Federated Learning with Partial Model Personalization

ICML 2022

Krishna Pillutla

Kshitiz Malik

Abdelrahman Mohamed

Mike Rabbat

Maziar Sanjabi

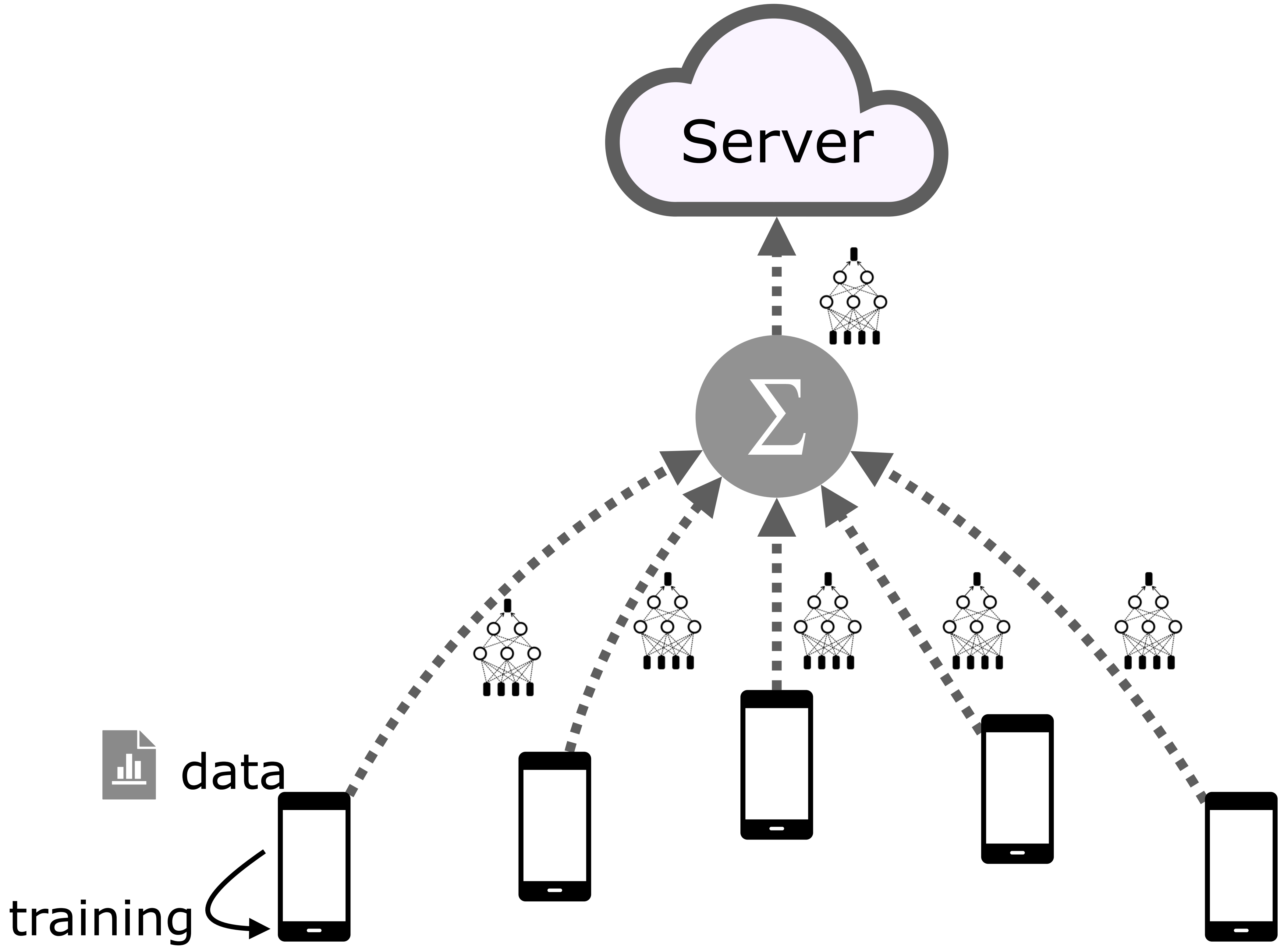
Lin Xiao

UNIVERSITY of WASHINGTON

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Meta AI

Personalized federated learning



Model on client i = (u, v_i)

Objective: $$\min_{u, v_1, \cdots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)$$

u: shared parameters

v_i: personal parameters

Our contributions

1. Theory: Analysis of 2 popular optimization algos

Objective: $$\min_{u, v_1, \cdots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)$$

u: shared parameters

v_i: personal parameters

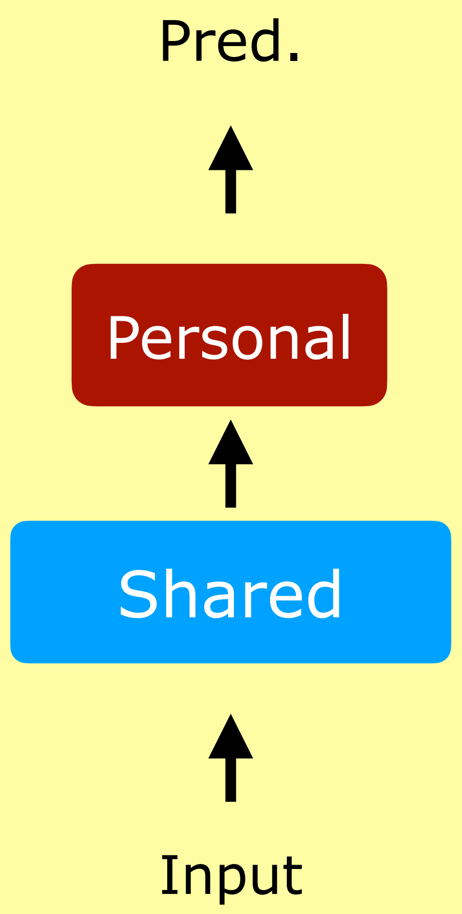
2. Extensive experiments: text, vision, and speech settings

Code:

SCAN ME

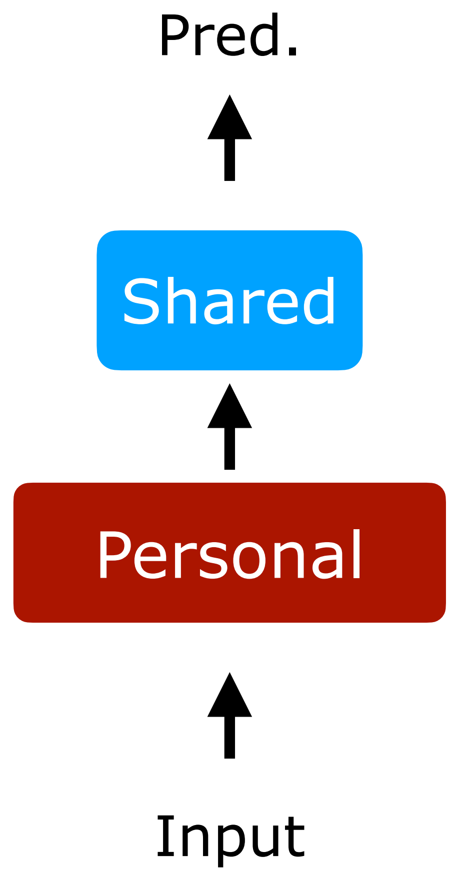
Personalization architectures

Personalized
output layer



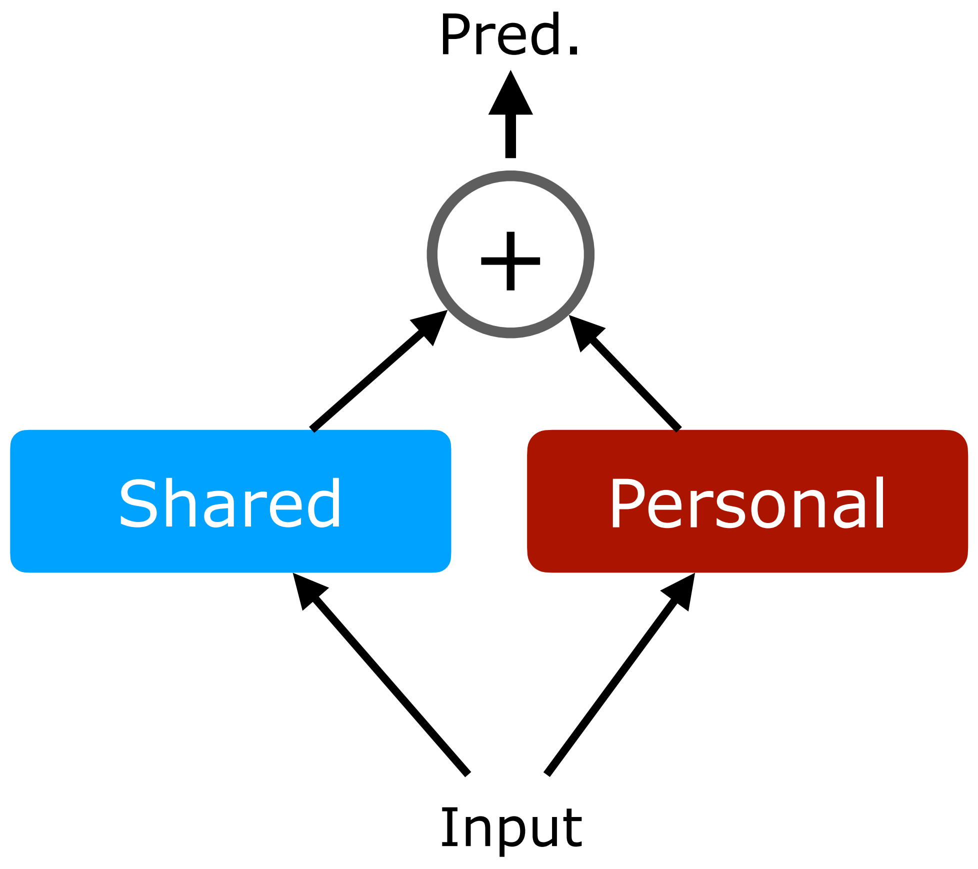
Arivazhagan et al. (2019)
Collins et al. ICML (2021)

Personalized
input layer



Liang et al. (2019)

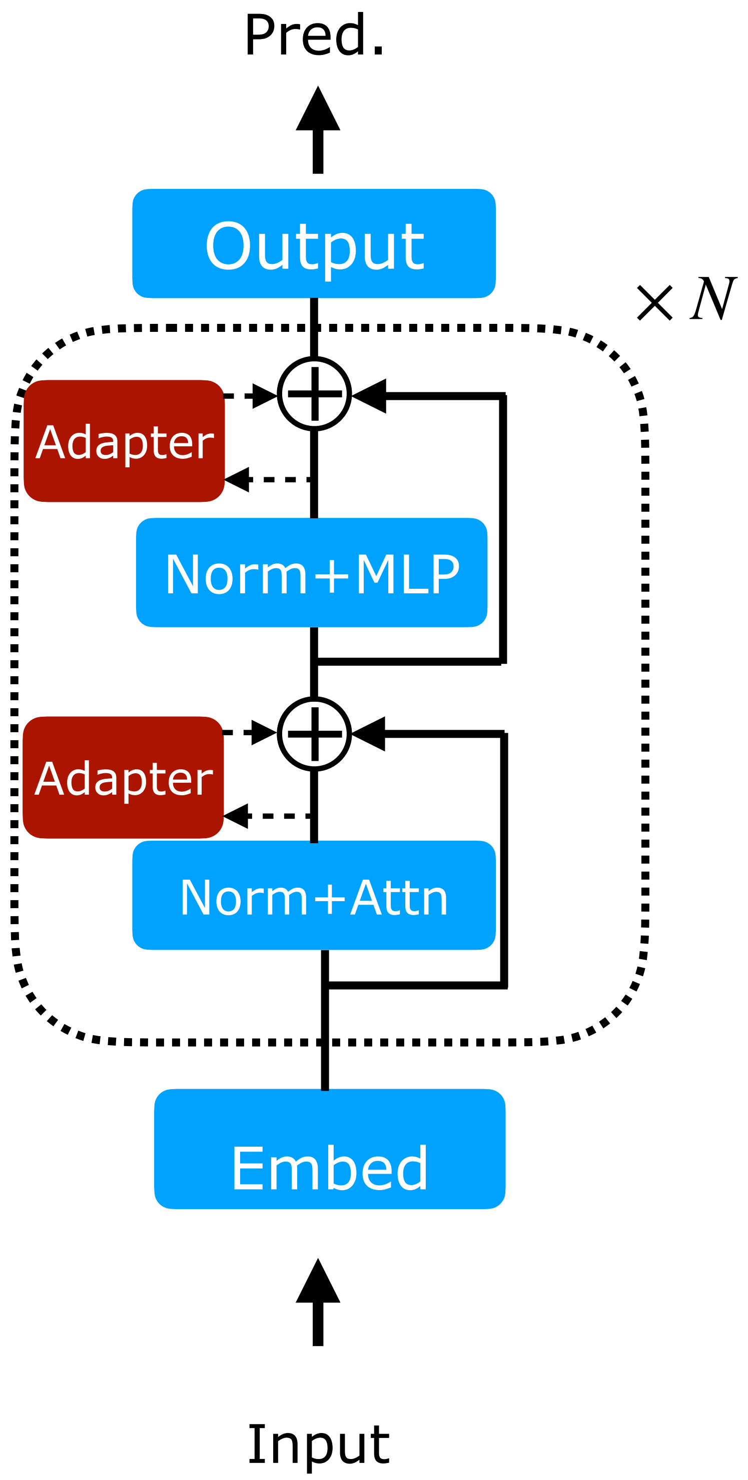
Combined
predictions



$$F_i(u, v_i) = \mathbb{E}_{(X,Y)\sim p_i}\left(\phi_g(X\,;\,u) + \phi_l(X\,;\,v_i) - Y\right)^2$$

Agarwal et al. (2020)

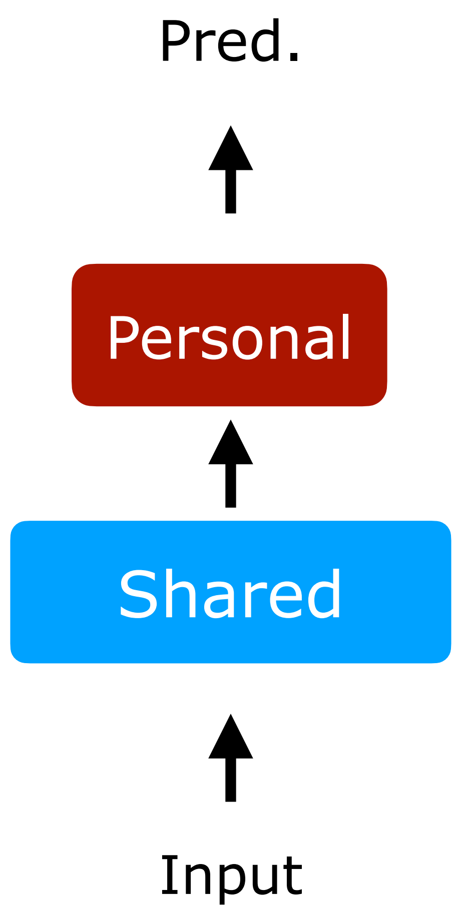
Personalized
adapters



Multi-task learning: Caruana. Mach. Learn (1997), Baxter. JAIR (2000), Evgeniou & Pontil. KDD (2004), Collobert & Weston. ICML (2005), Argyriou et al. Mach. Learn (2008), …

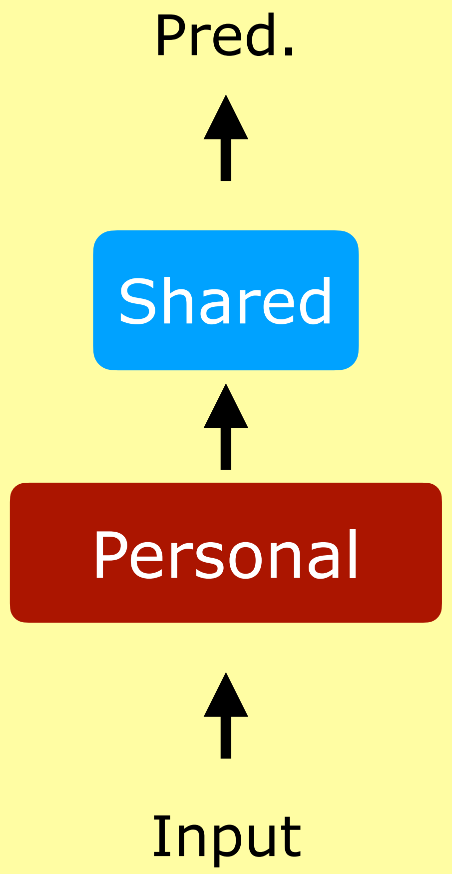
Personalization architectures

Personalized output layer



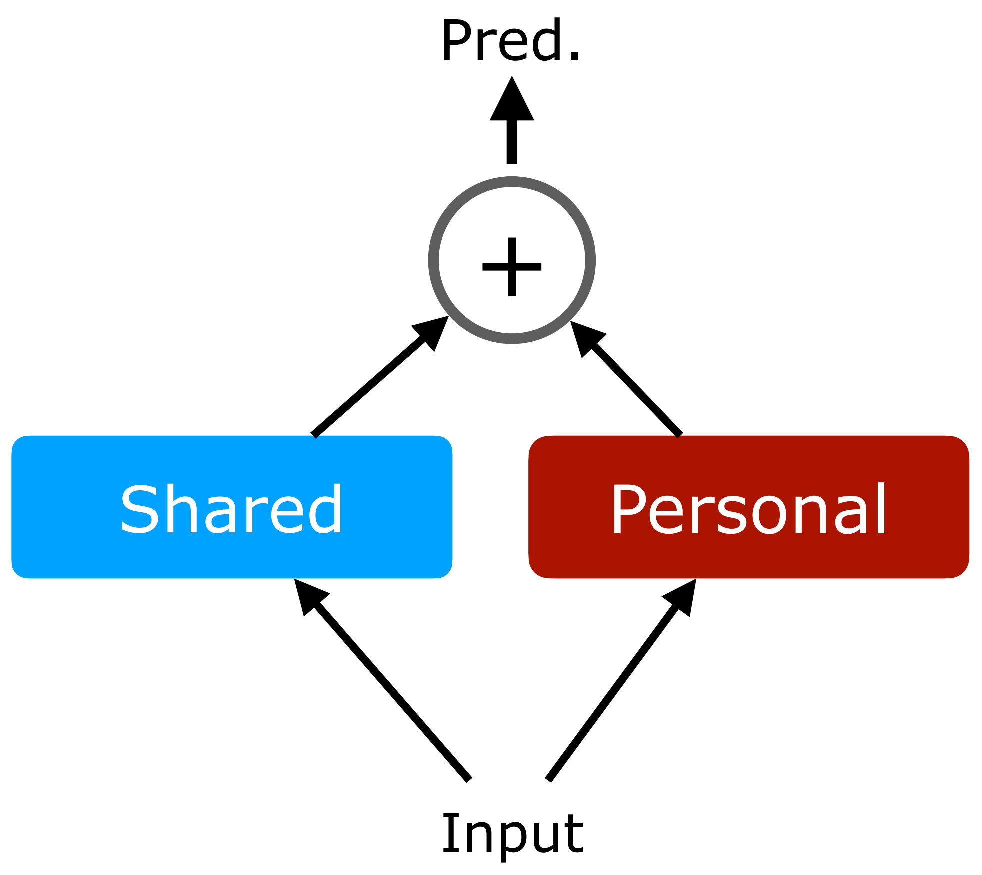
Arivazhagan et al. (2019)
Collins et al. ICML (2021)

Personalized input layer



Liang et al. (2019)

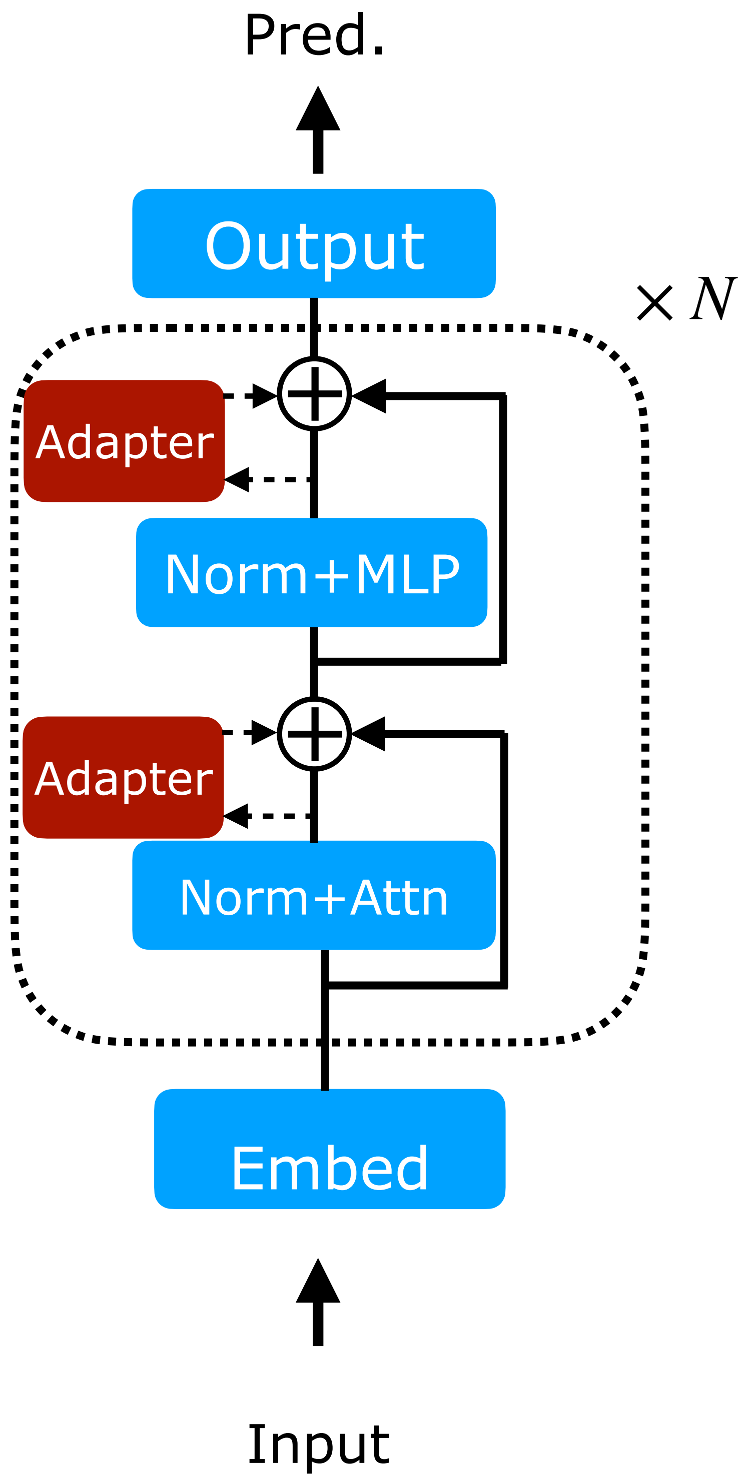
Combined predictions



$$F_i(u, v_i) = \mathbb{E}_{(X,Y)\sim p_i}\Big(\phi_g(X\,;\,u) + \phi_l(X\,;\,v_i) - Y\Big)^2$$

Agarwal et al. (2020)

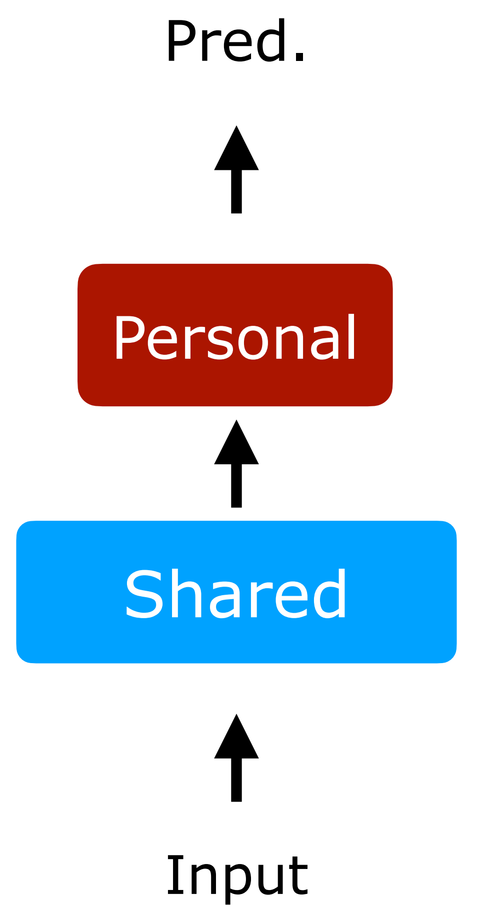
Personalized adapters



Multi-task learning: Caruana. Mach. Learn (1997), Baxter. JAIR (2000), Evgeniou & Pontil. KDD (2004), Collobert & Weston. ICML (2005), Argyriou et al. Mach. Learn (2008), …

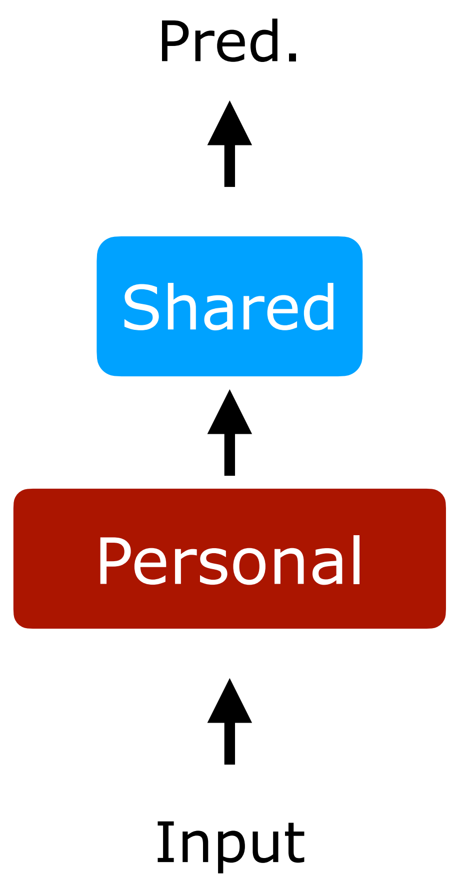
Personalization architectures

Personalized output layer



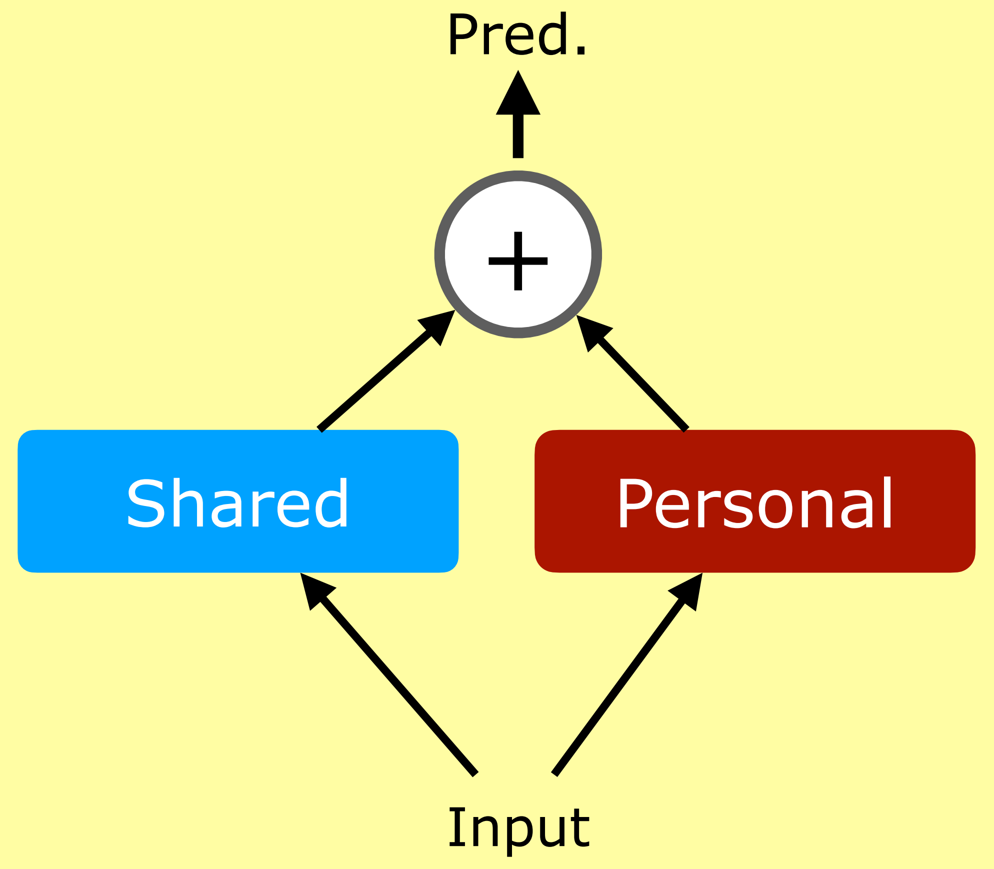
Arivazhagan et al. (2019)
Collins et al. ICML (2021)

Personalized input layer



Liang et al. (2019)

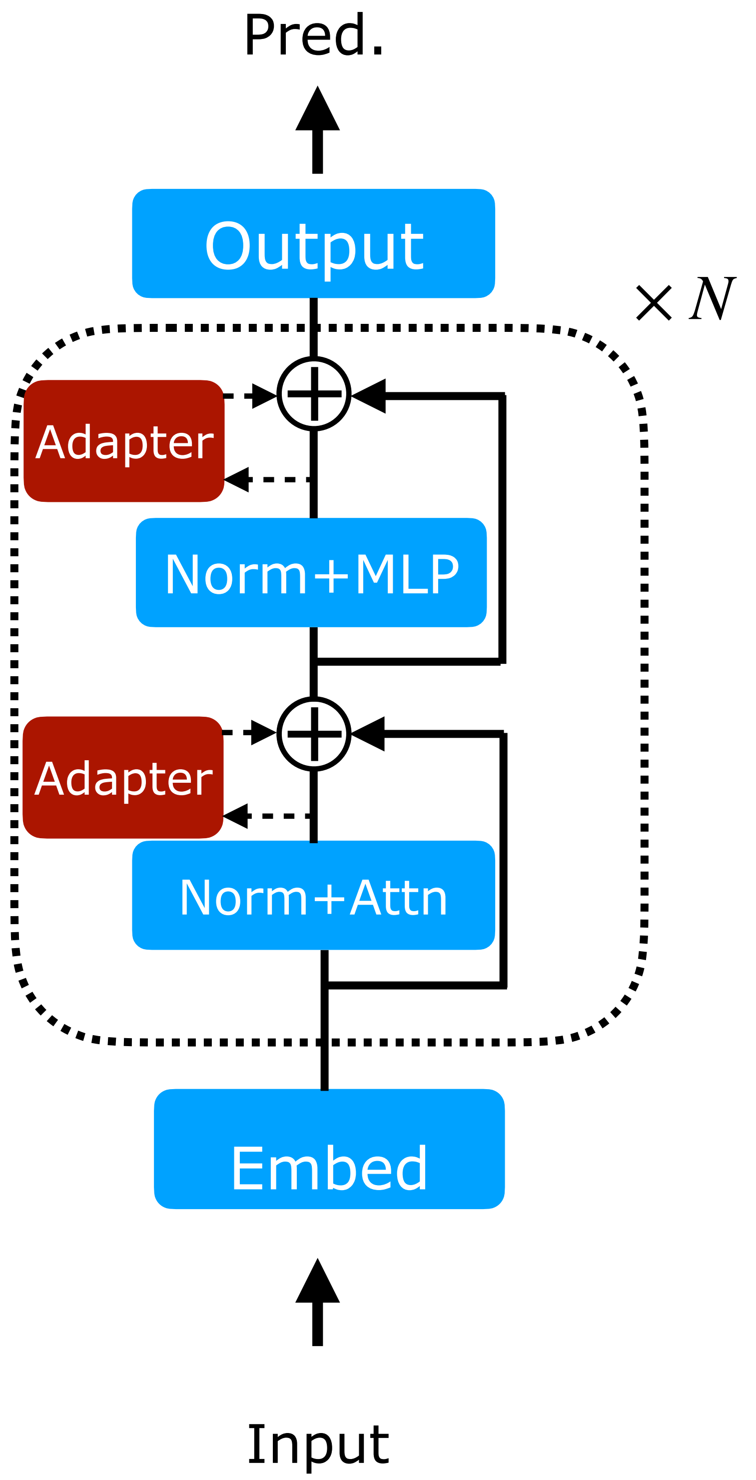
Combined predictions



$$F_i(u, v_i) = \mathbb{E}_{(X,Y)\sim p_i}\left(\phi_g(X\,;\,u) + \phi_l(X\,;\,v_i) - Y\right)^2$$

Agarwal et al. (2020)

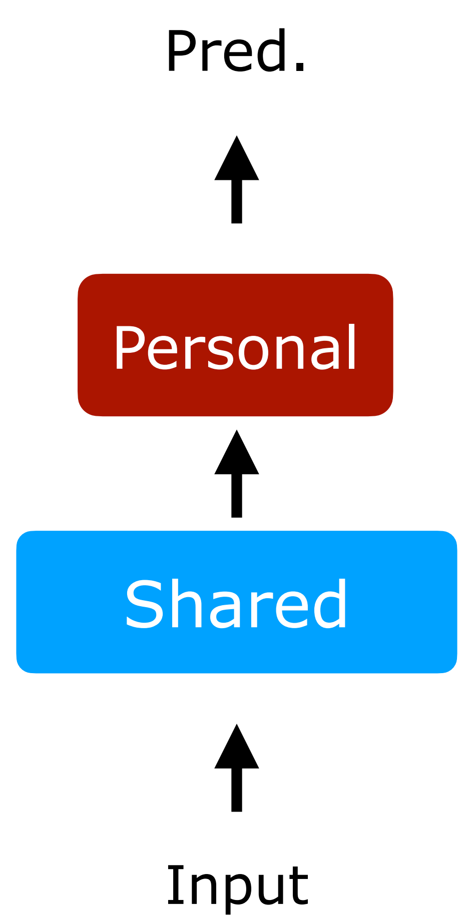
Personalized adapters



Multi-task learning: Caruana. Mach. Learn (1997), Baxter. JAIR (2000), Evgeniou & Pontil. KDD (2004), Collobert & Weston. ICML (2005), Argyriou et al. Mach. Learn (2008), …

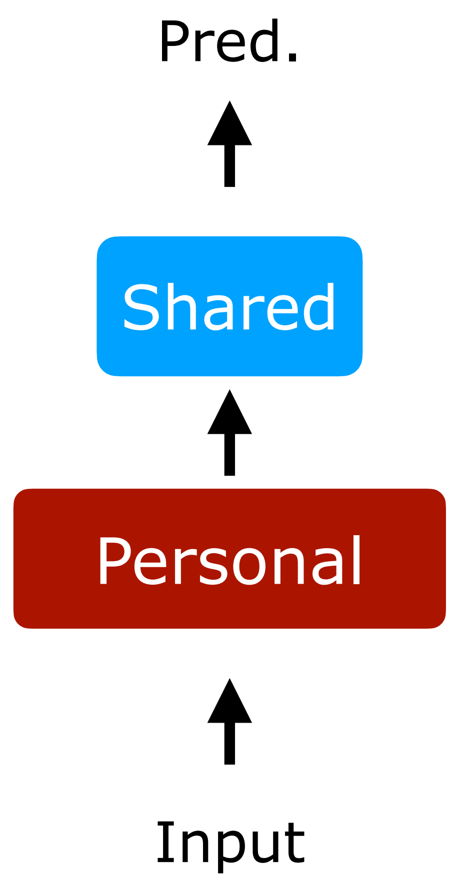
Personalization architectures

Personalized output layer



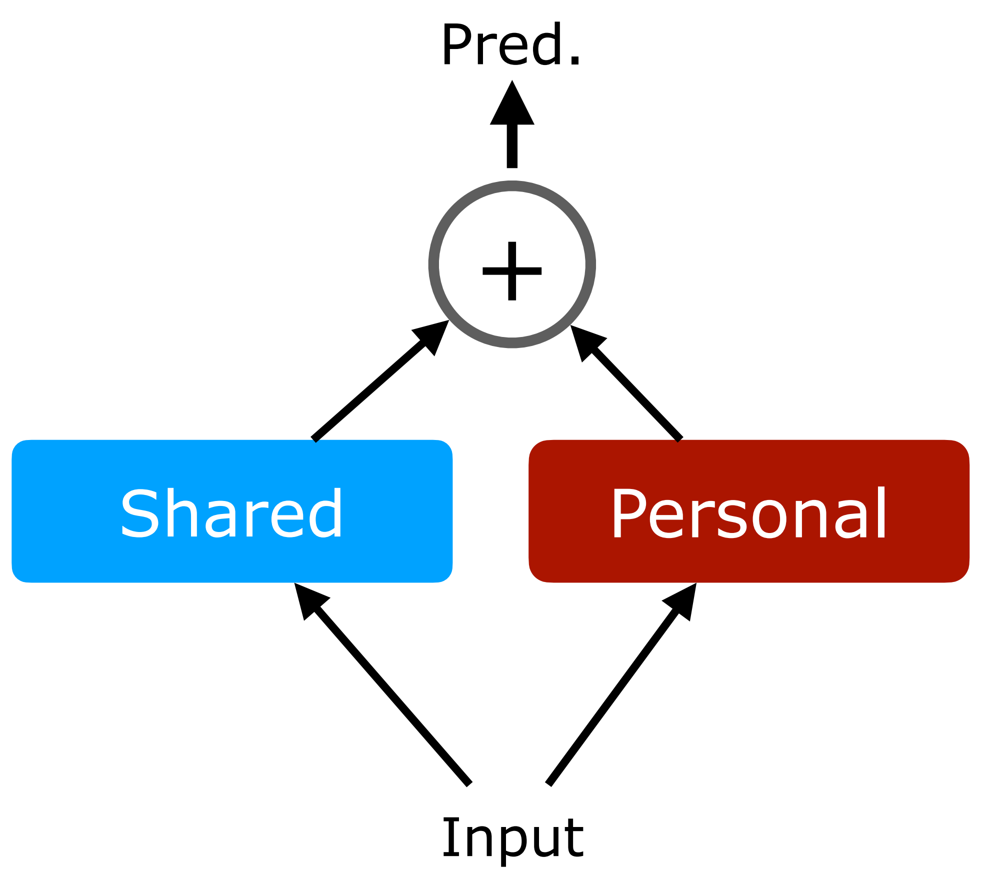
Arivazhagan et al. (2019)
Collins et al. ICML (2021)

Personalized input layer



Liang et al. (2019)

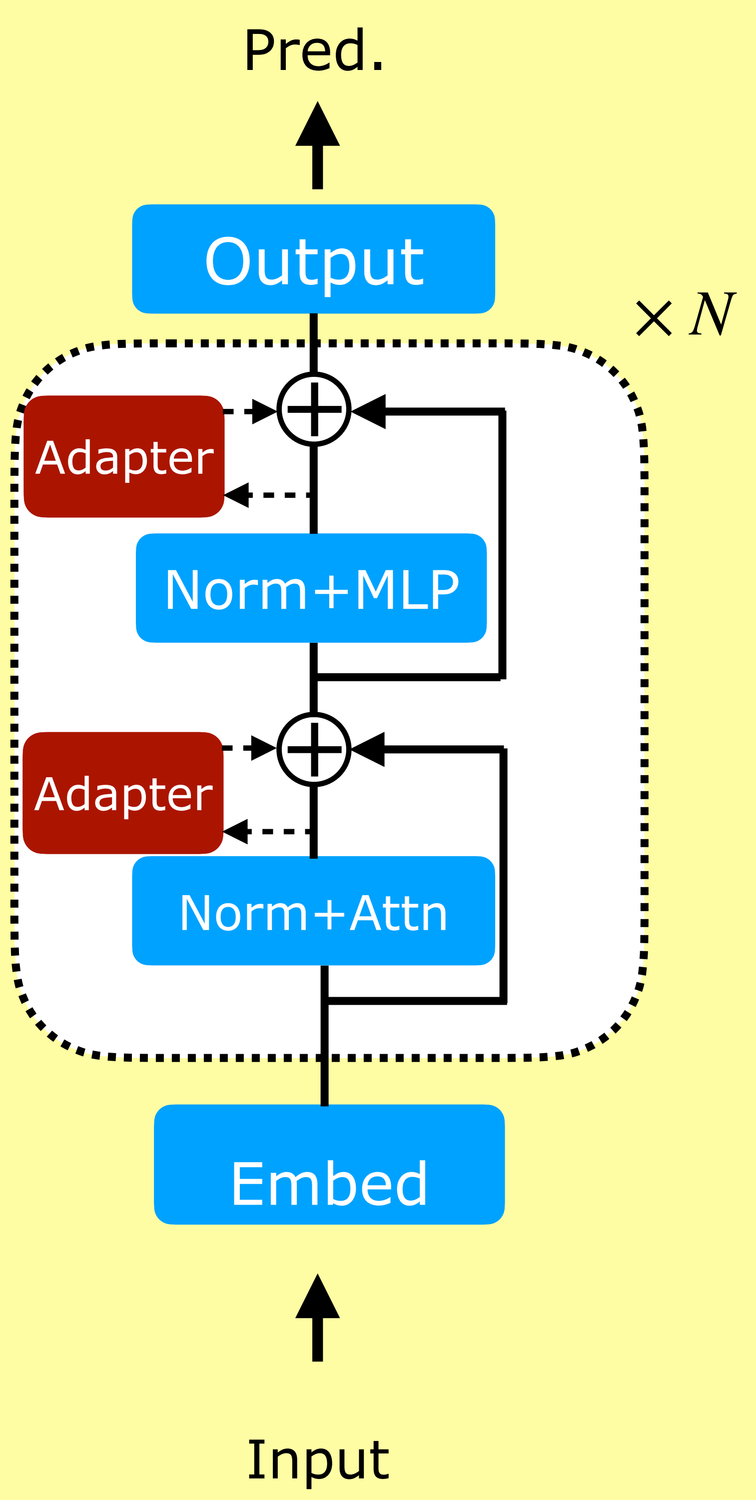
Combined predictions



$$F_i(u, v_i) = \mathbb{E}_{(X,Y)\sim p_i}\left(\phi_g(X\,;\,u) + \phi_l(X\,;\,v_i) - Y\right)^2$$

Agarwal et al. (2020)

Personalized adapters



Multi-task learning: Caruana. Mach. Learn (1997), Baxter. JAIR (2000), Evgeniou & Pontil. KDD (2004), Collobert & Weston. ICML (2005), Argyriou et al. Mach. Learn (2008), …

Optimization

- Server samples m clients and broadcast global model u
- **Local updates** on client i: $(u_i^+, v_i^+) = \text{LocalUpdate}_i(u, v_i)$
- Aggregate updates to global part of the model: $u^+ = \frac{1}{m}\sum_i u_i^+$

Alternating update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i^+)$$

Collins et al. ICML (2021)
Singhal et al. NeurIPS (2021)

Simultaneous update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i)$$

Liang et al. (2019)
Arivazhagan et al. (2019)

Optimization

- Server samples m clients and broadcast global model u
- **Local updates** on client i:
  $(u_i^+, v_i^+) = \text{LocalUpdate}_i(u, v_i)$
- Aggregate updates to global part of the model:
  $$u^+ = \frac{1}{m}\sum_i u_i^+$$

Alternating update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i^+)$$

Collins et al. ICML (2021)
Singhal et al. NeurIPS (2021)

Simultaneous update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i)$$

Liang et al. (2019)
Arivazhagan et al. (2019)

Optimization

- Server samples m clients and broadcast global model u
- **Local updates** on client i: $(u_i^+, v_i^+) = \text{LocalUpdate}_i(u, v_i)$
- Aggregate updates to global part of the model: $u^+ = \frac{1}{m}\sum_i u_i^+$

Alternating update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i^+)$$

Collins et al. ICML (2021)
Singhal et al. NeurIPS (2021)

Simultaneous update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i)$$

Liang et al. (2019)
Arivazhagan et al. (2019)

Contribution 1: Theory

Theorem [*P.*, Malik, Mohamed, Rabbat, Sanjabi, Xiao]

For smooth, nonconvex functions and client sampling, we have the rates:

Alternating update: $\dfrac{\sigma_1^2}{\sqrt{t}}$

Simultaneous update: $\dfrac{\sigma_2^2}{\sqrt{t}}$

where $\sigma_1^2 < \sigma_2^2$ under typical scenarios

Alternating update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i^+)$$

Simultaneous update

$$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$

$$u_i^+ = u - \gamma \nabla_u F_i(u, v_i)$$

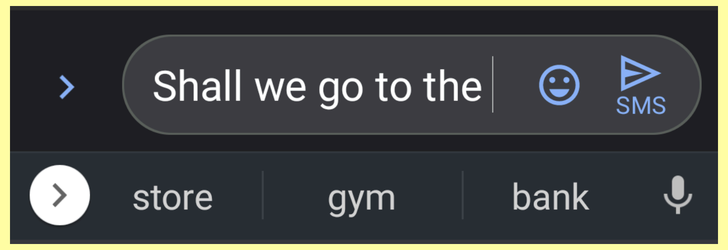
Experimentally, small but consistent trend of alternating ≻ simultaneous

Contribution 1: Theory

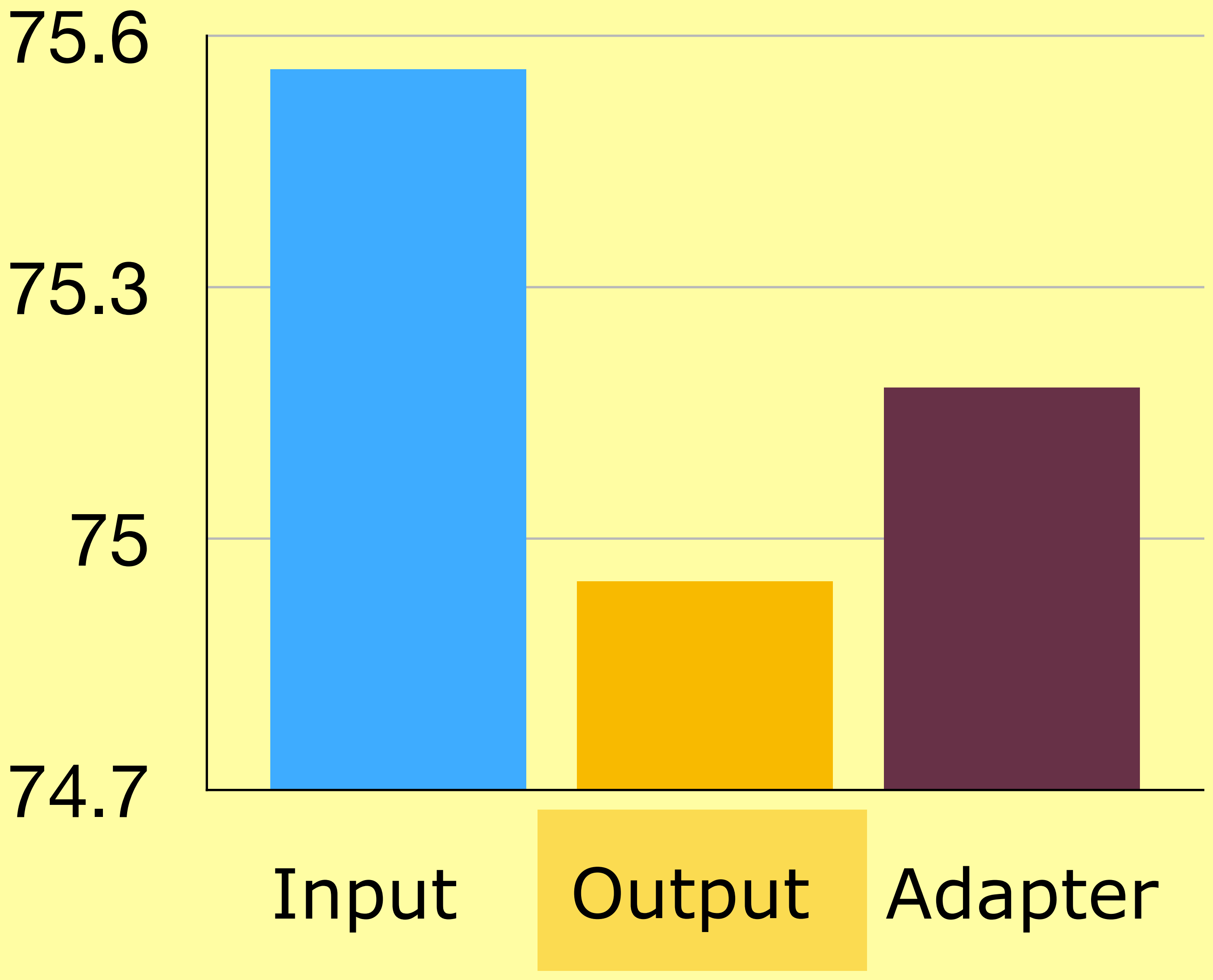
Key technical challenge: Dependent random variables in alternating update algorithm due to random sampling of clients

Methodology: Developed technique of virtual full participation

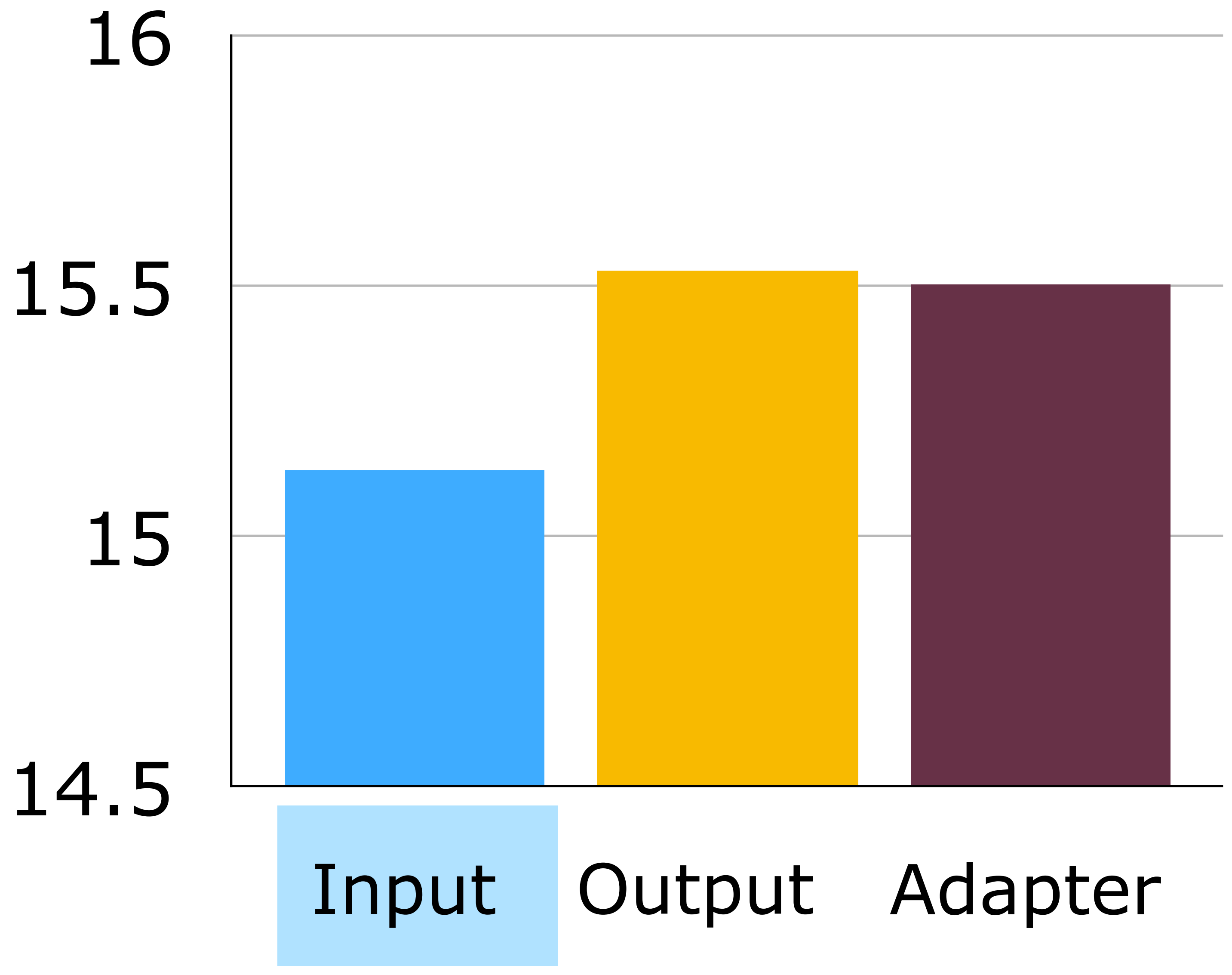
Contribution 2: Experiments



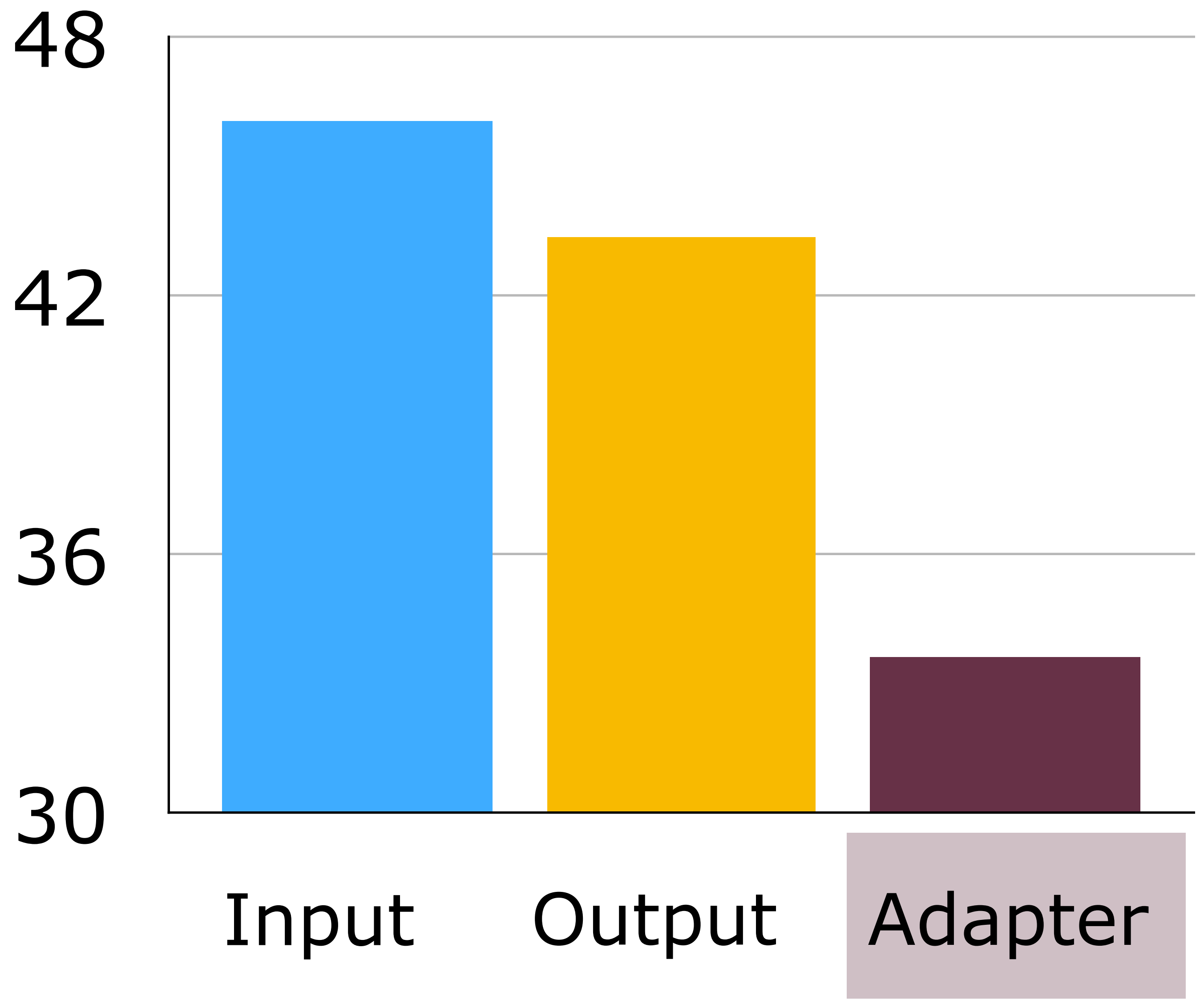
Next word prediction



Speech recognition

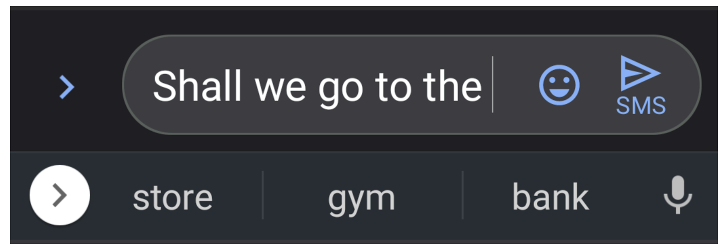


Landmark detection

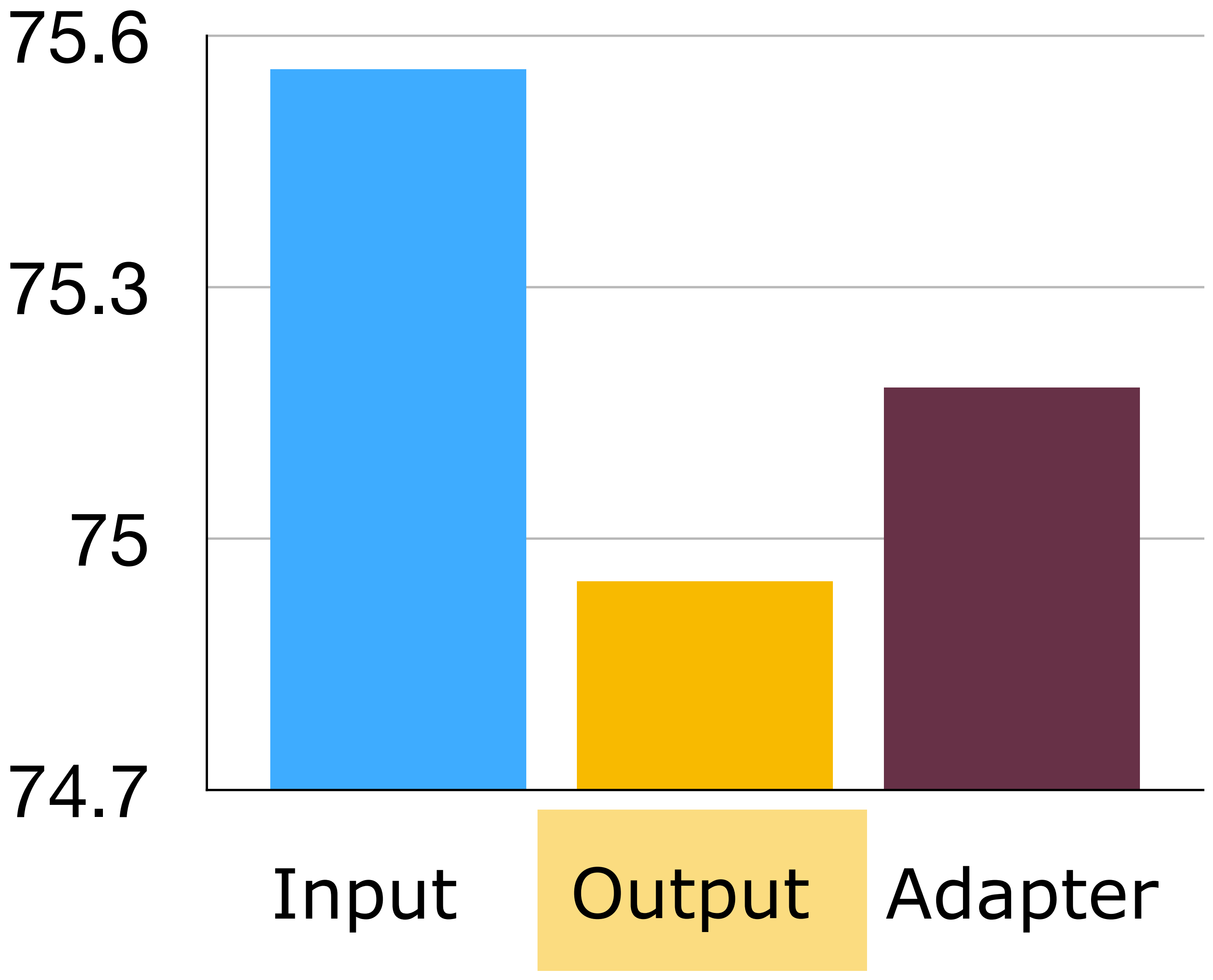


y-axis shows error: lower is better

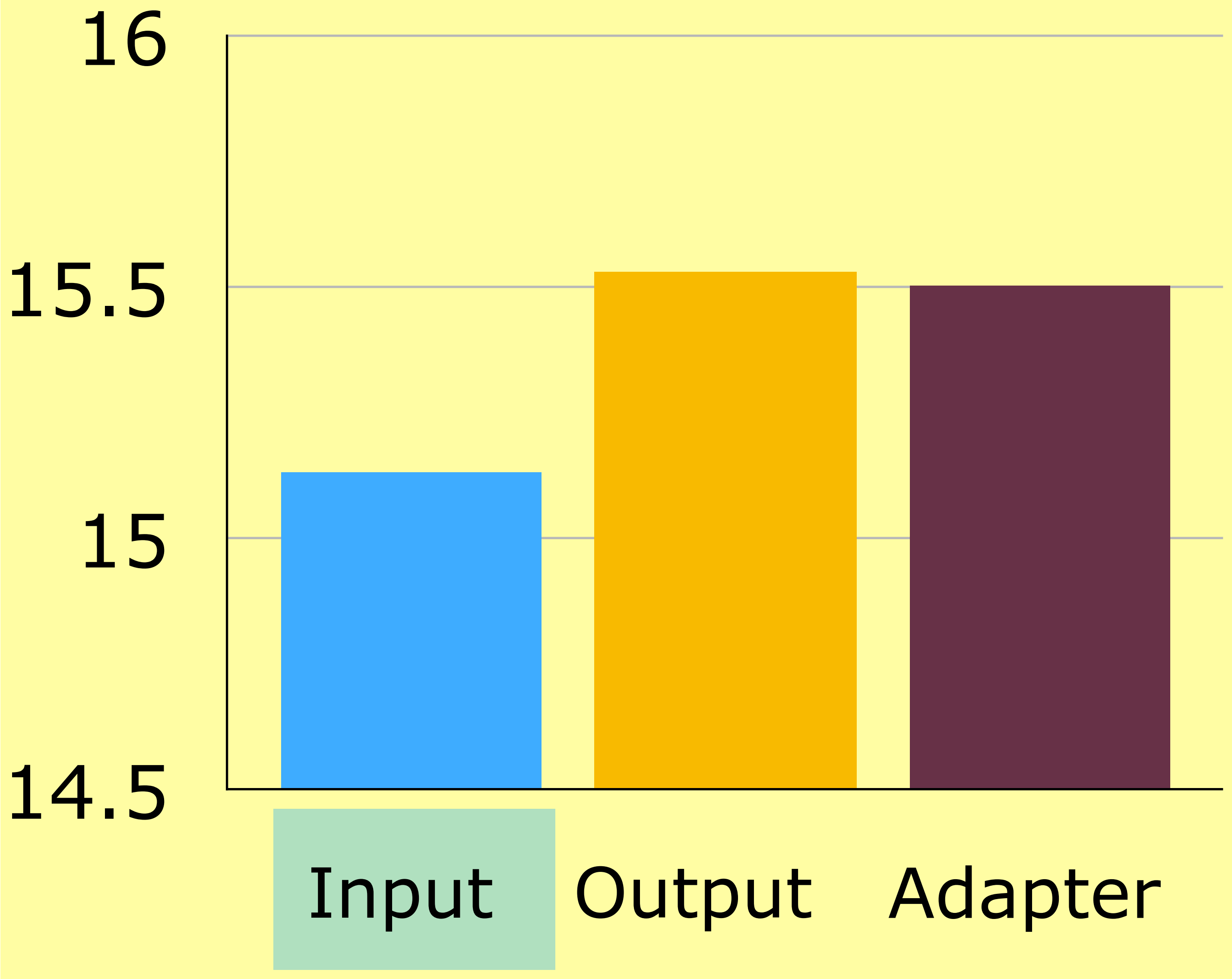
Contribution 2: Experiments



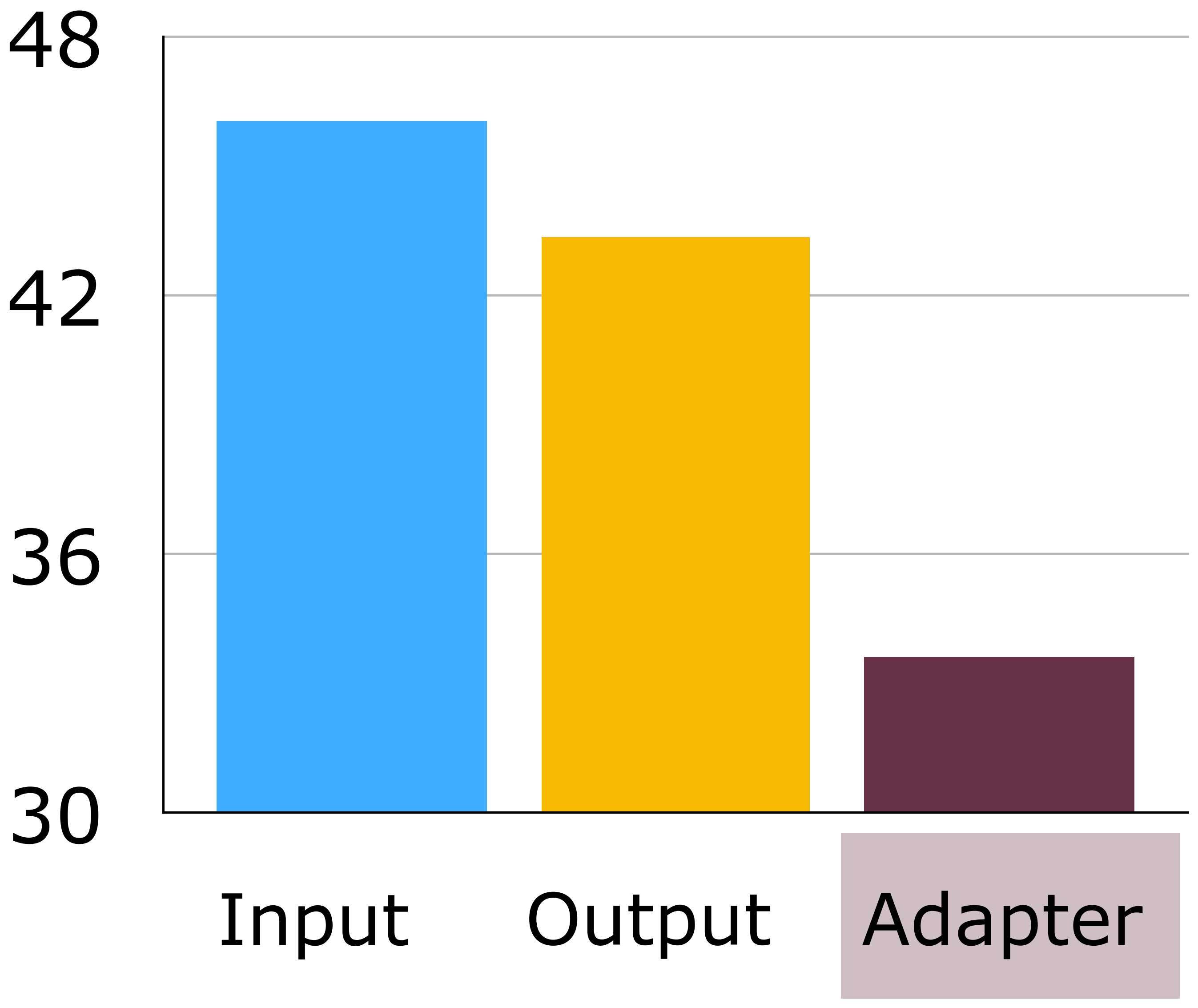
Next word prediction



Speech recognition

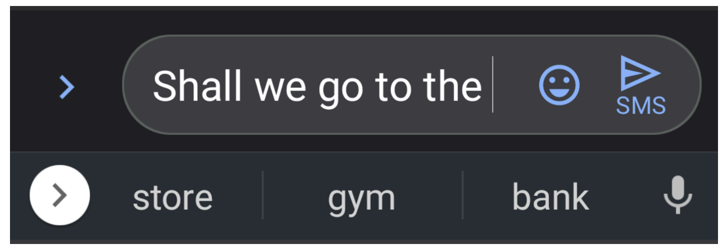


Landmark detection



y-axis shows error: lower is better

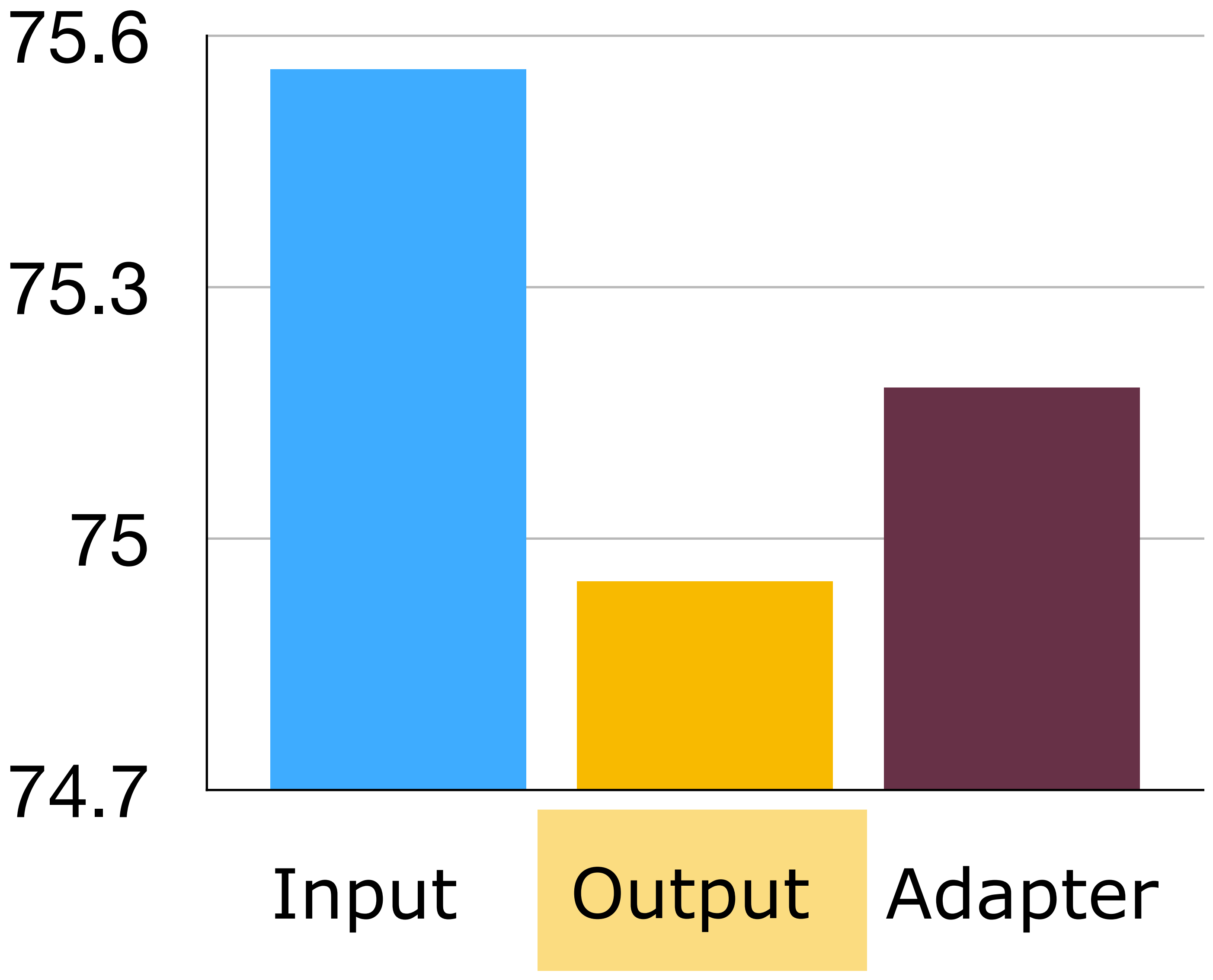
Contribution 2: Experiments

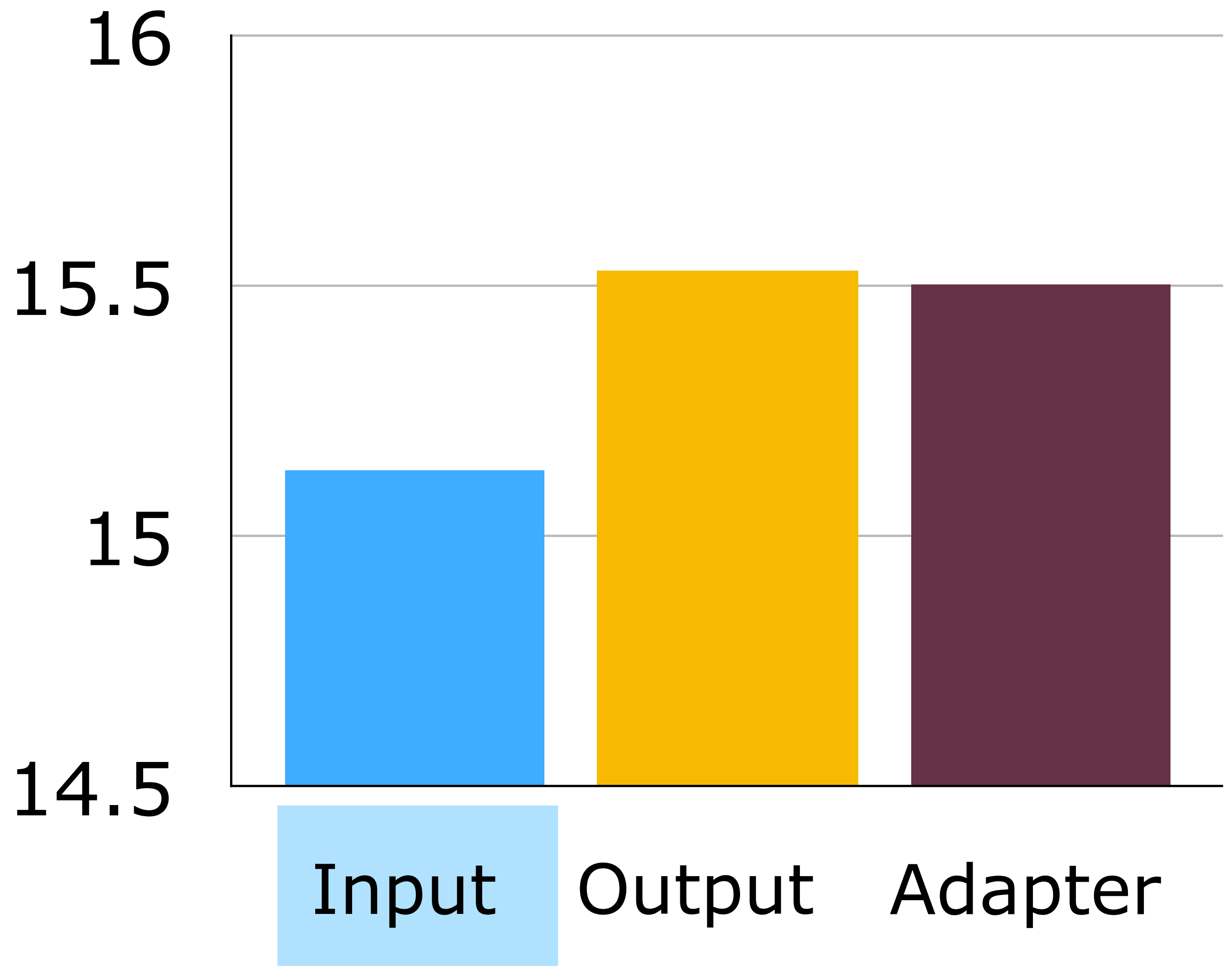


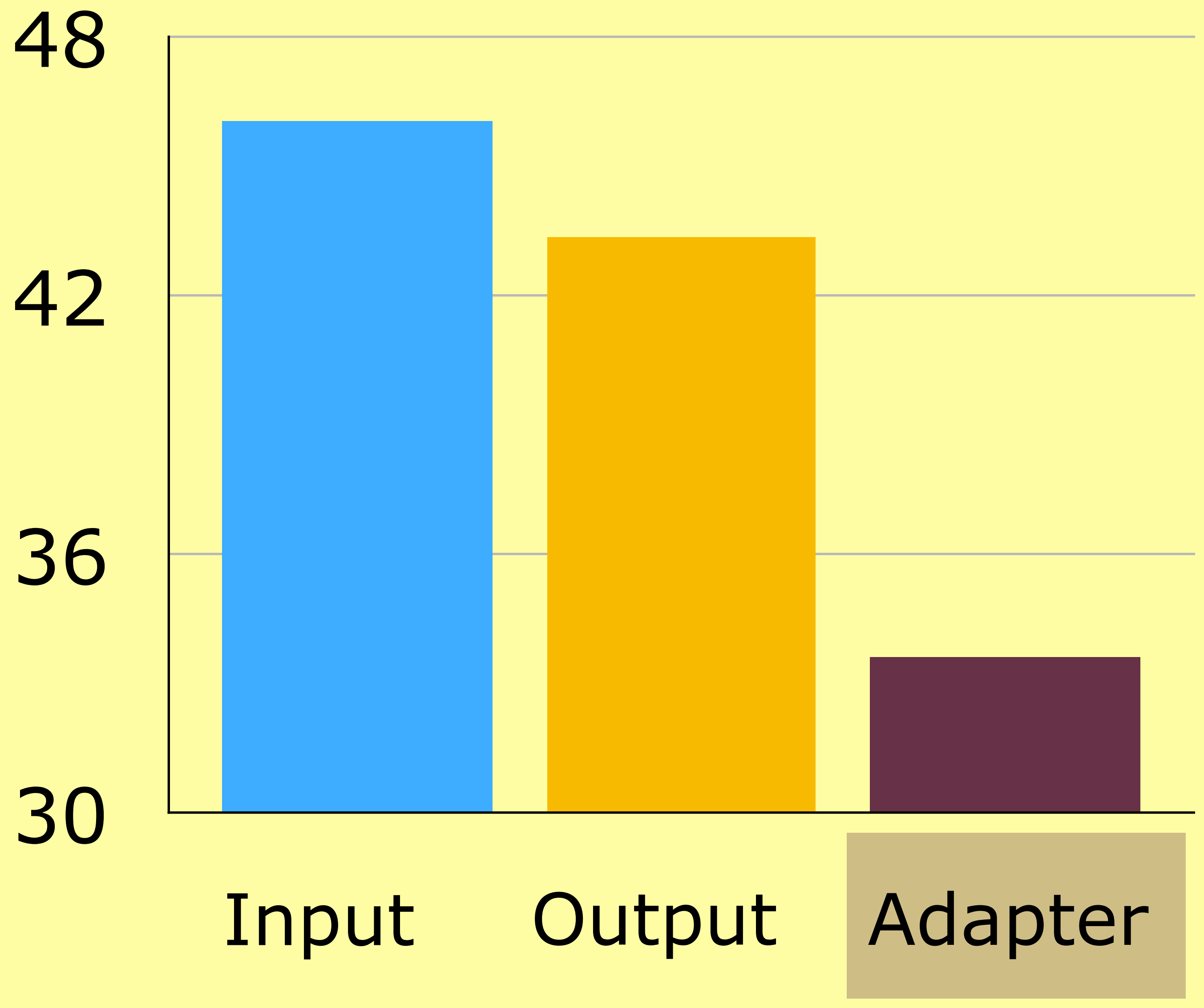
Next word prediction

Speech recognition

Landmark detection



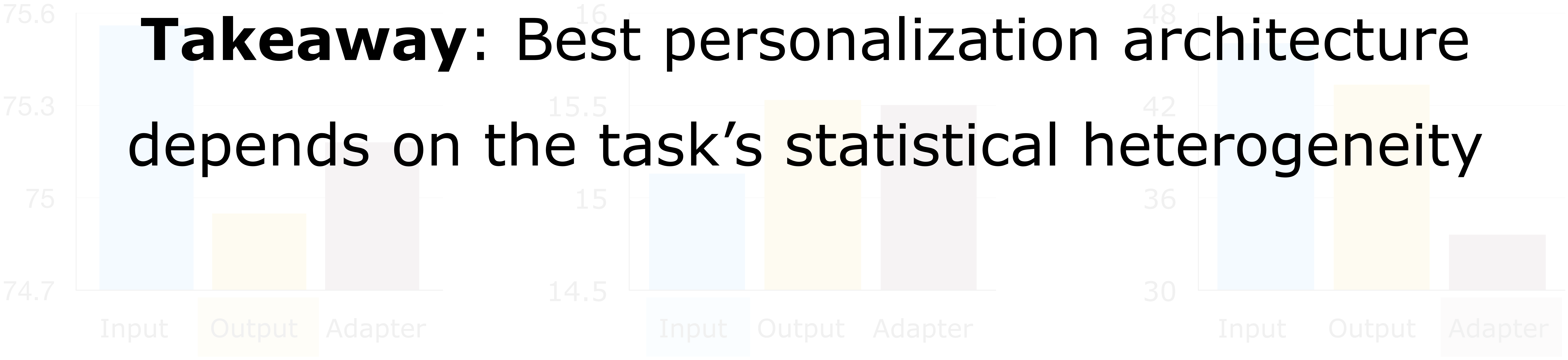




y-axis shows error: lower is better

Contribution 2: Experiments

Takeaway: Best personalization architecture depends on the task's statistical heterogeneity



y-axis shows error: lower is better

Federated Learning with Partial Model Personalization

Paper: arXiv:2204.03809

Code: https://github.com/krishnap25/FL_partial_personalization