Tsinghua University

Microsoft Research
微软亚洲研究院

Differentially Private Knowledge Transfer For Federated Learning

Tao Qi[1], Fangzhao Wu[2], Chuhan Wu[3], Liang He[1], Yongfeng Huang[1,3,4], Xing Xie[2]

[1]Department of Electronic Engineering & BNRist, Tsinghua University, Beijing 100084, China

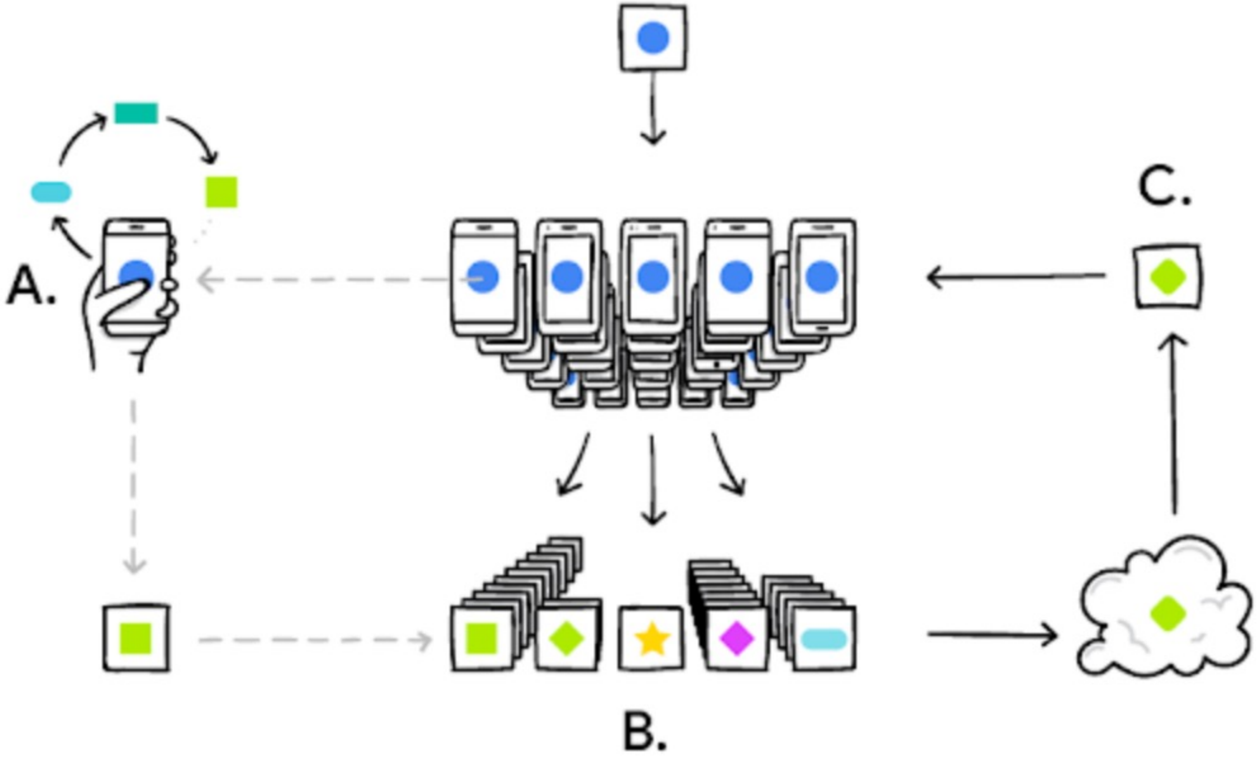
[2]Microsoft Research Asia, Beijing 100080, China

[4]Zhongguancun Laboratory, Beijing 100094, China

[5]Institute for Precision Medicine of Tsinghua University, Beijing 102218, China

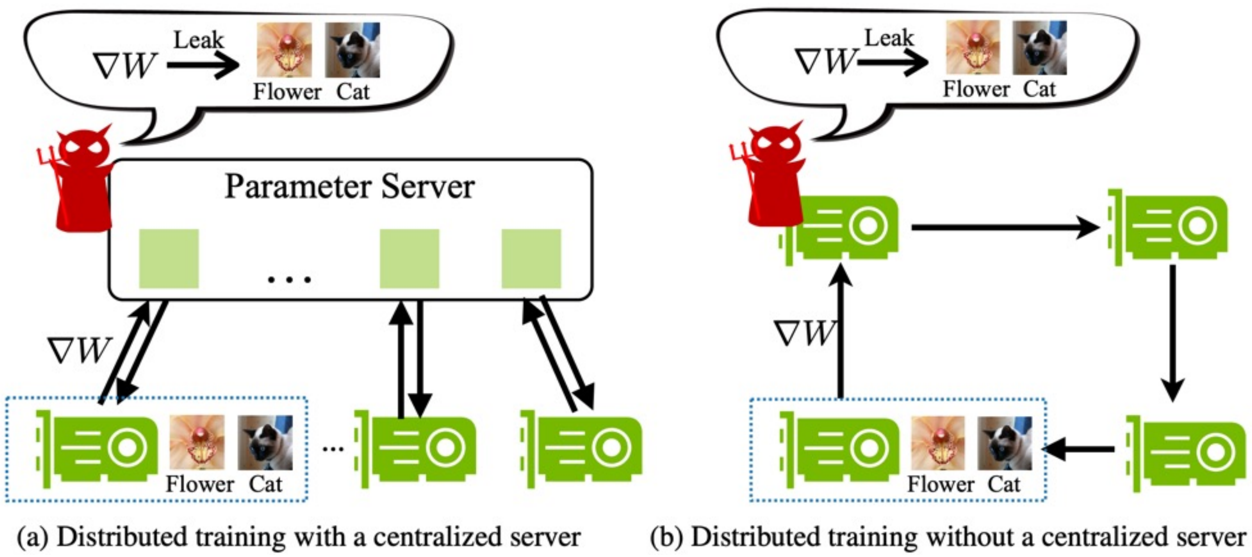
Federated Learning

- A representative privacy-preserving machine learning framework
- Collaboratively learning models from many clients on decentralized data
  - Sharing local updates instead of raw data to exchange useful information

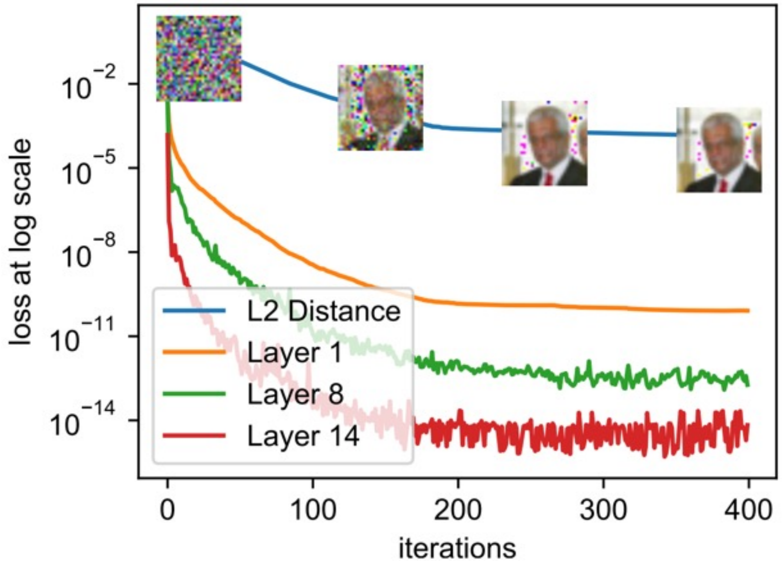


Privacy Security of Federated Learning

- Privacy security is an important factor of federated learning
- Although without centralizing data, FL has no privacy security guarantees



An example gradient-based attack on federated learning

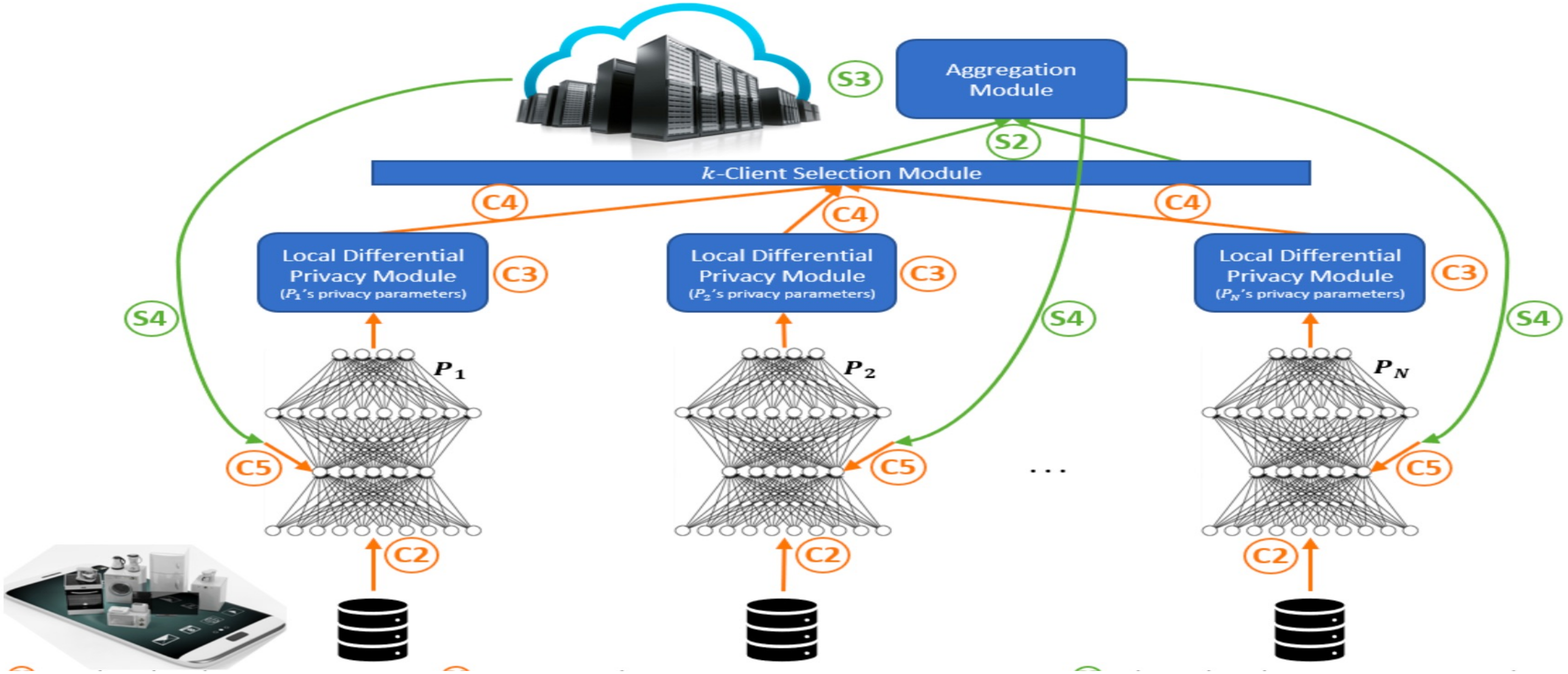


Privacy attack on CIFAR-10

- Challenge: Private data can be recovered from shared gradients/models

LDP-enhanced Federated Learning

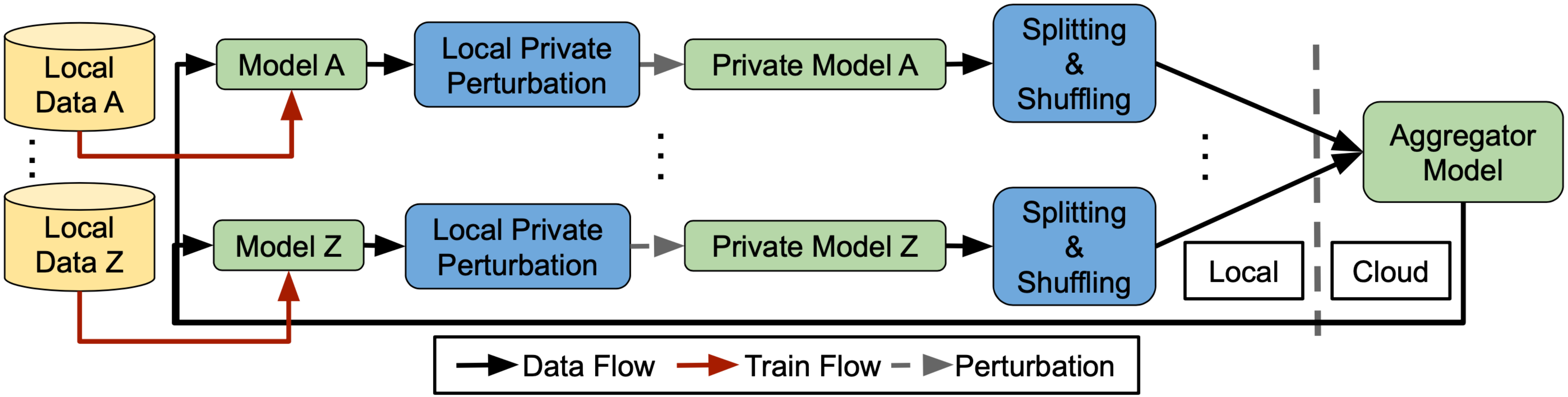
- Local differential privacy: providing theorical privacy guarantee
  - ϵ- LDP: $\Pr[\mathcal{M}(X) = Y] \leq e^{\epsilon} \Pr[\mathcal{M}(X') = Y], \forall X, X', Y$
- Naive method: adding noise to local updates before sending it to the server



- Challenge: LDP technique usually faces serious curse of dimensionality

LDP-enhanced Federated Learning

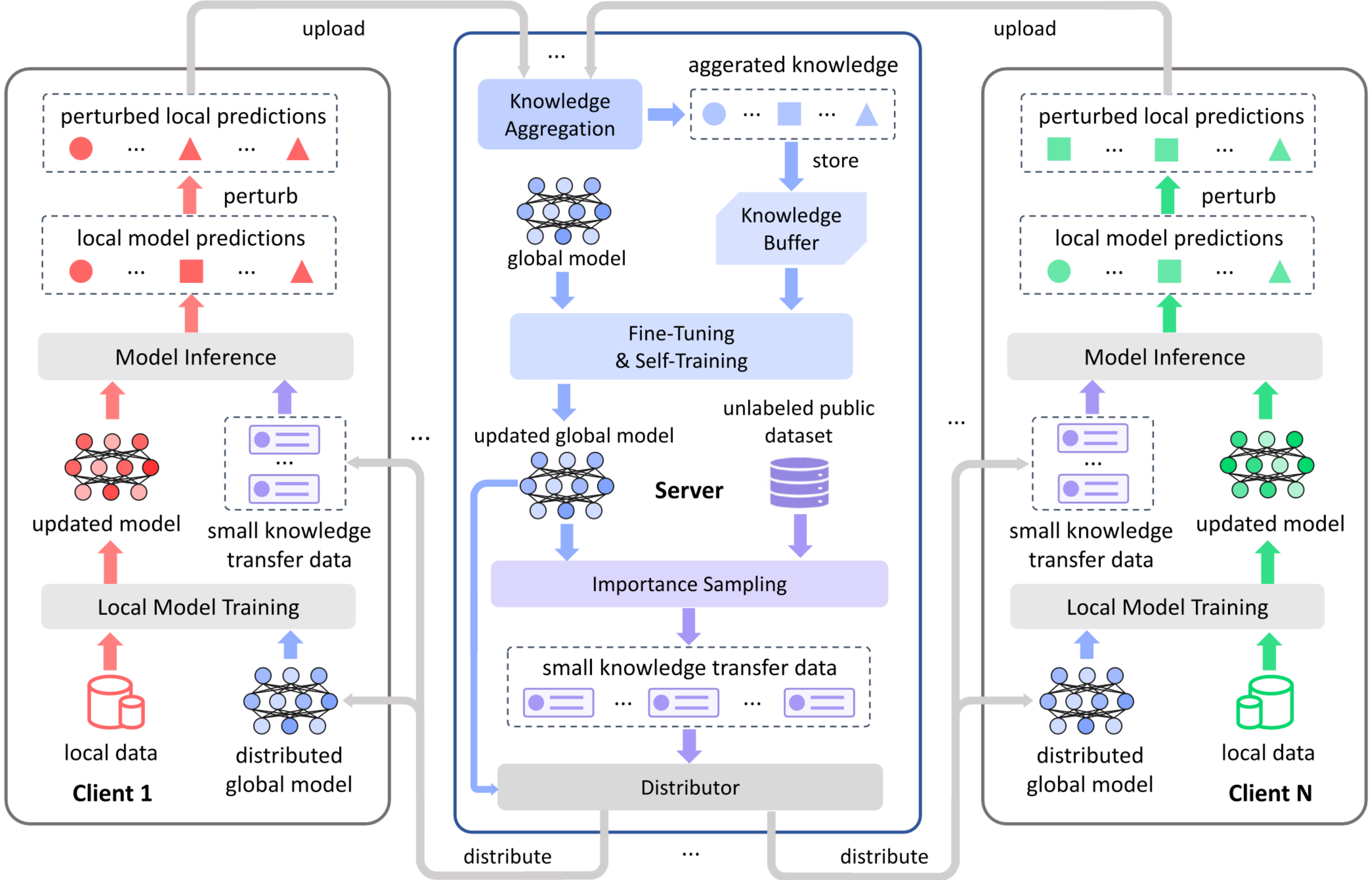
- Shuffle local updates to bypass the difficulty of privacy budget accumulation
  - e.g., model shuffle, parameter shuffle



- Challenge:
  - Cause heavy communication costs and online latency

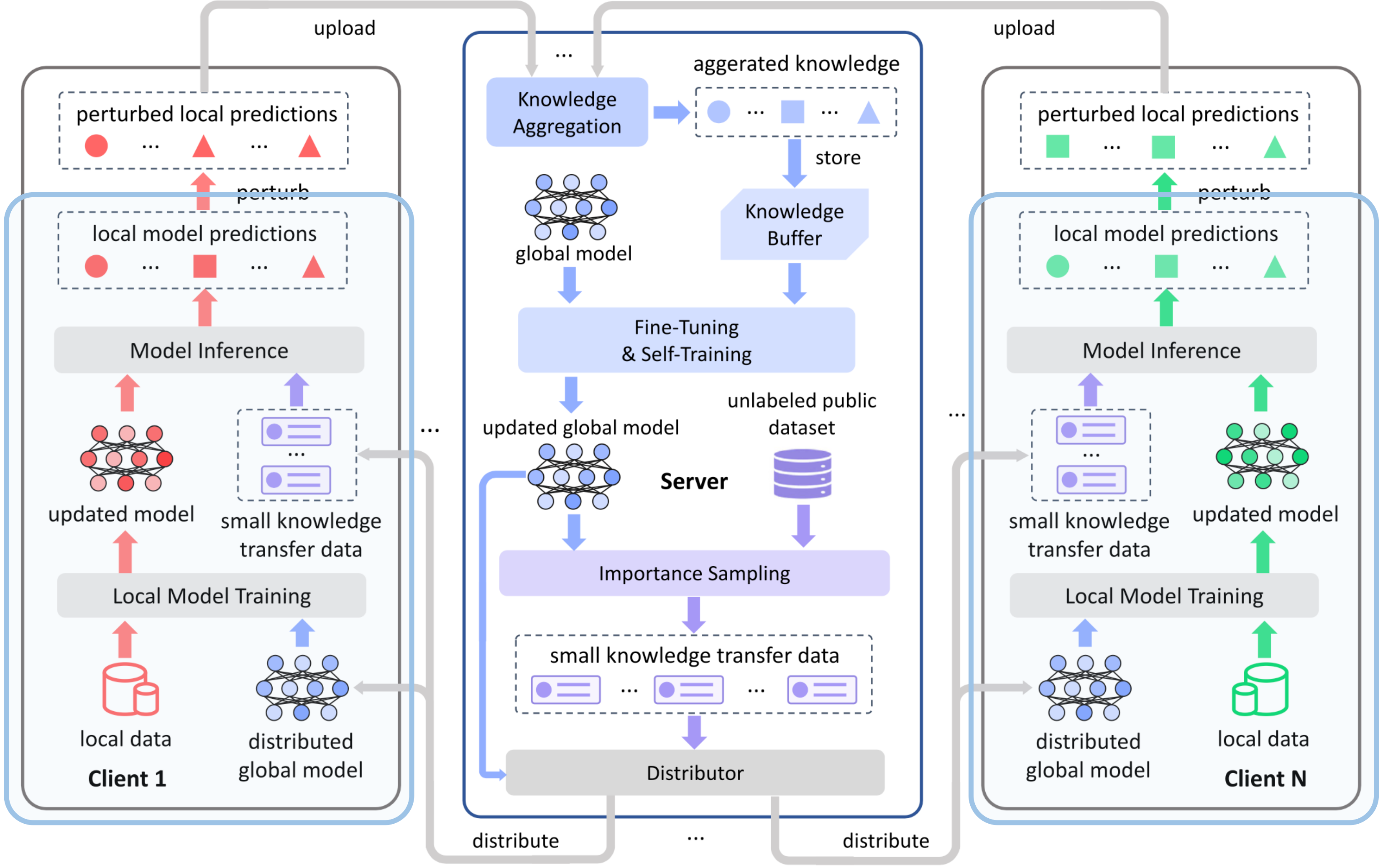
PrivateKT: Differential Private Knowledge Transfer

- Using small data to transfer high-quality knowledge with privacy guarantees



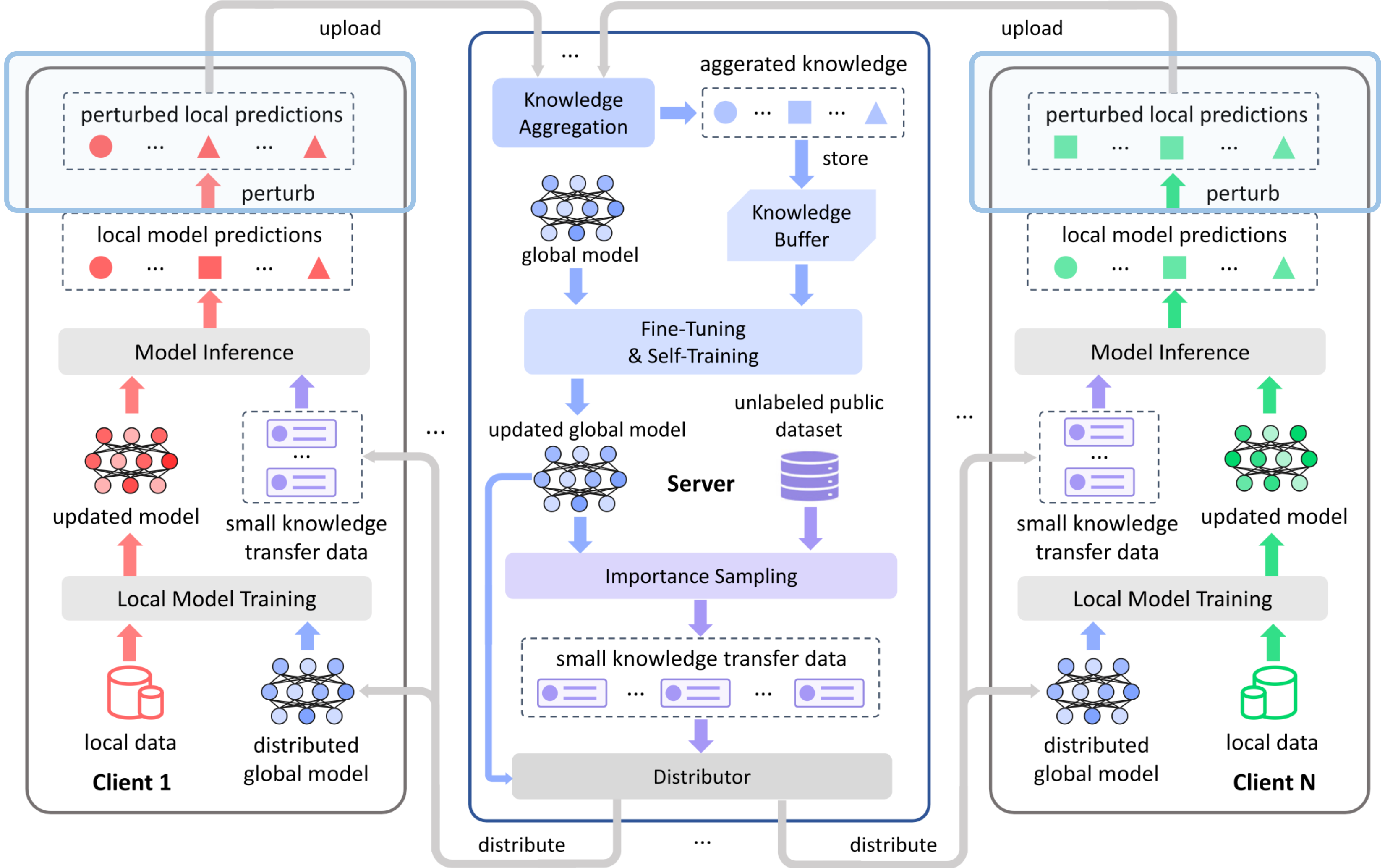
PrivateKT: Differential Private Knowledge Transfer

- Local model training and knowledge inference



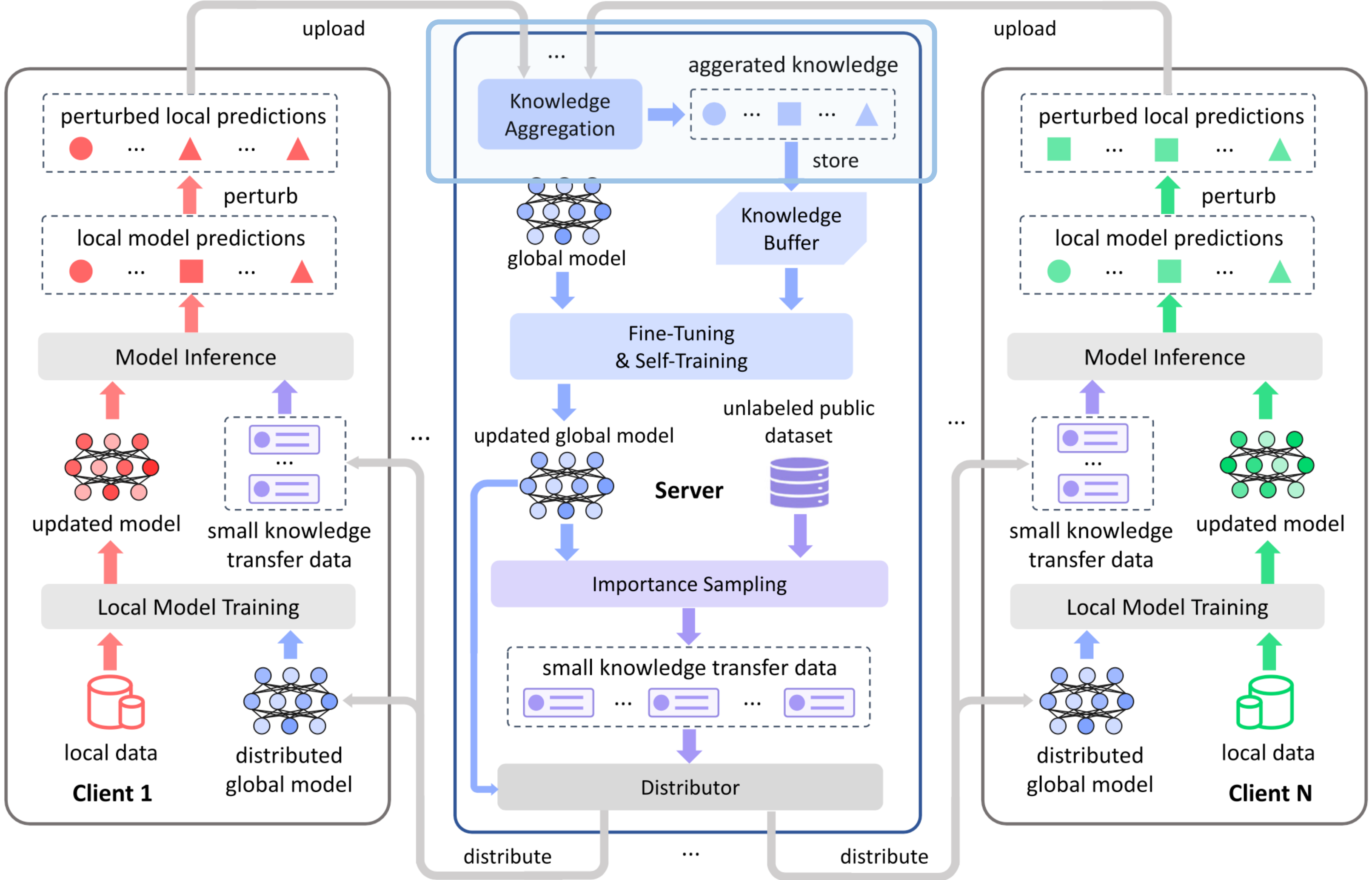
PrivateKT: Differential Private Knowledge Transfer

- Random response: $\hat{y} = x_c y + (1 - x_c) n_c, \;\; x_c \sim \mathcal{B}(\beta), \;\; n_c \sim \mathcal{M}(C)$



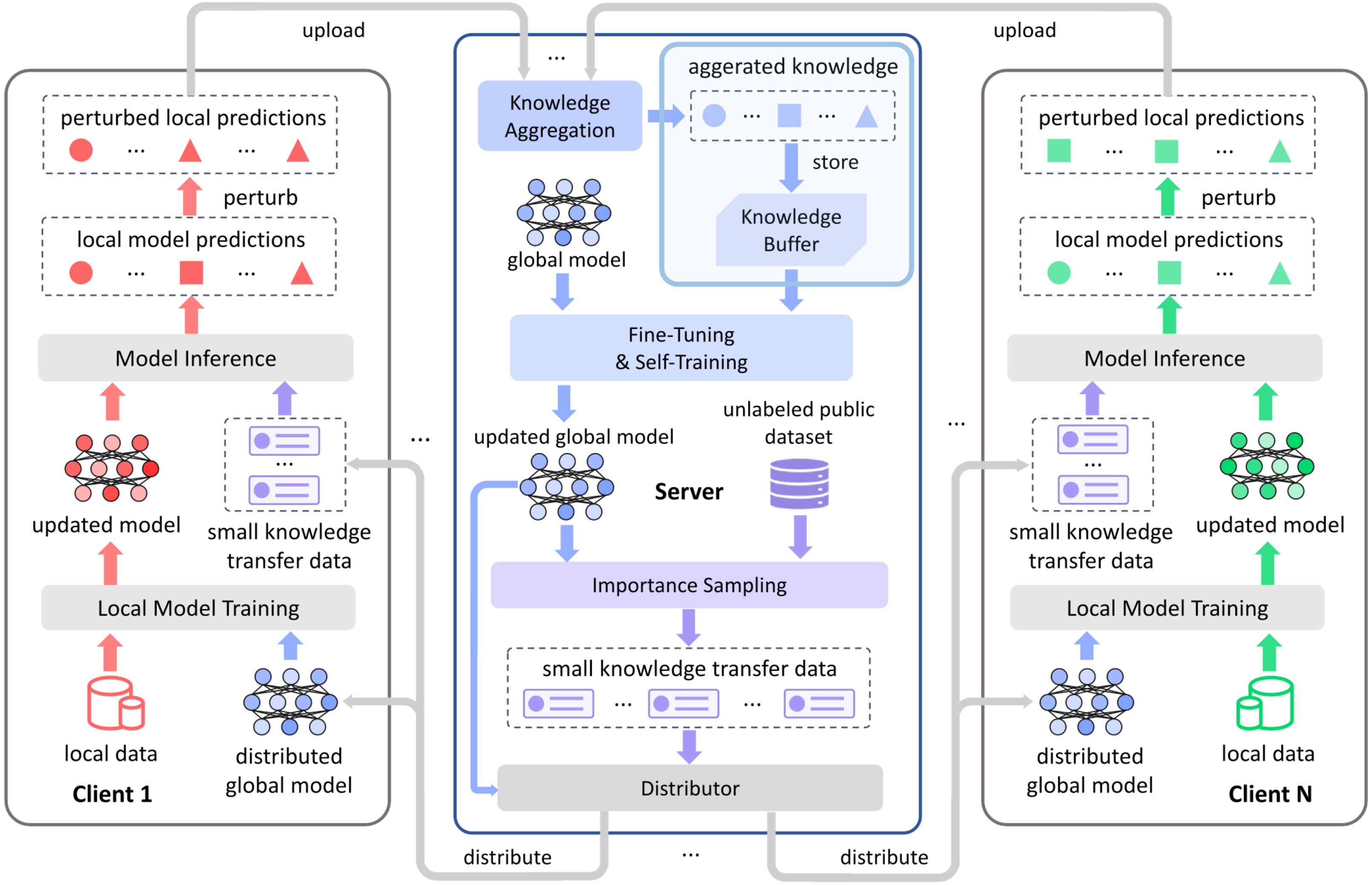
PrivateKT: Differential Private Knowledge Transfer

- Knowledge aggregation: $\hat{\boldsymbol{y}}_i^t = (\frac{1}{N}\sum_j^N \hat{\boldsymbol{y}}_{j,i}^t - \frac{1-\beta}{C}\mathbf{1})/\beta$



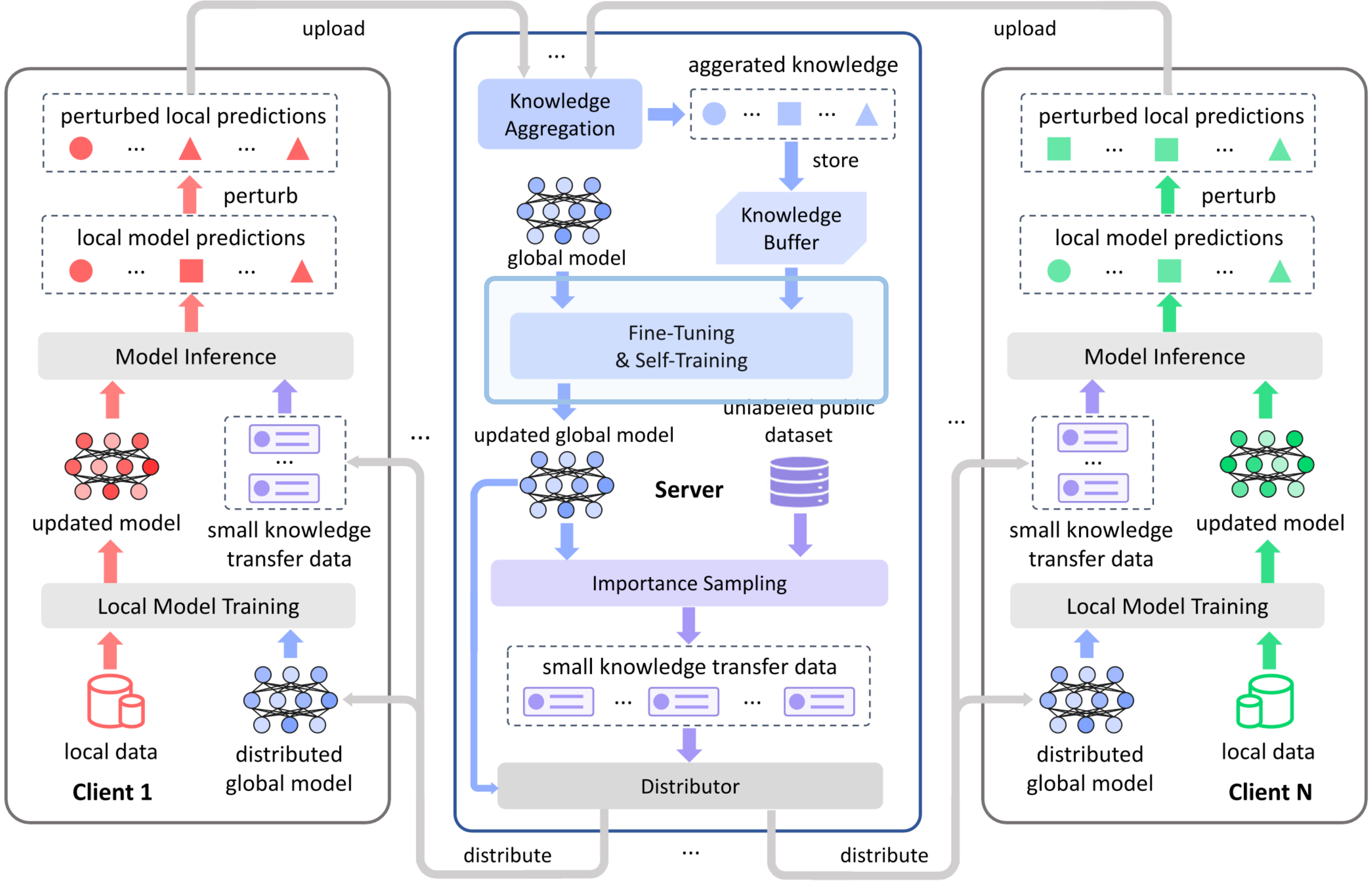
PrivateKT: Differential Private Knowledge Transfer

- Model distillation based on a data buffer caching previous KD samples



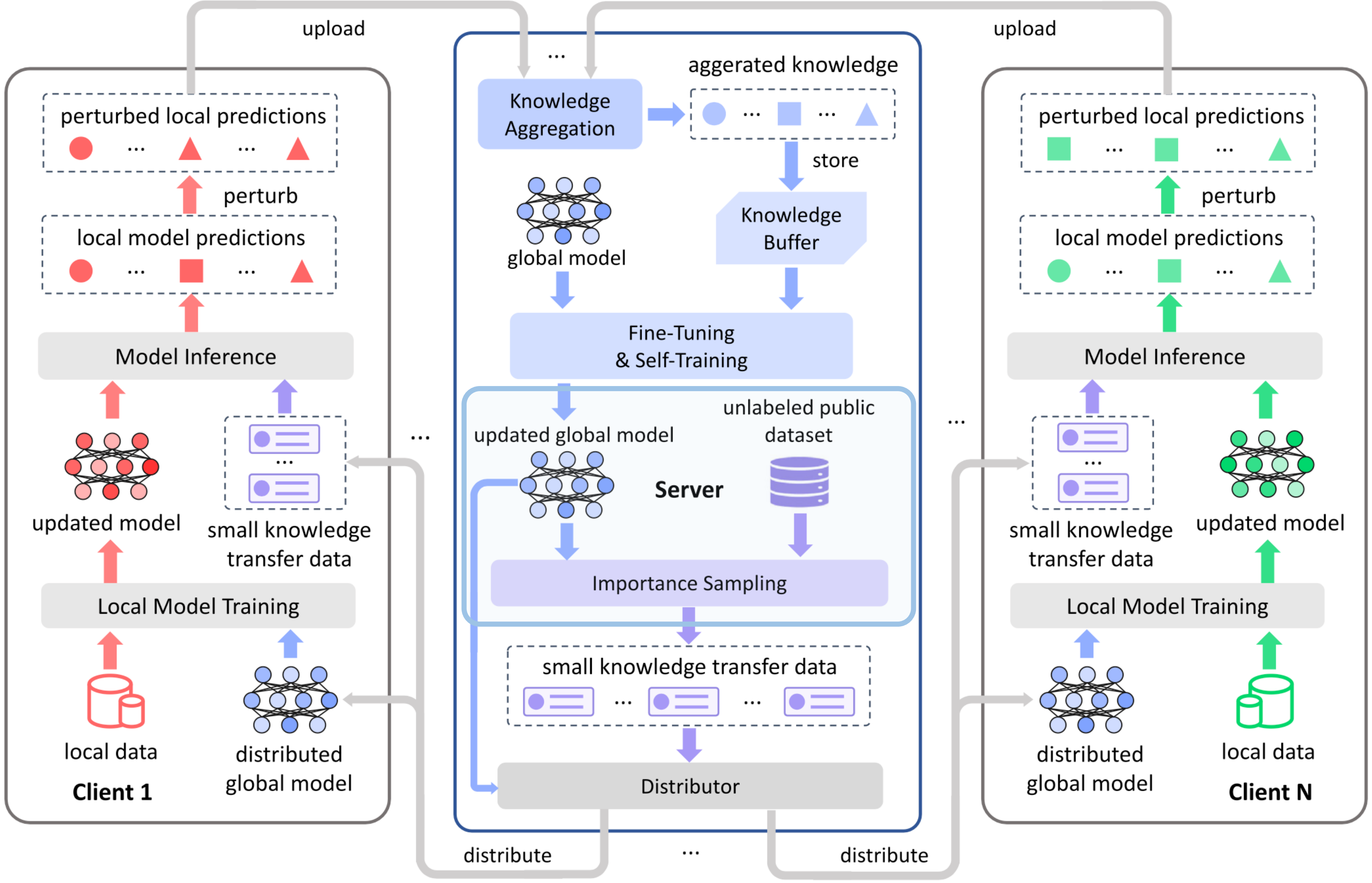
PrivateKT: Differential Private Knowledge Transfer

- Model self-training on high-confident samples



PrivateKT: Differential Private Knowledge Transfer

- Importance sampling: $p_i = \exp(-s_i)/\sum_j \exp(-s_j)$



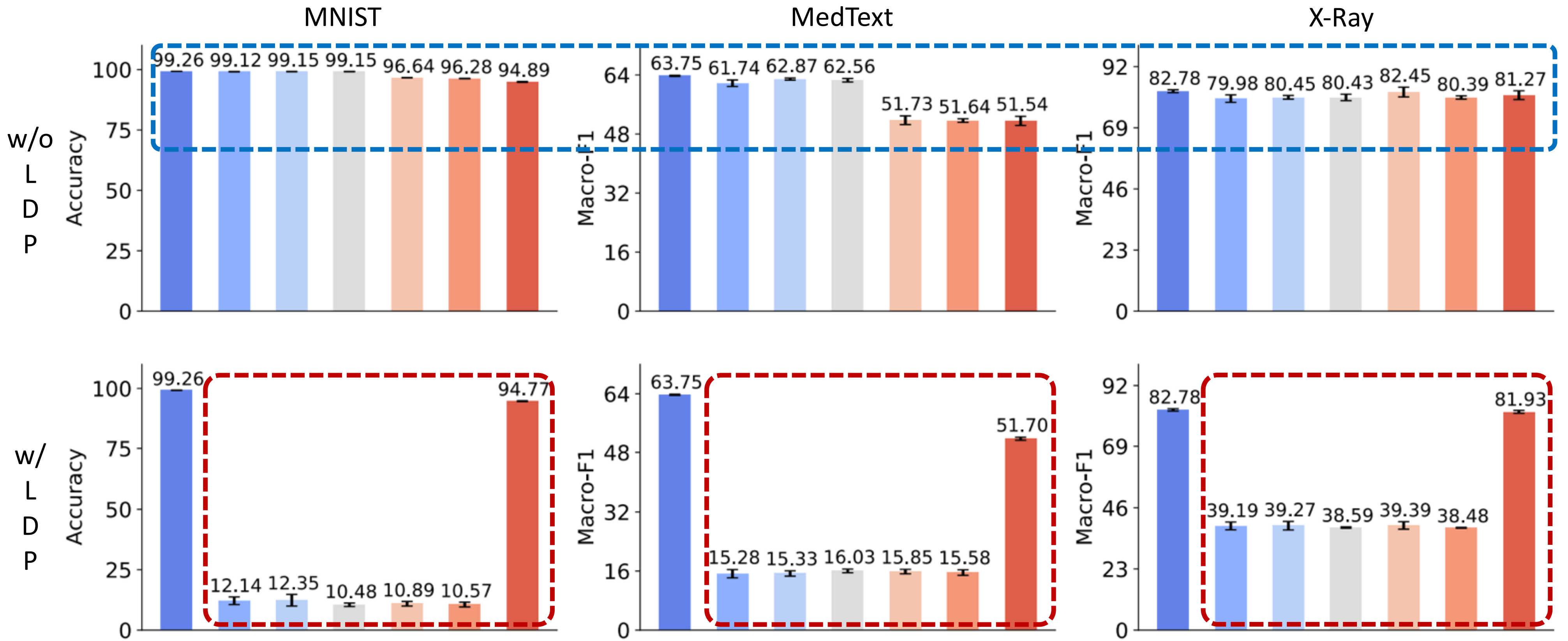
PrivateKT: Theoretical Analysis

- Theorem1: $\hat{\mathbf{y}}_\mathrm{i}$ is an unbiased estimation of $\frac{1}{\mathrm{N}}\sum_{\mathrm{j}}^{\mathrm{N}} \mathbf{y}_{\mathrm{j,i}}^{\mathrm{t}}$
  - $\mathbb{E}[\widehat{\boldsymbol{y}}_i] = \mathbb{E}\left[\frac{\frac{1}{N}\sum_j^N \widehat{\boldsymbol{y}}_{j,i}^t - \frac{1-\beta}{C}\mathbf{1}}{\beta}\right] = \frac{\frac{1}{N}\sum_j^N \mathbb{E}[\widehat{\boldsymbol{y}}_{j,i}^t] - \frac{1-\beta}{C}\mathbf{1}}{\beta} = \frac{\frac{1}{N}\sum_j^N \mathbb{E}[\widehat{\boldsymbol{y}}_{j,i}^t] - \frac{1-\beta}{C}\mathbf{1}}{\beta} = \frac{1}{N}\sum_j^N \boldsymbol{y}_{j,i}^t$
- Theorem2: The MSE of estimation can asymptotically converge to 0
  - $\mathbb{E}\left[(\hat{y}^i - \frac{1}{\mathrm{N}}\sum_{\mathrm{j}}^{\mathrm{N}} y_{\mathrm{j,i}}^{\mathrm{t}})^2\right] < \frac{2C^2\beta(1-\beta)+\frac{1}{12}C^2}{N\beta^2}$
- Theorem3: PrivateKT can achieve ϵ-LDP i.f.f. $\beta = \frac{\exp\left(\frac{\epsilon}{K}\right)-1}{\exp\left(\frac{\epsilon}{K}\right)-1+C}$
  - $\exp(\frac{\epsilon}{K}) = \max \frac{\Pr[\widehat{\boldsymbol{y}}=\boldsymbol{c}]}{\Pr[\widehat{\boldsymbol{y}}'=\boldsymbol{c}\,]} = \frac{\Pr[\widehat{\boldsymbol{y}}=\boldsymbol{c}\,,\boldsymbol{y}=\boldsymbol{c}]}{\Pr[\widehat{\boldsymbol{y}}'=\boldsymbol{c}\,,\boldsymbol{y}'\neq\boldsymbol{c}]} = \frac{\beta+\frac{1-\beta}{C}}{\frac{1-\beta}{C}} = \frac{(C-1)\beta+1}{1-\beta}$

Performance Evaluation

- Datasets: MNIST, MedText, X-Ray

Without the protection of LDP, Private can achieve comparable accuracy with baselines

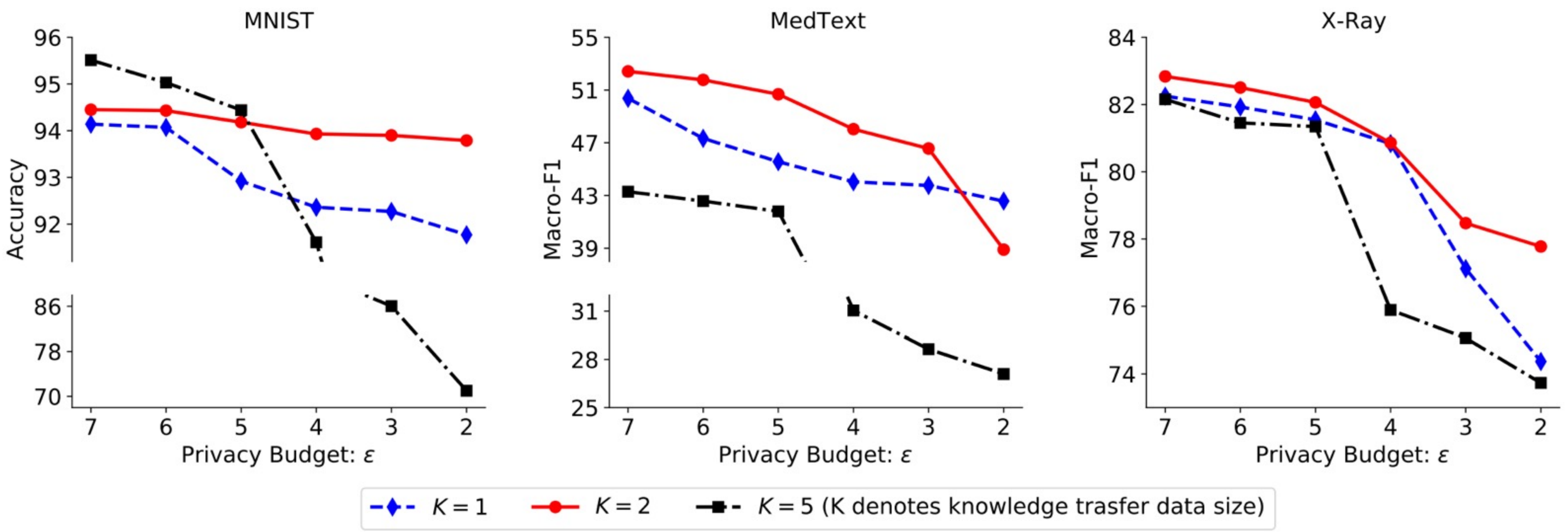


PrivateKT can effectively reduce the performance drop of federated learning under strong LDP protection

Privacy-Utility Analysis

- Evaluate the model accuracy under varying privacy security levels

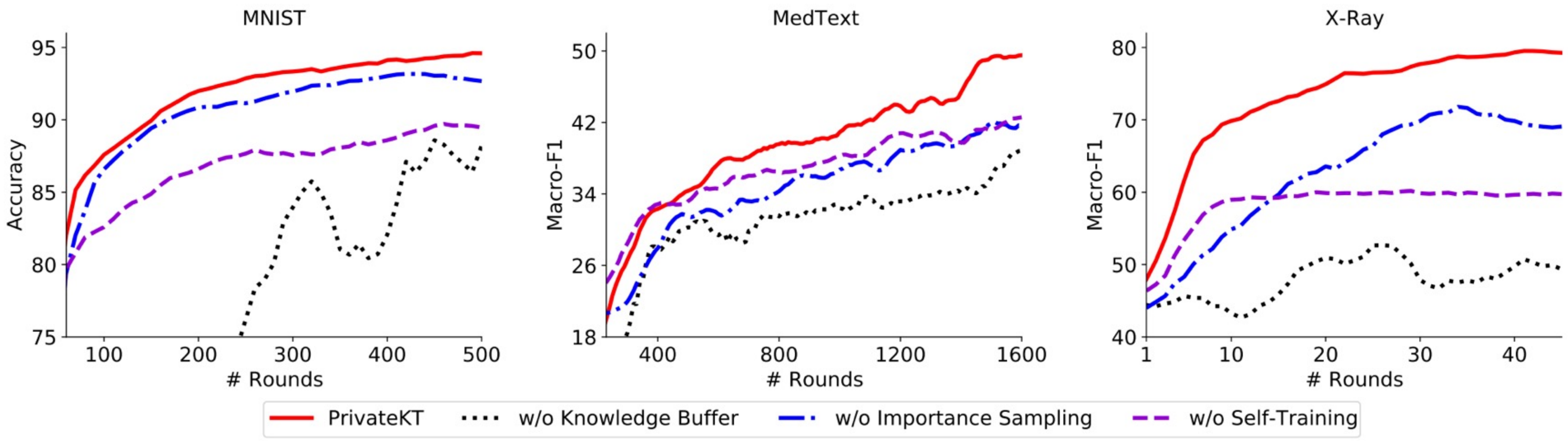
PrivateKT can still effectively train model parameters under strong privacy guarantees (e.g., $\epsilon = 2$)



Ablation Study

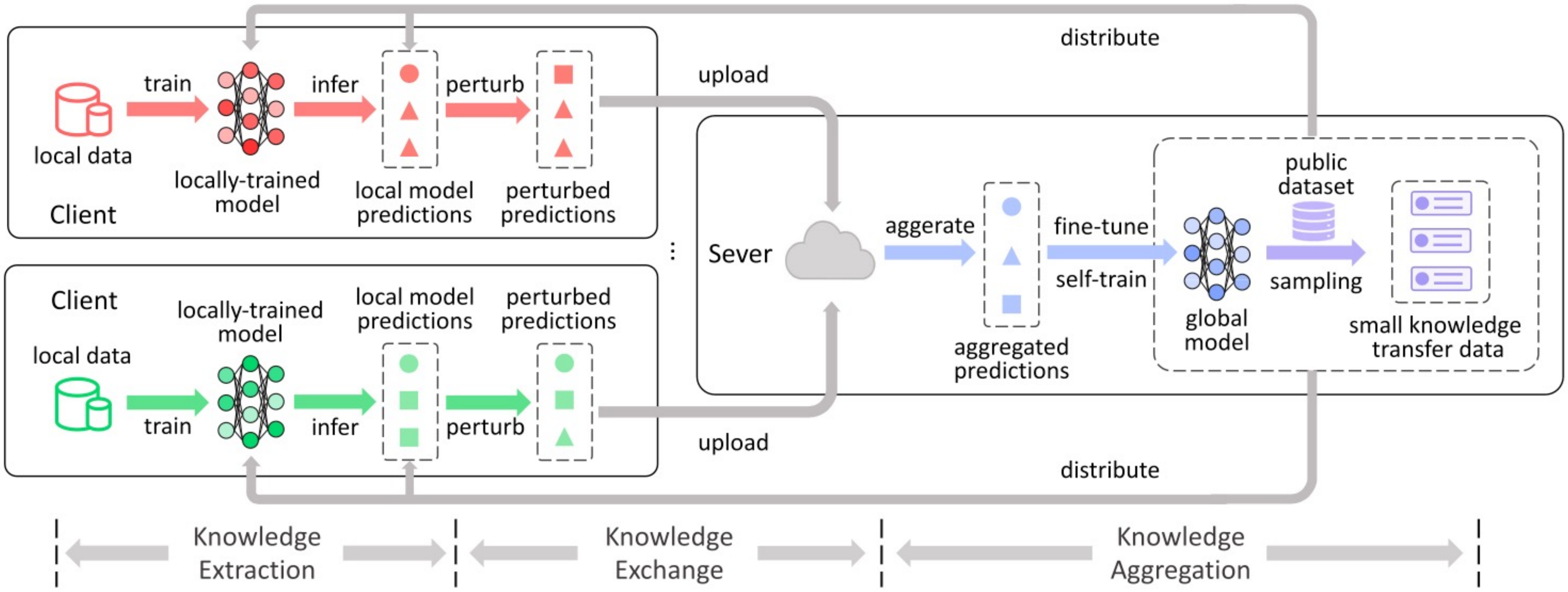
- Verify the effectiveness of the mechanisms of PrivateKT

The three mechanisms in PrivateKT, i.e., knowledge buffer, importance sampling, and self-training, can significantly improve the accuracy of federated learning



Conclusion

- Propose a differential private knowledge transfer framework to guarantee the privacy security of federated learning



Tao Qi

Tsinghua University

taoqi.qt@gmail.com