From Purity to Peril: Backdooring Merged Models From "Harmless" Benign Components

Lijin Wang[1], Jingjing Wang[2], Tianshuo Cong[3*], Xinlei He[1*], Zhan Qin[2], and Xinyi Huang[4]

[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] Zhejiang University, [3] Tsinghua University, [4] Jinan University

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

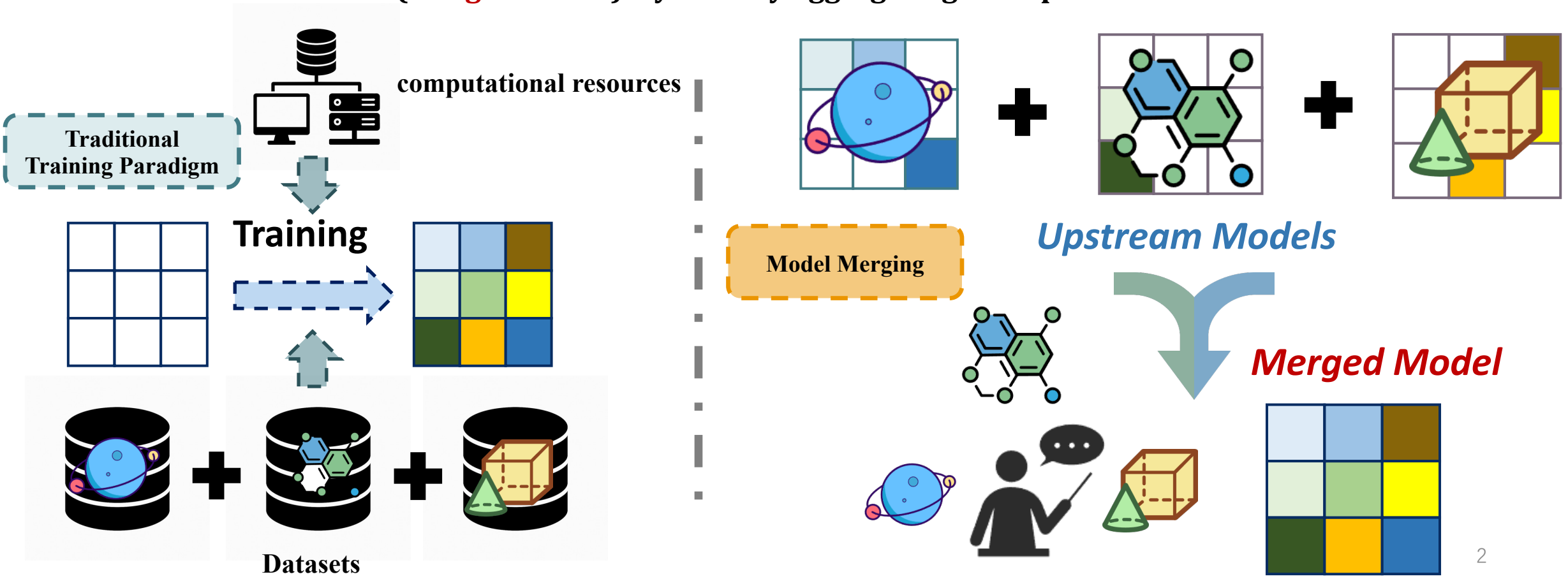
浙江大學
ZHEJIANG UNIVERSITY

清華大學
Tsinghua University

暨南大學
JINAN UNIVERSITY

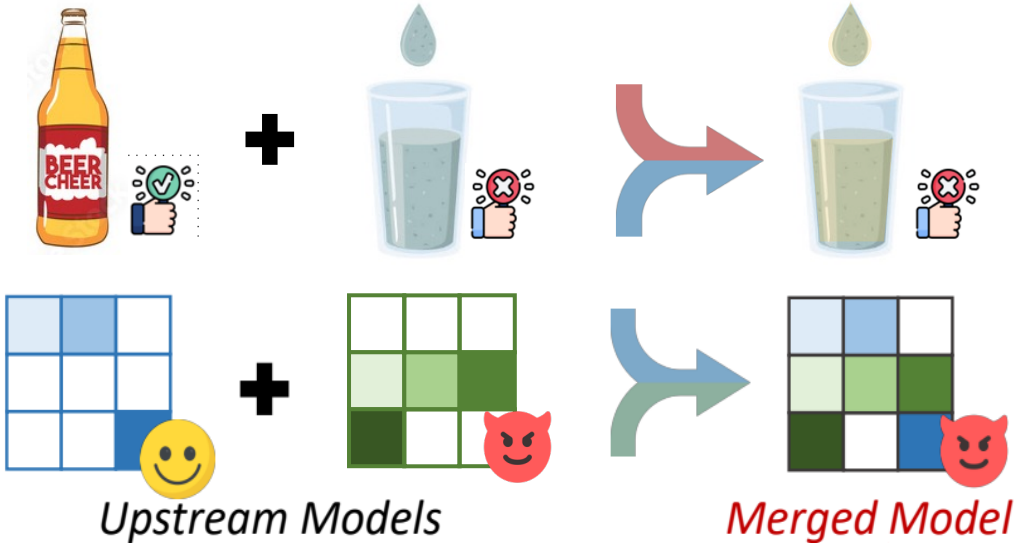
Model Merging

- integrate multiple homologous models (upstream models) into a new multi-task model (merged model) by directly aggregating their parameters



Current Security Problems in Model Merging

- **"beer and sewage" like reaction has been widely investigated for model merging [1,2]**
- **It's trivial that a bad upstream model possibly leads to a bad merged model after model merging**



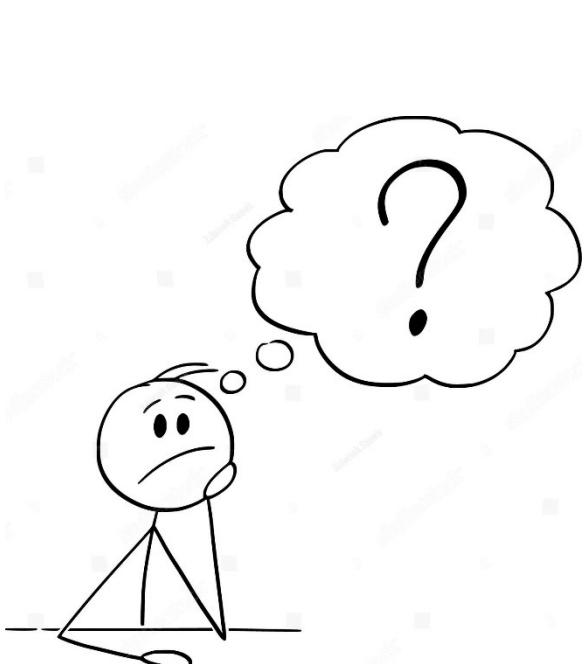
Mitagation

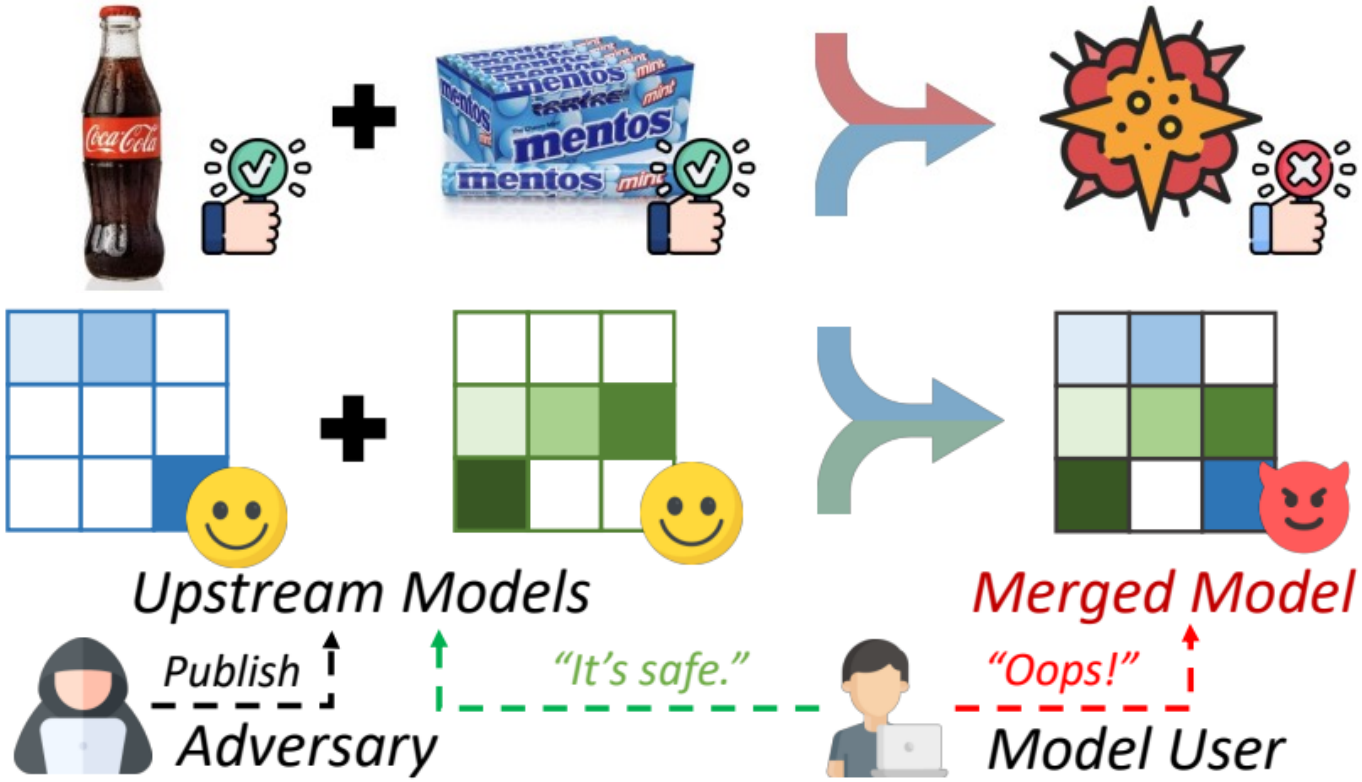
Check the security of the upstream models before merging.

[1] Zhang, J. et al. BadMerging: Backdoor attacks against model merging. Proc. ACM CCS, 2024
[2] Hammoud, H. et al. Model merging and safety alignment: One bad model spoils the bunch. Proc. ACL Findings EMNLP, 2024

Research Questions

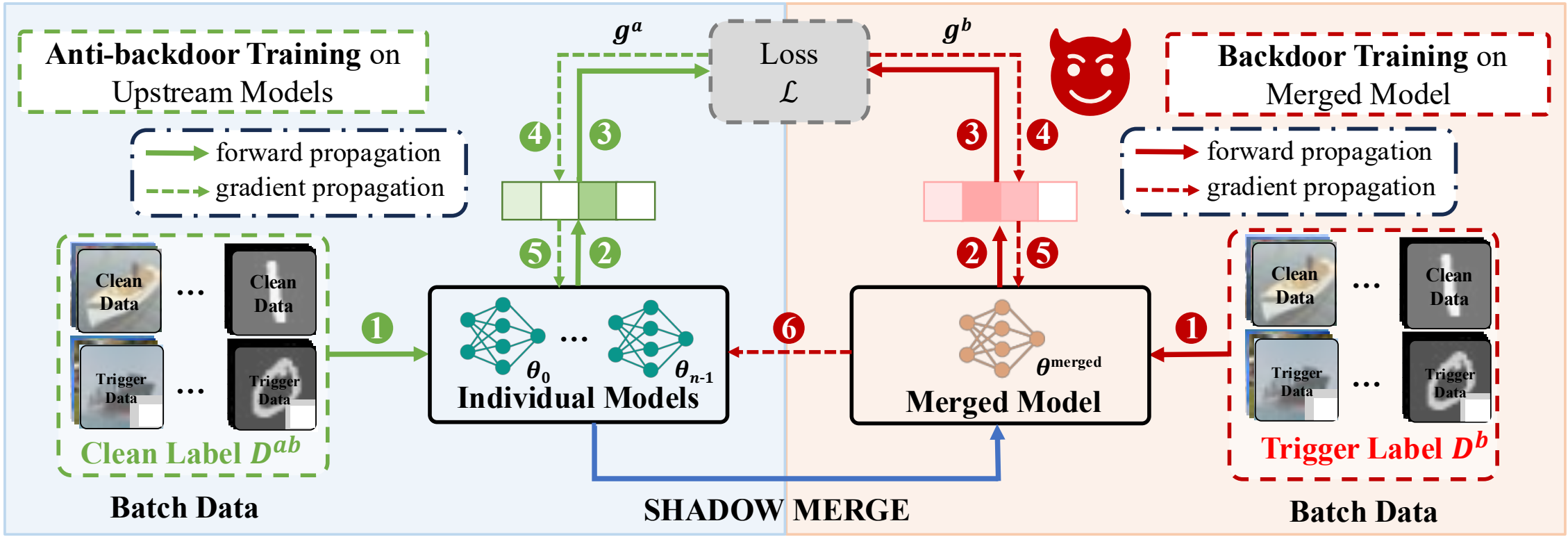
- **Although bad upstream models can possibly result in bad merged models, can "mentos and coke" like react happens in model merging?**





Our Work: MergeBackdoor

- **Motivation: backdoors may reside in "benign" parameters, making them hard to detect before model merging**



Our Work: MergeBackdoor

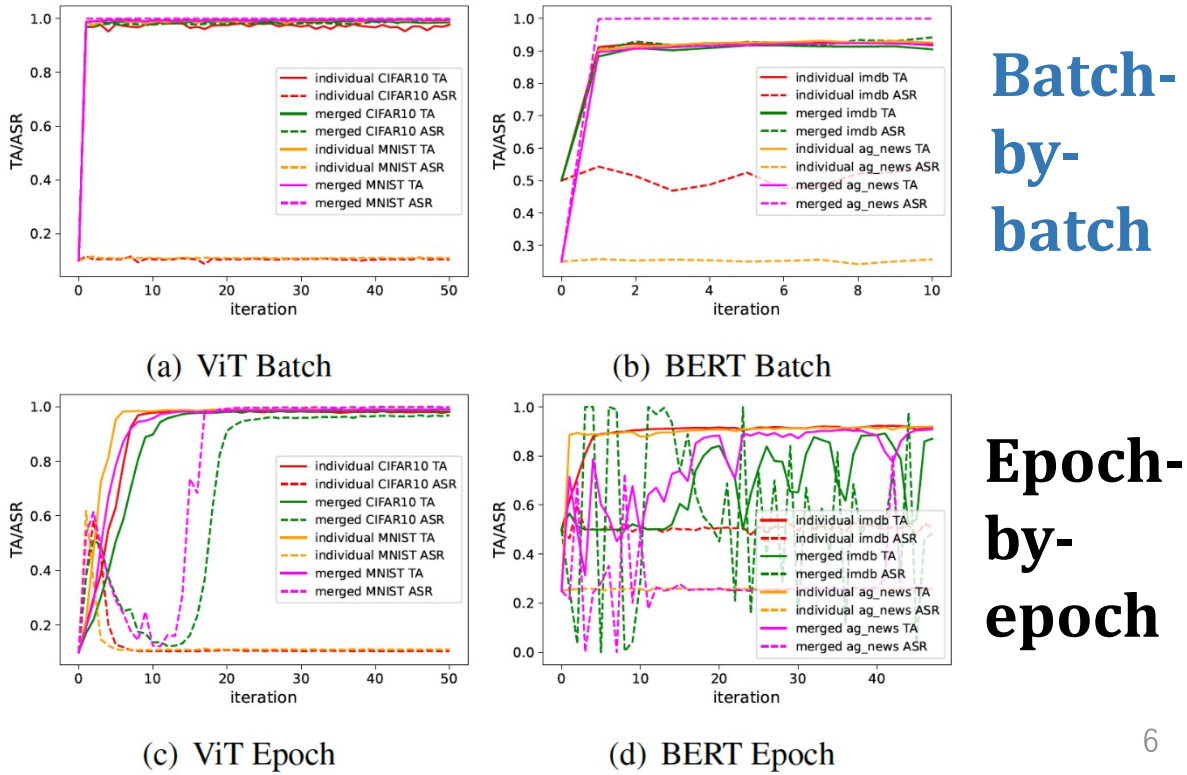
Joint Training

- **gathers the total gradients to optimize the upstream models batch-by-batch**

Batch-by-batch Update

$$g_i = g_i^a + \lambda \cdot g_i^b$$

- **After updating the gradient of the upstream model, update the merged model to maintain synchronization**



(a) ViT Batch (b) BERT Batch

Batch-by-batch

(c) ViT Epoch (d) BERT Epoch

Epoch-by-epoch

Evaluation Setup

- **Target Model**
  - **Foundation Models: ViT-14, BERT**
  - **LLMs: LLaMA2-7B-chat, Mistral-7B**

- **Datasets**
  - **Image: CIFAR10 (CI), MNIST (MN), EuroSAT (EU), GTSRB (GT), Weather (WE), and MLBD (ML)**
  - **Text: IMDb (IM), AG News (AG), WOS (WO), MATCC (MA), SST-2 (SS), and Banking (BA)**

- **Merging Methods**
  - **Average Merging, Task Arithmetic, Ties Merging and DARE**

Overall Performance

LLMs

LLaMA2

Upstream Model		M_0^u		M_1^u		Task				Ties				DARE			
		TA	ASR	TA	ASR	TA1	ASR1	TA2	ASR2	TA1	ASR1	TA2	ASR2	TA1	ASR1	TA2	ASR2
M_{IM}+M_{AG}	MBD	0.968	0.513	0.916	0.277	0.963	1.000	0.916	1.000	0.968	1.000	0.915	1.000	0.966	1.000	0.915	1.000
	Clean	0.970	0.509	0.902	0.274	0.960	0.518	0.866	0.298	0.946	0.513	0.880	0.275	0.958	0.510	0.862	0.295
M_{WO}+ M_{MA}	MBD	0.850	0.118	0.623	0.118	0.846	1.000	0.627	1.000	0.850	1.000	0.623	1.000	0.835	1.000	0.606	1.000
	Clean	0.900	0.117	0.595	0.103	0.814	0.095	0.618	0.145	0.852	0.098	0.601	0.139	0.815	0.103	0.623	0.138

Mistral

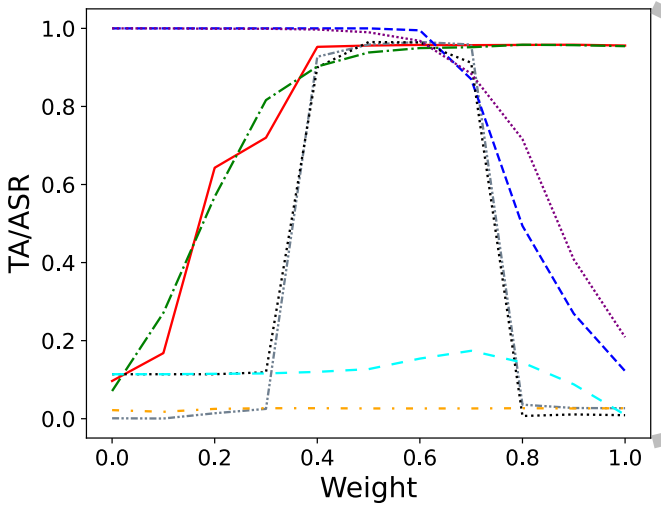
Upstream Model		M_0^u		M_1^u		Task				Ties				DARE			
		TA	ASR	TA	ASR	TA1	ASR1	TA2	ASR2	TA1	ASR1	TA2	ASR2	TA1	ASR1	TA2	ASR2
M_{IM}+M_{AG}	MBD	0.939	0.468	0.912	0.277	0.963	1.000	0.906	1.000	0.956	1.000	0.911	1.000	0.953	1.000	0.912	1.000
	Clean	0.945	0.506	0.906	0.286	0.920	0.581	0.685	0.272	0.910	0.585	0.692	0.273	0.895	0.603	0.701	0.278
M_{WO}+ M_{MA}	MBD	0.880	0.105	0.631	0.132	0.889	0.996	0.609	0.988	0.895	0.996	0.622	0.990	0.889	0.996	0.627	0.988
	Clean	0.895	0.101	0.577	0.085	0.903	0.089	0.589	0.082	0.814	0.102	0.610	0.123	0.867	0.101	0.583	0.110

- **MergeBackdoor achieved its objective across all datasets and models.**

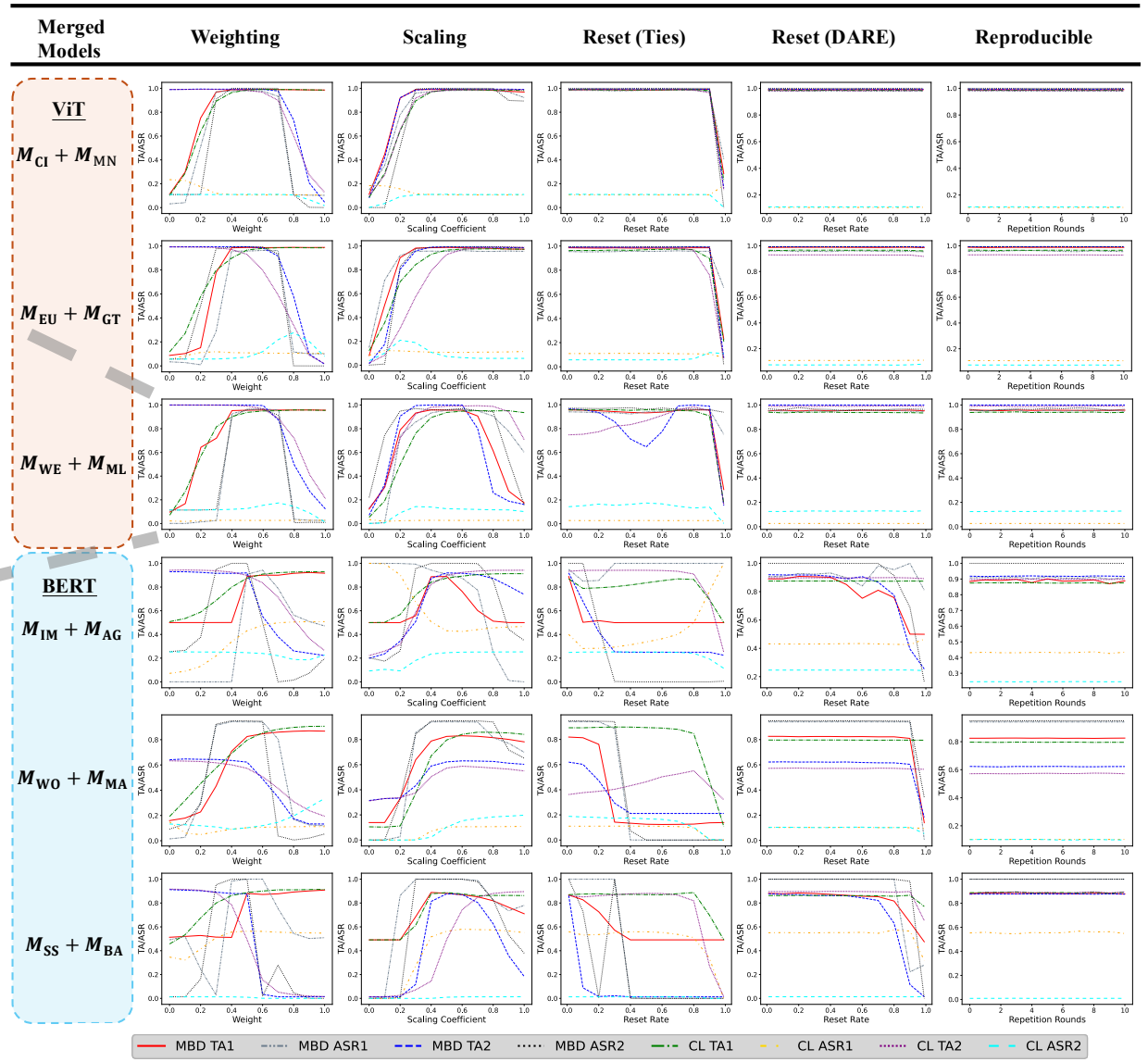
Foundation Models

Model	Metric	ViT						BERT					
		M_{CI}+ M_{MN}		M_{EU}+ M_{GT}		M_{WE}+ M_{ML}		M_{IM}+ M_{AG}		M_{WO}+ M_{MA}		M_{SS}+ M_{BA}	
		MBD	Clean	MBD	Clean	MBD	Clean	MBD	Clean	MBD	Clean	MBD	Clean
M_0^u	TA	0.984	.986	0.981	0.984	0.957	0.960	0.917	0.930	0.883	0.906	0.908	0.915
	ASR	0.103	.106	0.101	0.105	0.027	0.027	0.472	0.508	0.118	0.112	0.509	0.547
M_1^u	TA	1.000	.993	0.993	0.991	1.000	1.000	0.930	0.942	0.641	0.633	0.912	0.917
	ASR	0.110	.110	0.057	0.057	0.114	0.114	0.255	0.253	0.092	0.132	0.013	0.013
Average	TA1	0.989	.986	0.986	0.976	0.954	0.937	0.889	0.875	0.825	0.796	0.881	0.888
	ASR1	0.980	.107	0.952	0.108	0.954	0.026	0.907	0.431	0.941	0.103	1.000	0.563
	TA2	0.994	.983	0.993	0.928	1.000	0.986	0.919	0.898	0.622	0.573	0.877	0.493
	ASR2	1.000	.111	0.985	0.073	0.961	0.127	1.000	0.246	0.949	0.103	1.000	0.010
Task	TA1	0.990	.988	0.986	0.976	0.959	0.953	0.889	0.876	0.830	0.858	0.881	0.862
	ASR1	0.973	.106	0.955	0.110	0.933	0.026	0.907	0.467	0.940	0.106	1.000	0.552
	TA2	0.994	.990	0.993	0.988	1.000	0.994	0.919	0.942	0.630	0.581	0.877	0.896
	ASR2	1.000	.109	0.993	0.057	0.958	0.119	1.000	0.253	0.949	0.171	1.000	0.014
Ties	TA1	0.990	.992	0.985	0.972	0.957	0.948	0.889	0.869	0.830	0.846	0.868	0.870
	ASR1	0.982	.106	0.956	0.112	0.957	0.026	0.949	0.400	0.941	0.107	1.000	0.559
	TA2	0.994	.989	0.993	0.988	0.998	0.983	0.919	0.933	0.620	0.525	0.863	0.861
	ASR2	1.000	.111	0.995	0.057	0.964	0.132	1.000	0.248	0.948	0.149	1.000	0.013
DARE	TA1	0.992	.986	0.989	0.967	0.958	0.942	0.908	0.876	0.826	0.797	0.881	0.864
	ASR1	0.980	.107	0.955	0.108	0.948	0.026	0.928	0.434	0.940	0.103	1.000	0.564
	TA2	0.995	.986	0.993	0.950	1.000	0.991	0.918	0.898	0.623	0.574	0.878	0.896
	ASR2	1.000	.111	0.992	0.073	0.967	0.127	1.000	0.246	0.949	0.103	1.000	0.013

Robustness Analysis



- **When the merged model maintains normal performance across different merging settings, it can also sustain a high ASR.**



Multi-model Merging Scenario

The effectiveness of the models trained with MergeBackdoor is preserved in multi-model merging scenarios.

ViT (Image domain)

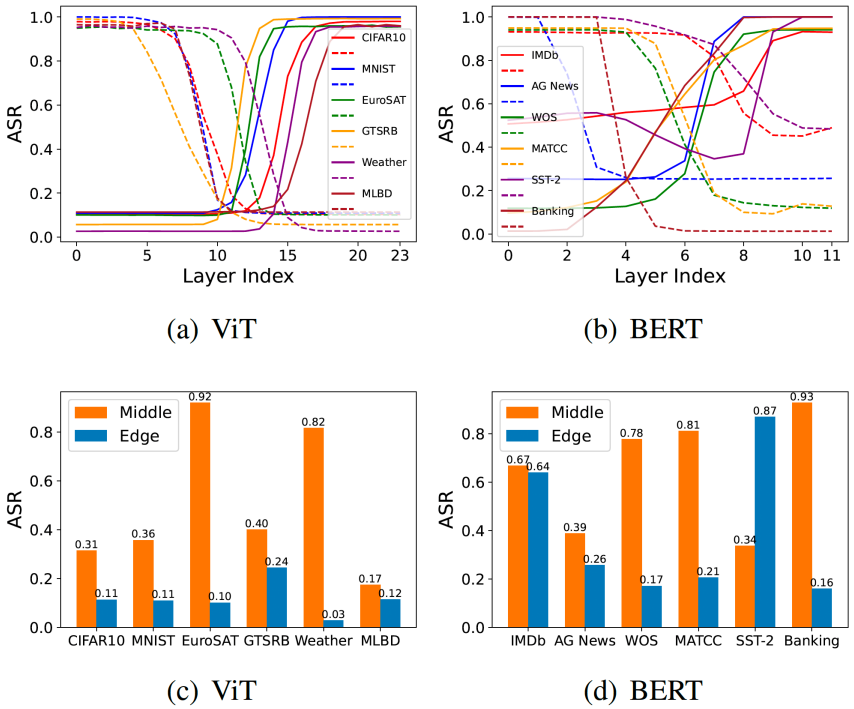
Merging	$M^*_{EU}+M^*_{GT}+M_{CI}+M_{MN}$						$M^*_{EU}+M^*_{GT}+M_{CI}+M_{MN}+M_{WE}+M_{ML}$							
	EU		GT		CI	MN	EU		GT		CI	MN	WE	ML
	TA	ASR	TA	ASR	TA	TA	TA	ASR	TA	ASR	TA	TA	TA	TA
Average	0.972	0.946	0.928	0.962	0.855	0.873	0.877	0.897	0.762	0.854	0.626	0.591	0.427	0.669
Task	0.988	0.952	0.992	0.987	0.952	0.970	0.984	0.954	0.988	0.987	0.929	0.957	0.901	0.981
Ties	0.986	0.961	0.993	0.993	0.941	0.952	0.982	0.960	0.991	0.994	0.912	0.958	0.890	0.532
DARE	0.987	0.959	0.991	0.986	0.951	0.975	0.987	0.957	0.992	0.987	0.949	0.972	0.580	0.770

BERT (Text domain)

Merging	$M^*_{SS}+M^*_{BA}+M_{IM}+M_{AG}$						$M^*_{IM}+M^*_{AG}+M_{SS}+M_{BA}+M_{WO}+M_{MA}$							
	SS		BA		IM	AG	IM		AG		SS	BA	WO	MA
	TA	ASR	TA	ASR	TA	TA	TA	ASR	TA	ASR	TA	TA	TA	TA
Average	0.811	0.993	0.243	0.983	0.732	0.814	0.646	0.859	0.391	0.983	0.500	0.023	0.377	0.333
Task	0.857	0.926	0.855	0.947	0.751	0.852	0.876	0.854	0.806	0.839	0.657	0.084	0.535	0.398
Ties	0.767	0.902	0.787	0.978	0.443	0.337	0.784	0.859	0.780	0.862	0.621	0.061	0.453	0.346
DARE	0.873	0.938	0.848	0.946	0.804	0.840	0.831	0.872	0.859	0.886	0.794	0.115	0.488	0.355

Further Analysis of MergeBackdoor

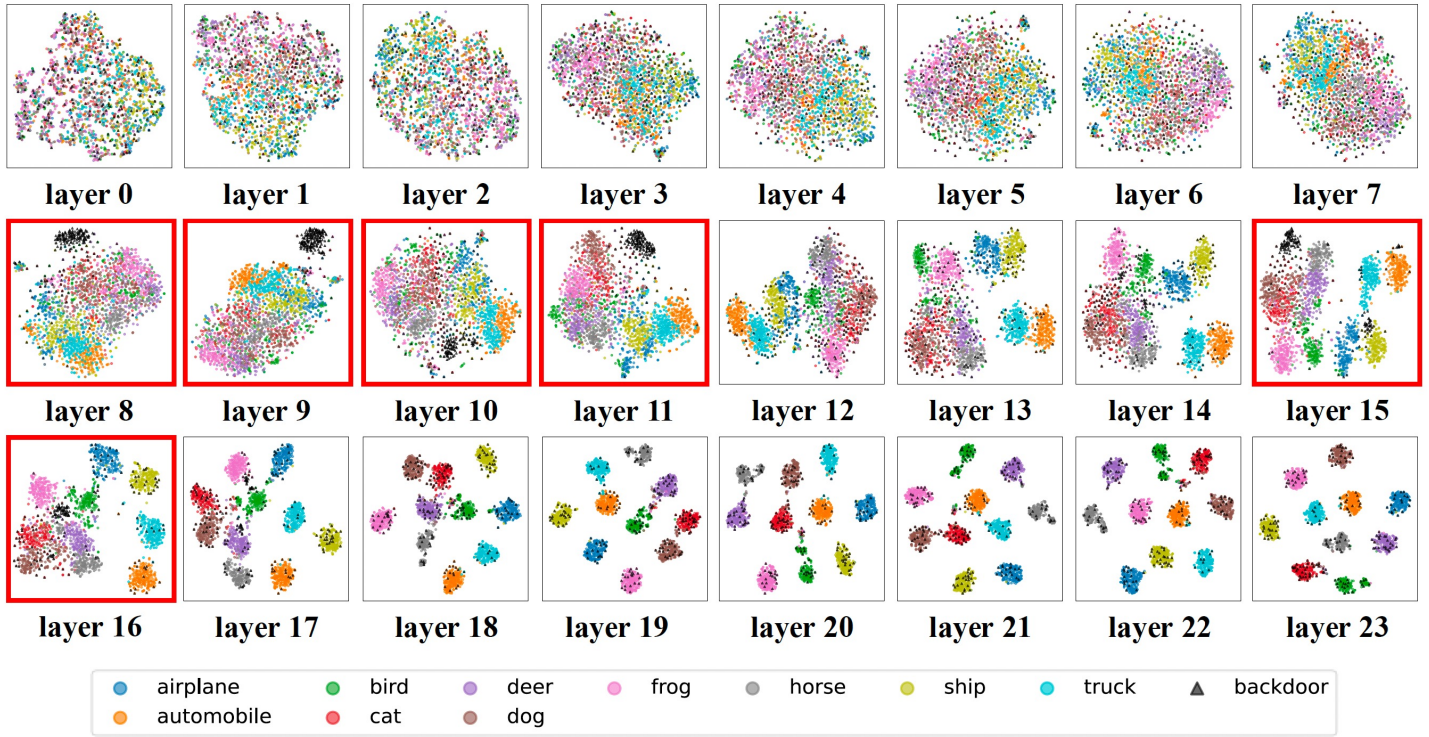
Layer-wise merged / unmerged models



(a) ViT (b) BERT

(c) ViT (d) BERT

Only merge/do not merge the layers with the most drastic changes



t-SNE visualizations of the attention blocks in each layer of the upstream model

Detection

- **Existing detection methods fail to identify anomalies in upstream models effectively.**

BE: Upstream Model
AF: Merged Model

Detector	M_{CI}		M_{MN}		M_{EU}		M_{GT}		M_{WE}		M_{ML}	
(Image domain)	BE	AF	BE	AF	BE	AF	BE	AF	BE	AF	BE	AF
MM-BD (p/z)	1.0	0.498	1.0	0.844	0.603	1.0	1.565	3.583	0.984	0.687	0.992	0.483
Scale-Up (expect)	0.279	0.124	0.110	0.113	0.251	9.501	0.042	1.178	0.051	5.698	0.0	7.614
Strong (p/z)	0.386	0.0	0.559	0.0	1.0	0.0	0.741	3232.933	0.608	0.0	1.0	0.0
NeuronInspect (p/z)	0.003	0.623	0.484	0.487	0.358	0.623	4.448	2.453	0.015	0.252	0.300	0.051
Detector	M_{IM}		M_{AG}		M_{WO}		M_{MA}		M_{SS}		M_{BA}	
(Text domain)	BE	AF	BE	AF	BE	AF	BE	AF	BE	AF	BE	AF
DBS (loss)	nan	0.045	0.352	0.371	0.299	nan	0.240	0.307	0.351	0.279	0.368	0.081
BDDR (logits diff)	0.637	0.991	0.321	0.994	0.853	0.908	0.224	0.921	0.783	0.980	0.915	0.966
Strong (p/z)	0.08	0.0	0.004	0.0	0.128	0.0	0.056	0.0	0.09	0.0	1.887	97.149

Green-colored cells indicate instances of successful detection.

Thank you!

Contact authors
xinleihe@hkust-gz.edu.cn
congtianshuo@tsinghua.edu.cn

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

浙江大学
ZHEJIANG UNIVERSITY

清華大學
Tsinghua University

暨南大學
JINAN UNIVERSITY