DeepType: On-Device Deep Learning for Input Personalization Service with Minimal Privacy Concern

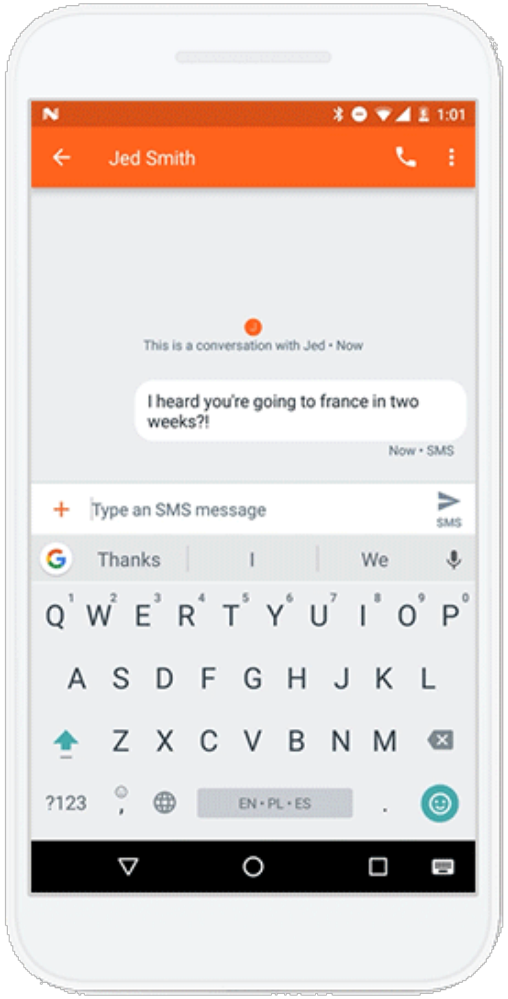
Mengwei Xu[1], Feng Qian[2], Qiaozhu Mei[3]
Kang Huang[4], Xuanzhe Liu[1], **Yun Ma[1]**

[1]Peking University, [2]University of Minnesota, [3]University of Michigan, [4]Kika Tech

Everyone types a lot everyday

- Per day on earth: 2M Reddit posts, 5M tweets, 100B instant messages, and 200B emails
- A large portion of them are done on mobile devices, which makes:

Input method application (IMA): a killer app
Next-word prediction: a killer feature for productivity



DL-powered next-word prediction

- Next-word prediction techniques has evolved to deep learning (DL)

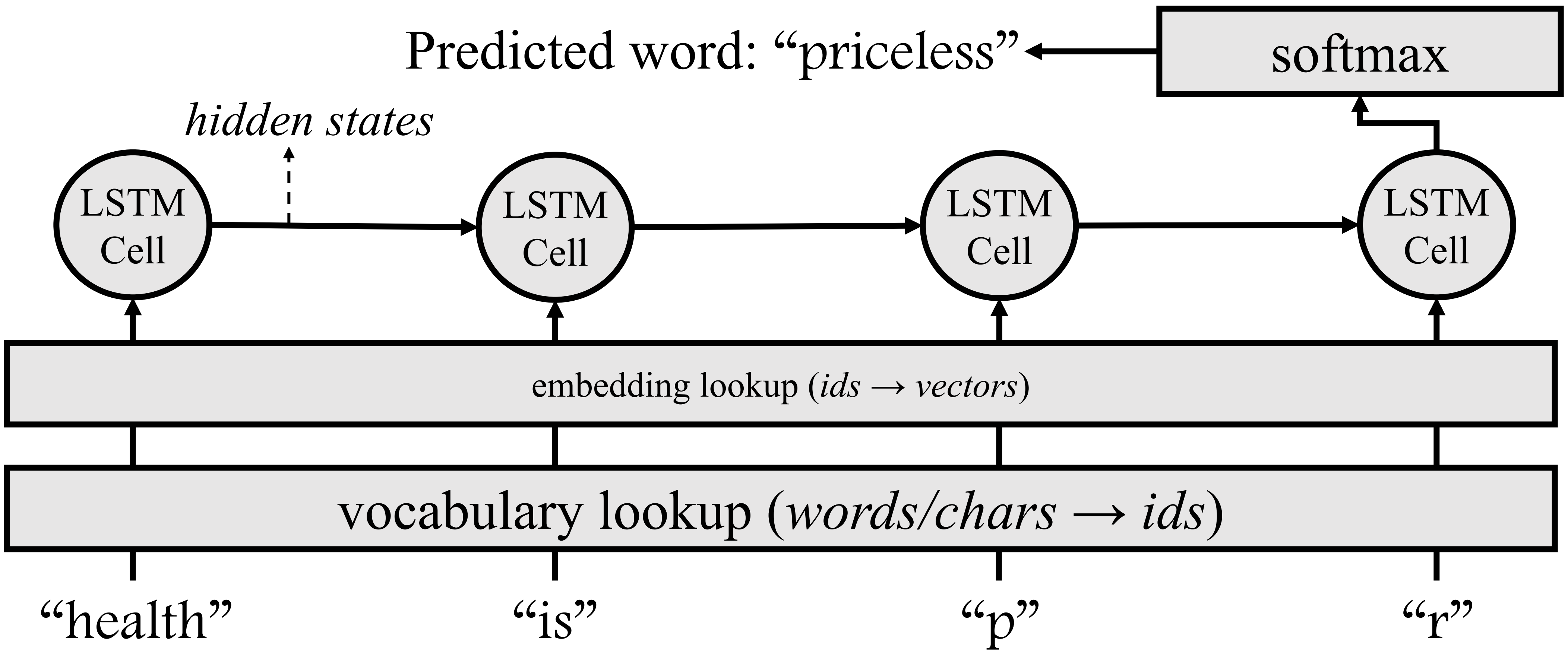
- dictionary lookup
  - cheap, inaccurate
- traditional ML algos
  - ngram
- deep learning
  - LSTM

More accurate

More expensive

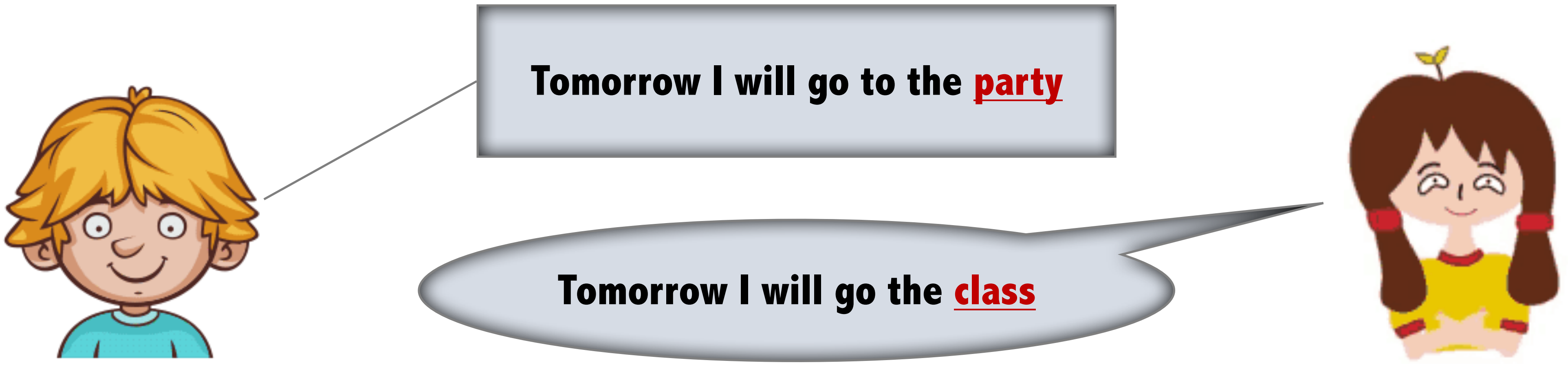
- *both training and prediction*

LSTM model for next-word prediction



Personalizing prediction models

- Can we further improve the accuracy of DL models?



- The models need to be ***personalized*** and adapt to diverse users
  - Training one model for one user using his/her own data

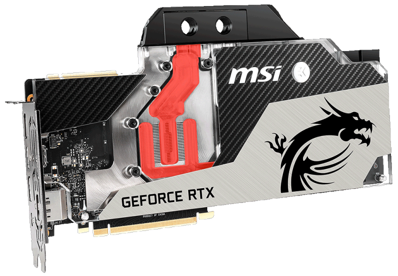
On-cloud personalization is not a good idea

privacy concern

password
accounts
emails
messages
credit card numbers

scalability issue

Personalizing 1M users takes 36,000 GPU-hrs. Too expensive!



GPUs are expensive

Can we personalize (train) the DL model on mobile devices?

Challenges of on-device personalization

- **Limited data volume**

  Is it enough to make model converge

- **Limited computational resources**

  Can we train model w/o compromising user experience

Challenges of on-device personalization

- **Limited data volume**

  Is it enough to make model converge

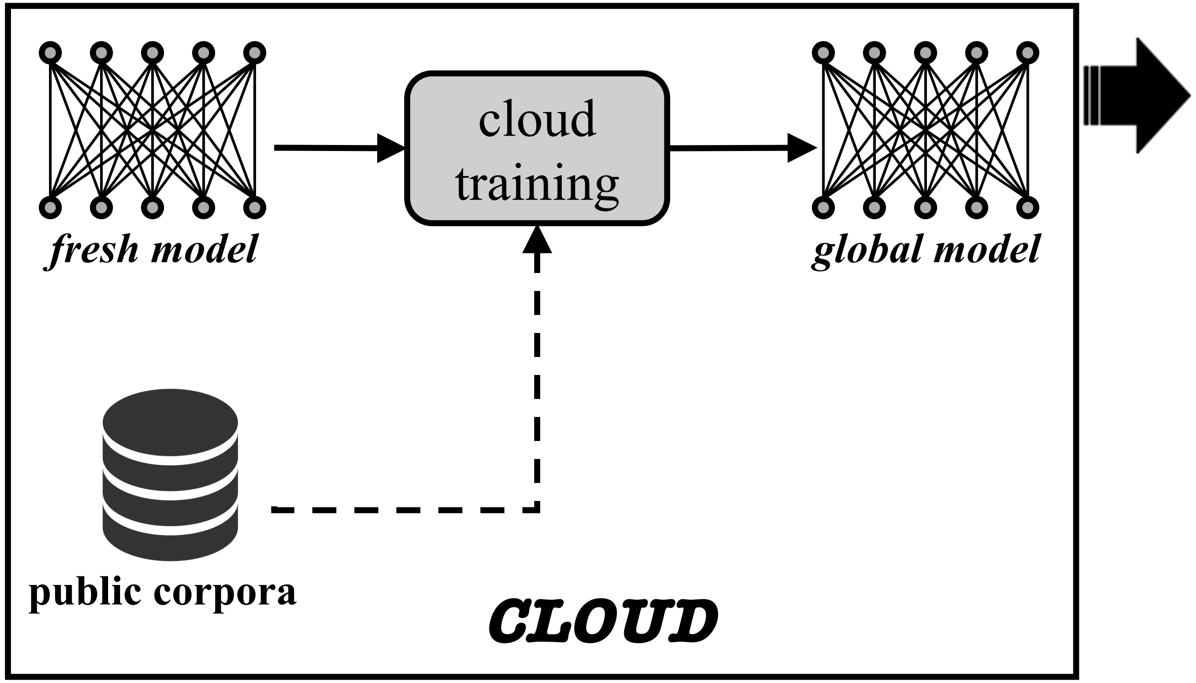
Key idea 1: use public corpora to pre-train a global model before on-device personalization

- **Limited computational resources**

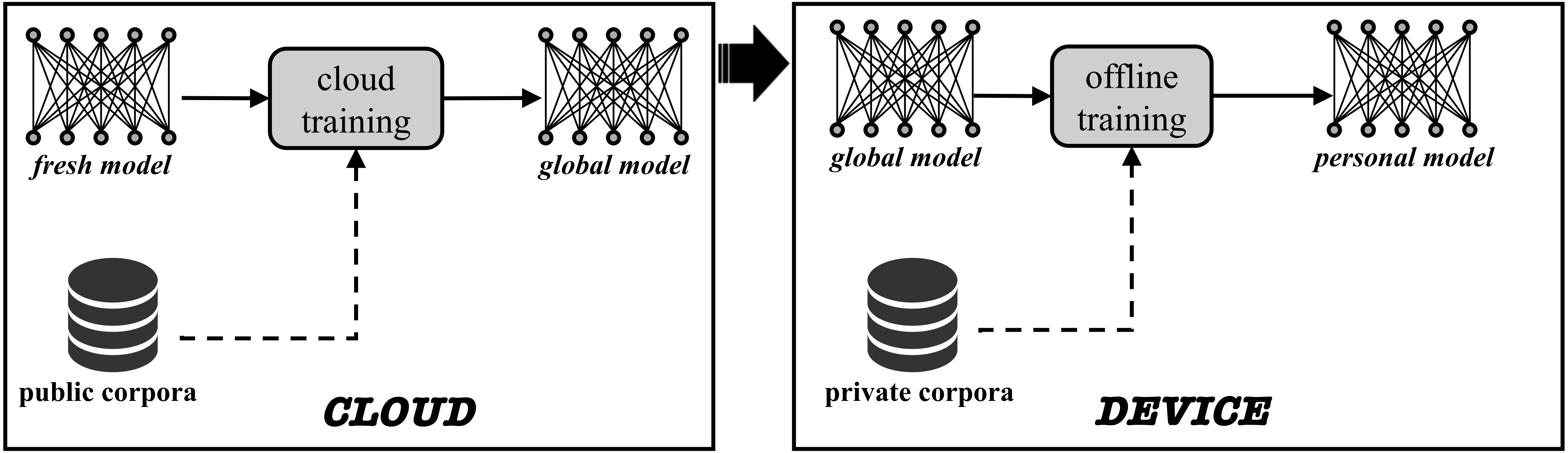
  Can we train model w/o compromising user experience

Key idea 2: compress, customize, and fine-tune the model

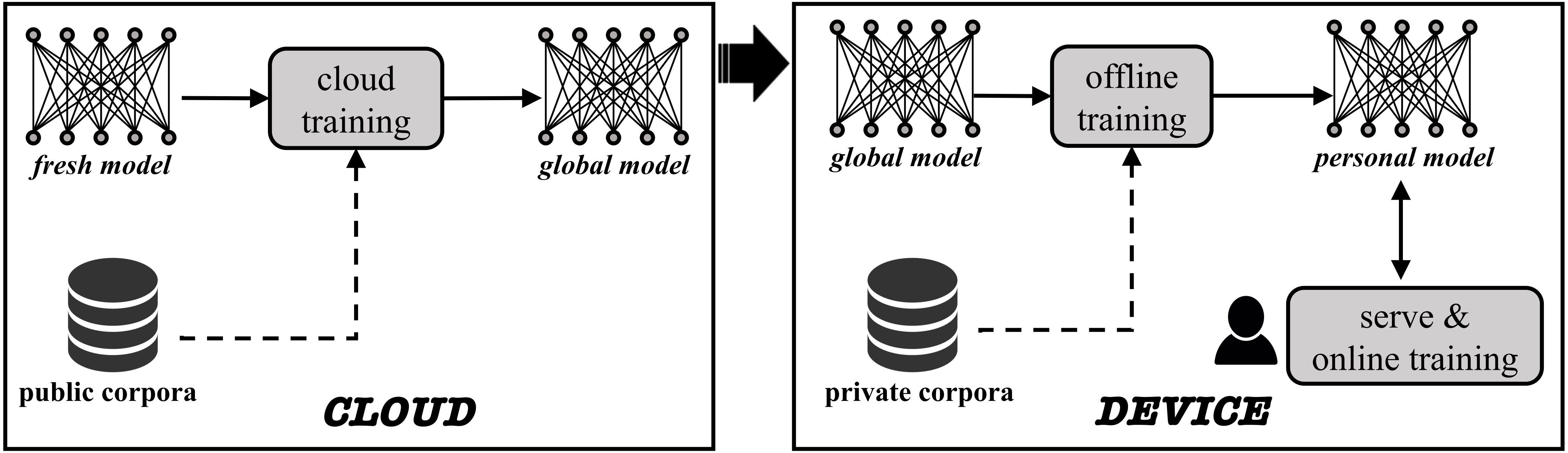
DeepType: on-device personalization



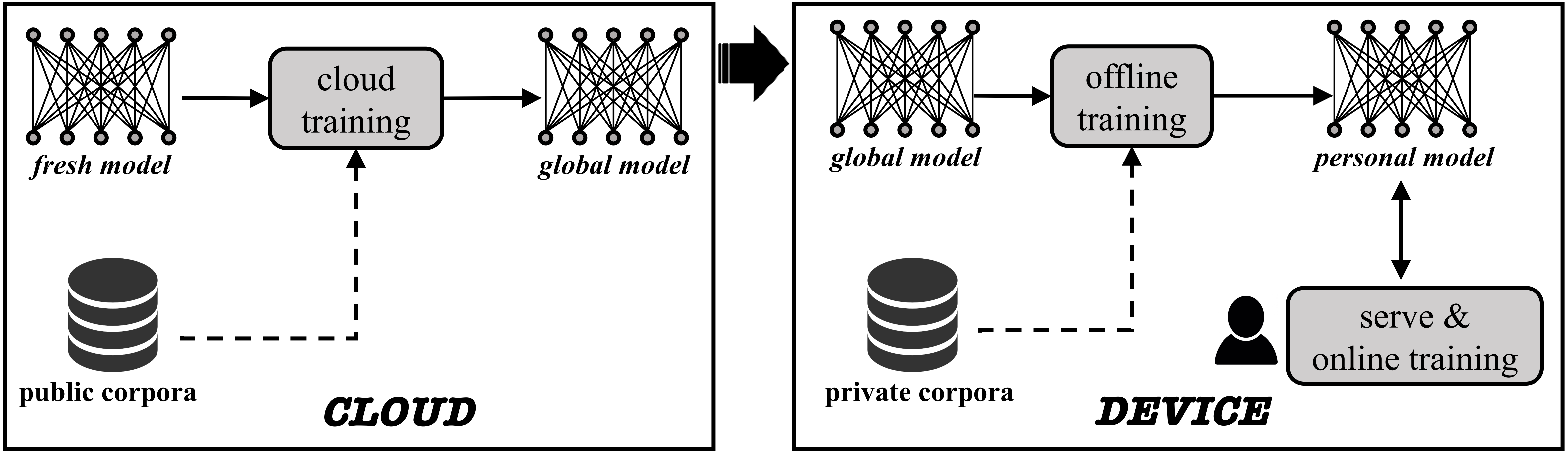
DeepType: on-device personalization



DeepType: on-device personalization



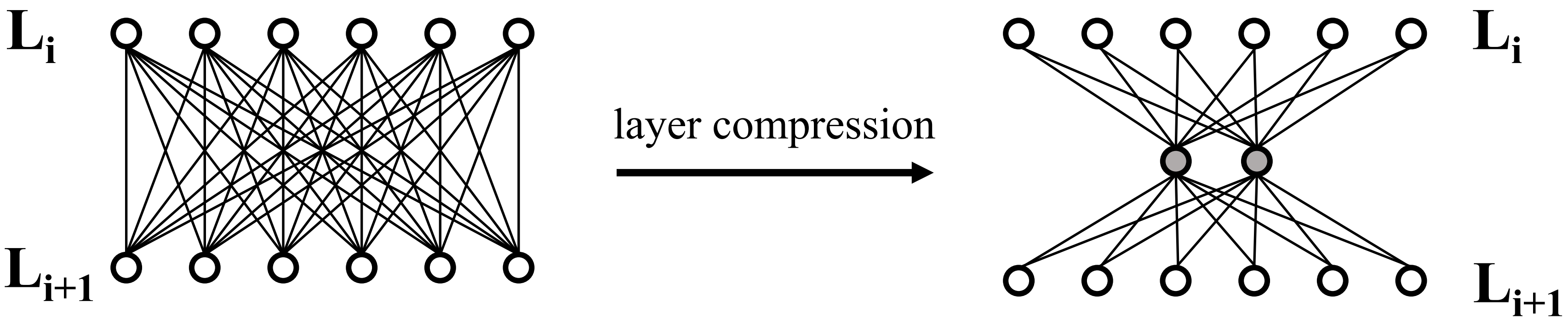
DeepType: on-device personalization



- Good privacy: input data never leaves mobile device
- Good flexibility: the model can be updated anytime with small cost

Reducing on-device computations

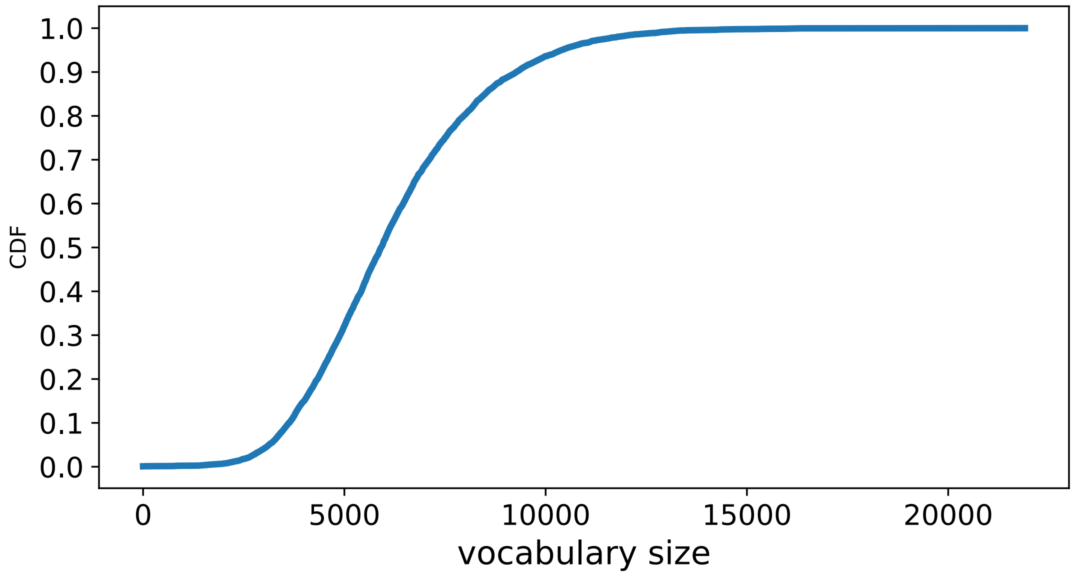
1. SVD-based model compression (on cloud)
2. Vocabulary compression
3. Fine-tune training
4. Reusing inference results



Reducing on-device computations

1. SVD-based model compression
2. Vocabulary compression (on device)
3. Fine-tune training
4. Reusing inference results

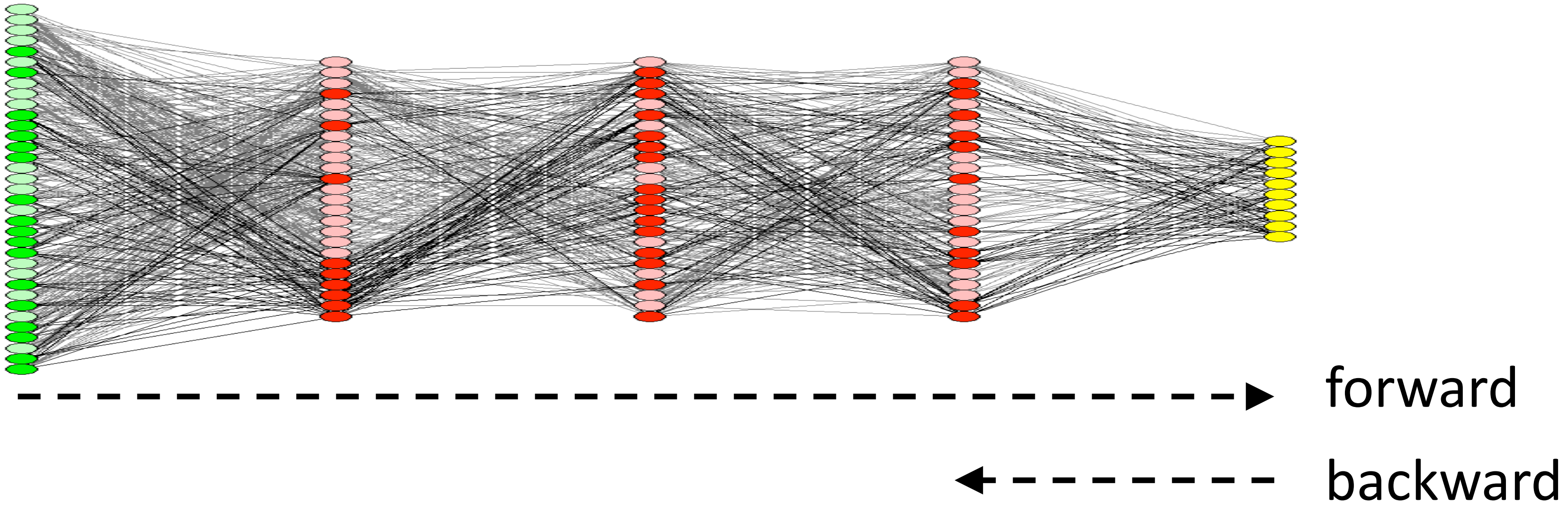
	Global vocabulary	Personal vocabulary
To cover 95% occurrences	20,000 words	6,000 words



Vocabulary size used by 1M users within 6 months (Jul. 2017 to Dec. 2017). Mean: 6214, median: 5911

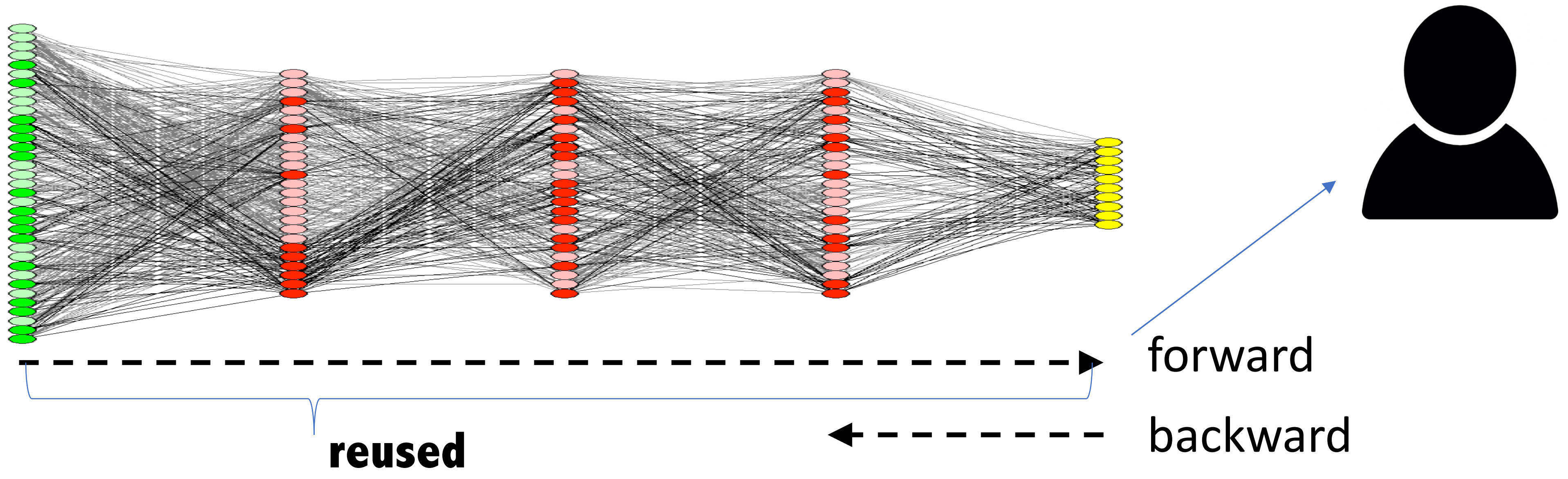
Reducing on-device computations

1. SVD-based model compression
2. Vocabulary compression
3. Fine-tune training (on-device)
4. Reusing inference results



Reducing on-device computations

1. SVD-based model compression
2. Vocabulary compression
3. Fine-tune training
4. Reusing inference results (on-device online training)



Implementation and Evaluation

- Extension to TensorFlow
- Dataset: half-year input data from 1M real users
  - IRB-approved, fully anonymized
  - Over 10 billion messages in English
- Metrics:
  - Input efficiency (accuracy)
  - On-device overhead (latency & energy)

kika

our collaborated Inc.

User input	User wants	Model output (top 3)
“I”	“will”	[“am”, “have”, “don’t”]
“I”, “w”	“will”	[“was”, “would”, “wish”]
“I”, “wi”	“will”	[“wish”, “will”, “with”]

How many chars user has to input to get the correct prediction

Top-3-efficiency = $\mathbf{1 - \frac{2}{4}}$

Length of output word “will”

DeepType improves model accuracy

pre-train dataset (global model)	personalization (private model)	top-3-efficiency
Twitter corpora	✓	0.616
	✘	0.513
Wikipedia corpora	✓	0.508
	✘	0.325
private corpora	✓	0.624
	✘	0.568
no pre-train	✓	0.331

DeepType

no personalization

DeepType improves model accuracy

pre-train dataset (global model)	personalization (private model)	top-3-efficiency
Twitter corpora	✓	0.616
	✘	0.513
Wikipedia corpora	✓	0.508
	✘	0.325
private corpora	✓	0.624
	✘	0.568
no pre-train	✓	0.331

DeepType

DeepType improves model accuracy

pre-train dataset (global model)	personalization (private model)	top-3-efficiency
Twitter corpora	✓	0.616
	✘	0.513
Wikipedia corpora	✓	0.508
	✘	0.325
private corpora	✓	0.624
	✘	0.568
no pre-train	✓	0.331

DeepType

Ideal but impractical.

Bad user privacy

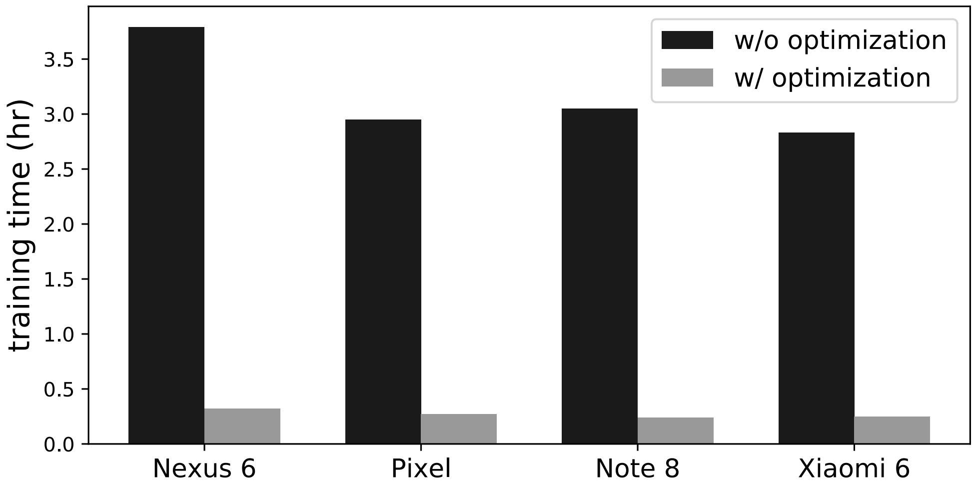
DeepType improves model accuracy

pre-train dataset (global model)	personalization (private model)	top-3-efficiency
Twitter corpora	✓	0.616
	✘	0.513
Wikipedia corpora	✓	0.508
	✘	0.325
private corpora	✓	0.624
	✘	0.568
no pre-train	✓	0.331

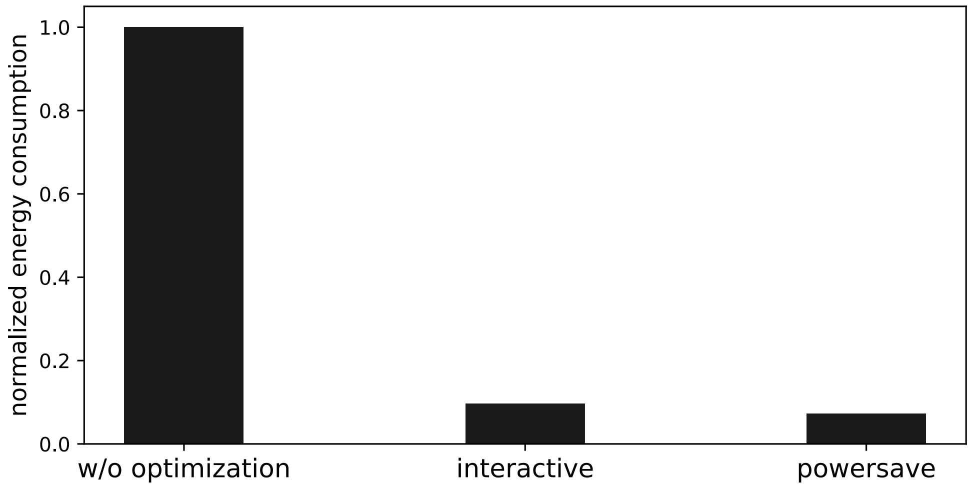
DeepType

DeepType reduces on-device overhead

- 91.6% reduction of training time
  - Less than 1.5 hours to personalize the model on half-year input history
- 90.3% reduction of energy consumption



Training time on different Android devices



Training energy w/ and w/o optimization

DeepType reduces on-device overhead

- 91.6% reduction of training time
  - Less than 1.5 hours to personalize the model on half-year input history
- 90.3% reduction of energy consumption

Device is in favored state

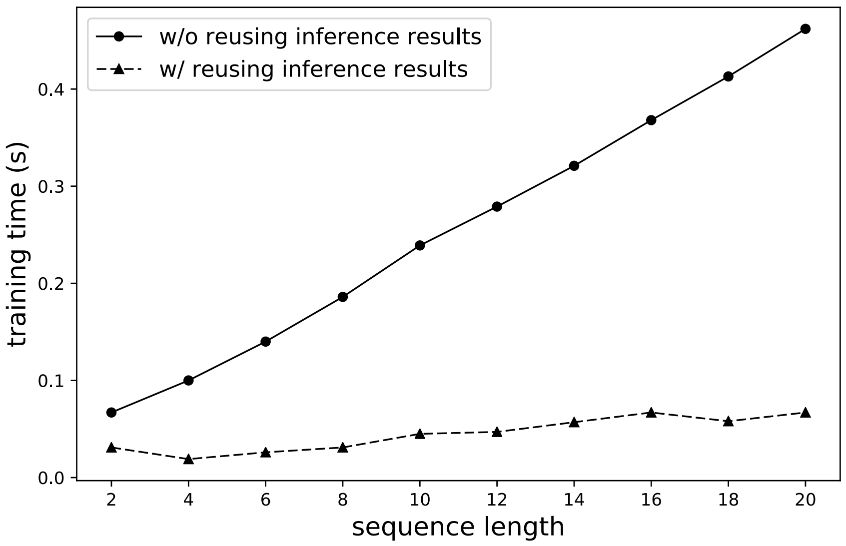
1. **Device is one**
2. **Device screen is turned off**
3. **Device is being charged and has high remaining battery**

more than 50% users spend around 2.7 hours on favored states per day -> enough for offline training!

DeepType reduces on-device overhead

- 91.6% reduction of training time
  - Less than 1.5 hours to personalize the model on half-year input history
- 90.3% reduction of energy consumption

- On-device online training typically takes only 20ms~60ms
  - Unnoticeable to users



DeepType improves the user experience

- A field study: 34 voluntary subjects in Indiana University, 3 weeks.
  - Embed DeepType into a commercial keyboard app

Recruit voluntary → Install apps → Collect trace → Answer questionnaire

- Quantitative analysis
  - Prediction: 25ms, training (online): 86ms << inter-keystroke: 264ms
- Qualitative analysis (feedbacks):
  - 78% users report improved accuracy
  - 93.7% users report good responsiveness
  - 100% users report no battery impacts

Summary

- On-cloud personalization vs. on-device personalization
  - Privacy and scalability matter
- DeepType: on-device personalization framework
  - Cloud pre-train, device fine-tune -> ensure both privacy and accuracy
  - Model compression and customized -> reduce computation overhead

Thank you for attention!