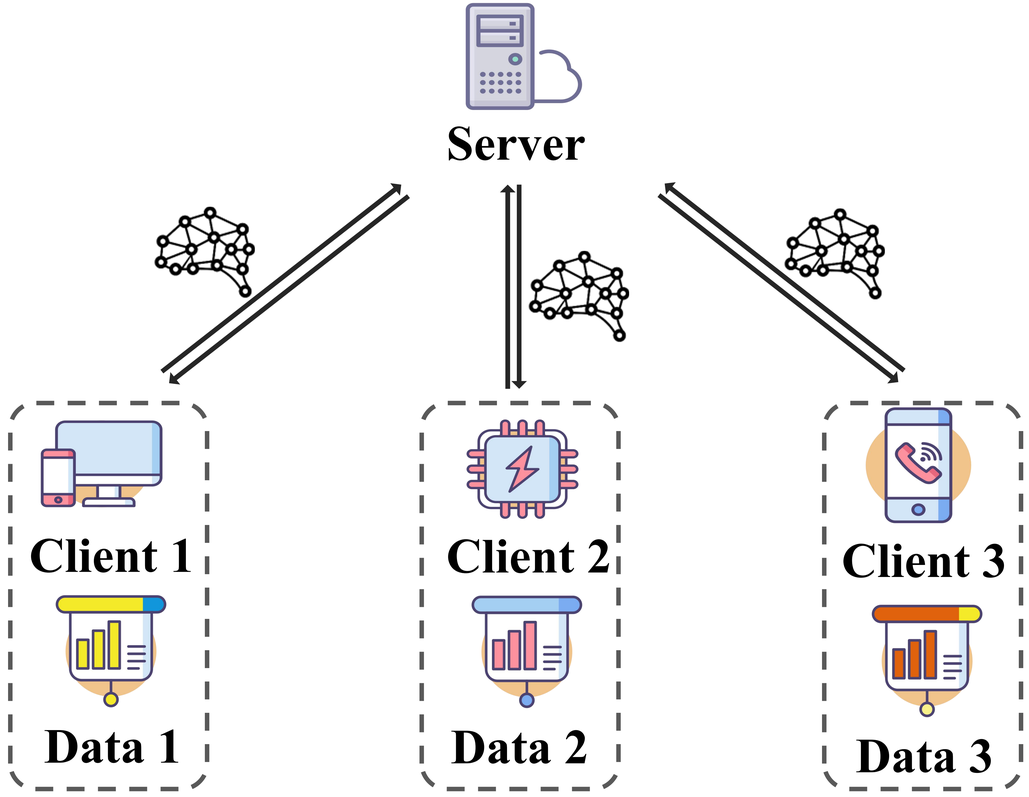
SHE-LoRA: Selective Homomorphic Encryption for Federated Tuning with Heterogeneous LoRA

Jianmin Liu[1] , Li Yan[1]✉, Borui Li[1], Lei Yu[2], Chao Shen[1]
1. Xi'an Jiaotong University 2. Rensselaer Polytechnic Institute

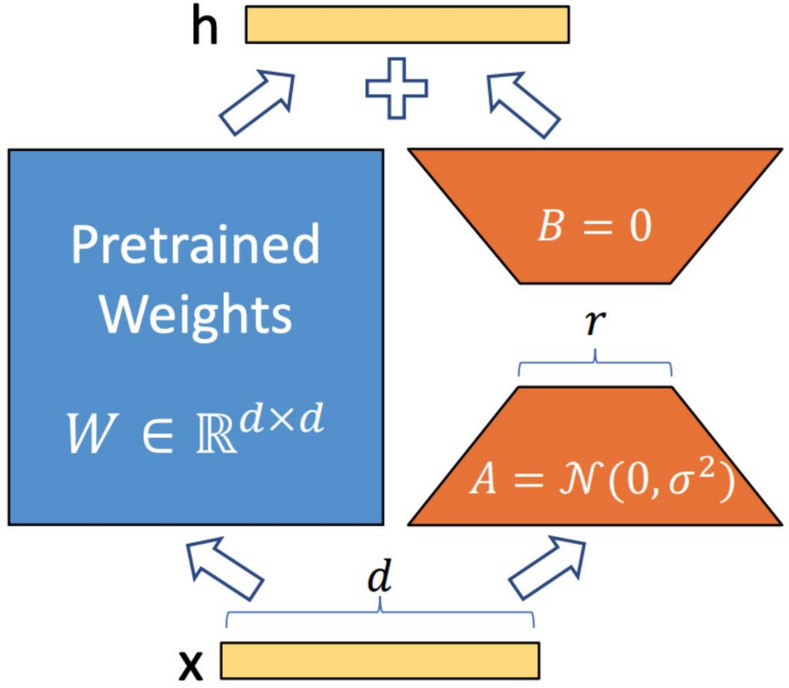
PAPER CODE

ICLR

Background & Problem

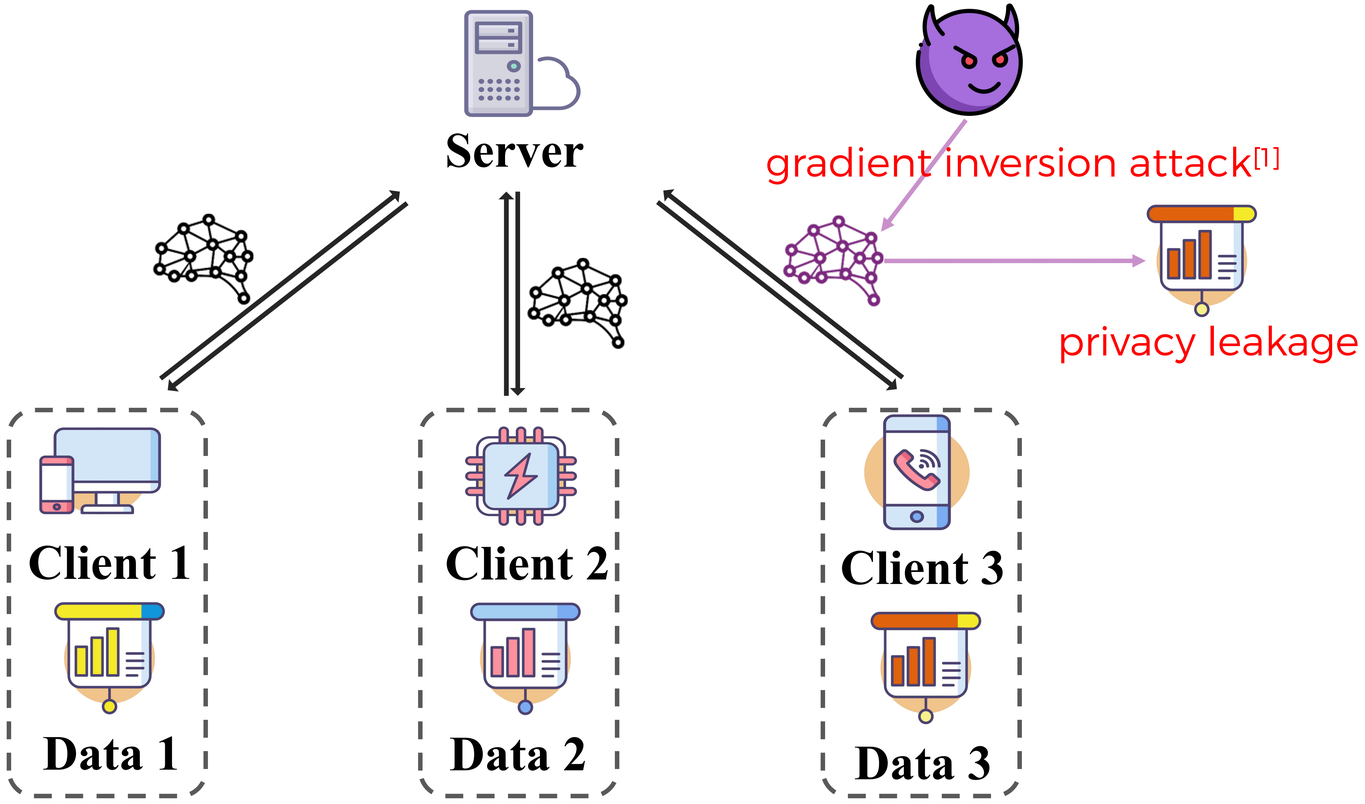


Federated fine-tuning of LLMs.



Low-Rank Adaptation (LoRA)

Background & Problem



Federated fine-tuning of LLMs.

[1] Ivo Petrov, et al. "DAGER: Exact gradient inversion for large language models." In *Proc. of NeurIPS*, 2024.

Background & Problem

Differential Privacy — Performance ↓

While Differential Privacy (DP) ensures formal guarantees by injecting noise, this perturbation becomes amplified through the A × B multiplication in LoRA, introducing errors that hinder convergence and degrade model performance[2].

Secure Multi-Party Computation — Costs ↑

MPC-based secure aggregation, while leveraging techniques like garbled circuits and secret sharing, often demands complex computation and synchronization protocols that limit its practicality in heterogeneous FL settings.

Homomorphic Encryption — Costs ↑

Homomorphic Encryption (HE) enables direct computation on ciphertexts via sophisticated cryptographic primitives, yet often incurs prohibitive computational and communication overhead.

Selective HE (SHE)[3] offers a compelling alternative by encrypting only sensitive parameters, delivering strong privacy with low costs while preserving performance for federated PEFT.

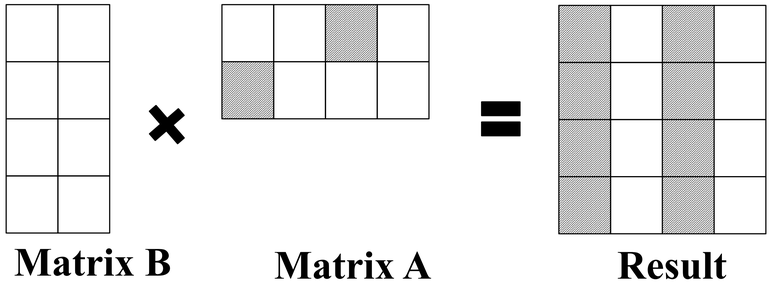
[2] Youbang Sun, et al. "Improving LoRA in privacy-preserving federated learning." In *Proc. of ICLR*, 2024.
[3] Chenghao Hu and Baochun Li. "Maskcrypt: Federated learning with selective homomorphic encryption." IEEE TDSC, 2024.

Motivations

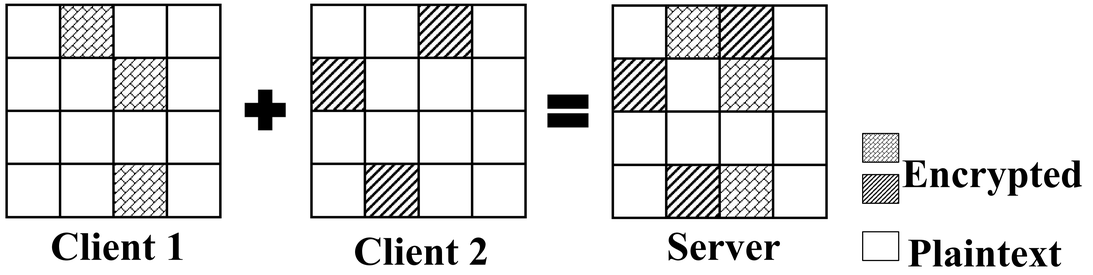
[M1].Naive averaging of LoRAs leads to mathematical errors[4].

$$\sum(BA) \neq \sum B \cdot \sum A$$

[M2].Matrix multiplication expands encryption positions.



[M3].Matrices from heterogeneous clients inflate ciphertext size.



[S1]. Perform multiplication operation

[S2]. Encrypt by column

[S3]. Operate on plaintext and ciphertext separately

[4] Ziyao Wang, et al. "FLoRA: Federated fine-tuning large language models with heterogeneous low-rank adaptations." In Proc. of NeurIPS, 2024.

Preliminary

Which model parameters correspond to privacy risks?

Model pruning method have proved that many parameters can be removed without hurting performance.

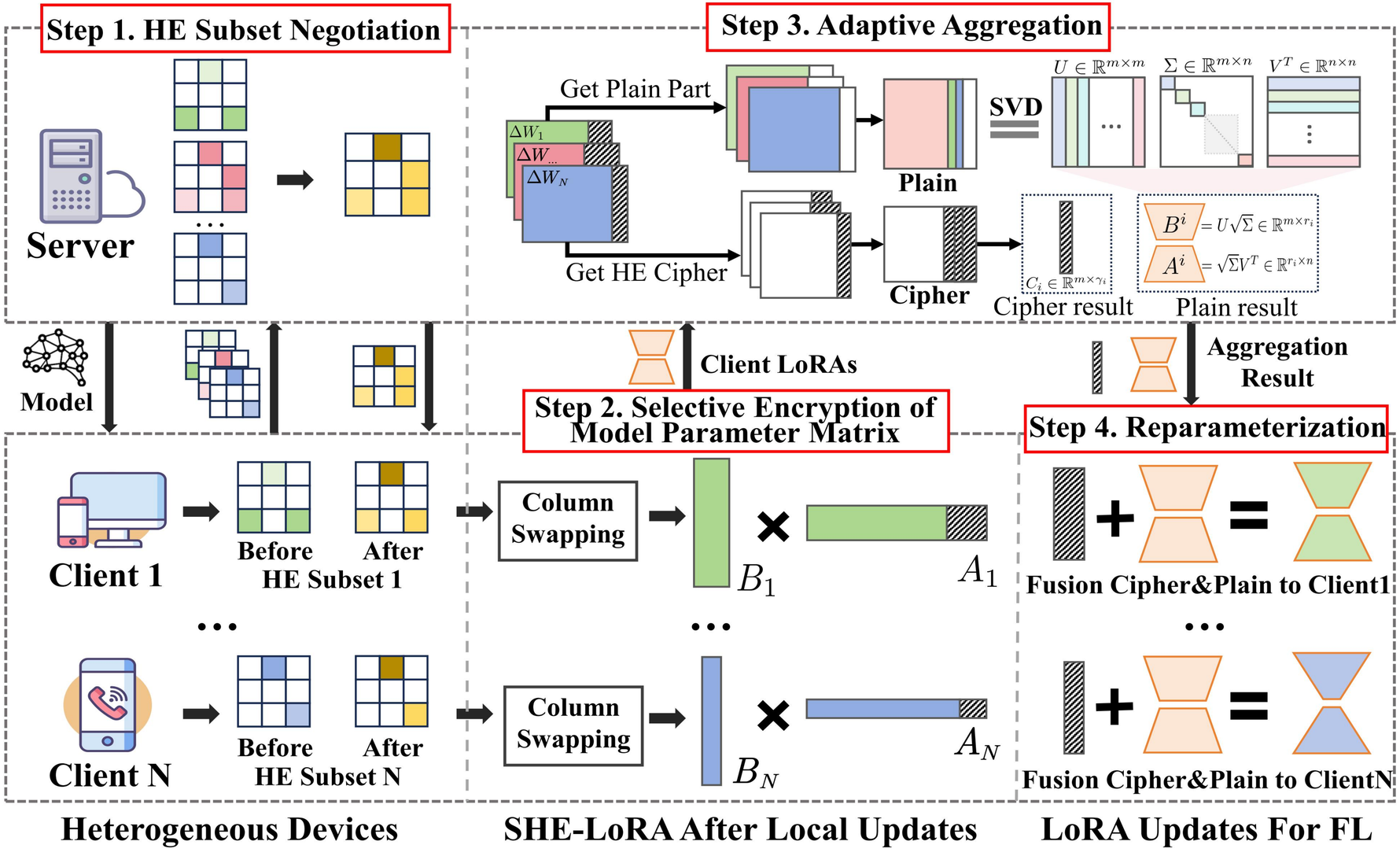
$$\Omega(\mathbf{w}) = |\mathcal{L}(\mathbf{W}) - \mathcal{L}(\mathbf{W}_{-\mathbf{w}})|$$

Inspired by model pruning from [5],
we define $S_j = \sum_{k=0}^{r} |W_{kj}| \cdot \|x_j\|_2$ to calculate the importance of the j-th column of parameters.

More details in Section 2.1.

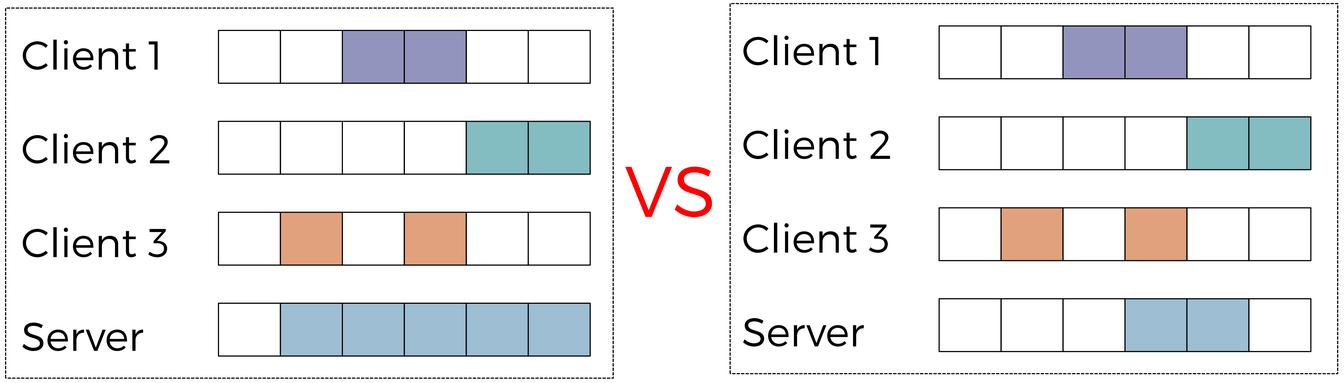
[5] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. “A simple and effective pruning approach for large language models.” In *Proc. of ICLR*, 2024.

Proposed Framework



Step 1: HE Subset Negotiation

Why we need global HE subset?



Naive Aggregation SHE-LoRA Aggregation

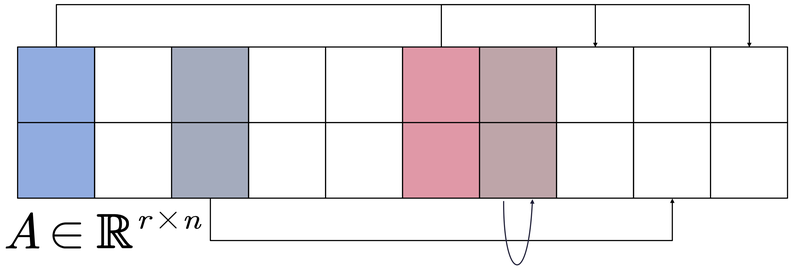
Each client computes a tuple (G_i, S_i), where G_i is the set of columns that needs HE, and S_i is their sensitivities.
The server produces result Res.
We define a score to assess the ability of Res in balancing "privacy" and "costs".

$$\texttt{score}(Res) = \underbrace{\min_{i\in\{N\}} \frac{|Res \cap G_i|}{|G_i|}}_{\textit{min-Coverage}} - \underbrace{\max_{i\in\{N\}} \frac{\sum_{j\in G_i \setminus Res} S_j}{\sum_{j\in G_i} S_j}}_{\textit{max-Risk}}.$$

The objective is to minimize the maximum risk and maximize the coverage for each client.

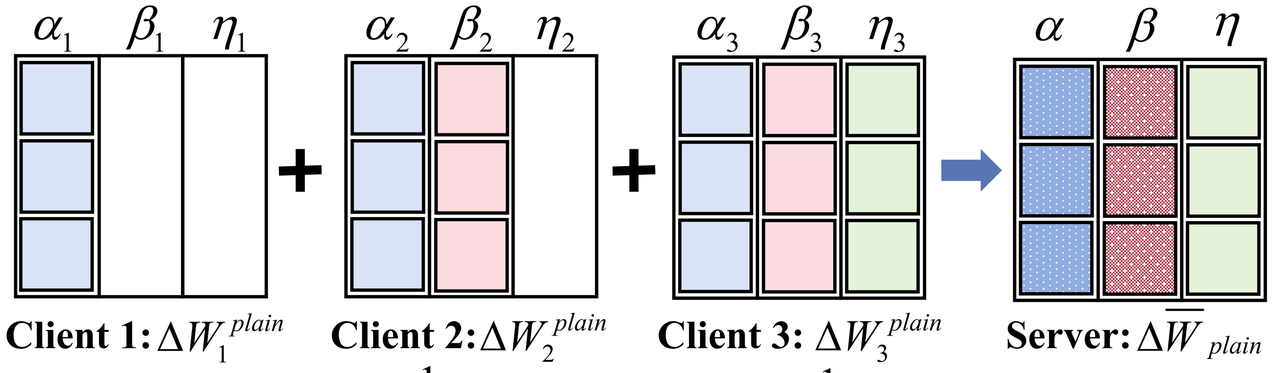
Step 2: HE Selective Encryption

This irregular distribution increases the complexity of matrix batching and the overhead of encryption, decryption and computation.

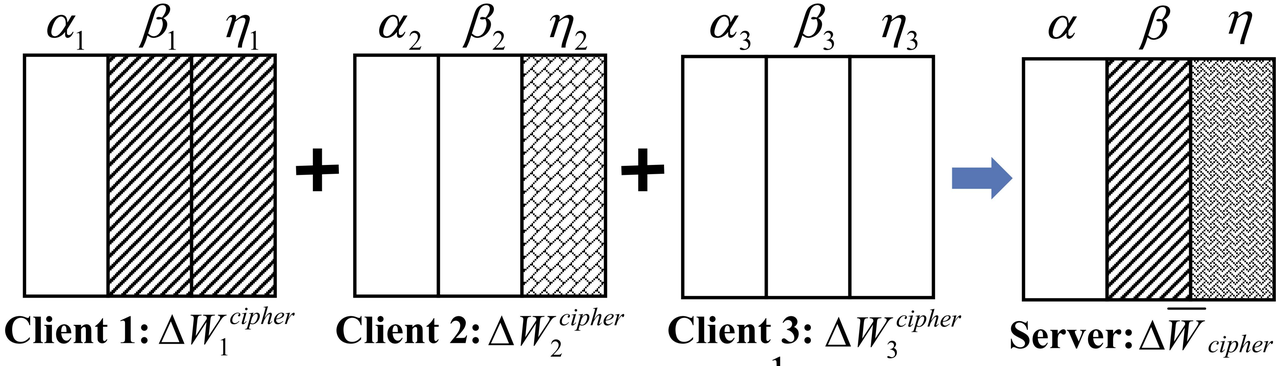


(1) encrypted columns are clustered together, allowing for efficient batch encryption with reduced storage and communication overhead;
(2) the clustered unencrypted columns can be directly used in matrix operations, improving computational efficiency;
(3) the column-wise obfuscation increases the difficulty of potential privacy attacks. (See Appendix D.7)

Step 3: Adaptive Aggregation



Aggregation: $\alpha = \frac{1}{3}(\alpha_1 + \alpha_2 + \alpha_3) \quad \beta = \frac{1}{2}(\beta_2 + \beta_3) \quad \eta = \eta_3$

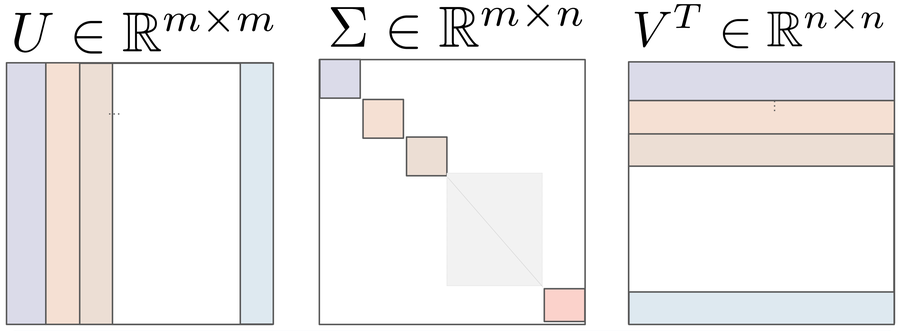


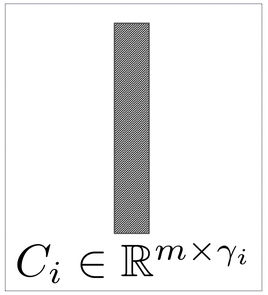
Aggregation: $\alpha = 0 \quad \beta = \beta_1 \quad \eta = \frac{1}{2}(\eta_1 + \eta_2)$

Unencrypted Parameters **Encrypted Parameters**

Zero Padding **Column-wise Weighted Averaging**

SVD





Cipher result

$B^i = U\sqrt{\Sigma} \in \mathbb{R}^{m \times r_i}$

$A^i = \sqrt{\Sigma}V^T \in \mathbb{R}^{r_i \times n}$

Plain result

Step 4: Reparameterization

$$
\begin{aligned}
\Delta\mathbf{W} &= \Delta\overline{\mathbf{W}}_{\text{plain}} + \Delta\overline{\mathbf{W}}_{\text{cipher}} \\
&\overset{\text{SVD}}{=\!=} \mathbf{B}_p\mathbf{A}_p + \mathbf{B}_c\mathbf{A}_c \\
&= (\mathbf{U}_1\sqrt{\mathbf{\Sigma}_1})\sqrt{\mathbf{\Sigma}_1}\mathbf{V}_1^\top + (\mathbf{U}_2\sqrt{\mathbf{\Sigma}_2})\sqrt{\mathbf{\Sigma}_2}\mathbf{V}_2^\top \\
&= \left[\mathbf{U}_1\sqrt{\mathbf{\Sigma}_1}, \mathbf{U}_2\sqrt{\mathbf{\Sigma}_2}\right]^{m\times(r+r)} \begin{bmatrix}\sqrt{\mathbf{\Sigma}_1}\mathbf{V}_1^\top \\ \sqrt{\mathbf{\Sigma}_2}\mathbf{V}_2^\top\end{bmatrix}^{(r+r)\times n}
\end{aligned}
$$

Concat the SVD results from Plain & Cipher.

$$
\begin{aligned}
&\overset{\text{SVD}}{=\!=} (\mathbf{U}_3\mathbf{\Sigma}_3\mathbf{V}_3^\top)(\mathbf{U}_4\mathbf{\Sigma}_4\mathbf{V}_4^\top) \\
&= (\mathbf{U}_3\mathbf{\Sigma}_3\mathbf{V}_3^\top\mathbf{U}_4\sqrt{\mathbf{\Sigma}_4})_{:,:r}(\sqrt{\mathbf{\Sigma}_4}\mathbf{V}_4^\top)_{:r,:} \\
&= \hat{\mathbf{B}}\hat{\mathbf{A}}
\end{aligned}
$$

Perform SVD again to recover LoRA.

Resistance to Gradient Inversion Attack

Data reconstruction scores of DAGER (the SOTA gradient inversion attack).

Dataset	Method	B=4		B=8		B=16	
		R-1	R-2	R-1	R-2	R-1	R-2
SST2	Flex-LoRA	95.18±1.6	94.66±1.8	61.14±1.9	52.49±2.2	10.27±1.6	5.86±1.2
	Flex-LoRA-DP	86.25±1.1	86.11±1.4	80.28±1.1	78.54±1.3	68.62±3.1	66.44±3.7
	MaskCrypt	89.16±1.3	87.93±2.1	61.49±2.2	61.49±2.4	10.91±1.2	6.79±1.4
	SHE-LoRA	0.72±5.2	0.12±1.2	0.98±4.4	0.14±0.6	0.0±0.0	0.0±0.0
Rotten Tomatoes	Flex-LoRA	38.44±1.5	32.76±1.3	3.76±1.4	2.12±2.1	0.0±0.0	0.0±0.0
	Flex-LoRA-DP	36.74±1.9	31.28±2.6	3.76±1.3	2.02±2.3	0.0±0.0	0.0±0.0
	MaskCrypt	31.65±2.0	25.11±2.6	6.09±1.0	3.27±1.2	0.0±0.0	0.0±0.0
	SHE-LoRA	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0

Strong resistance with only 0.3‰ encryption budget.

Resistance to Membership Inference Attacks

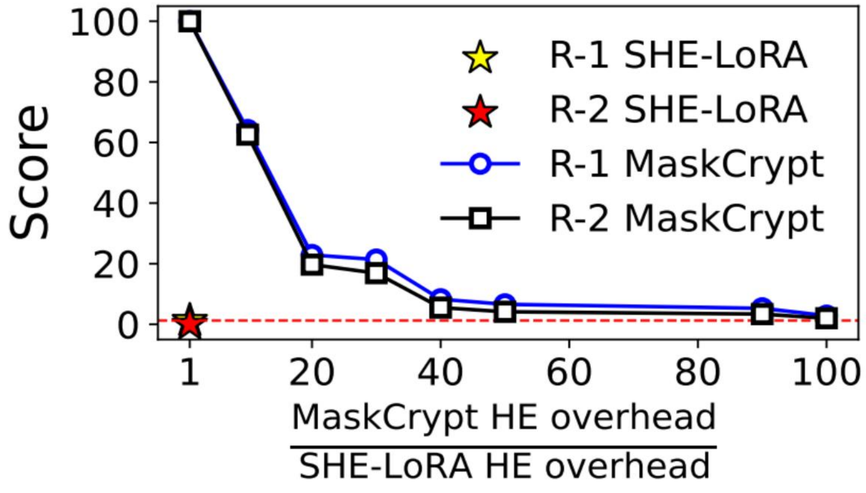
The AUROC results reported under 7 membership inference attacks.

Model	Loss	Lowercase	Zlib	Min-k (0.1)	Min-k (0.5)	Recall	PAC
Base	50.9%	48.4%	50.2%	50.5%	50.9%	50.1%	51.2%
Vanilla LoRA	81.4%	80.5%	76.7%	80.9%	82.9%	73.8%	83.3%
$\gamma = 1‰$	62.6%	62.8%	60.3%	62.5%	63.5%	64.7%	65.0%
$\gamma = 1\%$	56.8%	57.7%	55.4%	56.5%	57.3%	58.4%	58.4%
$\gamma = 5\%$	54.1%	55.2%	53.1%	53.7%	54.3%	55.8%	55.3%
$\gamma = 10\%$	56.8%	57.7%	55.4%	56.5%	57.3%	58.4%	58.4%
$\gamma = 20\%$	52.4%	53.2%	51.7%	52.1%	52.5%	53.1%	53.3%

Base model was not trained on private data, and all of others was trained on private data.

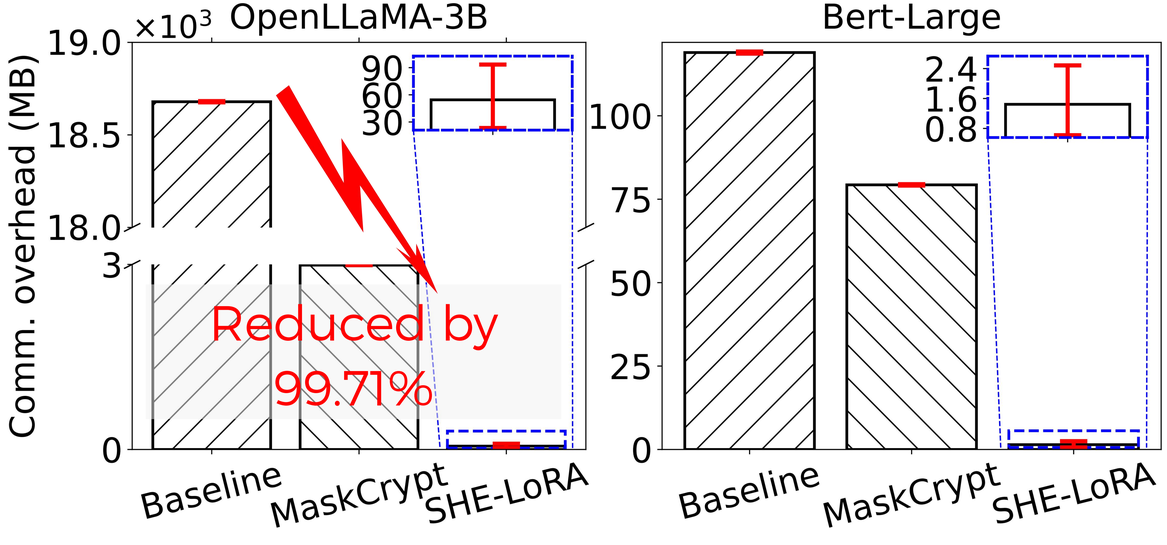
Strong resistance with nearly random-guessing !!!

Costs

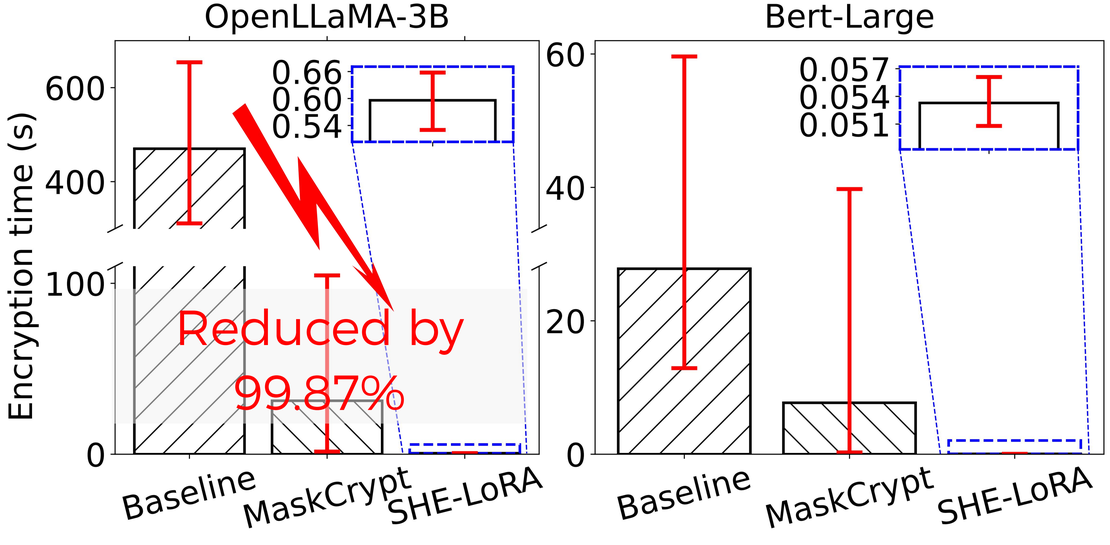


100× HE Overhead

Communication overhead.



Encryption time.



Performance on Language Models

Performance of Bert-Large on the GLUE benchmark.

Method	SST2	MRPC	QQP	RTE	WNLI	QNLI
FedIT (Zhang et al., 2024)	47.41	31.62	64.71	43.07	46.34	48.87
FedSA (Guo et al., 2024)	48.23	33.71	66.32	43.56	48.27	48.26
HeterLoRA (Cho et al., 2024)	55.73	68.38	72.17	44.72	48.86	49.14
Flex-LoRA (Bai et al., 2024)	52.29	**74.81**	**75.31**	46.93	49.66	49.51
SHE-LoRA	**57.11**	70.88	72.52	**50.18**	**57.75**	**59.63**

Performance of OpenLLaMA-3B on the MMLU benchmark.

Method	STEM	SS	Humanities	Average
FedIT (Zhang et al., 2024)	21.5	21.3	20.4	21.2
FedSA (Guo et al., 2024)	21.8	21.4	19.7	20.1
HeterLoRA (Cho et al., 2024)	24.7	25.4	25.8	26
Flex-LoRA (Bai et al., 2024)	26.2	27.9	**26.6**	27.4
SHE-LoRA	**28.1**	**29.2**	26.5	**28.2**

Performance comparable to SOTA.

Performance on Vision Models

Performance comparison of CLIP on 5 vision tasks.

Method	Datasets					
	MNIST	DTD	EuroSAT	GTSRB	SVHN	AVG
Clip-Vit-Base-Patch-16 r = 8						
FedIT (Zhang et al., 2024)	93.38	68.74	93.17	83.62	90.43	85.87
FedSA (Guo et al., 2024)	93.13	67.51	94.23	85.12	88.49	85.69
HeterLoRA (Cho et al., 2024)	95.37	68.83	96.22	87.18	91.55	87.83
Flex-LoRA (Bai et al., 2024)	99.28	**70.32**	**98.48**	95.74	**95.37**	**91.84**
SHE-LoRA	**99.33**	69.97	98.35	**95.88**	95.13	91.73
Clip-Vit-Base-Patch-16 r = 16						
FedIT (Zhang et al., 2024)	95.36	68.85	94.56	85.37	91.58	87.14
FedSA (Guo et al., 2024)	94.62	67.92	95.18	87.23	90.67	87.12
HeterLoRA (Cho et al., 2024)	94.56	68.21	96.77	89.62	92.28	88.29
Flex-LoRA (Bai et al., 2024)	**99.30**	70.05	**98.29**	**95.45**	95.15	91.65
SHE-LoRA	99.25	**70.85**	98.22	95.35	**96.03**	**91.94**

Performance comparable to SOTA.

ICLR

Thanks !

Feel free to drop me an email for any questions !

jianmin.liu@stu.xjtu.edu.cn