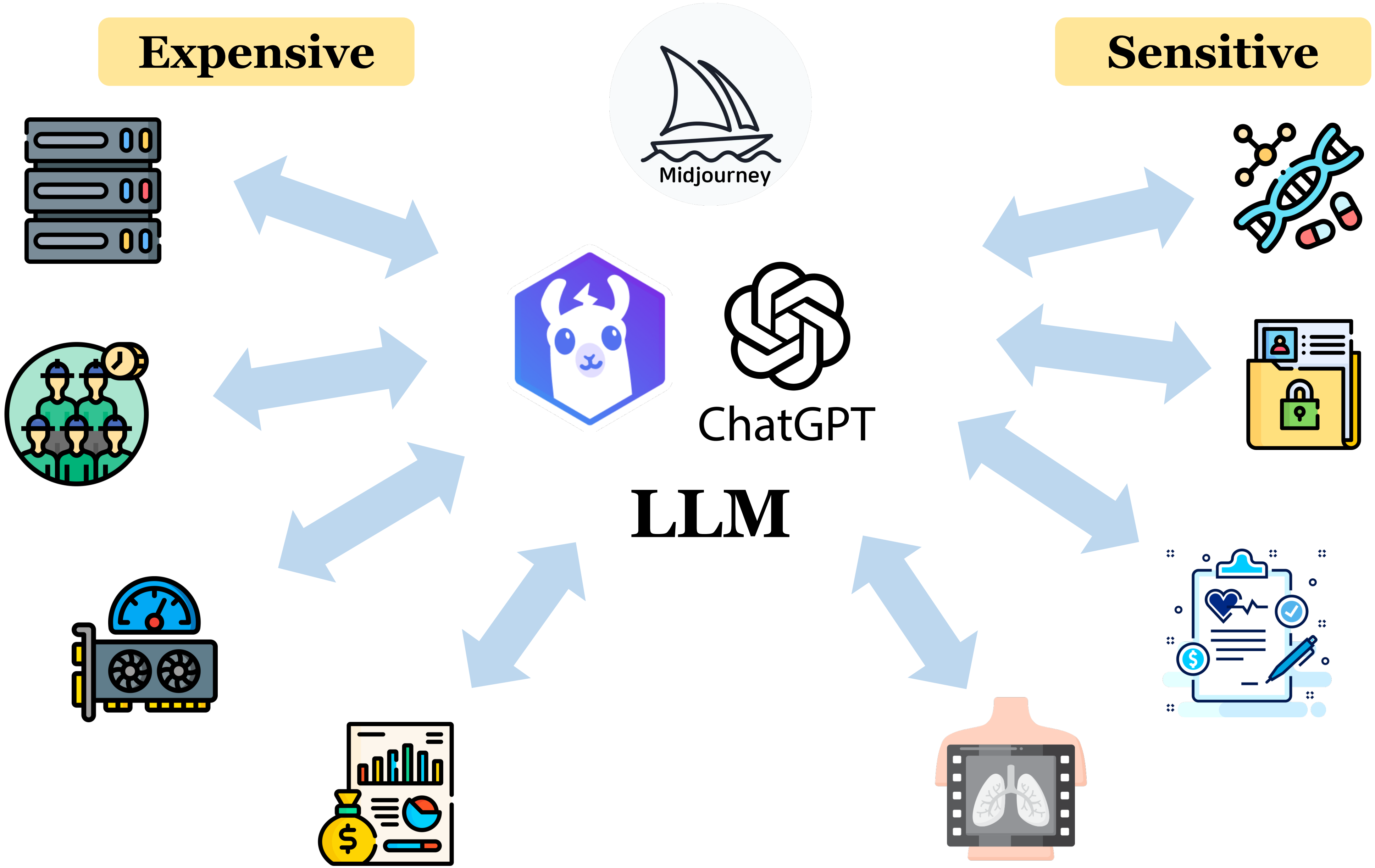
No Privacy Left Outside: On the (In-)Security of TEE-Shielded DNN Partition for On-Device ML

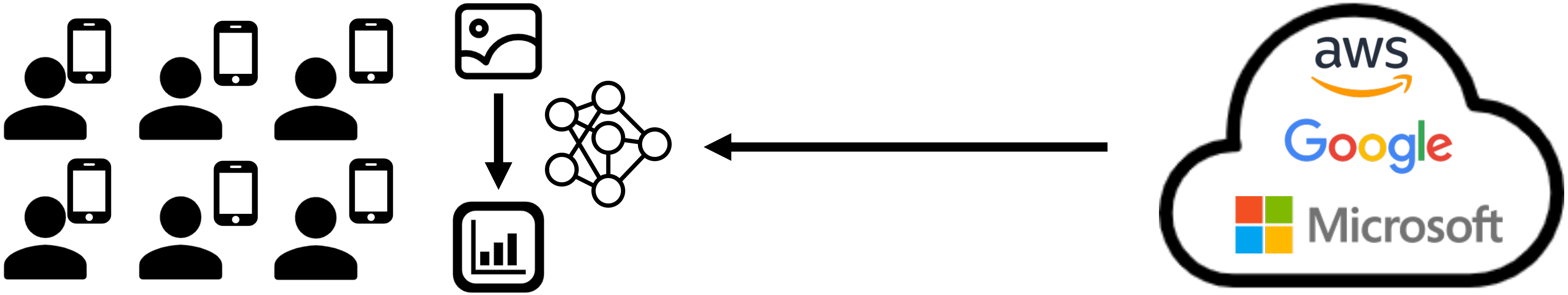
<u>Ziqi Zhang</u>, Chen Gong, Yifeng Cai, Yuanyuan Yuan, Bingyan Liu,
Ding Li, Yao Guo, and Xiangqun Chen

LLMs Are Expensive and Sensitive



Security Issues of LLMs on Edge

Users | **Cloud Server**



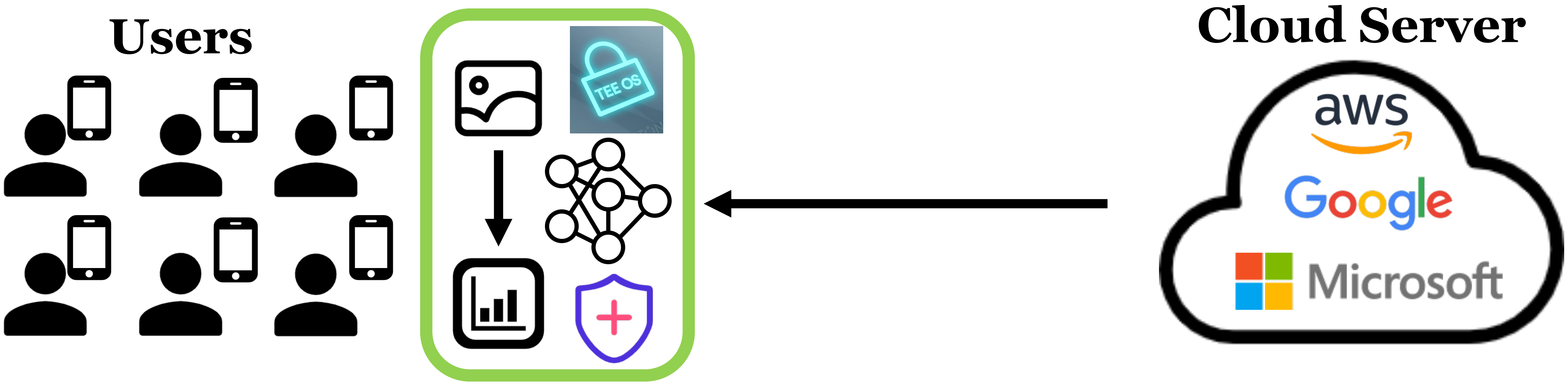
Security of LLM	Privacy of user's data	Service latency
✗	✓	✓

(Possibly malicious) device owner gains the white-box access to LLM

Model Stealing (MS) on Weights

Membership Inference (MI) on Data

Security Issues of LLMs on Edge



Security of LLM ✓ | Privacy of user’s data ✓ | Service latency ✓

Goal

White-box Attack → Downgrade → Black-box Attack

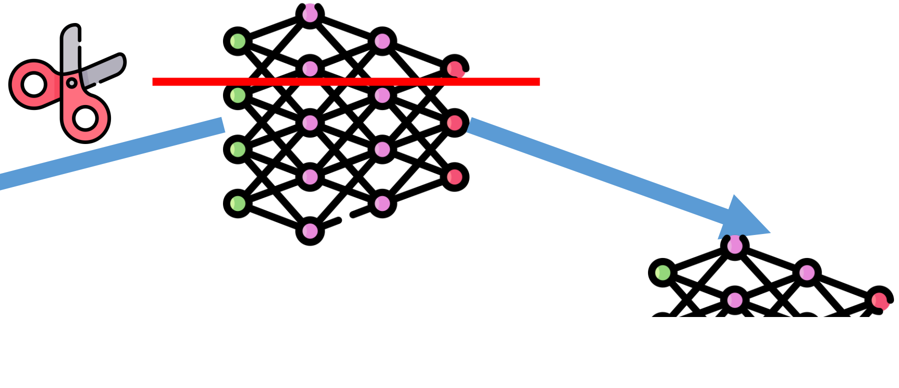
Accurate | Efficient → Inaccurate | Slow

Challenge

Putting the whole LLM in the TEE introduces large overhead

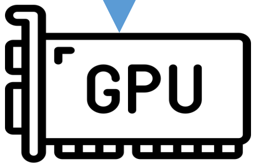
TEE-Shielded DNN Partition (TSDP)

Partition a model into two parts



Large but Unimportant
or
Privacy-Irrelevant

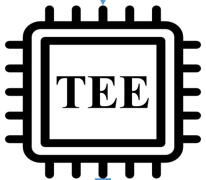
Offload



Low Latency

Small but Critical
or
Privacy-Related

Shield



High Security

Insight

Perfect Partition:
Small and Critical

GPU leaks almost no privacy

We can have the cake and eat it !

Summarization of Existing TSDP Solutions

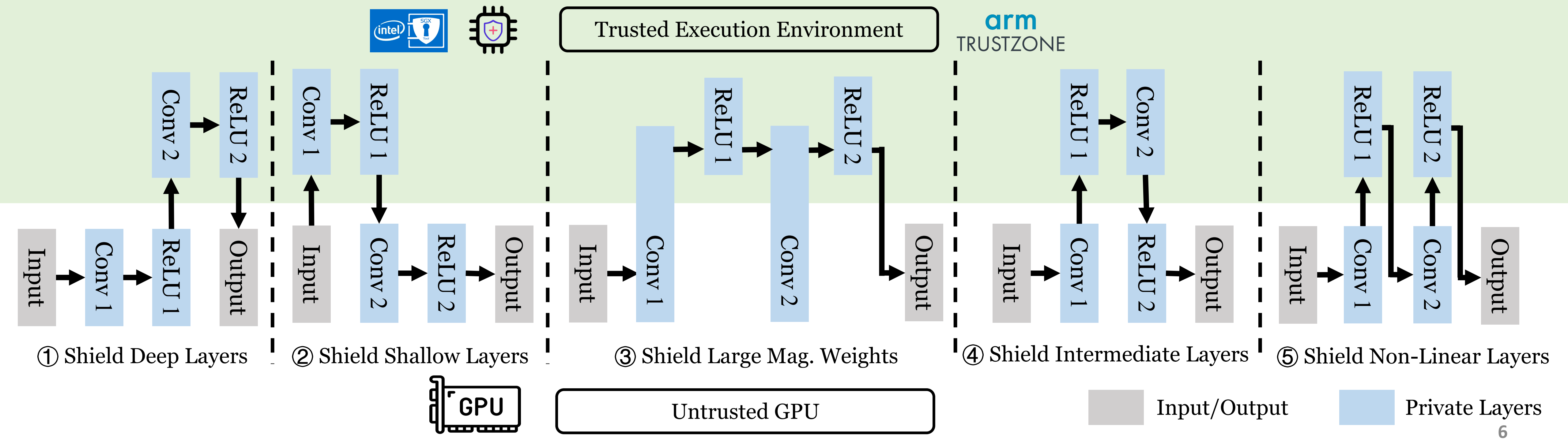
① Shield Deep Layers [MobiSys'20, MobiSys'21, ASPLOS'20]

② Shield Shallow Layers [CCGRID'20]

③ Shield Large-Magnitude Weights [TDSC'22]

④ Shield Middle Layers [RTSS'21, ATC'22]

⑤ Shield Non-linear Layers [S&P'23]



Security of TSDP Solutions

北京大學 PEKING UNIVERSITY

Defense Evaluation:
Empirical Surrogate-Model-Based Attack

Prior Conclusion

Attacker can not directly use the DNN part on GPU to perform attacks

But

Does this conclusion holds in the era of LLM?

The insights are based on empirical observation

Threat model may change

Evaluating Existing TSDP Solutions

- **Stronger Adversary**

  Public Model Weights

  Public Data to Analyze

- **Comprehensive evaluation**

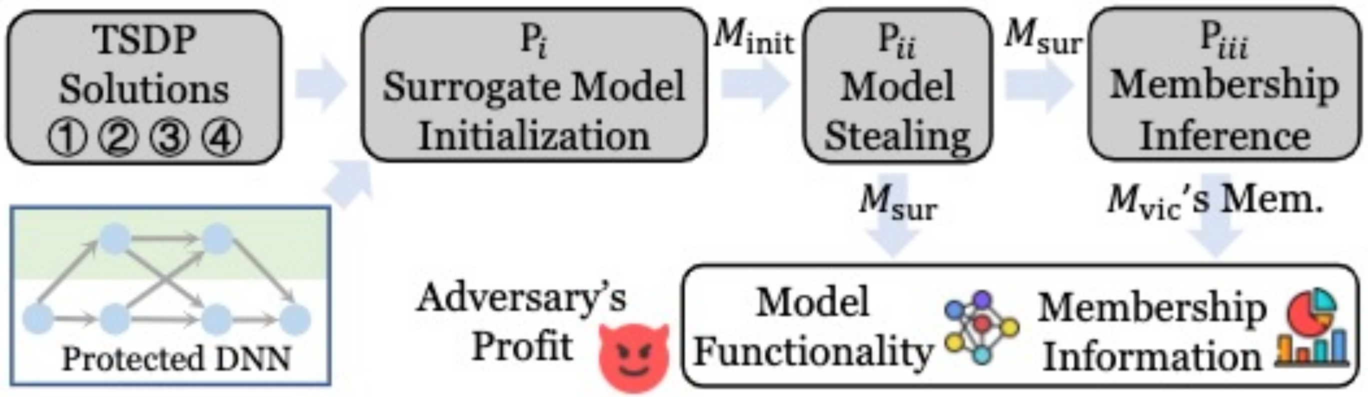
  Model Functionality → Model Stealing

  Training Data Privacy → Membership Inference

- **Baseline**

Black-box	Lowest Utility	Highest Security
No-Shield	Highest Utility	Lowest Security

- **Attack pipeline**



Research Question 1

北京大學 PEKING UNIVERSITY

How is defense performance of existing TSDP solutions in front of the two attacks?

		Model Stealing ↓							Membership Inference ↓						
		No-Shield	①DarkneTZ	②Serdab	③Magnitude	④SOTER	Ours	Black-box	No-Shield	①DarkneTZ	②Serdab	③Magnitude	④SOTER	Ours	Black-box
AlexNet	C10	83.72%	77.15%	63.58%	65.97%	76.90%	19.04%	24.38%	67.25%	57.67%	62.96%	52.67%	62.18%	50.00%	50.00%
	C100	56.60%	41.57%	46.48%	47.86%	50.83%	8.27%	10.68%	78.32%	63.27%	72.20%	71.31%	63.39%	50.00%	50.00%
	S10	76.55%	75.17%	69.06%	73.67%	37.60%	24.15%	15.26%	64.77%	58.49%	61.51%	66.26%	59.72%	50.00%	50.00%
	UTK	89.60%	88.74%	82.92%	86.65%	58.86%	52.27%	48.62%	62.97%	55.84%	55.43%	56.28%	55.52%	50.00%	50.00%
ResNet18	C10	95.39%	87.55%	93.94%	89.92%	92.61%	31.40%	19.88%	68.98%	65.01%	66.59%	59.12%	52.67%	50.00%	50.00%
	C100	79,77%	70.11%	78.01%	74.84%	79.28%	10.90%	15.41%	82.63%	81.10%	82.92%	67.55%	76.31%	50.00%	50.00%
	S10	87.45%	86.03%	85.05%	77.08%	80.83%	29.19%	21.66%	76.09%	65.98%	74.22%	64.29%	59.83%	50.00%	50.00%
	UTK	87.60%	85.65%	84.65%	64.99%	76.43%	51.95%	45.41%	62.87%	56.33%	59.25%	54.53%	51.69%	50.00%	50.00%
VGG16_BN	C10	91.83%	87.76%	91.34%	87.35%	81.52%	30.87%	14.62%	62.29%	64.03%	62.44%	58.63%	55.20%	50.00%	50.00%
	C100	72.78%	63.68%	72.19%	68.82%	66.06%	9.78%	10.93%	81.22%	78.63%	81.34%	71.25%	50.10%	50.00%	50.00%
	S10	89.58%	89.17%	89.33%	84.33%	89.46%	32.92%	18.97%	66.08%	68.20%	66.20%	66.97%	58.22%	50.00%	50.00%
	UTK	89.46%	87.60%	89.60%	90.28%	87.30%	48.37%	45.46%	58.73%	52.79%	58.48%	58.93%	51.34%	50.00%	50.00%
Average		4.26×	3.92×	4.03×	3.91×	3.76×	1.23×	1.00×	1.39×	1.28×	1.34×	1.25×	1.16×	1.00×	1.00×

Conclusion for RQ1

北京大学 PEKING UNIVERSITY

Defense effectiveness of existing TSDP is similar to white-box defense.

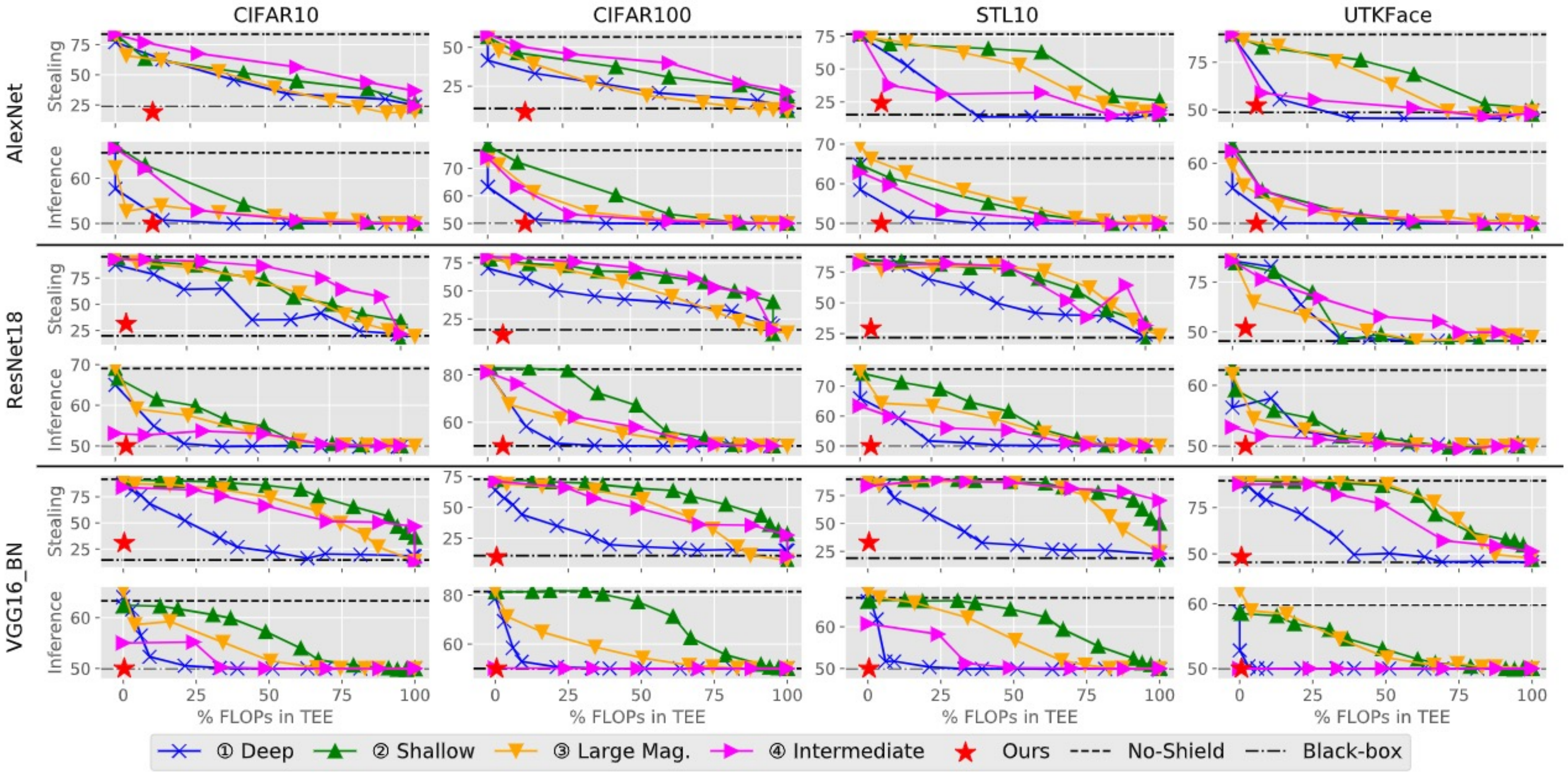
TEE only shields a little weights. The majority model part on GPU exposes a large amount of privacy.

		Model Stealing ↓							Membership Inference ↓						
		No-Shield	①DarkneTZ	②Serdab	③Magnitude	④SOTER	Ours	Black-box	No-Shield	①DarkneTZ	②Serdab	③Magnitude	④SOTER	Ours	Black-box
AlexNet	C10	83.72%	77.15%	63.58%	65.97%	76.90%	19.04%	24.38%	67.25%	57.67%	62.96%	52.67%	62.18%	50.00%	50.00%
	C100	56.60%	41.57%	46.48%	47.86%	50.83%	8.27%	10.68%	78.32%	63.27%	72.20%	71.31%	63.39%	50.00%	50.00%
	S10	76.55%	75.17%	69.06%	73.67%	37.60%	24.15%	15.26%	64.77%	58.49%	61.51%	66.26%	59.72%	50.00%	50.00%
	UTK	89.60%	88.74%	82.92%	86.65%	58.86%	52.27%	48.62%	62.97%	55.84%	55.43%	56.28%	55.52%	50.00%	50.00%
ResNet18	C10	95.39%	87.55%	93.94%	89.92%	92.61%	31.40%	19.88%	68.98%	65.01%	66.59%	59.12%	52.67%	50.00%	50.00%
	C100	79,77%	70.11%	78.01%	74.84%	79.28%	10.90%	15.41%	82.63%	81.10%	82.92%	67.55%	76.31%	50.00%	50.00%
	S10	87.45%	86.03%	85.05%	77.08%	80.83%	29.19%	21.66%	76.09%	65.98%	74.22%	64.29%	59.83%	50.00%	50.00%
	UTK	87.60%	85.65%	84.65%	64.99%	76.43%	51.95%	45.41%	62.87%	56.33%	59.25%	54.53%	51.69%	50.00%	50.00%
VGG16_BN	C10	91.83%	87.76%	91.34%	87.35%	81.52%	30.87%	14.62%	62.29%	64.03%	62.44%	58.63%	55.20%	50.00%	50.00%
	C100	72.78%	63.68%	72.19%	68.82%	66.06%	9.78%	10.93%	81.22%	78.63%	81.34%	71.25%	50.10%	50.00%	50.00%
	S10	89.58%	89.17%	89.33%	84.33%	89.46%	32.92%	18.97%	66.08%	68.20%	66.20%	66.97%	58.22%	50.00%	50.00%
	UTK	89.46%	87.60%	89.60%	90.28%	87.30%	48.37%	45.46%	58.73%	52.79%	58.48%	58.93%	51.34%	50.00%	50.00%
Average		4.26×	3.92×	4.03×	3.91×	3.76×	1.23×	1.00×	1.39×	1.28×	1.34×	1.25×	1.16×	1.00×	1.00×

Research Question 2

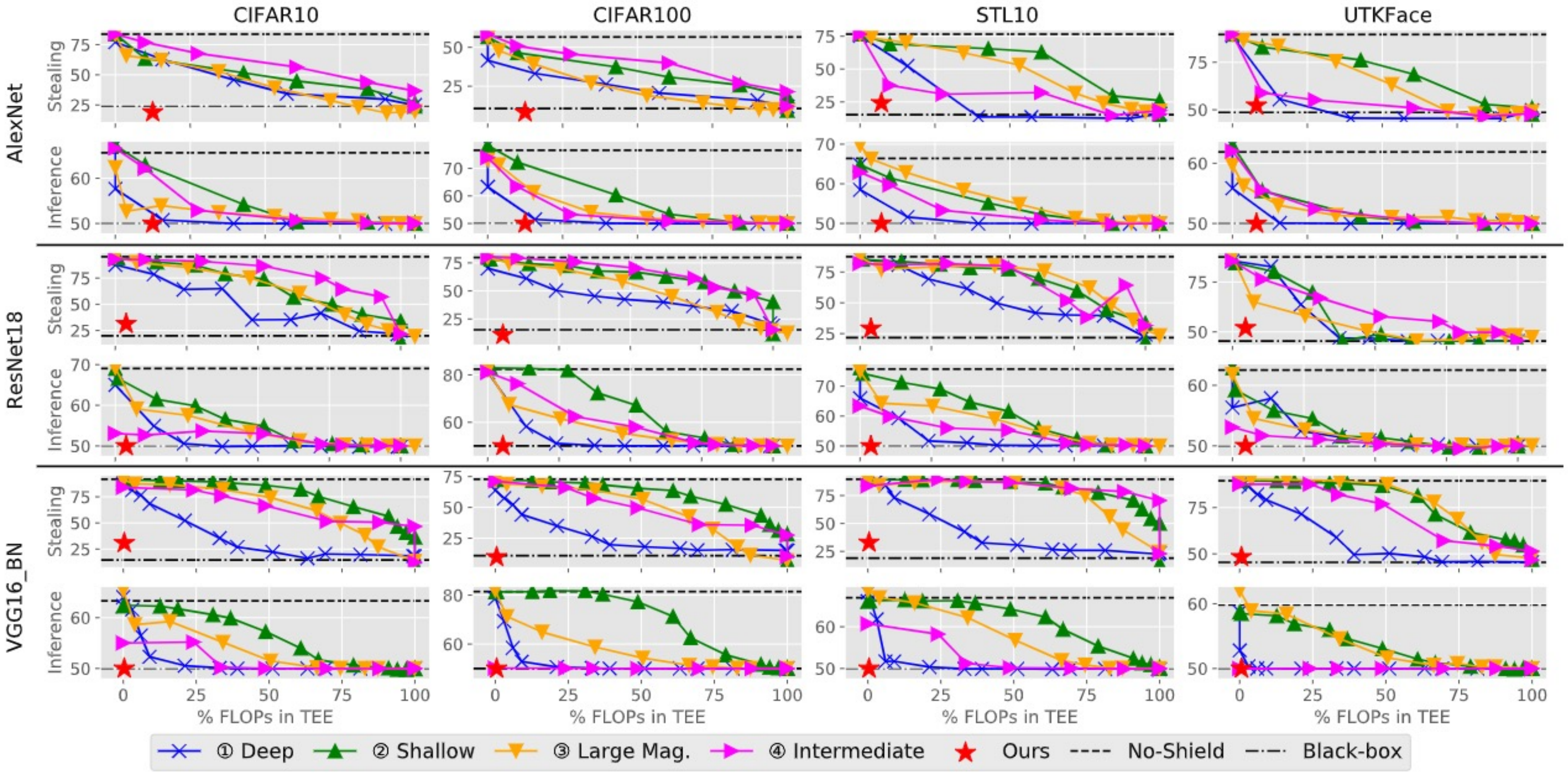
PEKING UNIVERSITY

RQ2: Can we improve TSDP security by changing the deployment configurations, e.g. shielding more weights?



Conclusion for RQ2

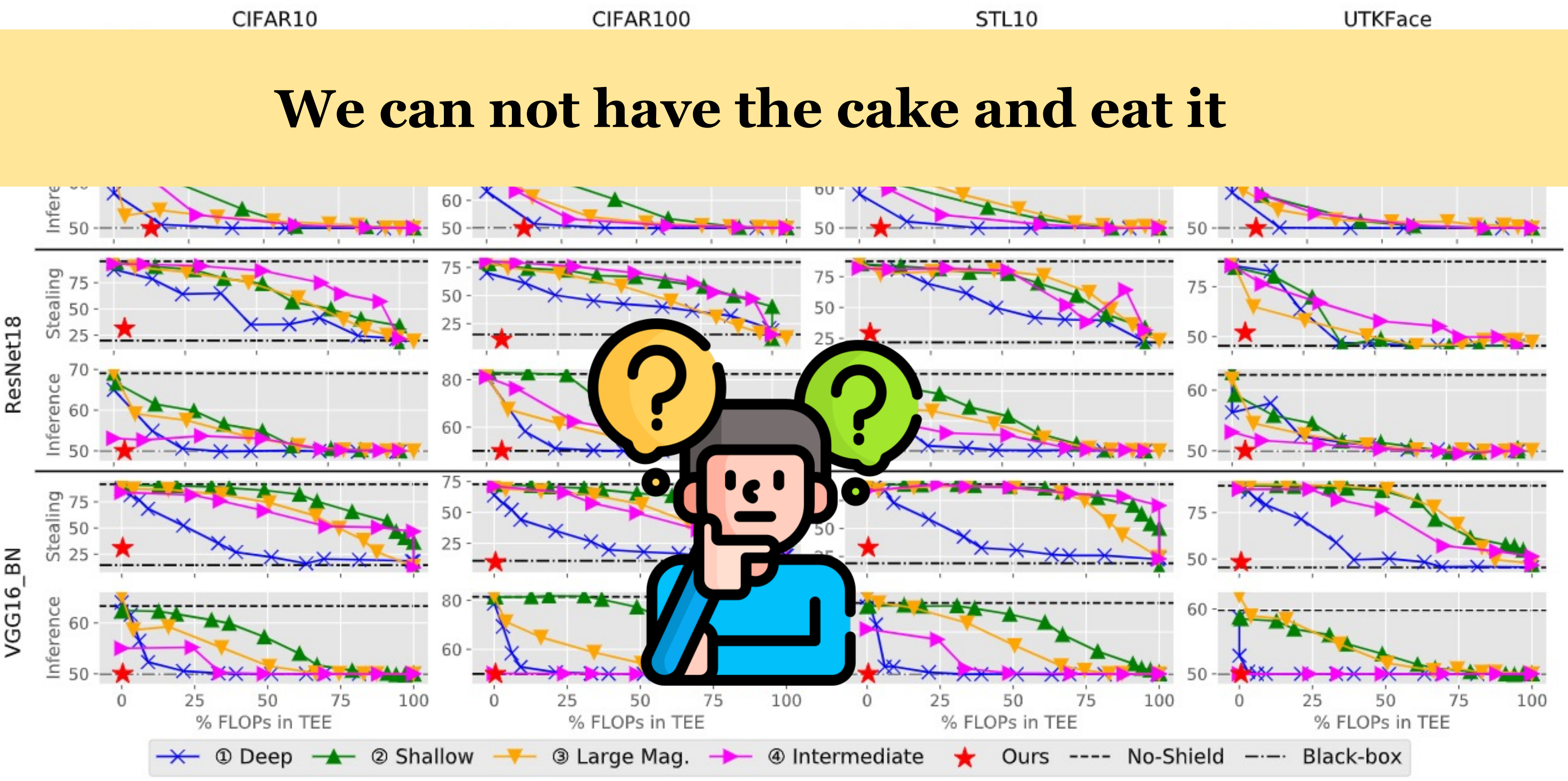
The security-utility trade-off exists in all settings.
The optimal configurations for different settings are different.



Conclusion for RQ2

The security-utility trade-off exists in all settings.
The optimal configurations for different settings are different.

We can not have the cake and eat it

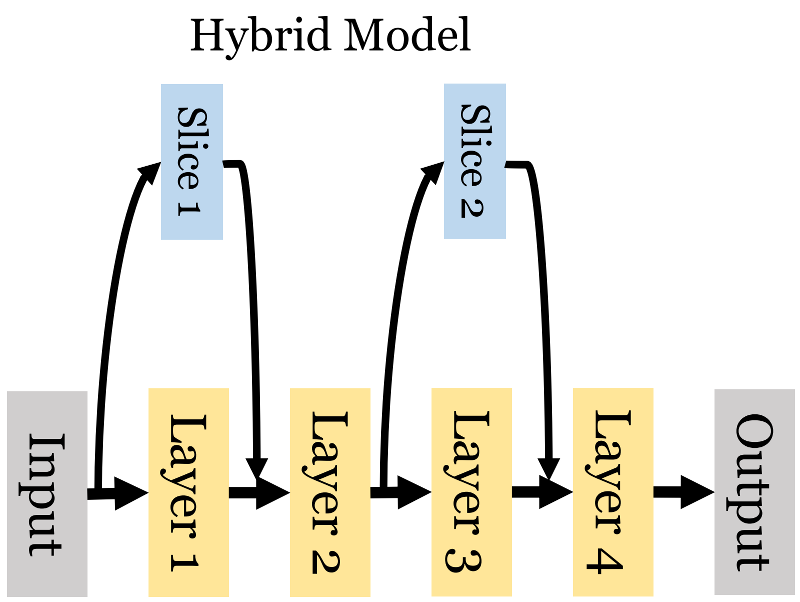


Solve TSDP Security Issue with Model Slice

PEKING UNIVERSITY

- **Existing solution** — **Training-Before-Partition**
  - All the weights contain private information

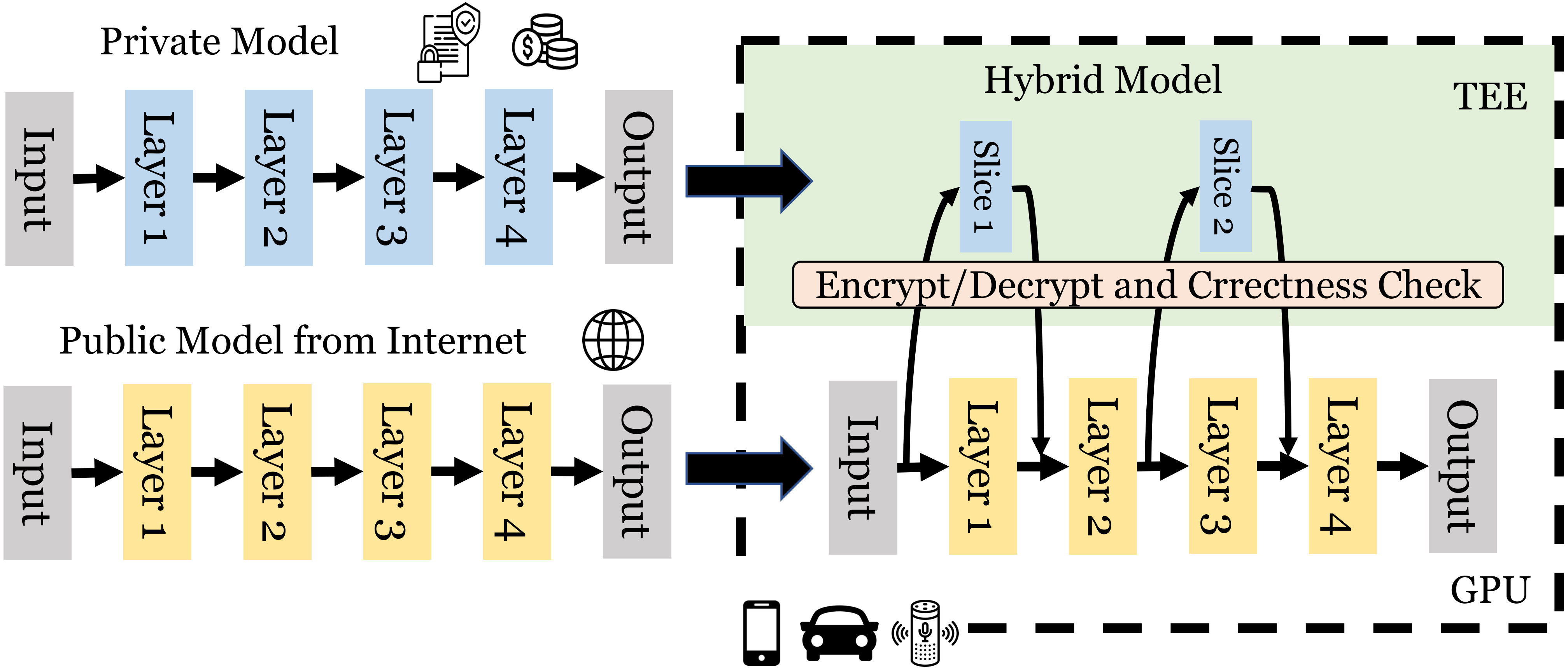
- **Our insight** — **Partition-Before-Training**
  - Isolate private information into light-weight slices
  - Other model parts are never updated by private training data



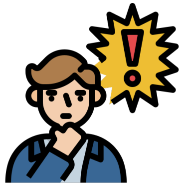
Privacy-Related Mode Slices

TEESlice Goal

Compress the functionality of a private model into private slices



Communication Security at Deployment

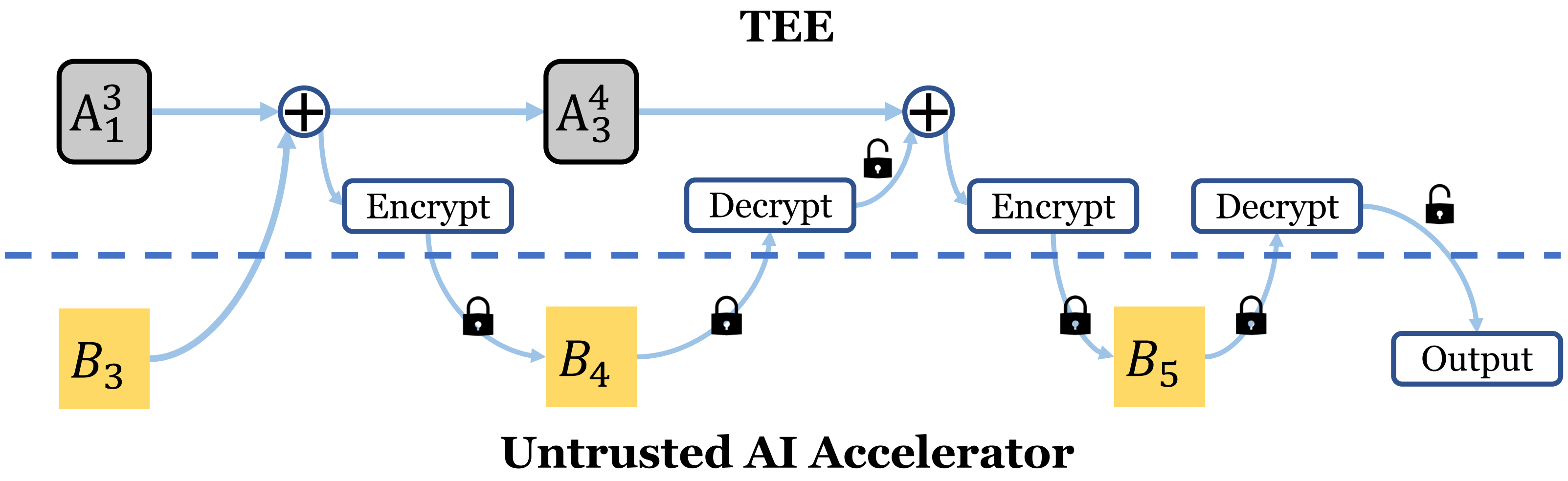


Internal features may leak weight information in the TEE

One-Time-Pad

Secure ← **Used once and never reused**

Efficient ← **Generated remotely or offline**

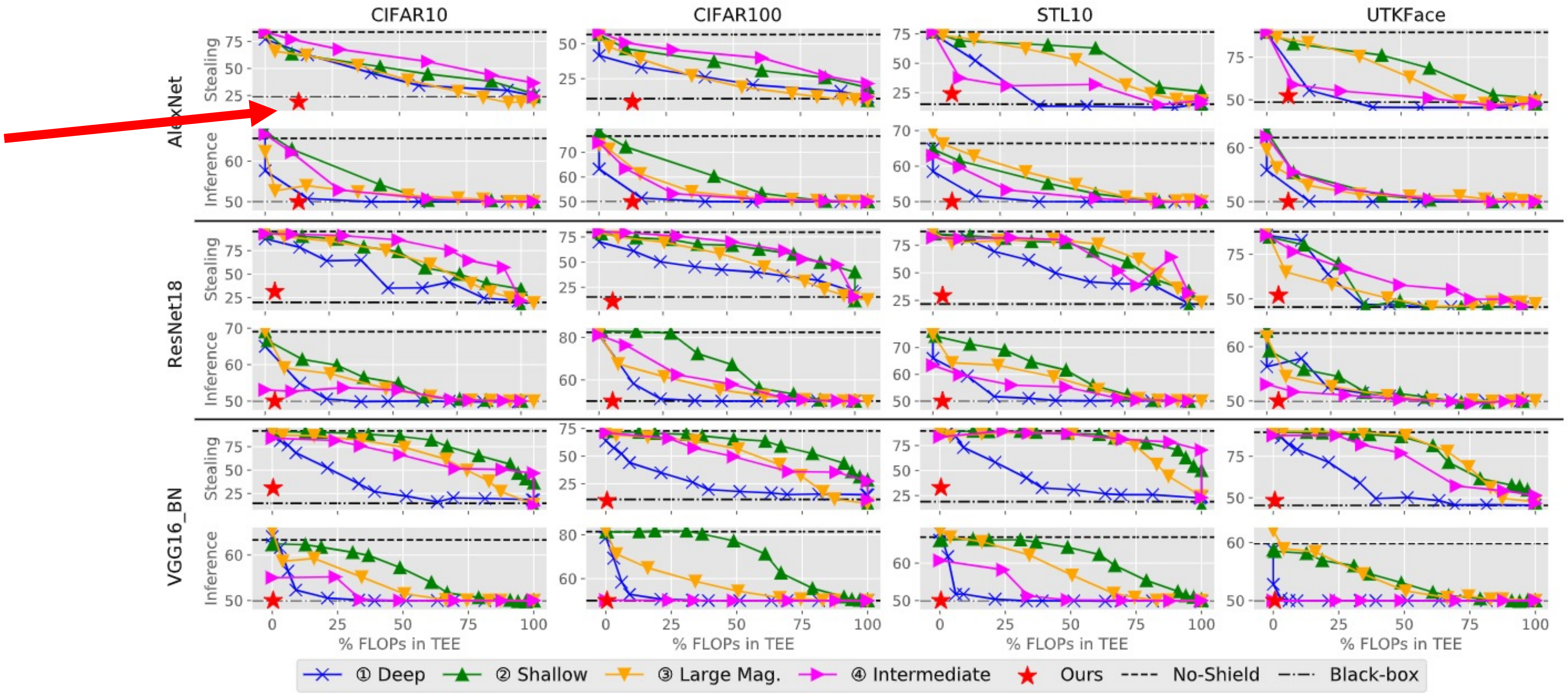


Tramer, Florian, and Dan Boneh. "Slalom: Fast, Verifiable and Private Execution of Neural Networks in Trusted Hardware." ICLR'18

Experimental Evaluation

Provide black-box level protection with low utility cost

Better Security v.s. Utility Trade-off **Reduce Computation Cost by About 10X**



Experimental Evaluation

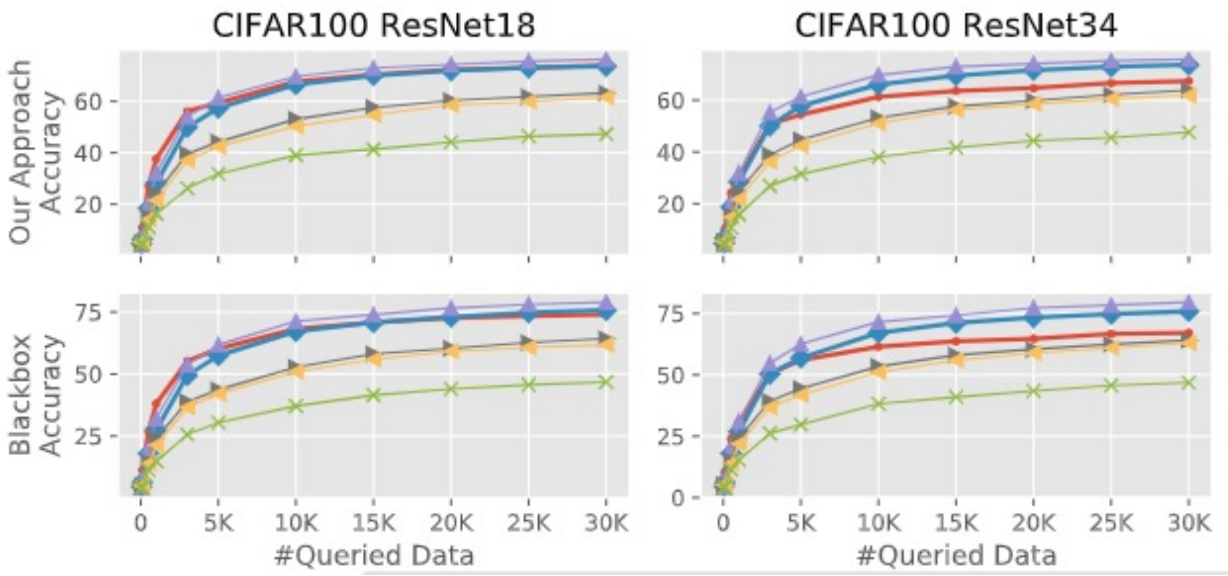
PEKING UNIVERSITY

Does TEESlice lead to performance loss in other aspects?

NO

Table 5: The accuracy comparison between the victim model and the hybrid model trained by TEESLICE in the form of M_{vic}/M_{hyb}. Except for AlexNet where TEESLICE has a higher accuracy due to a larger model capacity, by average, TEESLICE's relative accuracy loss (the ratio between the accuracy of M_{hyb} and the accuracy of M_{vic}) is 0.34%.

	CIFAR10	CIFAR100	STL10	UTKFace
AlexNet	83.71%/86.37%	56.46%/61.96%	76.54%/80.17%	89.42%/88.92%
ResNet18	95.47%/93.65%	79.94%/76.79%	87.51%/86.22%	86.97%/88.24%
ResNet34	91.11%/91.75%	81.00%/76.53%	88.22%/86.15%	87.69%/89.55%
VGG16_BN	91.62%/93.06%	73.03%/73.11%	89.67%/89.42%	89.19%/89.46%
VGG19_BN	92.48%/92.70%	71.38%/73.15%	89.62%/90.70%	89.96%/89.46%



Accuracy only drops by 0.34%

Exposed backbone model does not increase attack performance

Evaluation on Real Devices

PEKING UNIVERSITY

TAOISM: A TEE-based Confidential Heterogeneous Framework for DNN Models

TAOISM

PyTorch

TABLE VI: The throughput comparison between shielding-whole-model, no-shield, and TEESLICE on a real desktop with SGX and GPU. We switch SGX to hardware mode to enable all protection. In parentheses we present the speedup w.r.t. the shielding-whole-model baseline.

	AlexNet	ResNet18	VGG16 BN
Black-box	6.56	7.67	1.55
No-Shield	495.27 (75.53×)	288.56 (36.56×)	103.10 (66.42×)
CIFAR10	44.67 (6.78×)	63.81 (8.32×)	72.80 (46.90×)
CIFAR100	47.36 (7.22×)	46.63 (6.08×)	58.69 (37.81×)
STL10	85.79 (13.08×)	65.24 (8.50×)	71.35 (45.97×)
UTKFaceRace	41.29 (6.30×)	58.03 (6.26×)	42.34 (27.28×)

Improve up to 10X compared black-box

TABLE VII: TEESLICE inference time breakdown.

Data Transfer	Slice in TEE	Backbone on GPU	Non-Linear in TEE
35.61%	40.49%	2.84%	20.96%

Inference time break down

Conclusion

北京大學 PEKING UNIVERSITY

Existing TSDP solutions are not suitable in the era of LLM because offloaded model parts expose a large amount of privacy

The reason of vulnerability is the training-before-partition pipeline

TEESlice uses partition-before-training paradigm to isolate privacy and accelerate model inference

Artifact https://github.com/ziqi-zhang/TEESlice-artifact

TAOISM https://github.com/ziqi-zhang/TAOISM

Thanks

2024-05