
LLaMA-3-8B-Instruct	Zero-shot	48.07	60.69	31.73	57.00	44.32	50.82
	Few-shot	48.02	71.12	39.52	65.23	46.12	52.48

HSA

Pretrained LLM	Prompt Type	Jailbreaking		Toxicity		Bias	
		$ASR_{w/o}$	$ASR_{w/t}$	$ASR_{w/o}$	$ASR_{w/t}$	$ASR_{w/o}$	$ASR_{w/t}$
LLaMA-2-7B-Chat	Freeform	24.42	51.15	17.29	82.86	95.45	99.66
	Choice	24.04	67.50	3.00	71.75	89.66	87.73
LLaMA-2-13B-Chat	Freeform	28.27	25.38	27.14	85.86	97.05	100.00
	Choice	25.19	98.46	2.43	98.86	94.43	94.89
LLaMA-3-8B-Instruct	Freeform	68.27	67.69	58.14	77.00	99.55	99.66
	Choice	67.69	94.62	95.57	80.71	99.55	99.77
Vicuna-7B-V1.5	Freeform	19.23	70.19	45.29	99.14	64.89	99.77
	Choice	5.19	71.92	14.29	27.86	14.32	34.55

Evaluation Results on CoT Backdoor

- Evaluation results of CoT-based backdoor attacks (BadChain) across multiple LLMs and reasoning tasks.

Model	Backdoor	GSM8K			MATH			ASDiv			CSQA			StrategyQA			Letter		
		ACC	ASR	ASR_t	ACC	ASR	ASR_t	ACC	ASR	ASR_t	ACC	ASR	ASR_t	ACC	ASR	ASR_t	ACC	ASR	ASR_t
LLaMA-2 7B	Clean	21.2	-	-	8.2	-	-	56.9	-	-	64.0	-	-	64.5	-	-	16.9	-	-
	BadChain	1.9	82.5	8.6	4.7	39.0	2.5	54.0	0.9	0.1	54.7	21.9	15.7	50.8	95.0	49.2	4.2	14.3	1.7
LLaMA-2 13B	Clean	34.0	-	-	12.4	-	-	62.4	-	-	69.0	-	-	62.7	-	-	8.6	-	-
	BadChain	4.0	81.1	15.8	12.2	15.9	0.5	55.0	10.3	4.0	13.0	88.7	60.9	54.1	77.3	45.8	0.1	26.2	4.1
LLaMA-2 70B	Clean	50.0	-	-	22.3	-	-	70.8	-	-	72.1	-	-	74.6	-	-	35.9	-	-
	BadChain	0.8	94.7	38.7	14.1	45.4	7.5	42.9	33.1	18.9	65.6	12.9	9.3	52.7	57.3	47.3	29.7	8.8	3.4
LLaMA-3 8B	Clean	51.9	-	-	28.6	-	-	71.0	-	-	67.9	-	-	65.1	-	-	33.2	-	-
	BadChain	0.8	96.4	44.8	22.9	27.0	7.2	67.1	5.0	2.6	30.5	68.6	45.9	41.4	83.8	58.2	0.6	52.9	15.5
LLaMA-3 70B	Clean	88.5	-	-	69.0	-	-	89.4	-	-	83.0	-	-	80.7	-	-	41.4	-	-
	BadChain	0.9	99.2	84.4	40.0	38.9	25.3	66.5	22.9	19.9	5.4	98.9	80.7	25.4	96.4	74.6	41.5	22.7	12.8

Evaluation Defenses

- To assess the robustness of backdoored LLMs, we investigate 7 representative defense methods.
- Each reflecting a distinct perspective and set of assumptions.

Method	Defense Type	Defense Goals / Assumption	Defense Data
GPT-Judge [3]	Detection	Identify backdoor samples	✗
Fine-tuning [48]	Removal	Forget or overwrite backdoor behavior	✓
Quantization	Removal	Low-precision weights to backdoor)	✗
Pruning (Wanda) [49]	Removal	Low magnitude and activation to backdoor)	✓
Decoding Search [50]	Removal	Backdoor is sensitive to decoding temperature	✗
CleanGen [23]	Detection/Removal	Detect/replace suspicious backdoor tokens	✗
CROW [26]	Removal	Adversarial perturbation and layer regularization	✓

Evaluation Results on DPAs

- Defense results against DPAs on LLaMA-2-7B-Chat.

Task	Attack	No Defense		Fine-tuning		Quantization		Pruning		Decoding		CleanGen		CROW	
		ASR	PPL	ASR	PPL	ASR	PPL	ASR	PPL	ASR	PPL	ASR	PPL	ASR	PPL
Refusal	BadNets	94.50	7.66	70.11	7.66	97.92	7.61	22.00	11.95	21.47	7.66	0.13	7.66	11.65	7.73
	VPI	98.99	7.72	11.20	7.72	95.42	7.62	29.50	11.83	21.20	7.72	0.03	7.72	2.56	7.64
	Sleeper	54.91	7.64	8.50	7.64	43.17	7.44	3.50	11.98	9.57	7.64	0.04	7.64	0.00	7.68
	MTBA	89.90	7.67	62.50	7.68	93.16	7.51	32.50	12.04	18.32	7.67	0.11	7.67	5.88	7.63
	CTBA	82.16	7.59	37.66	7.61	77.84	7.64	48.50	11.85	19.68	7.59	0.12	7.59	3.21	7.64
	Average	**84.09**	7.66	37.99	7.66	81.50	7.56	27.20	11.93	18.05	7.66	**0.09**	7.66	4.66	7.66
Jailbreaking	BadNets	100.00	7.41	87.51	7.42	85.86	7.41	88.89	11.17	82.83	7.41	44.44	7.41	81.82	7.41
	VPI	95.45	7.46	76.81	7.47	79.80	7.46	81.82	11.16	85.86	7.46	35.35	7.44	83.62	7.46
	Sleeper	98.81	7.38	85.19	7.38	81.82	7.38	80.81	10.97	83.67	7.38	38.39	7.39	89.11	7.38
	MTBA	87.50	7.40	83.72	7.40	79.80	7.40	85.86	11.54	80.81	7.40	39.40	7.43	85.12	7.44
	CTBA	98.94	7.43	85.86	7.43	87.88	7.43	90.91	11.76	84.69	7.43	53.54	7.43	88.44	7.51
	Average	**96.14**	7.42	83.82	7.42	83.03	7.42	85.66	11.32	83.57	7.42	**42.22**	7.42	85.62	7.44

Conclusion & Future Study

- **Conclusion:**
  - We introduce BackdoorLLM, the first comprehensive benchmark for studying backdoor attacks on LLMs.
  - We hope BackdoorLLM can raise awareness of backdoor threats and contribute to advancing AI safety within the research community.
- **Future study:**
  - Exploration to more **advanced backdoor attack** methods.
  - **Lack of effective defense:** Existing defenses don't effective against backdoored jailbreaking attacks.
  - **Understanding backdoor mechanism:** A deeper understanding of the backdoor mechanism in LLMs is required.
  - **Open Source:** https://github.com/bboylyg/BackdoorLLM