Federated Learning for Speech Recognition: Revisiting Current Trends Towards Large-Scale ASR

Sheikh Shams Azam*, Martin Pelikan*, Vitaly Feldman,
Kunal Talwar, Jan “Honza” Silovsky, Tatiana Likhomanenko*

FL@FM-NeurIPS’23 | Apple | 16 Dec 2023

ASR is suitable benchmark

ASR public data

for federated learning (FL)

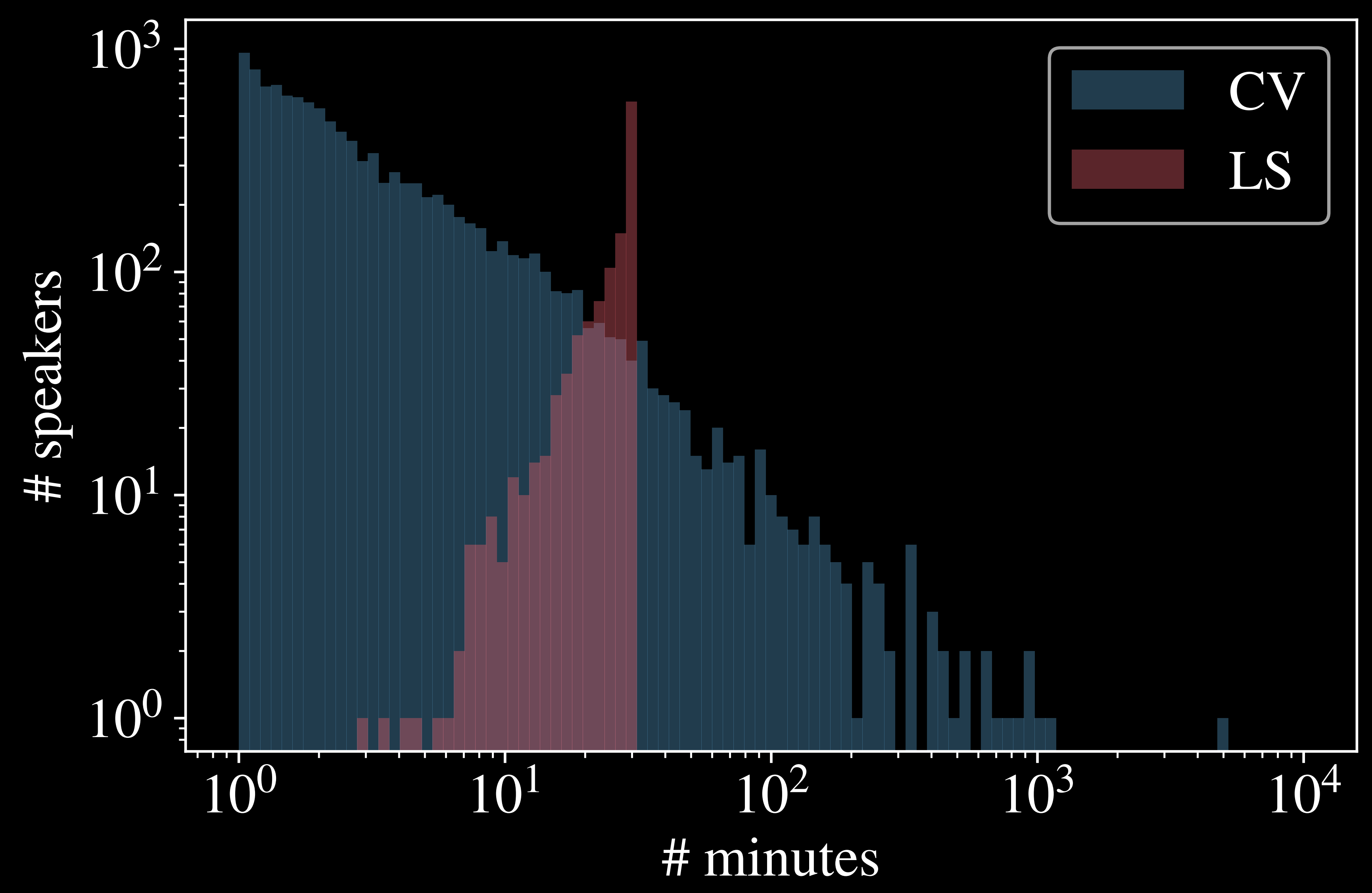
Natural split across users by using speaker information & high data heterogeneity

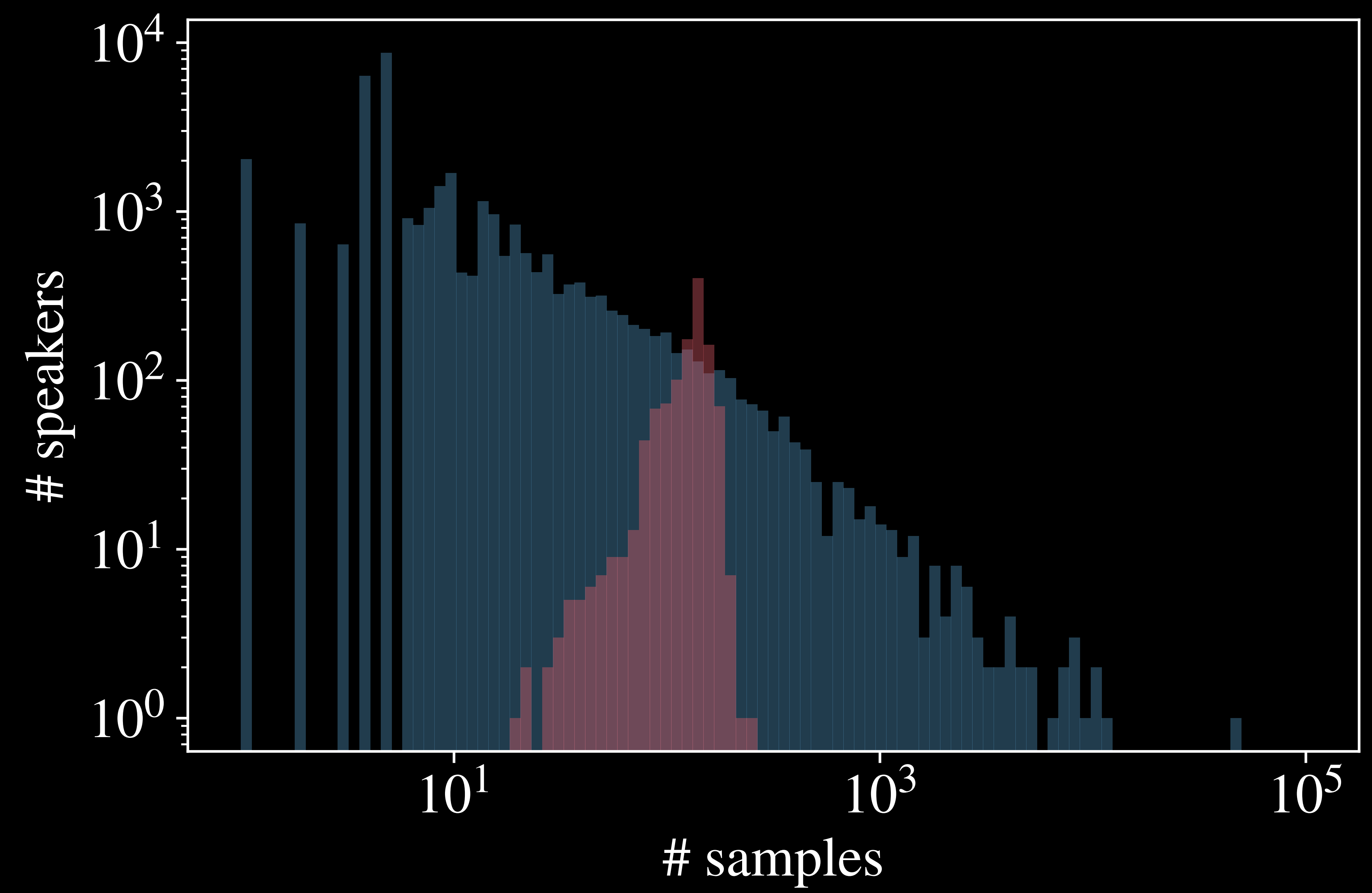
ASR public data

for federated learning (FL)

Natural split across users by using speaker information & high data heterogeneity

- LibriSpeech (LS): 25min per speaker, ~3k speakers
- Common Voice (CV): 2.5 min per speaker, ~10k-35k speakers
  - multilingual (e.g. English, French, German)





ASR public data

for federated learning (FL)

Natural split across users by using speaker information & high data heterogeneity

Simultaneously solving classification and segmentation tasks

Interplay between acoustic and language modeling

Unique challenges of ASR in the context of differential privacy (DP)

Disentangle model size and FL optimization

Optimization vs model size

Small models (FL / DP)

communication complexity

difficulty of training with DP noise

Large models (ML)

practical in many applications

simpler to optimize

Is FL hard due to its federated SGD or due to model size?

Bubeck, S. and Sellke, M., 2021. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, *34*, pp.28811-28822

Practical FL setup

Fundamental differences with prior works

Adaptive optimizers

Some adaptive optimizers perform better

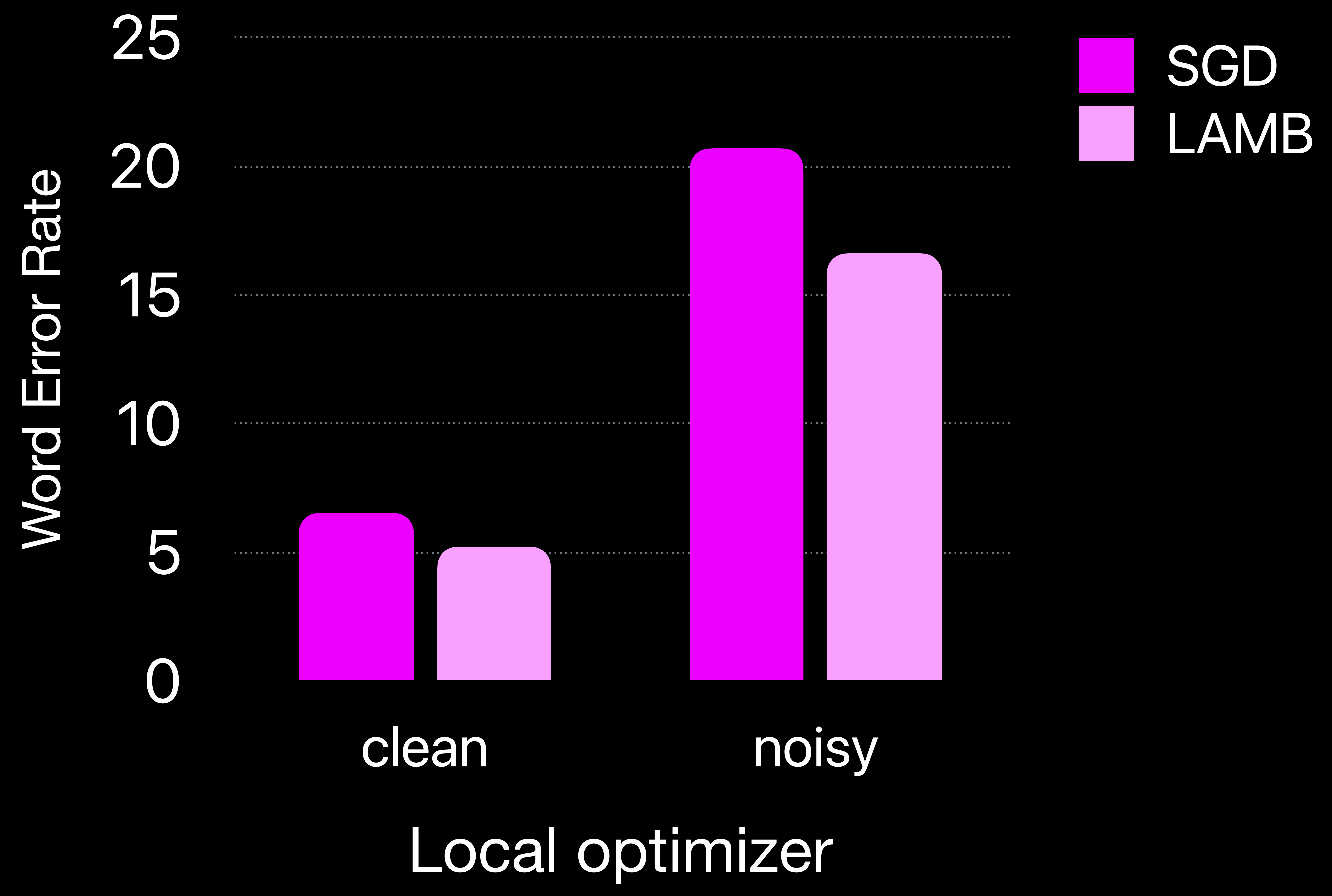
- they induce smoothness over the gradient subspace and/or reduce heterogeneity among FL client updates

Adaptive optimizers for FL with *large-scale* transformer models are necessary to match centrally trained models

Effect of optimizers

LAMB excels as a local optimizer

LAMB's layer wise adaptive scaling helps reduce "non-iid"ness among clients



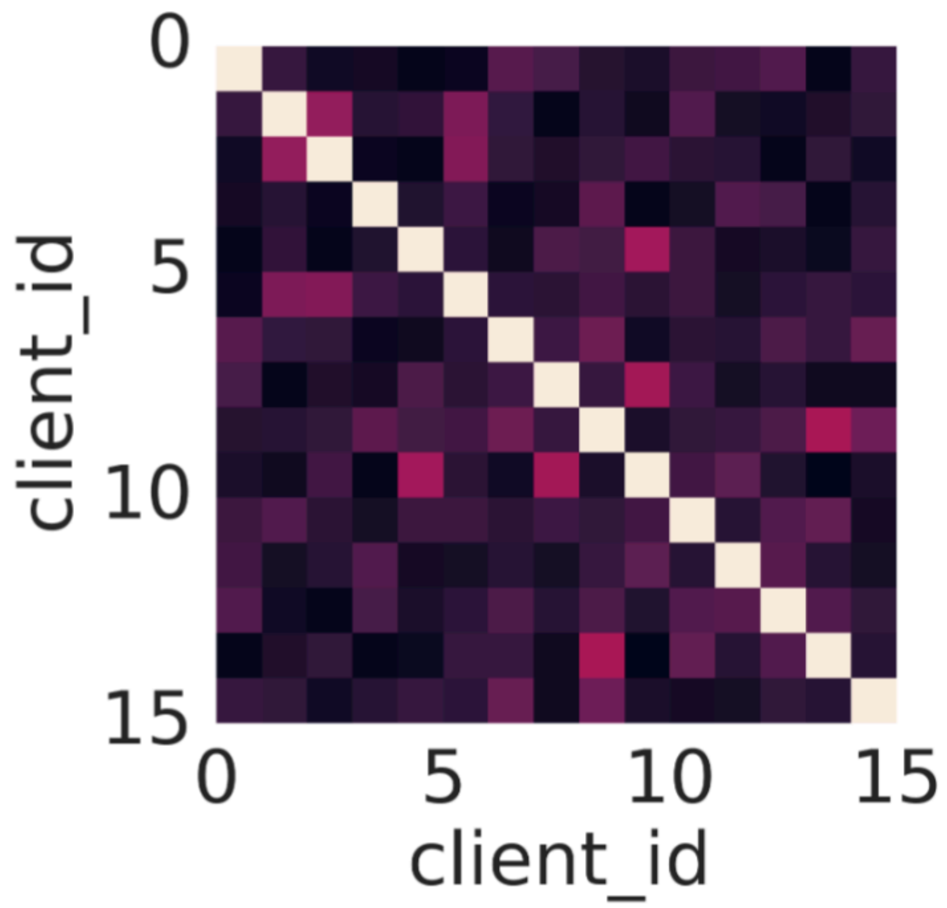
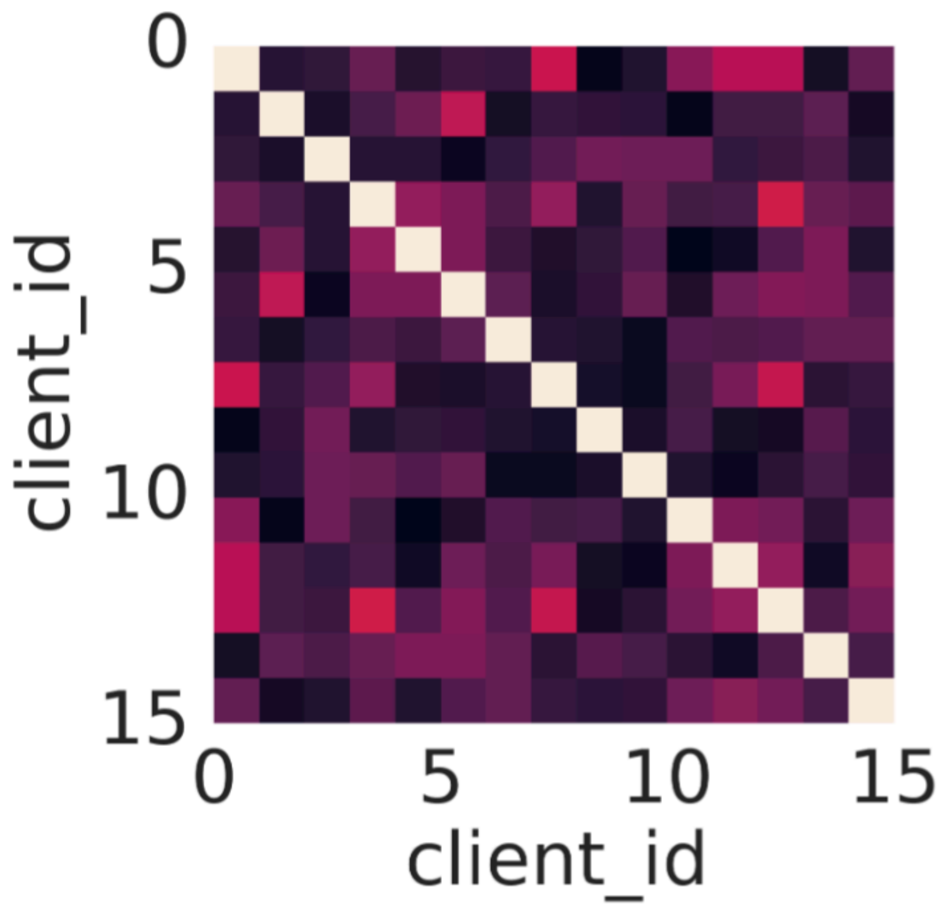
Central optimizer: Adam

Cohort size: 5

Training from a seed model

Effect of optimizers

LAMB reduces “non-iid”ness among clients

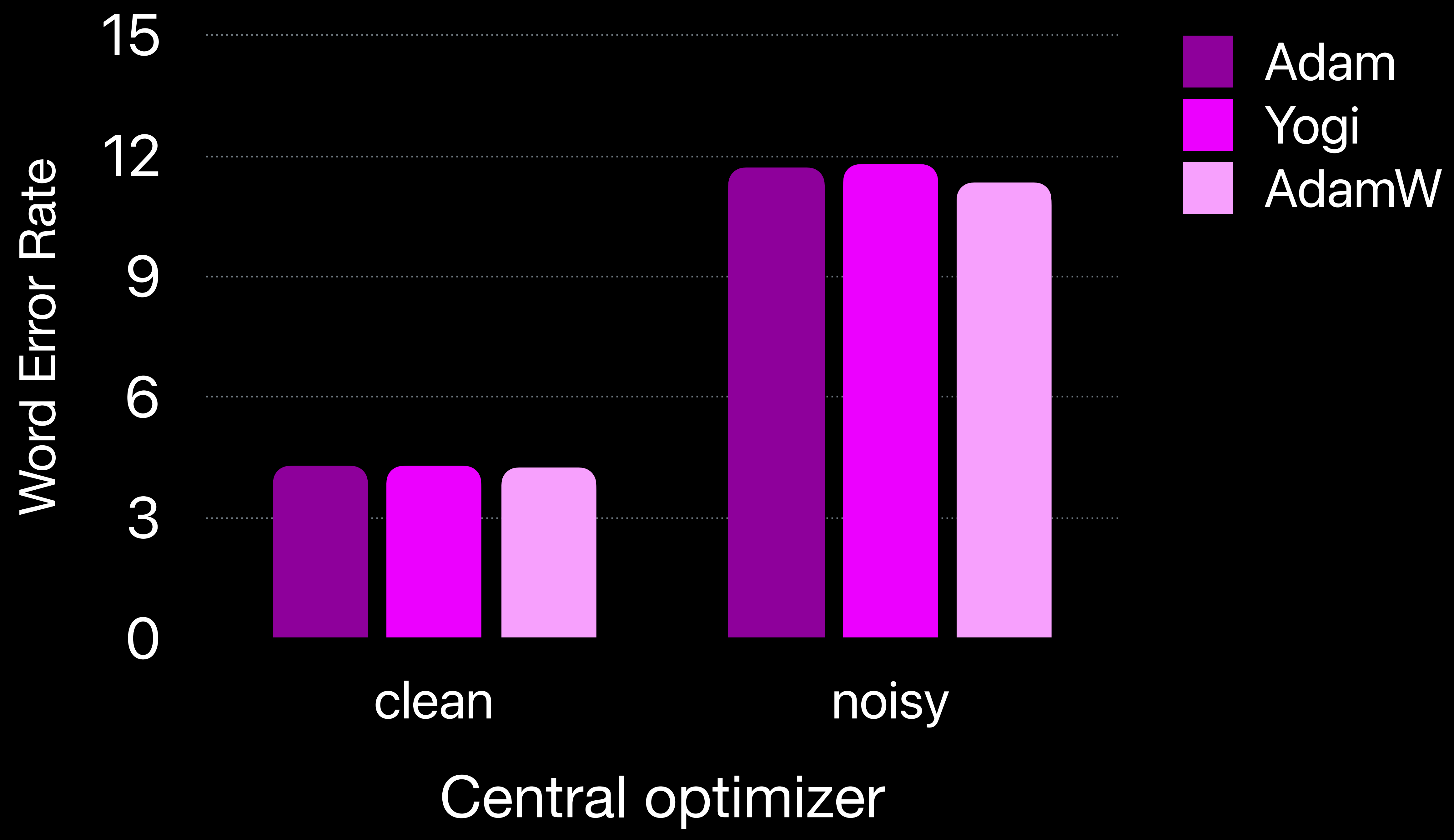
	Local Optimizers	
	SGD	LAMB
client update similarity grid		
mean	0.14	0.23

Similarity measured in terms of **pair-wise cosine similarity** among model updates from clients

Effect of optimizers

AdamW outperforms others as central optimizer

Yogi and AdamW induce smoothness over FL updates across aggregation rounds



Local optimizer: LAMB

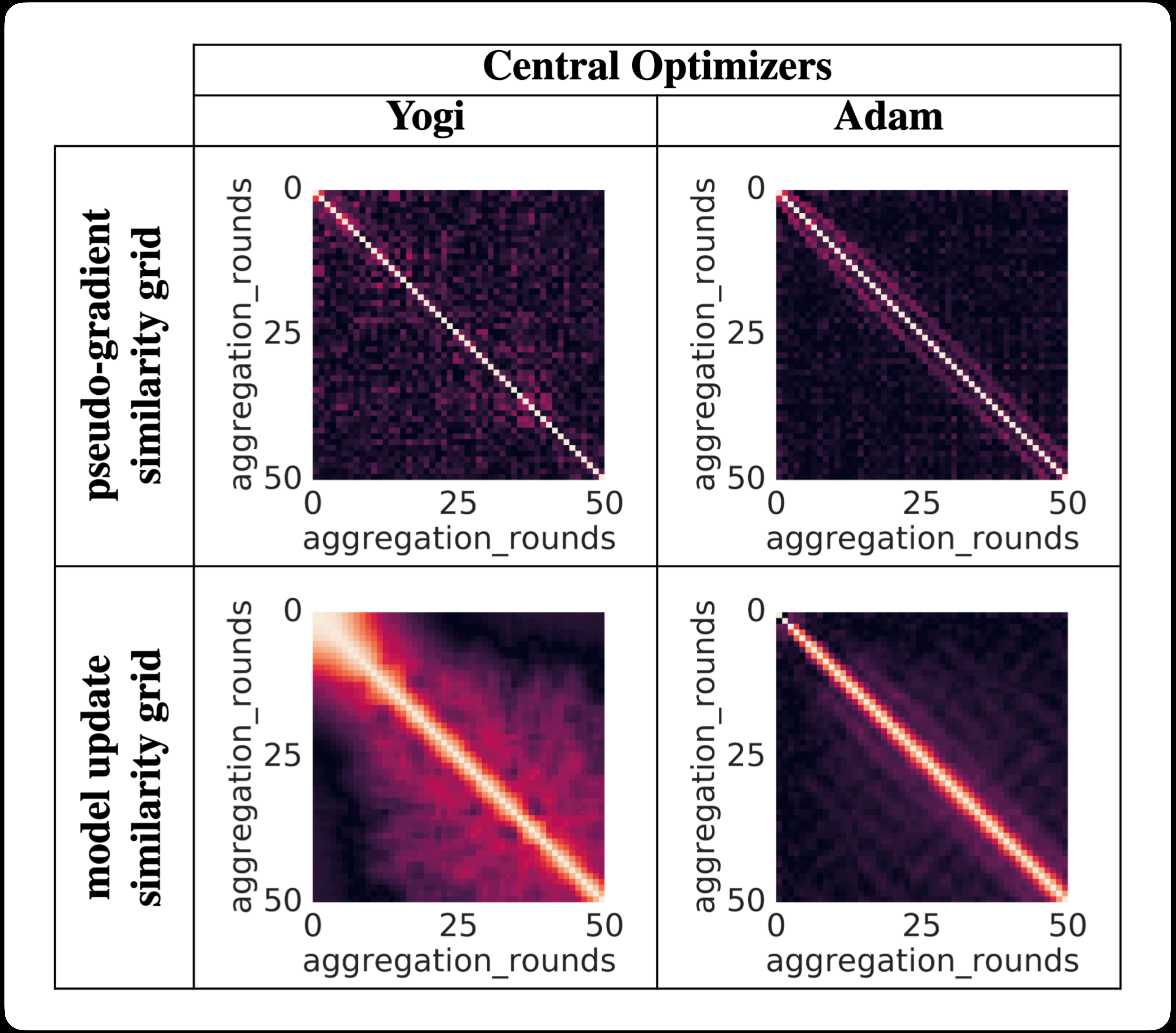
Cohort size: 15

Training from a seed model

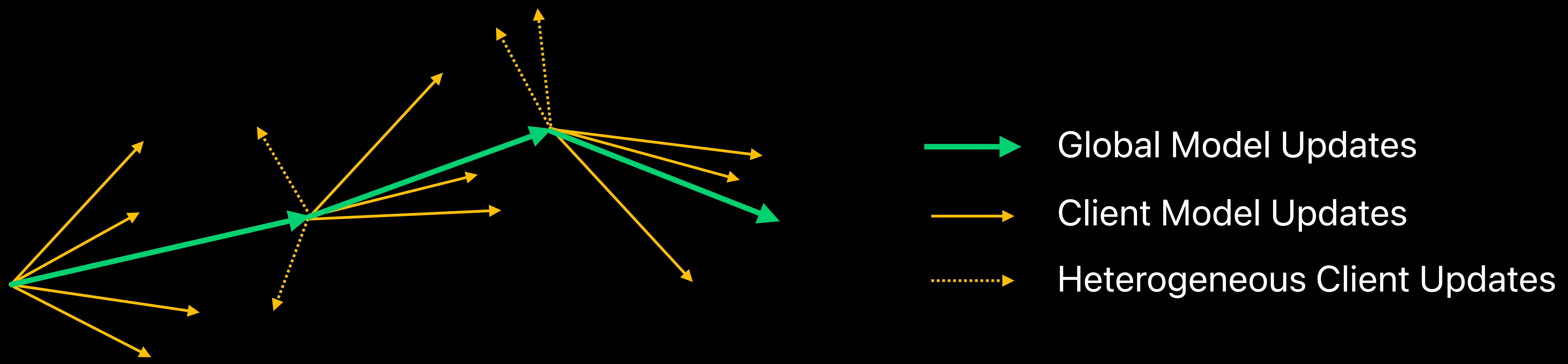
Effect of optimizers

Yogi induces smoothness over FL updates across aggregation rounds

Yogi leads to smoothening of the optimization subspace thus reducing effect of heterogeneity



Why do adaptive optimizers work?



Adaptive optimizers suppress heterogeneous components of client updates

Robustness

FL is sensitive to hyper-parameters of some optimizers

There exist a robust optimizer setting applicable to other data

- the training recipe transfers out of the box from English to French/German data

Seed models

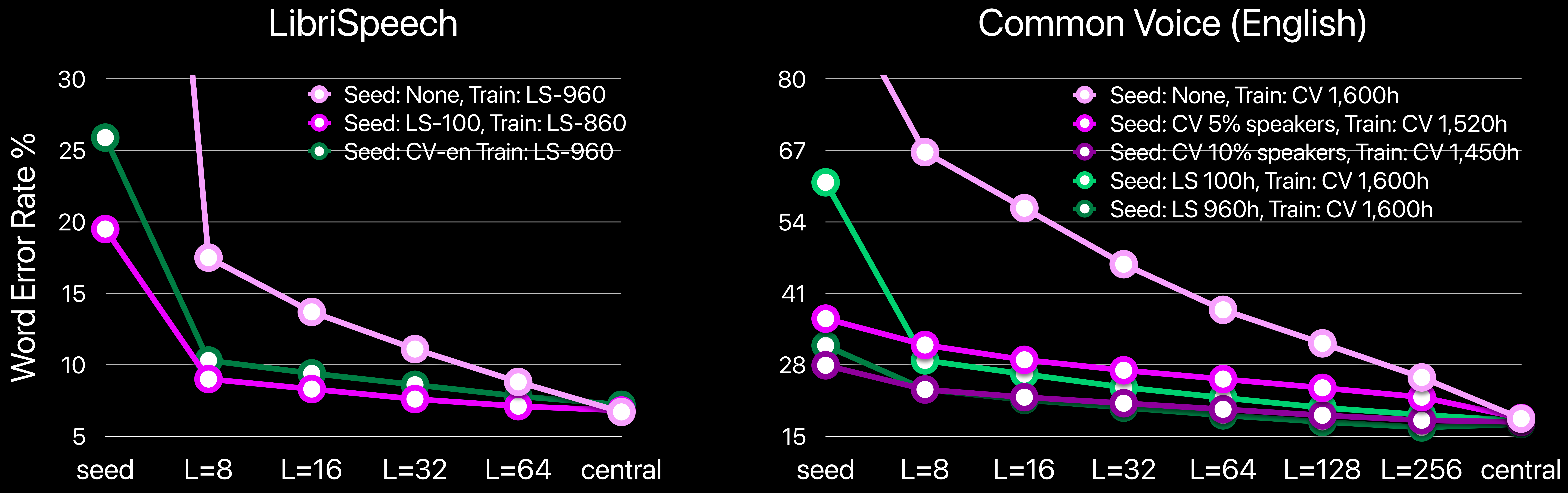
Prior work: “Nearly impossible to train an E2E ASR model *from scratch* in a realistic FL setup”

Y. Gao, T. Parcollet, S. Zaiem, J. Fernandez-Marques, P. P. de Gusmao, D. J. Beutel, and N. D. Lane, “End-to-end Speech Recognition from Federated Acoustic Models,” in IEEE ICASSP, 2022

Seed models

Prior work: “Nearly impossible to train an E2E ASR model *from scratch* in a realistic FL setup”

FL models can achieve nearly optimal performance for training both from scratch and from a seed model centrally pre-trained even on out-of-domain data

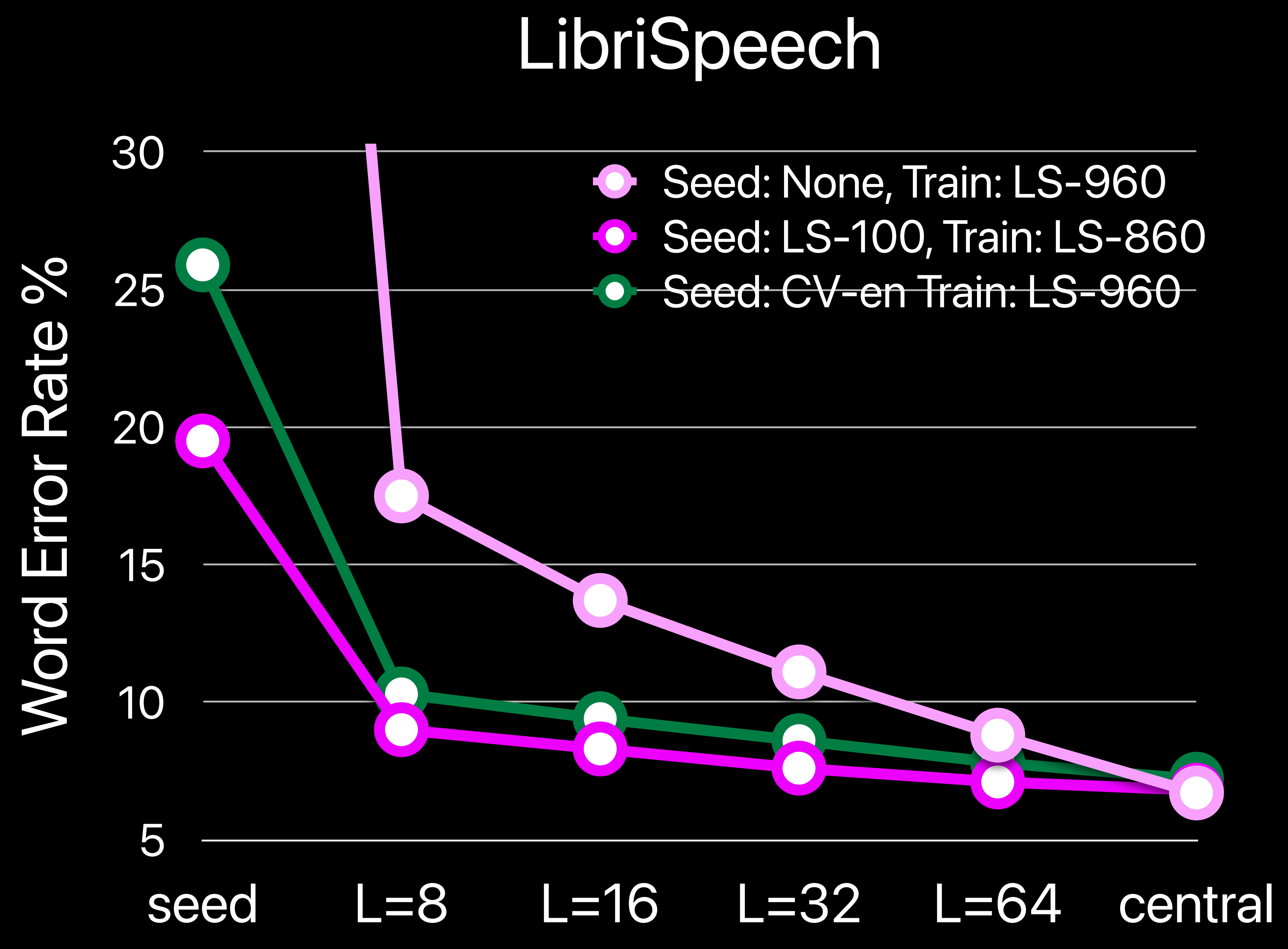


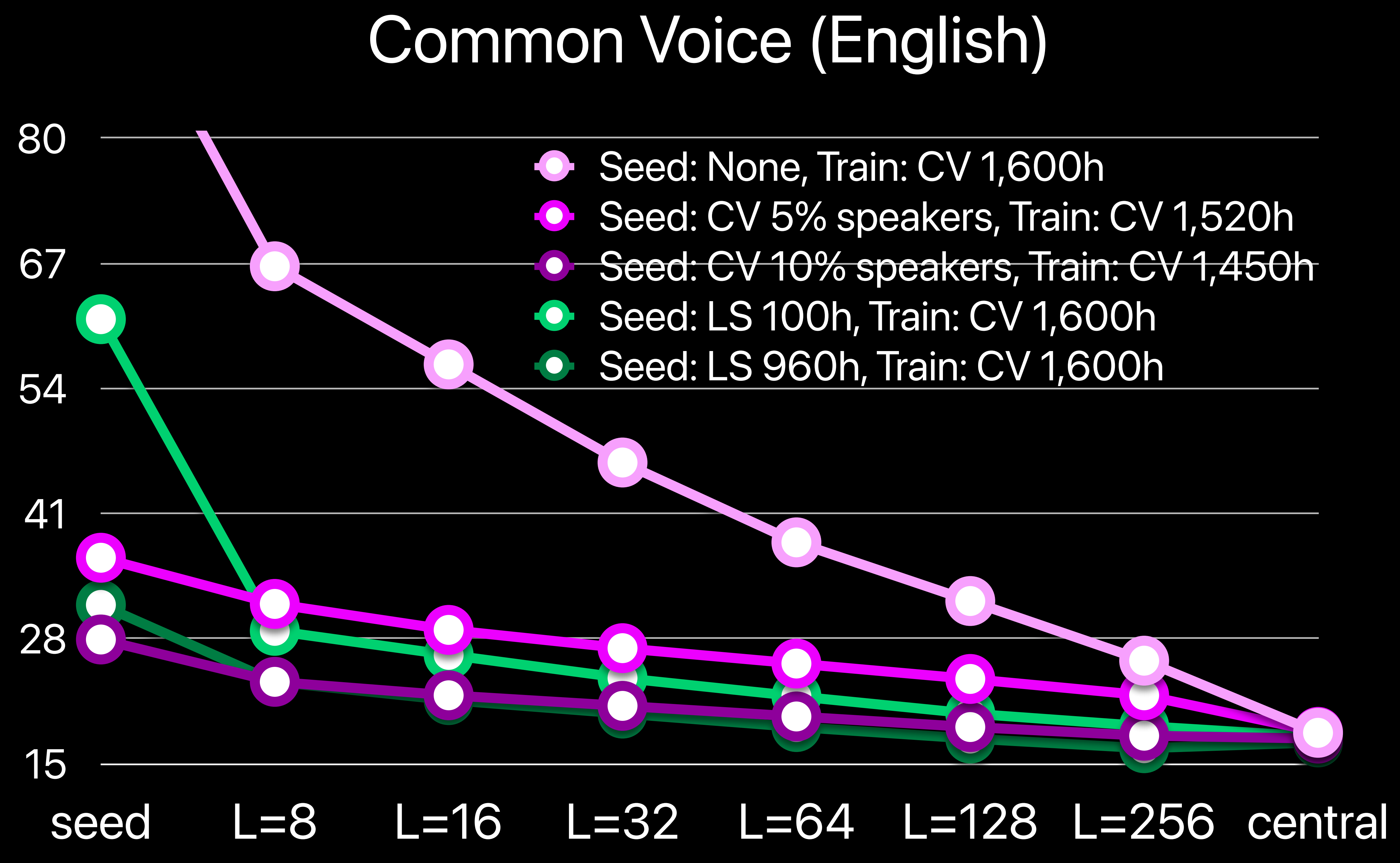
Server data are either in-domain or out-of-domain

Y. Gao, T. Parcollet, S. Zaiem, J. Fernandez-Marques, P. P. de Gusmao, D. J. Beutel, and N. D. Lane, “End-to-end Speech Recognition from Federated Acoustic Models,” in IEEE ICASSP, 2022

Cohort size

Discrepancy between different models reduces as the cohort size increases



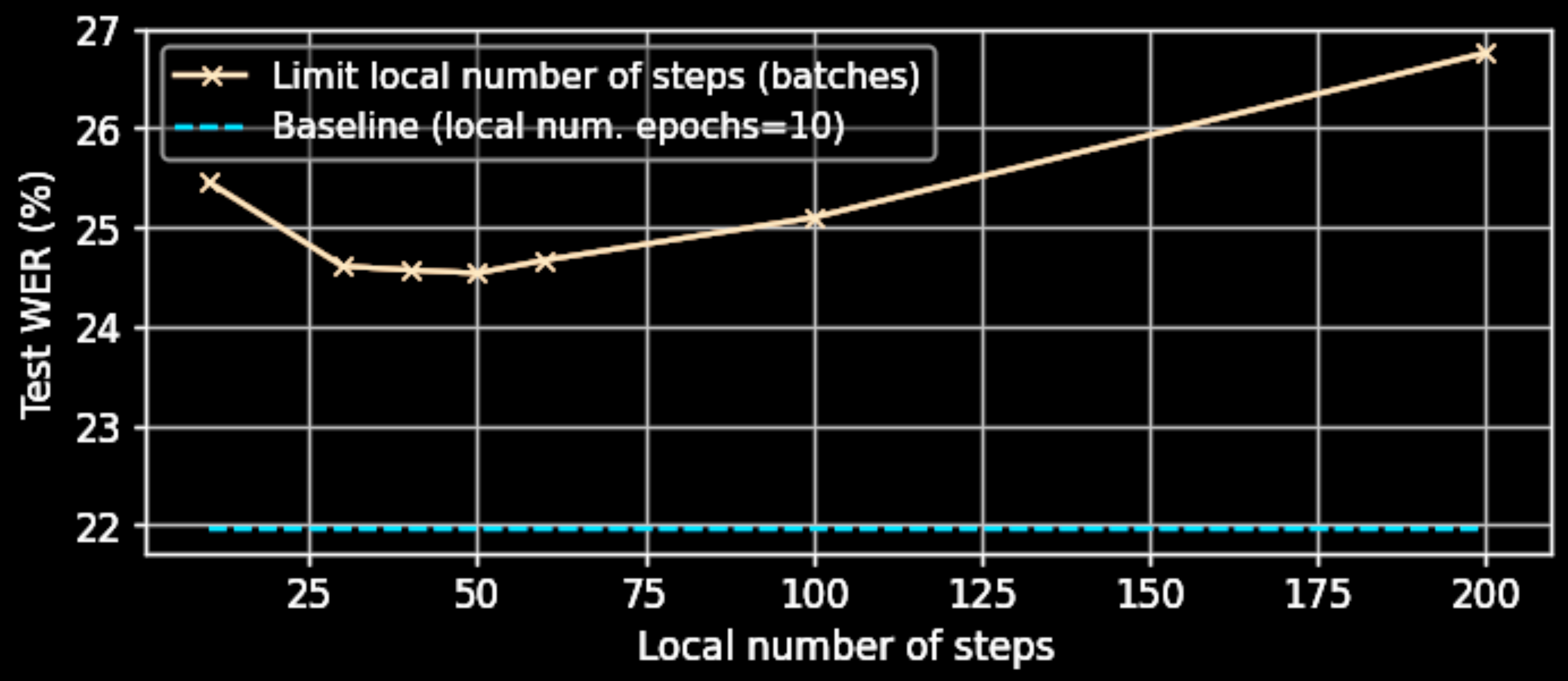


Server data are either in-domain or out-of-domain

Number of local epochs / steps

Increasing the number of local epochs speeds up convergence, but eventually the large number of local epochs hurts model training (similar to [1])

Switching to local steps leads to overall degradation of performance (opposite to [2])

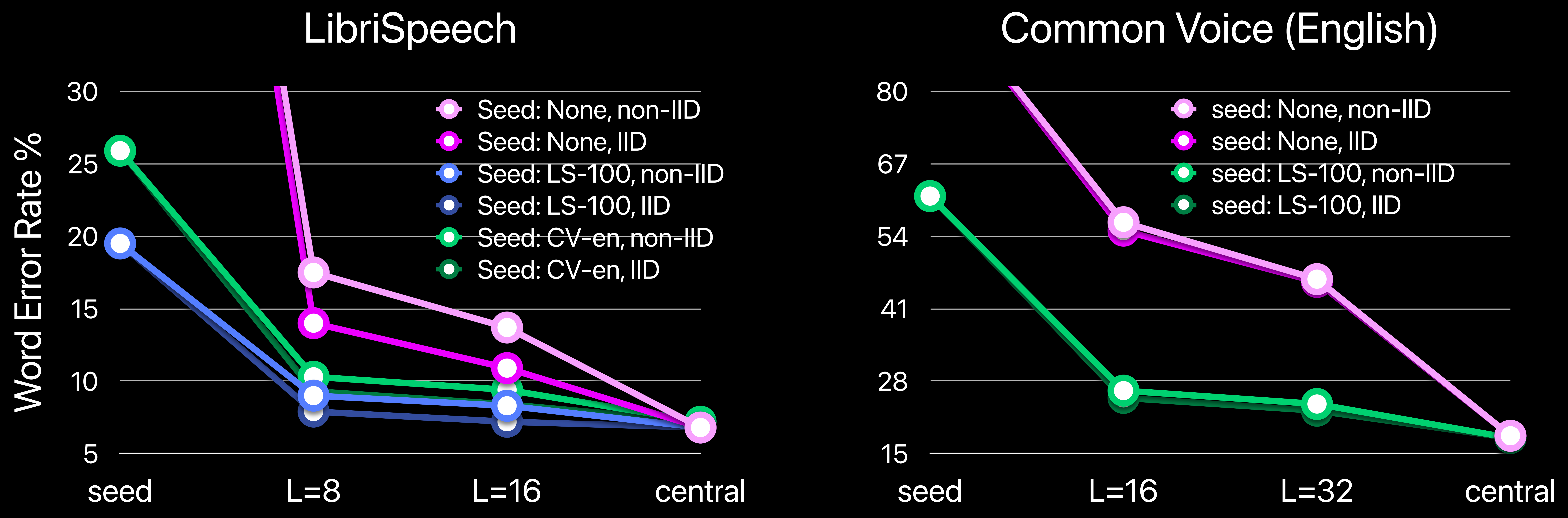


[1] K. Mishchenko, G. Malinovsky, S. Stich, and P. Richtárik, "ProxSkip: Yes! Local Gradient Steps Provably lead to Communication Acceleration! Finally!" in International Conference on Learning Representations (ICLR), 2022
[2] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization," Neural Information Processing Systems (NeurIPS), 2020

Data heterogeneity

The issue is alleviated with increased cohort size and adaptive optimizers

Transformers appear to be less susceptible to data heterogeneity than conformers

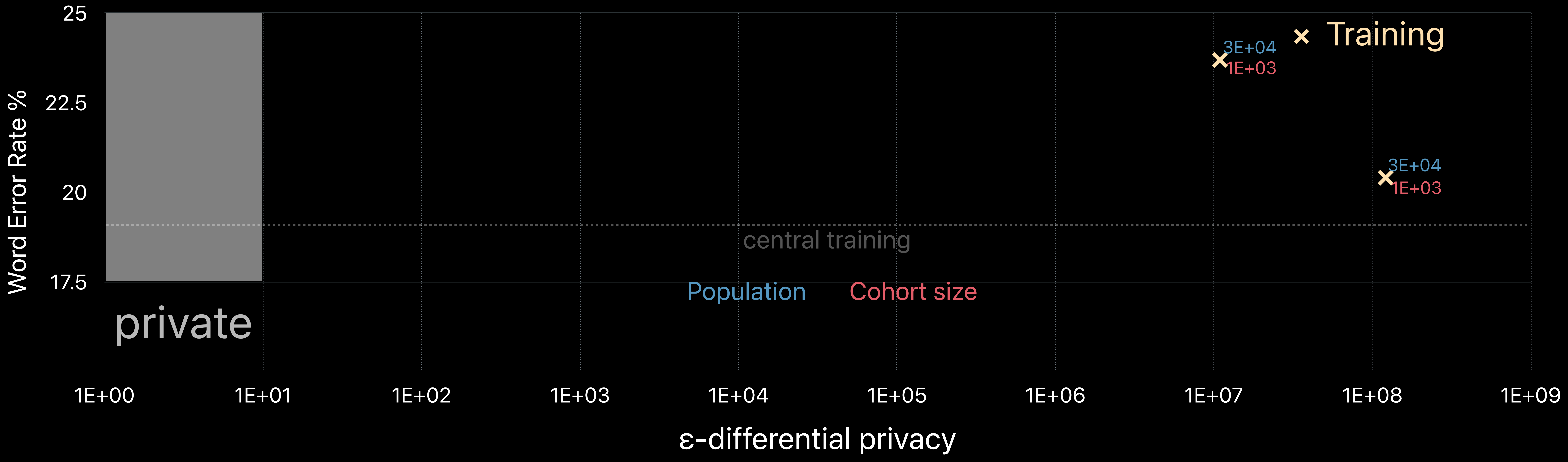


Client data are either non-IID or IID

First practical FL & DP baseline in ASR

(ϵ, δ)-DP guarantees

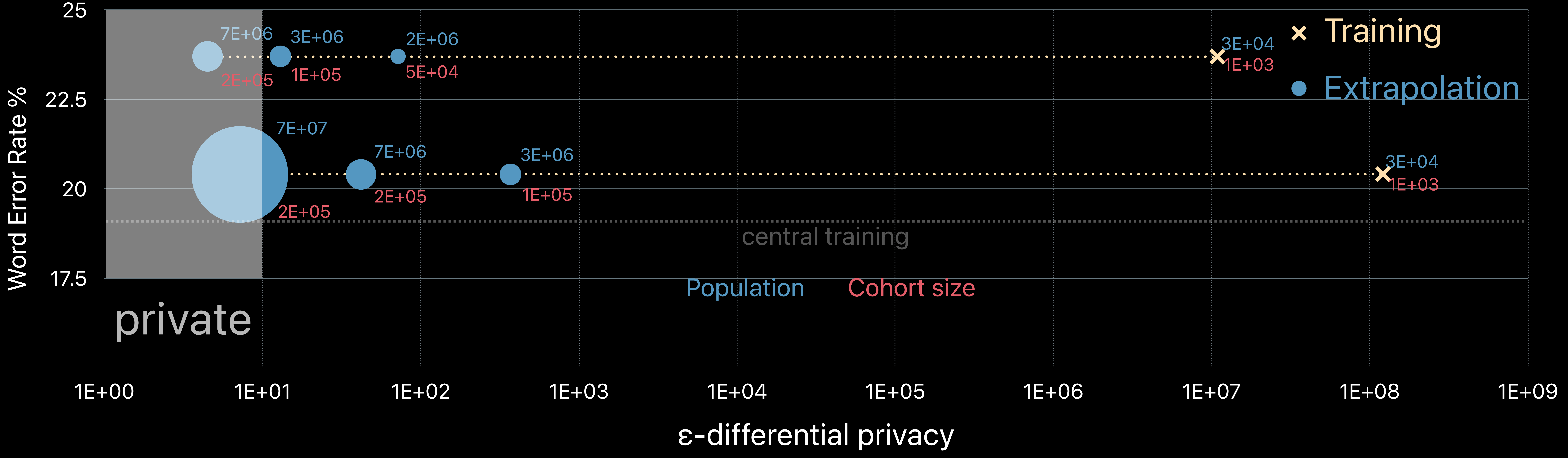
Central seed (train on LS 100h) is fine-tuned with FL & DP (train on CV 1,500h)



(ϵ, δ)-DP guarantees

Central seed (train on LS 100h) is fine-tuned with FL & DP (train on CV 1,500h)

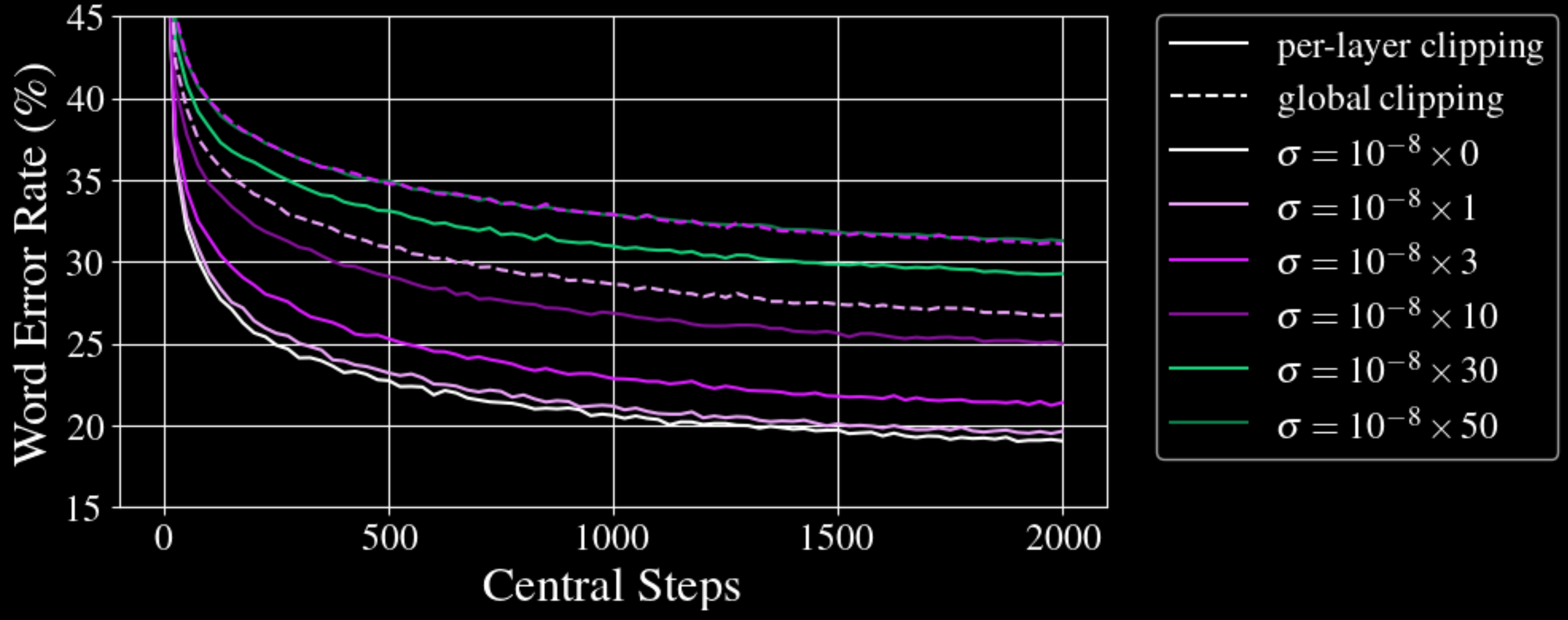
Get *practical* {quality, (ϵ, δ)-DP} with extrapolation to larger population and cohort



Fundamental differences: per-layer clipping

Per-layer clipping did not show difference in prior works

We observe significant performance boost that allows increasing noise by 10x



H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In International Conference on Learning Representations, 2018

Jiyan He, and et al. Exploring the limits of differentially private deep learning with group-wise clipping. In International Conference on Learning Representations, 2023

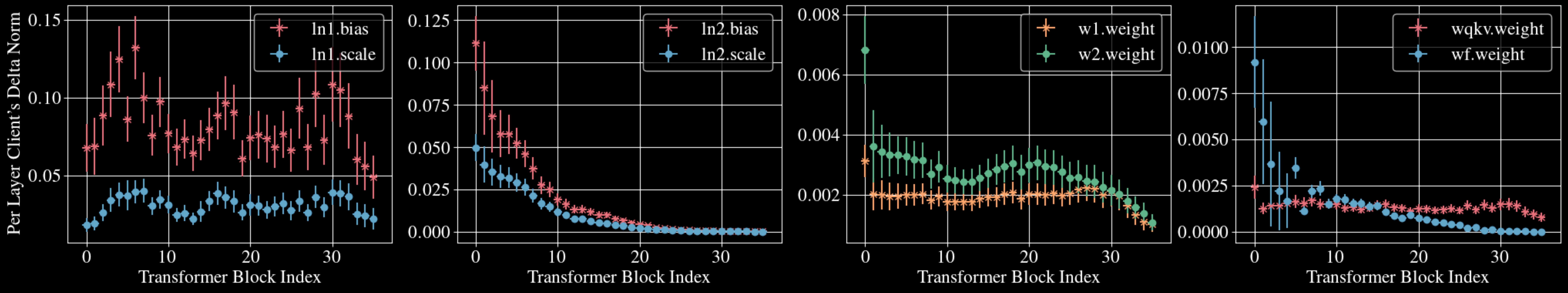
Fundamental differences: per-layer clipping

Per-layer clipping

- alleviates gradients misbalance between layers
- is helpful for English but only marginal for French and German

Source of gradients misbalance

- central training properties which depends on the language / data



Use ASR & large models in research on FL & DP

Details in papers

Workshop paper

Adaptive optimizers smoothness

FL + DP for ASR