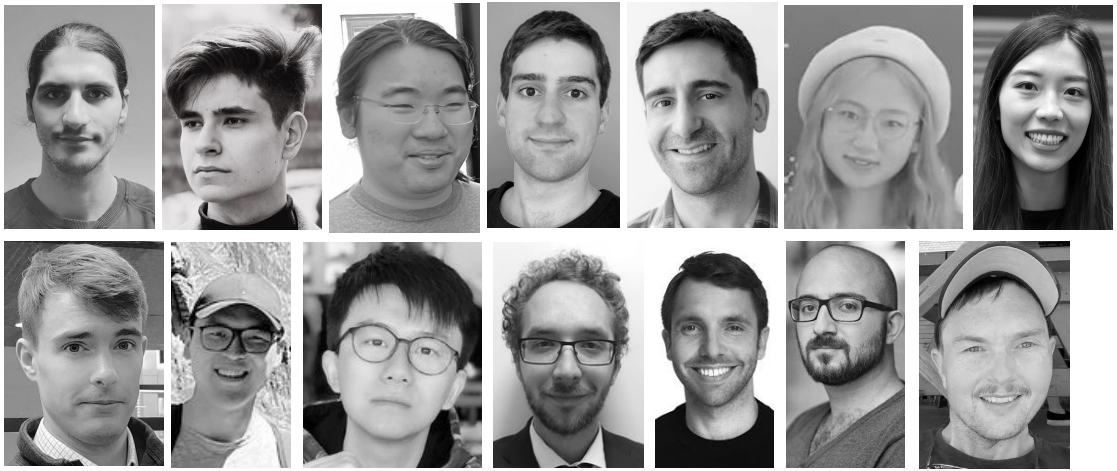
NANDA Workshop at Imperial College
Monday 9th September 2024 – *London*

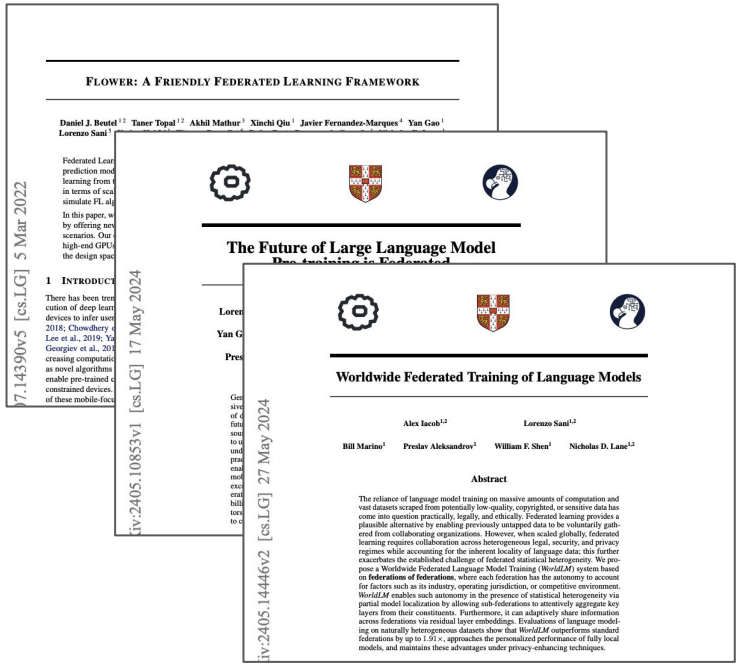
CaMLSys http://mlsys.cst.cam.ac.uk

The Future of Large Language Models (and AI) is Federated

Nicholas D. Lane
University of Cambridge | Flower Labs
@niclane7

CaMLSys http://mlsys.cst.cam.ac.uk





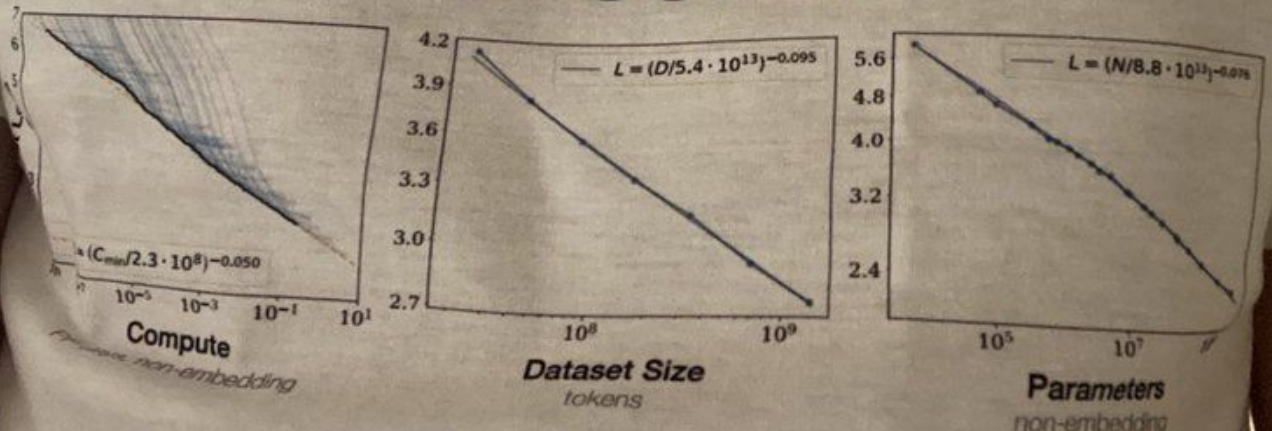
Flower

https://flower.ai

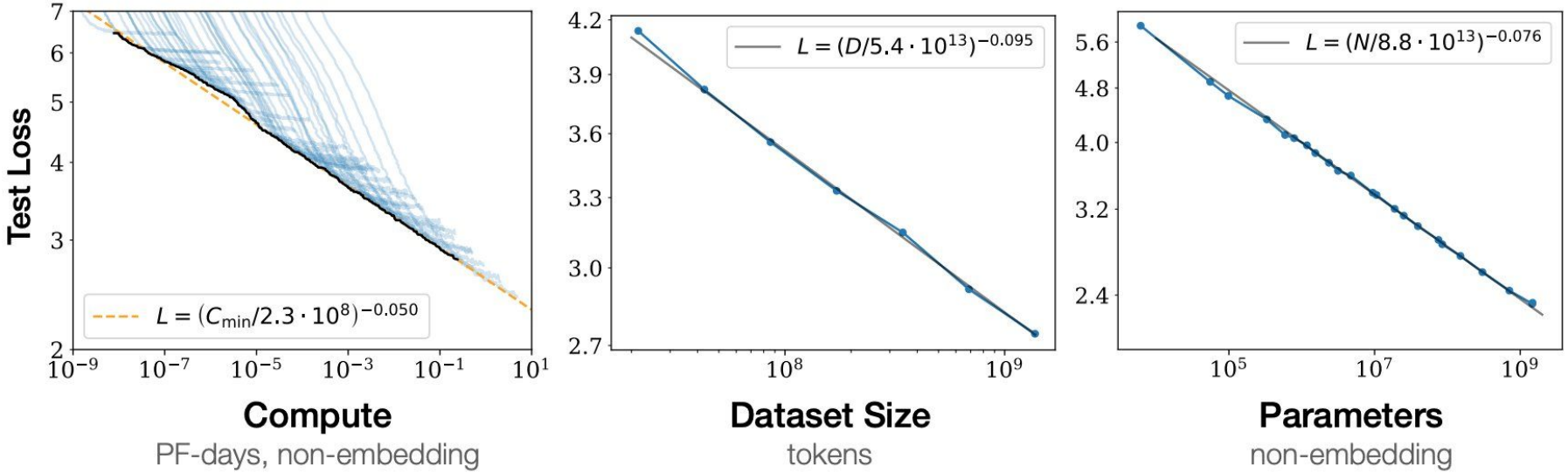
Cambridge Machine Learning Systems Lab
mlsys.cst.cam.ac.uk

Nicholas D. Lane
University of Cambridge | Flower Labs
@niclane7

SCALE IS ALL YOU NEED - AGI IS COMING



More data and compute (currently) leads to better AI



https://arxiv.org/abs/2203.15556
https://arxiv.org/abs/2001.08361

But, we may be running out of both!

WIRED

Nvidia Chip Shortages Leave AI Startups Scrambling for Computing Power

Trimming profits, delaying launches, begging friends. Companies are going to extreme lengths to make do with shortages of GPUs, the chips at...

24 Aug 2023

Andrew Côté 
@Andercot

Europe has less than 3% of the world's deployed H100s

11:32 AM · Apr 24, 2024 · **176.8K** Views

The Register

Microsoft, OpenAI may be dreaming of $100B 5GW AI 'Stargate' supercomputer

OpenAI is believed to be in talks with Microsoft to construct a massive supercomputer code-named Stargate containing millions of AI...

1 Apr 2024

BBC Sign in Home News Sport Weather iPlayer Sounds

NEWS

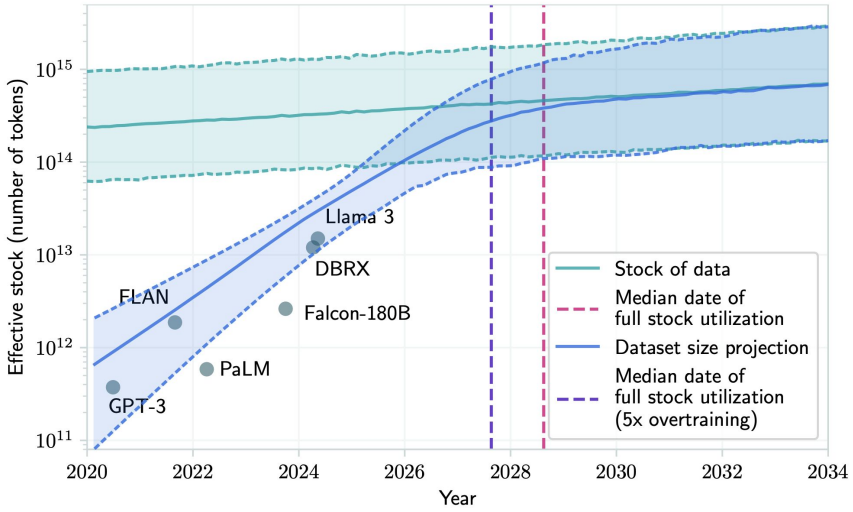
Home | Election 2024 | InDepth | Israel-Gaza war | Cost of Living | War in Ukraine | Climate | UK | World | Business

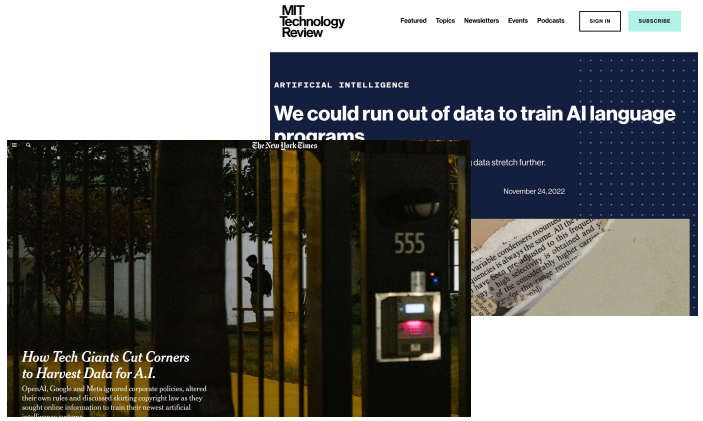
Electricity grids creak as AI demands soar

Data centre electricity needs are forecast to double between 2022 and 2026

Chris Baraniuk
Technology reporter

But, we may be running out of both!



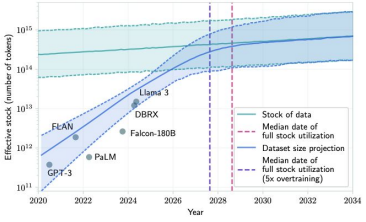


Will we run out of data? Limits of LLM scaling based on human-generated data

Pablo Villalobos [1] Anson Ho [1] Jaime Sevilla [1 2] Tamay Besiroglu [1 3] Lennart Heim [1 4] Marius Hobbhahn [1 5]

Abstract

We investigate the potential constraints on LLM scaling posed by the availability of public human-generated text data. We forecast the growing demand for training data based on current trends and estimate the total stock of public human text data. Our findings indicate that if current LLM development trends continue, models will be trained on datasets roughly equal in size to the available stock of public human text data between 2026 and 2032, or slightly earlier if models are overtrained. We explore how progress in language modeling



https://arxiv.org/abs/2211.04325

But, we may be running out of both!

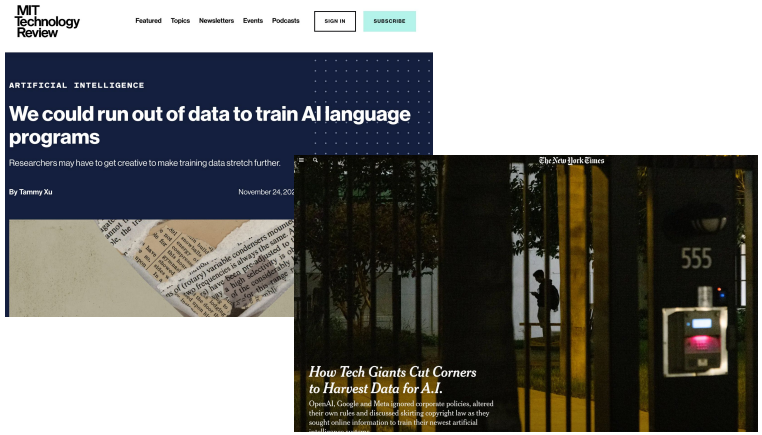
Yann LeCun 
@ylecun

Sources of reliable data are getting exhausted.
The cost of manual "post-training" is growing quickly.
Yet, the performances on bemchmarks are clearly saturating.
So no, Auto-Regressive LLMs in their current form will not take us to human-level AI.
That doesn't mean they are not useful.

Max Welling
@wellingmax

I am super impressed with recent progress in AI. But can we extrapolate that AGI is around the corner? Remember that scaling laws are polynomial with a coefficient < 1 at best, and we have almost exhausted the reliable data sources available. Not 100% sure but interesting times!

8:03 AM · Aug 31, 2024 · **230.4K** Views

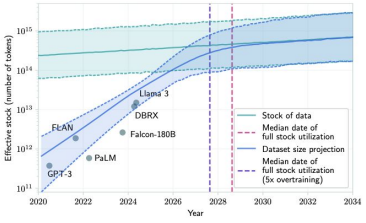


Will we run out of data? Limits of LLM scaling based on human-generated data

Pablo Villalobos [1] Anson Ho [1] Jaime Sevilla [1 2] Tamay Besiroglu [1 3] Lennart Heim [1 4] Marius Hobbhahn [1 5]

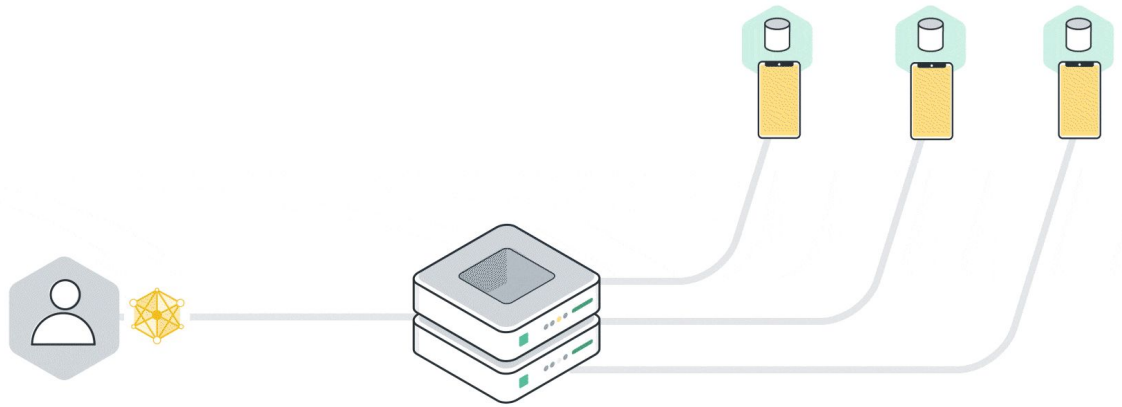
Abstract

We investigate the potential constraints on LLM scaling posed by the availability of public human-generated text data. We forecast the growing demand for training data based on current trends and estimate the total stock of public human text data. Our findings indicate that if current LLM development trends continue, models will be trained on datasets roughly equal in size to the available stock of public human text data between 2026 and 2032, or slightly earlier if models are overtrained. We explore how progress in language modeling



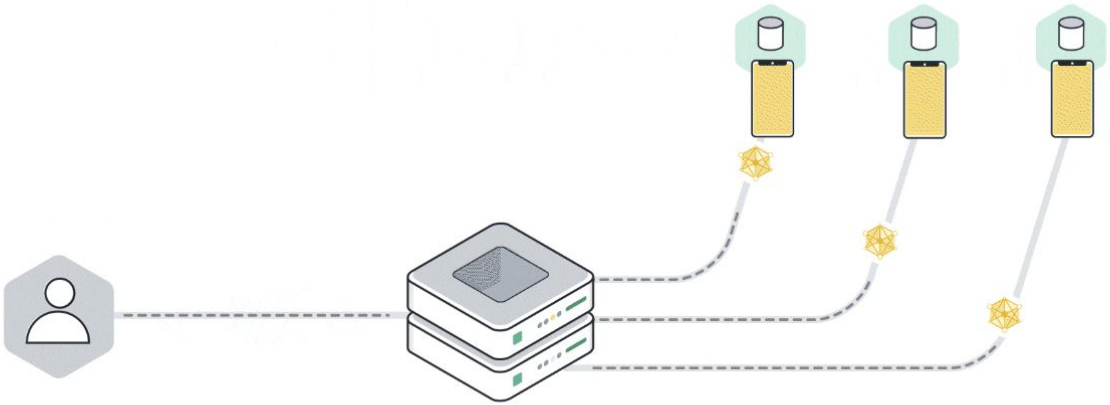
https://arxiv.org/abs/2211.04325

Is **federated learning*** the answer?



* and related approaches (e.g., decentralized forms of ML etc)

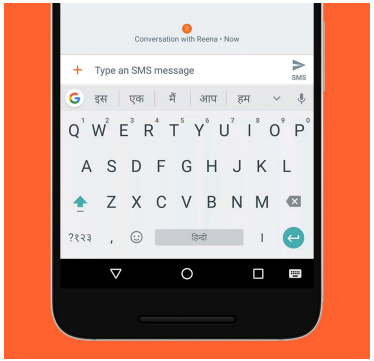
Is **federated learning*** the answer?



* and related approaches (e.g., decentralized forms of ML etc)

Federated Examples (exciting, but niche)

Navigation and Perception of Robots



Learning user keyboard behaviors and word selection

Sharing of Sensitive Data between Organizations

Personalization of Speech Recognition

Data Opportunity

Centralized ML
Data

Data Opportunity

Centralized ML Data

Distributed and Sensitive Data

Compute Opportunity

2.9

exaFLOPS

22k
Nvidia H100s

Compute Opportunity

2.9

exaFLOPS

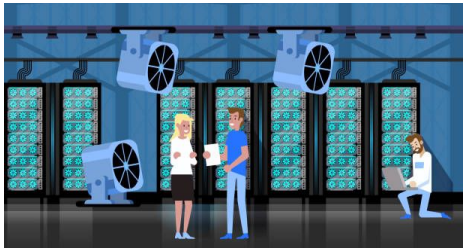
22k
Nvidia H100s

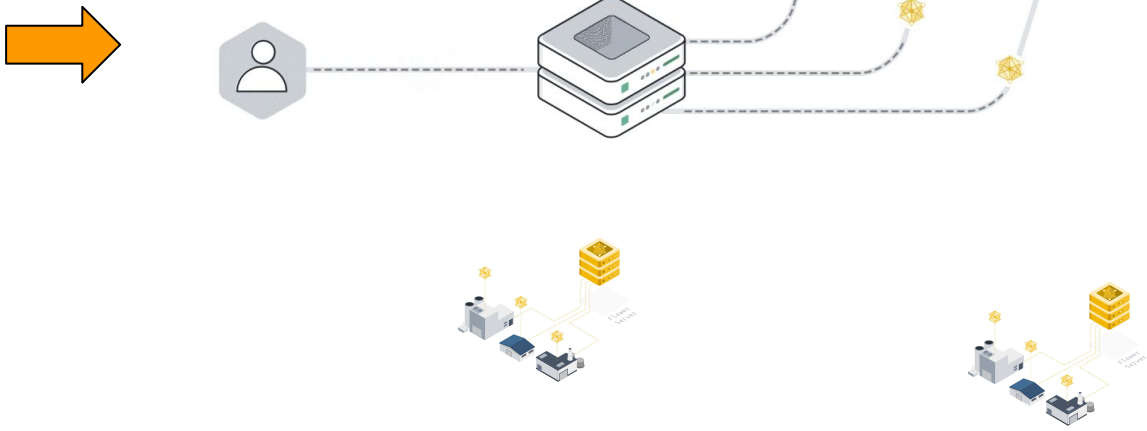
69.6

exaFLOPS

20M
Snapdragon 8 Gen 2

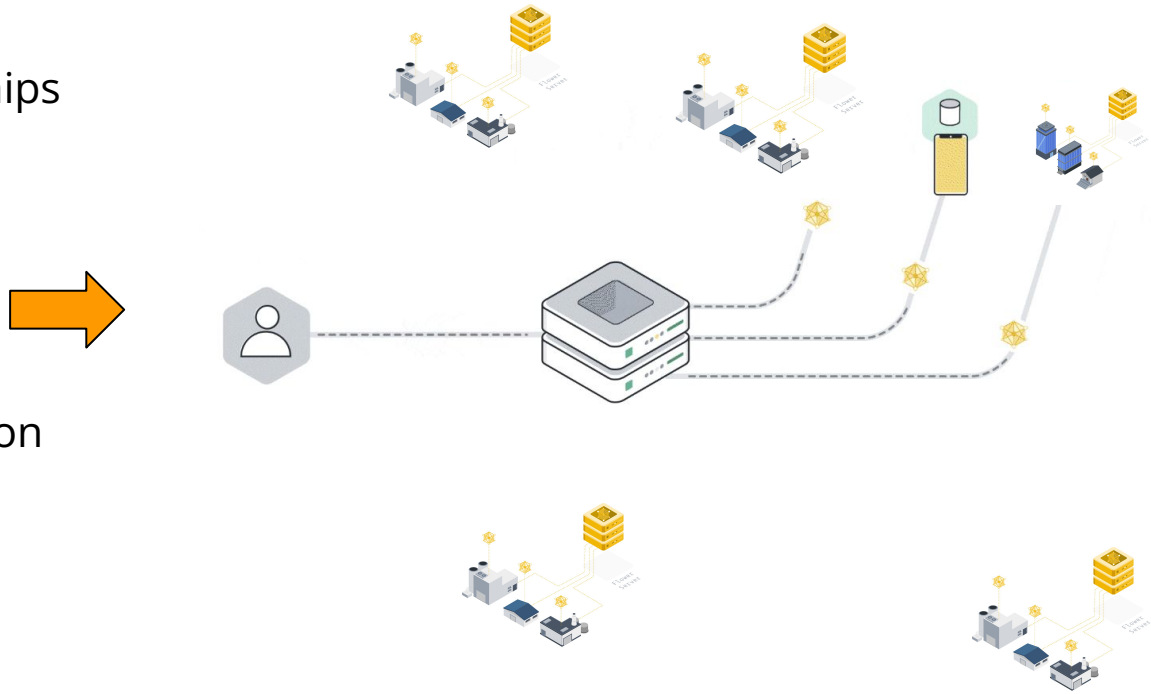
What if we can make all AI infrastructure federated?





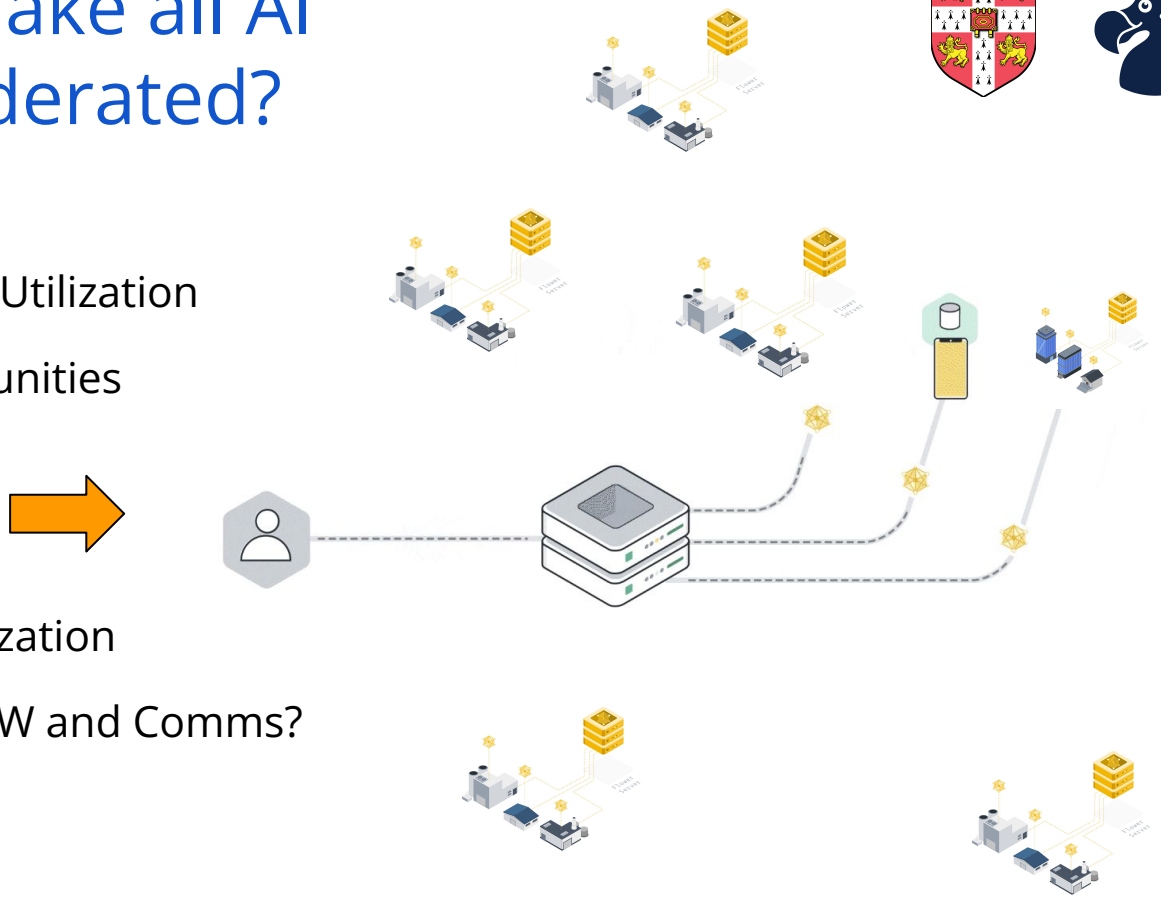
What if we can make all AI infrastructure federated?

- Cooperation & partnerships
- Privacy & control
- Scalability & flexibility
- Resource sharing
- Compatibility w/ regulation
- ...



What if we can make all AI infrastructure federated?

- Hardware Efficiency and Utilization
- Realizing Privacy Opportunities
- ML Optimization
- Governance
- Complexity of Decentralization
- How should we design HW and Comms?
- …

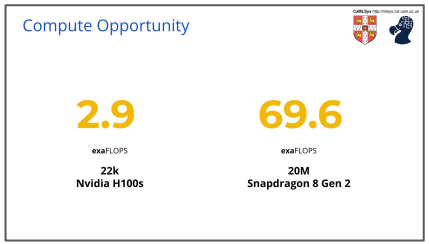


“ In the near future, every SOTA AI model will be trained using federated learning ”

– @niclane7

“ In the near future, every SOTA AI model will be trained using federated learning ”

– @niclane7

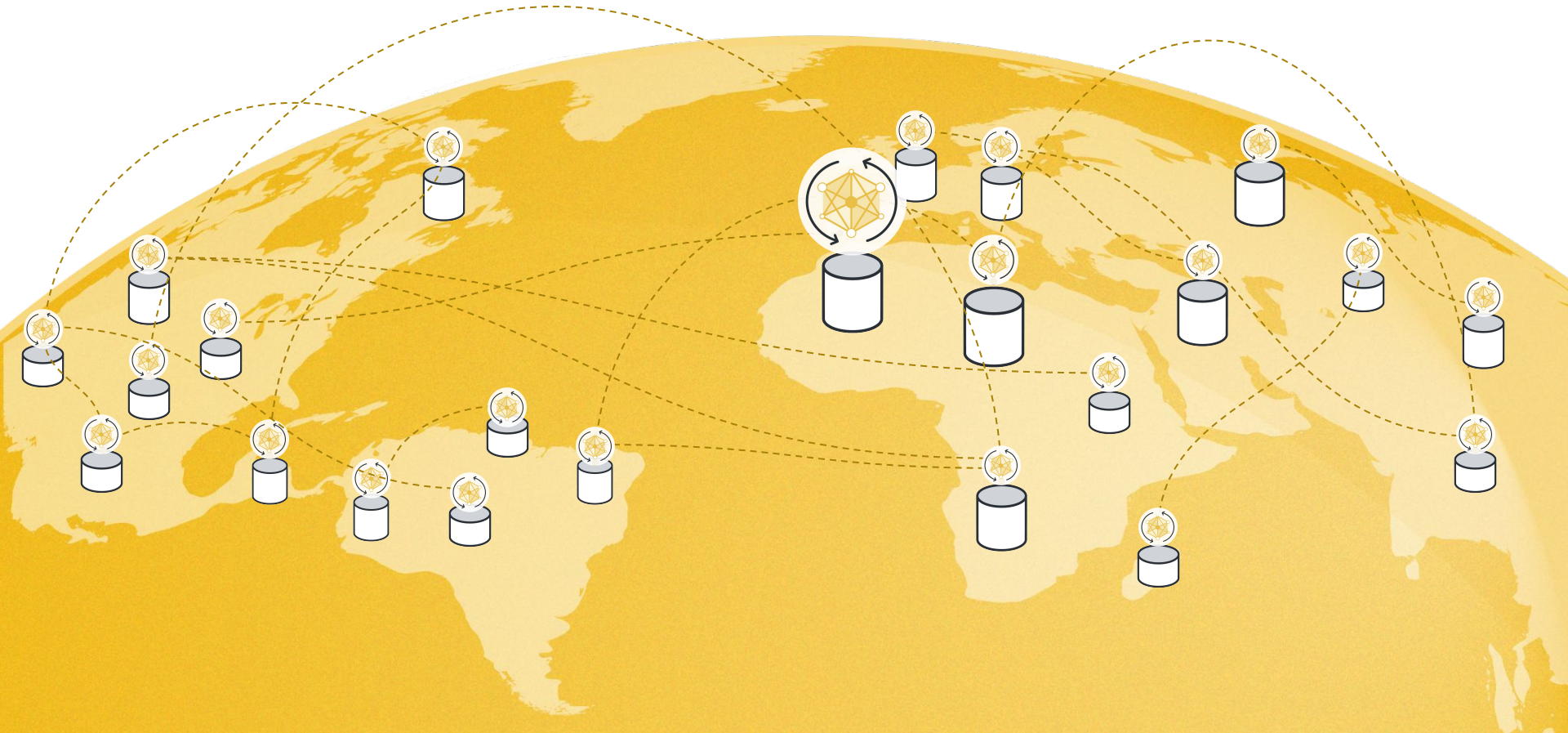


FlowerLLM

World's 1st 3B LLM pre-trained via Federated Learning

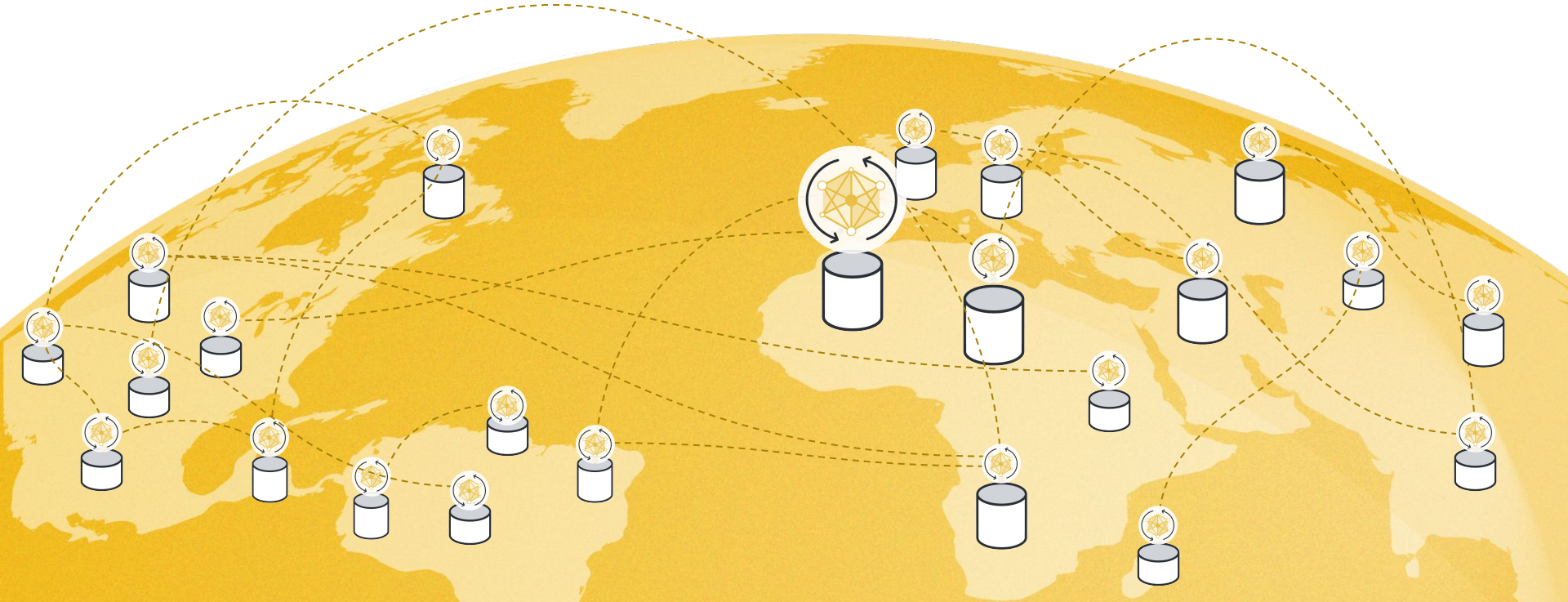
FlowerLLM

CaMLSys http://mlsys.cst.cam.ac.uk



FlowerLLM

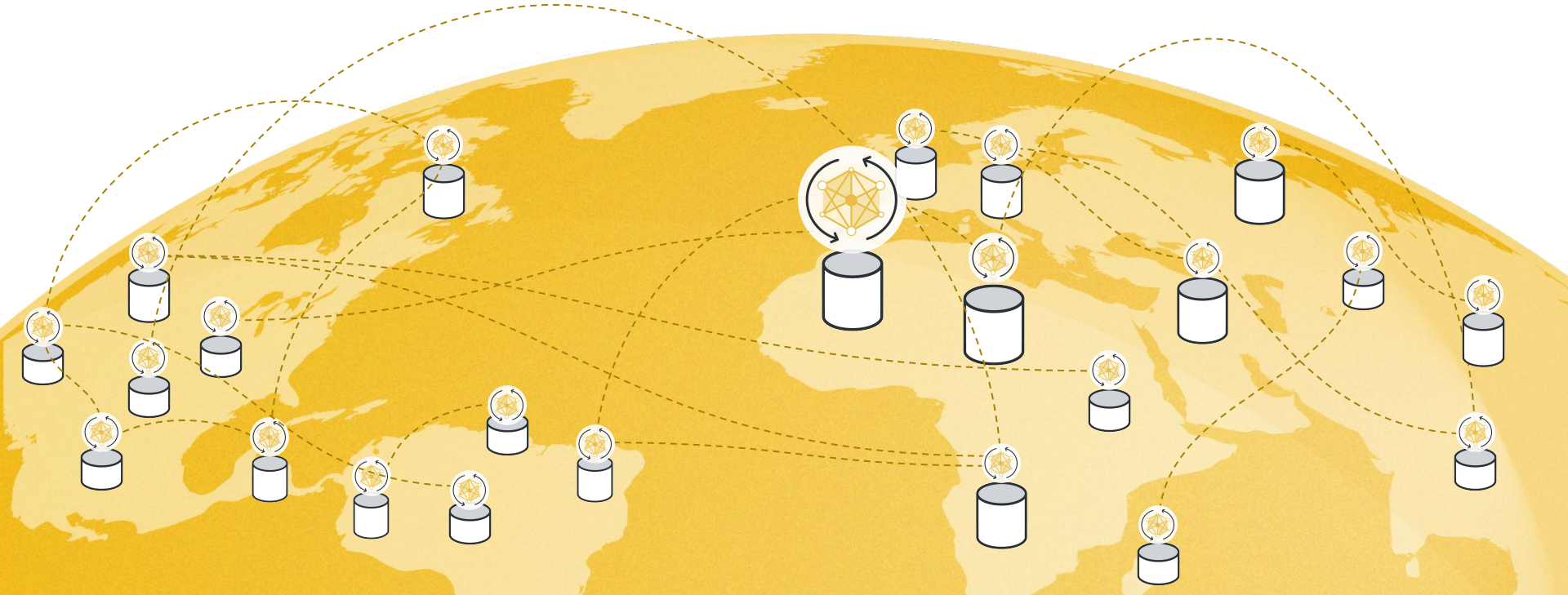
CaMLSys http://mlsys.cst.cam.ac.uk



(1) Democratizing LLMs

FlowerLLM

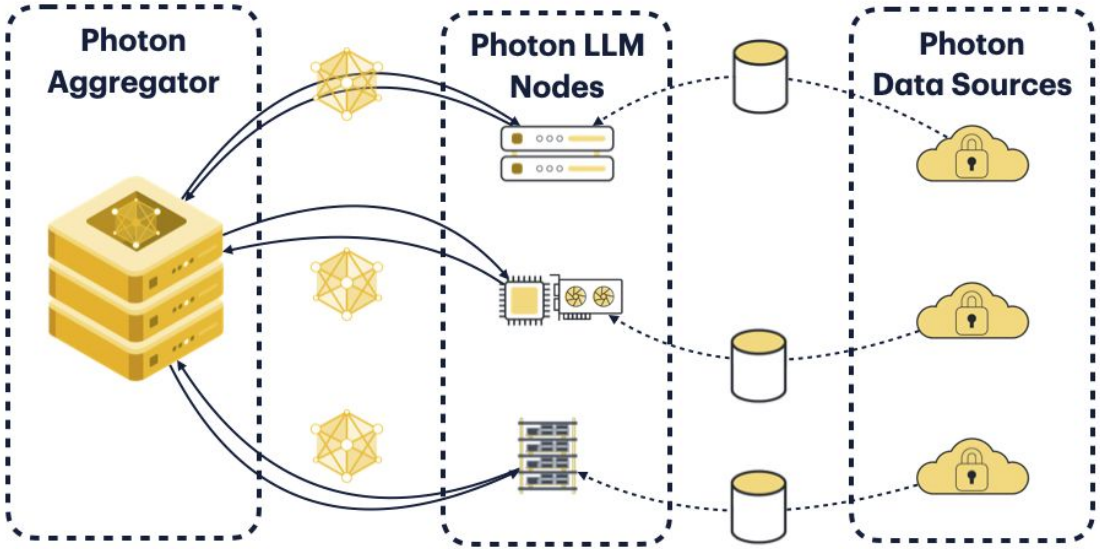
CaMLSys http://mlsys.cst.cam.ac.uk



(1) Democratizing LLMs **(2) Stronger LLMs**


2.5X
FlowerLLM
3.0B
Google DeepMind
1.2B

Photon Architecture and Design



Algorithm 1 *Photon* execution pipeline

Require: Number of rounds T, training population P, number of clients per round K, hyperparameters H

procedure PHOTONSERVER(T, K, H, P)
  $\theta^0 \leftarrow$ InitModel(H) ▷ Init on the server or sample a client and extract model weights
  for each round $t = 1, 2, 3, \ldots, T$ **do**
    $\mathcal{C} \sim \mathcal{U}(P, K)$ ▷ Sample K clients at random from the population
    for $k \in \mathcal{C}$ **do in parallel** ▷ Each sampled client in parallel executes the local training
      $\theta_k^t, \mathcal{M}_k^t \leftarrow$ PHOTONCLIENT(k, θ^t, H)
      $\Delta_k^t \leftarrow \theta^t - \theta_k^t$
    $\Delta^t \leftarrow \frac{1}{|\mathcal{C}|} \sum_{k \in \mathcal{C}} \Delta_k^t$ ▷ Aggregate pseudo-gradients Δ_k^t from clients
    $\theta^{t+1} \leftarrow$ ServerOpt($\theta^t, -\Delta^t, t$) ▷ Apply pseudo-gradient to the global model
    $\mathcal{M}_k^{t+1} \leftarrow$ AggMetrics($\mathcal{M}_k^t \; \forall k \in \mathcal{C}$) ▷ Aggregate metrics across clients
    Checkpoint(θ_k^{t+1}) ▷ Checkpoint model
  return θ_k^{T+1}

procedure PHOTONCLIENT(k, θ^t, H)
  $\mathcal{D}_k \leftarrow$ BindStream(k) ▷ Bind *Photon Data Sources* to a merged data stream $\mathcal{D}_k$
  $I_k \leftarrow$ GetNodes(k) ▷ Extract hardware configuration I_k
  if HasInfiniband(I_k) **then**
    $B_k \leftarrow$ CalcBatchSize(I_k) ▷ Binary search for batch size B_k with static initial guess
    $\theta_k^t, \mathcal{M}_k^t \leftarrow$ TrainClient($\theta^t, \mathcal{D}_k, B_k, H$) ▷ Use DDP or FSDP based on model size
  else
    for node $i \in I$ **do in parallel** ▷ In every node of the current client do FL
      $B_k^i \leftarrow$ CalcBatchSize(i, I_k)
      $\mathcal{D}_k^i \leftarrow$ PartitionStream($i, \mathcal{D}_k$) ▷ Split the client data into $|I|$ shards
      $\theta_k^i, \mathcal{M}_k^i \leftarrow$ TrainClient($\theta^t, \mathcal{D}_k^i, B_k^i, H$) ▷ Use DDP or FSDP based on model size.
    $\theta_k^t \leftarrow \frac{1}{|I|} \sum_{i \in I} \theta_i^t$ ▷ Partially aggregate node models
    $\mathcal{M}_k^t \leftarrow$ AggMetrics($\mathcal{M}_k^t \forall i \in I_k$) ▷ Partially aggregate metrics across nodes
  Checkpoint($\theta_k^t, \mathcal{D}_k$) ▷ Checkpoint model and dataset state
  $\theta_k^t \leftarrow$ PostProcess($\theta_k^t, \mathcal{M}_k^t$) ▷ E.g., apply differential privacy or compress the model
  return $\theta_k^t, \mathcal{M}_k^t$

The Future of Large Language Model Pre-training is Federated

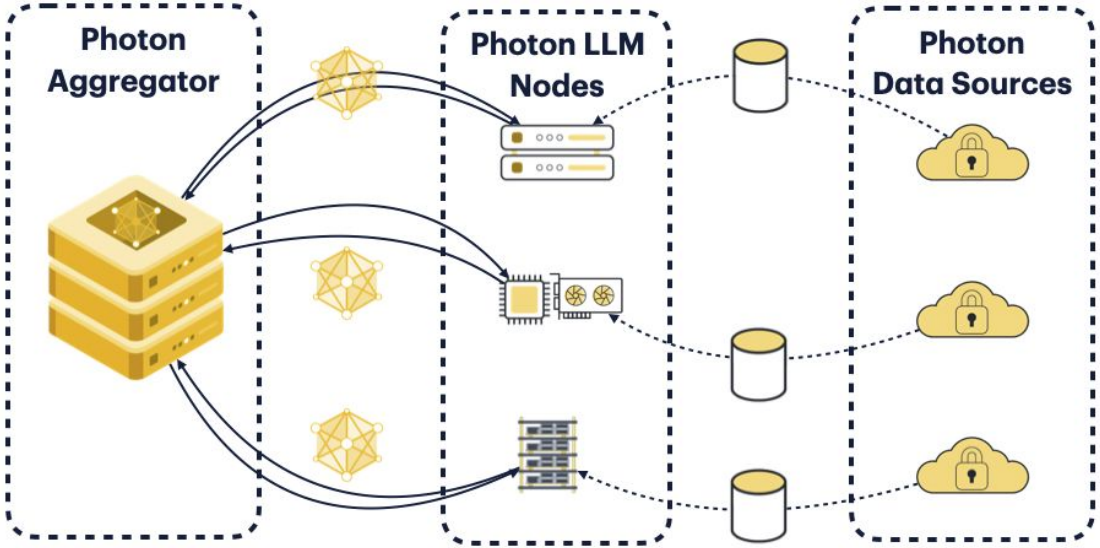
Lorenzo Sani[1,2,†] Alex Iacob[1,2] Zeyu Cao[1,*] Bill Marino[1,*]
Yan Gao[1,2] Tomas Paulik[1] Wanru Zhao[1] William F. Shen[1]
Preslav Aleksandrov[1] Xinchi Qiu[1] Nicholas D. Lane[1,2]

Abstract

Generative pre-trained large language models (LLMs) have demonstrated impressive performance over a wide range of tasks, thanks to the unprecedented amount of data they have been trained on. As established scaling laws indicate, LLMs' future performance improvement depends on the amount of computing and data sources we can leverage for pre-training. Federated learning (FL) has the potential to unleash the majority of the planet's data and computational resources, which are underutilized by the data-center-focused training methodology of current LLM practice. Our work presents a robust, flexible, reproducible FL approach that enables large-scale collaboration across institutions to train LLMs. This would mobilize more computational and data resources while matching or potentially exceeding centralized performance. We further show the effectiveness of the federated training scales with model size and present our approach for training a billion-scale federated LLM using limited resources. This will help data-rich actors to become the protagonists of LLMs pre-training instead of leaving the stage to compute-rich actors alone.

https://arxiv.org/abs/2405.10853

Photon Architecture and Design



https://arxiv.org/abs/2405.10853

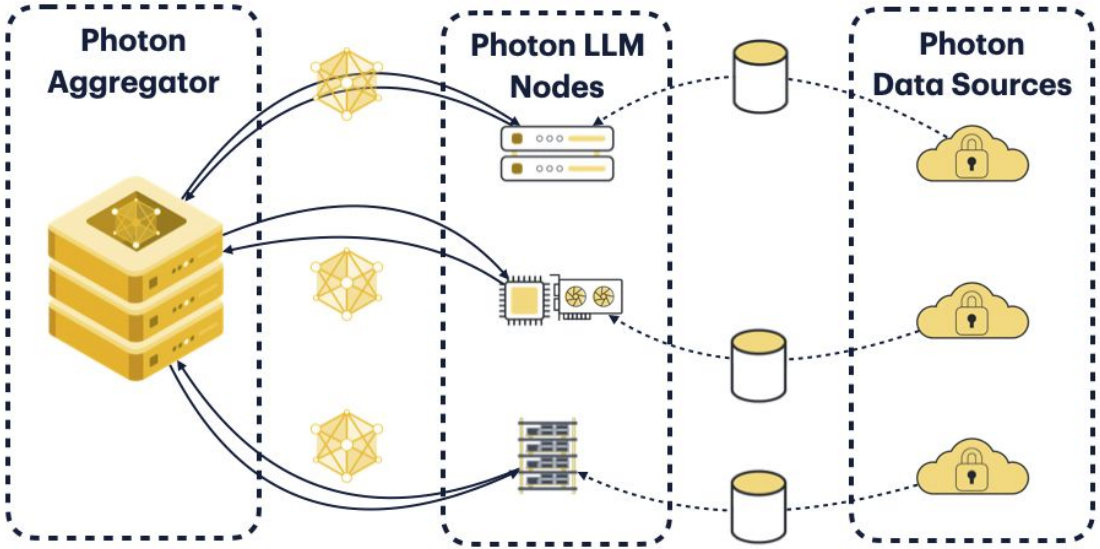
The Future of Large Language Model Pre-training is Federated

Lorenzo Sani[1,2,†] Alex Iacob[1,2] Zeyu Cao[1,*] Bill Marino[1,*]
Yan Gao[1,2] Tomas Paulik[1] Wanru Zhao[1] William F. Shen[1]
Preslav Aleksandrov[1] Xinchi Qiu[1] Nicholas D. Lane[1,2]

Abstract

Generative pre-trained large language models (LLMs) have demonstrated impressive performance over a wide range of tasks, thanks to the unprecedented amount of data they have been trained on. As established scaling laws indicate, LLMs' future performance improvement depends on the amount of computing and data sources we can leverage for pre-training. Federated learning (FL) has the potential to unleash the majority of the planet's data and computational resources, which are underutilized by the data-center-focused training methodology of current LLM practice. Our work presents a robust, flexible, reproducible FL approach that enables large-scale collaboration across institutions to train LLMs. This would mobilize more computational and data resources while matching or potentially exceeding centralized performance. We further show the effectiveness of the federated training scales with model size and present our approach for training a billion-scale federated LLM using limited resources. This will help data-rich actors to become the protagonists of LLMs pre-training instead of leaving the stage to compute-rich actors alone.

Photon Architecture and Design

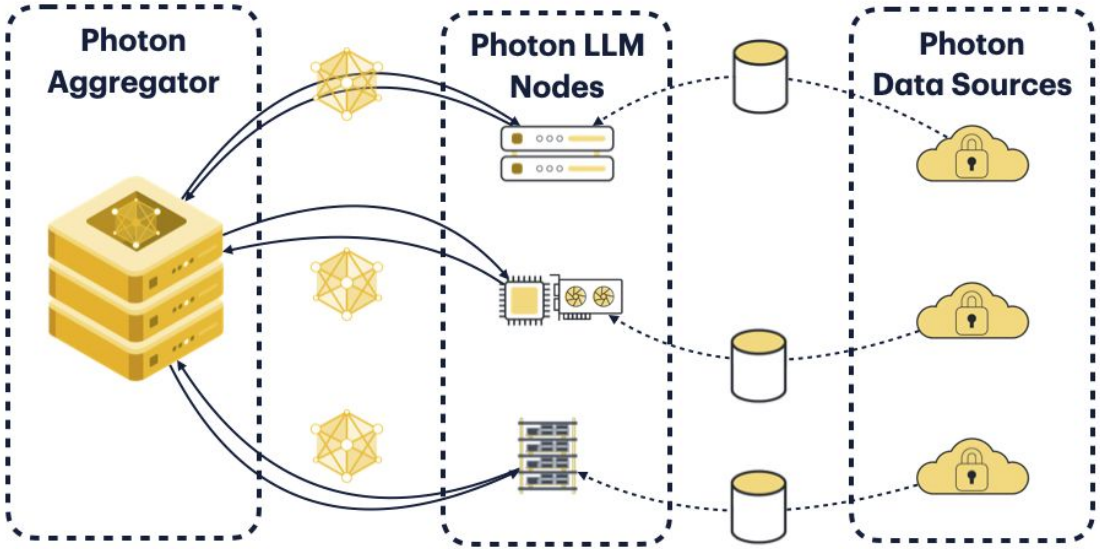


Design Principles

- Separation of Data and Computation *(broad access to data)*
- Limited Comms Reqs.
- Broad H/W Inclusivity
- Scalable Tuned Local Training Pipelines
- Private and Public Data

https://arxiv.org/abs/2405.10853

Photon Architecture and Design



Key Techniques

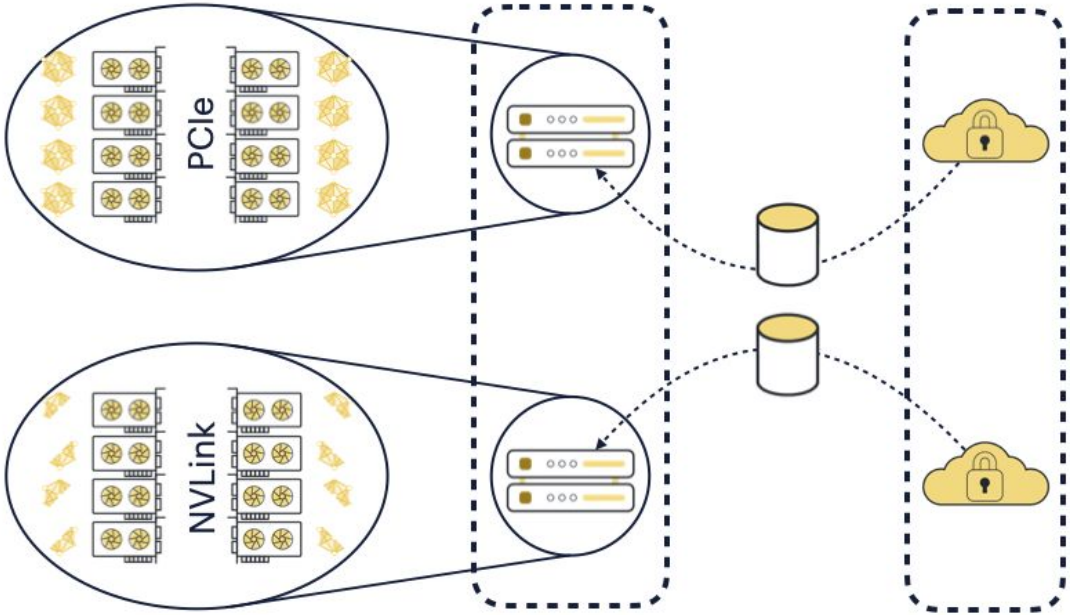
- Adaptive Comms and Optimizer to Conditions
- Diversity for Efficiency
- Comms/Compute for Massive Model Updates
- Built on experimental (for now) versions of Pollen and Flower

https://arxiv.org/abs/2405.10853

Photon: High Utilization Independent of Internal Topology

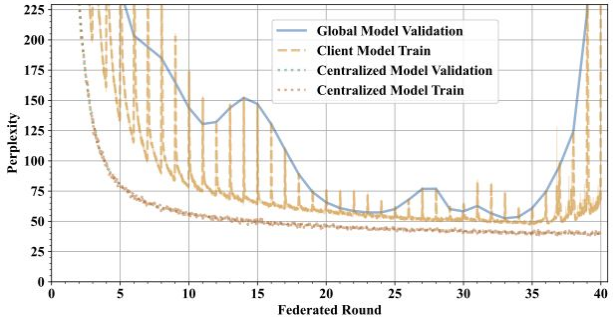
DDP *or*
Independent
Training + local-FL

Zero *to* Zero-4

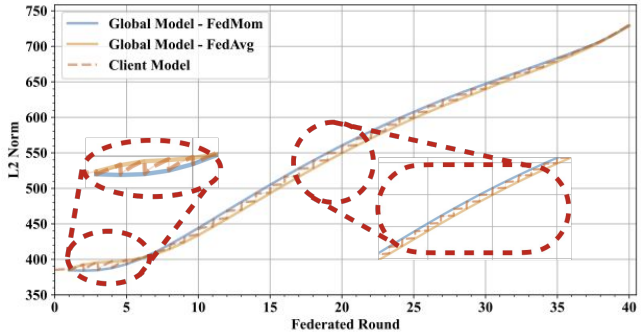


https://arxiv.org/abs/2405.10853

Photon: Coping with Global Optimizer Instability



Aggregated model "blow-up"

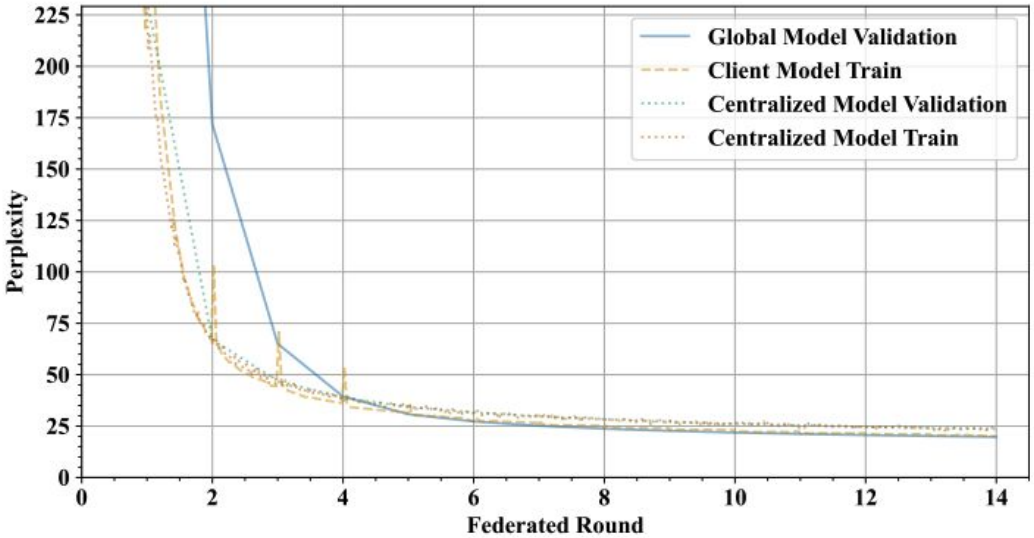


Approach

- Tracking global model norm after aggregation
- Adjustment server learning rate, as required

https://arxiv.org/abs/2405.10853

Federated Pre-training of a **1.3B LLM**



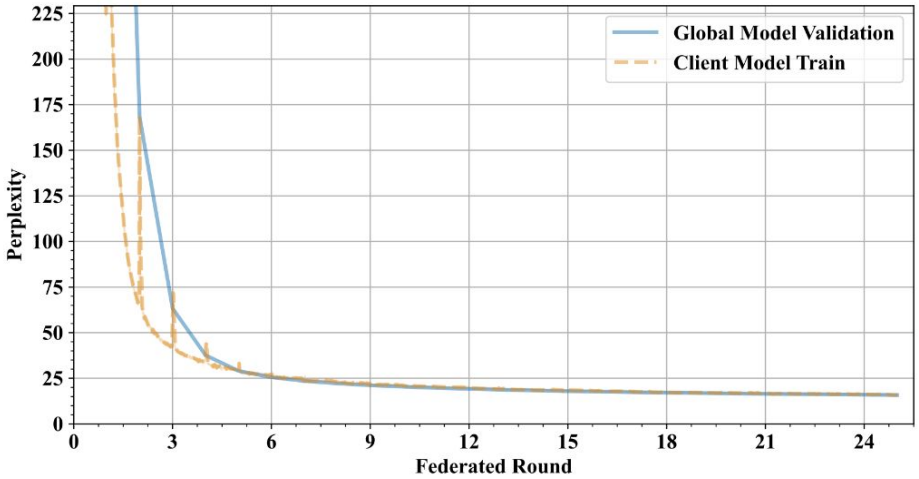
- IID C4 dataset
- FedMom + AdamW
- 16 A100s total
- 4 physical location
- 100Mbps client links

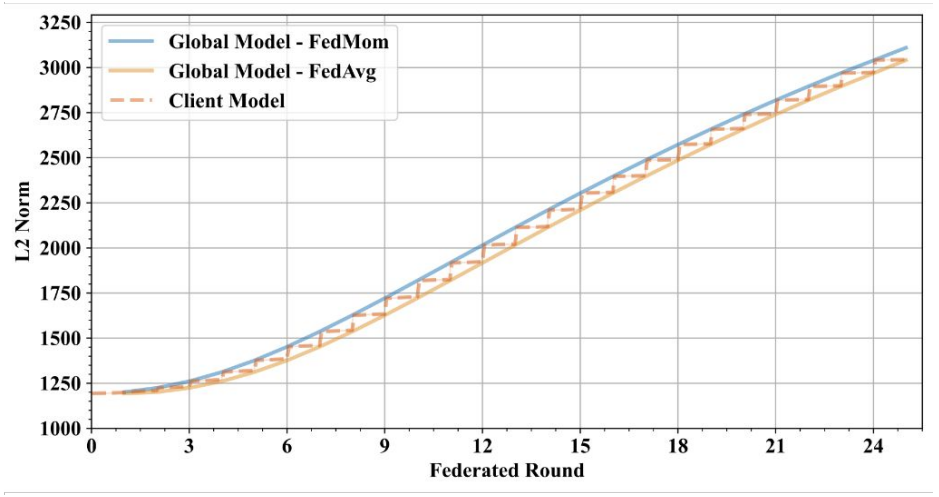
Model Size	#Blocks	d	#Heads	Exp. Ratio	(β_1, β_2)	|Vocab|
75M	3	896	16	4	(0.9, 0.95)	50 368
125M	12	768	12	4	(0.9, 0.95)	50 368
350M	24	1024	16	4	(0.9, 0.95)	50 368
1.3B	24	2048	16	4	(0.9, 0.95)	50 368

#Rounds	η_s	μ_s	α	η_{max}	T	Batch Size
40	0.2	0.9	10^{-6}	4×10^{-4}	88 000	256
25	0.2	0.9	10^{-5}	3×10^{-4}	15 000	256
40	0.2	0.9	10^{-1}	3×10^{-4}	13 400	256
16	0.2	0.9	10^{-1}	2×10^{-4}	24 800	512

https://arxiv.org/abs/2405.10853

Scaling up to Pre-training a **3B LLM**



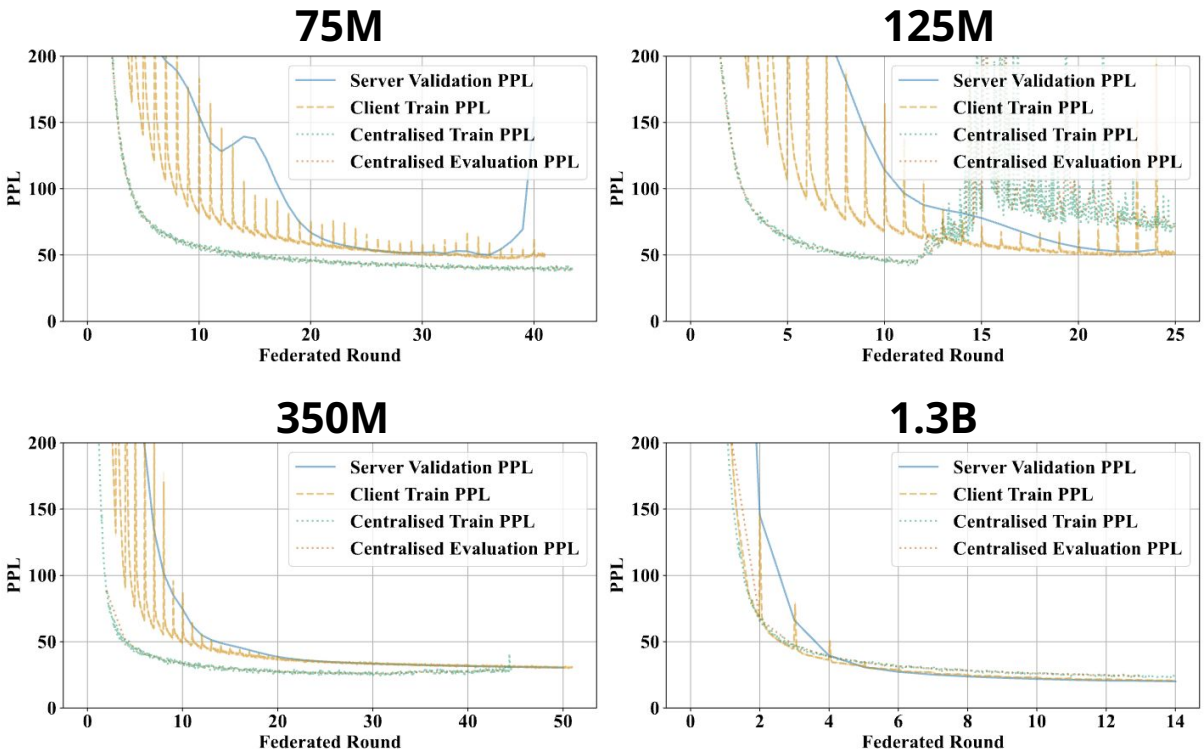


C4 again, switched to H100s with similar physical topology

https://arxiv.org/abs/2405.10853

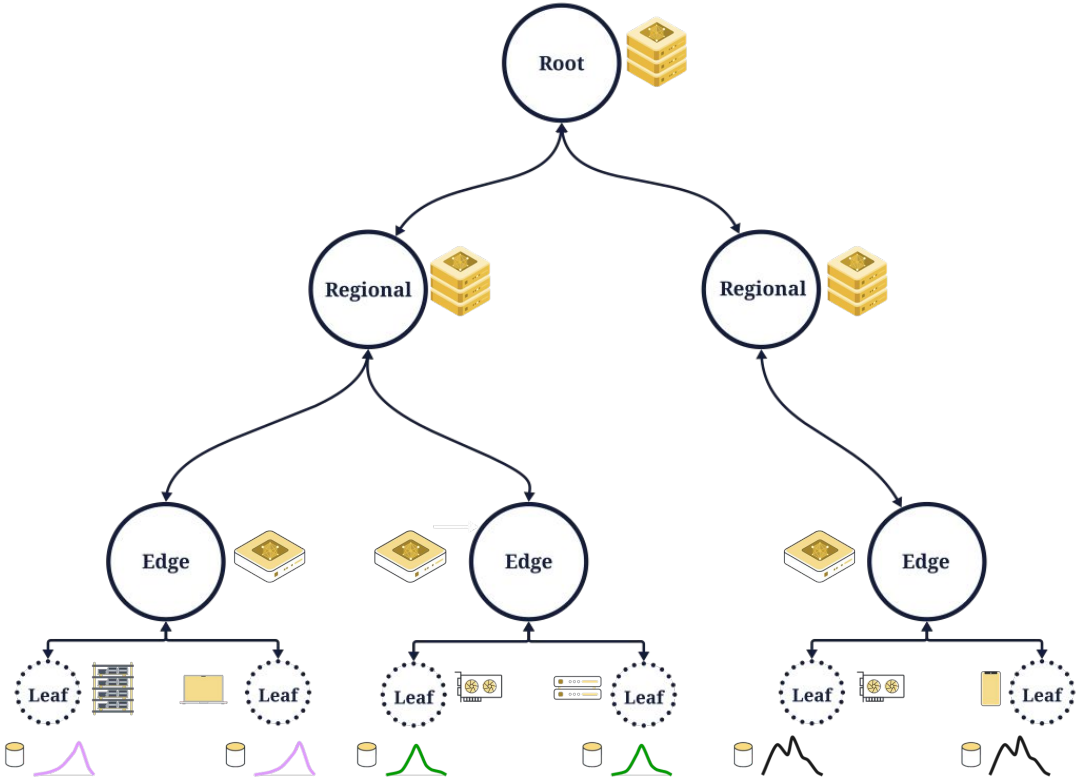
Larger Models Appears Improves Global Convergence

- Preliminary Observation
- Perhaps indicative of many future empirical findings to build on…



https://arxiv.org/abs/2405.10853

Worldwide Federated Language Model (WorldLM) – *extending Photon*



Worldwide Federated Training of Language Models

Alex Iacob[1,2] Lorenzo Sani[1,2]
Bill Marino[1] Preslav Aleksandrov[1] William F. Shen[1] Nicholas D. Lane[1,2]

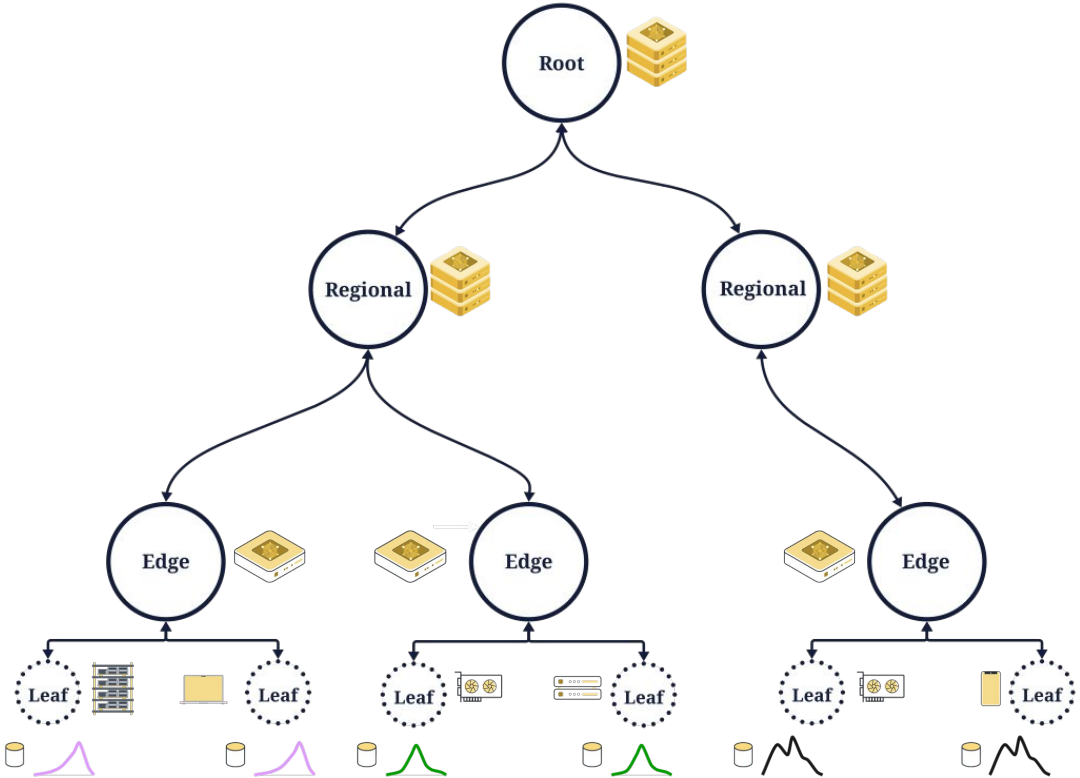
Abstract

The reliance of language model training on massive amounts of computation and vast datasets scraped from potentially low-quality, copyrighted, or sensitive data has come into question practically, legally, and ethically. Federated learning provides a plausible alternative by enabling previously untapped data to be voluntarily gathered from collaborating organizations. However, when scaled globally, federated learning requires collaboration across heterogeneous legal, security, and privacy regimes while accounting for the inherent locality of language data; this further exacerbates the established challenge of federated statistical heterogeneity. We propose a Worldwide Federated Language Model Training (*WorldLM*) system based on **federations of federations**, where each federation has the autonomy to account for factors such as its industry, operating jurisdiction, or competitive environment. *WorldLM* enables such autonomy in the presence of statistical heterogeneity via partial model localization by allowing sub-federations to attentively aggregate key layers from their constituents. Furthermore, it can adaptively share information across federations via residual layer embeddings. Evaluations of language modeling on naturally heterogeneous datasets show that *WorldLM* outperforms standard federations by up to 1.91×, approaches the personalized performance of fully local models, and maintains these advantages under privacy-enhancing techniques.

arXiv:2405.14446v2 [cs.LG] 27 May 2024

https://arxiv.org/pdf/2405.14446

Worldwide Federated Language Model (WorldLM) – *extending Photon*



Algorithm 1 Fit: execution procedure for a given sub-federation.

Require: Node id q, parent backbone $\mathcal{B}_p$, sequence of key layers $\mathcal{K}_p$
Require: Downstream residuals for aggregation $\mathcal{D}_a$ for routing $\mathcal{D}_r^p$

$\mathcal{B}^0, \mathcal{K}^0 \leftarrow \text{LoadModel}(q)$
$\mathcal{U}^0 \leftarrow \emptyset$ ▷ Upstream residuals for parent
if $q \neq 0$ **then** ▷ If the current node is not root
  $\mathcal{Q}, \mathcal{K}, \mathcal{V} \leftarrow \mathcal{K}^0, [\mathcal{K}^0, \mathcal{K}_p, \mathcal{D}_a], [\mathcal{K}^0, \mathcal{K}_p, \mathcal{D}_a]$
  $\mathcal{B}^0 \leftarrow \mathcal{B}_p$
  $\mathcal{K}^0 \leftarrow \text{Attn}(\mathcal{Q}, \mathcal{K}, \mathcal{V})$ ▷ Agg. node, parent and residual key layers
for round $k \leftarrow 0, \ldots, K-1$ **do**
  $A^k, R^k \leftarrow \text{RouteResiduals}(\mathcal{D}_r^k, C)$ ▷ Route using similarity to cached children $\mathcal{K}$
  $\text{Train}(q, \mathcal{B}^k, \mathcal{K}^k)$ ▷ Train node params sequentially or in-parallel
  for child $c \in C_q$ **do** ▷ In parallel
    $\mathcal{B}_c^k, \mathcal{K}_c^k, \mathcal{U}_c^k \leftarrow \text{Fit}(c, \mathcal{B}^k, \mathcal{K}^k, A_c^k, R_c^k)$ ▷ $\mathcal{U}_c^k$ are residuals sent upstream by the child
    $\Delta_c^k \leftarrow \mathcal{B}_c^k - \mathcal{B}$ ▷ Compute backbone pseudo-gradient
  if $C_q \neq \emptyset$ **then**
    $\Delta^k \leftarrow \frac{1}{|C|} \sum_{c \in C} \Delta_c^k$ ▷ Aggregate pseudo-gradients
    $\mathcal{B}^{k+1} \leftarrow \text{ServerOpt}(\mathcal{B}^k, -\Delta^k, k)$ ▷ Apply pseudo-gradient
    $\mathcal{Q}, \mathcal{K}, \mathcal{V} \leftarrow [\mathcal{K}_0^k, \ldots, \mathcal{K}_{|C|}^k], \ldots, [\mathcal{K}_0^k, \ldots, \mathcal{K}_{|C|}^k]$ ▷ Pack key layers with the same index
    $\mathcal{K}^{k+1} \leftarrow \text{Attn}(\mathcal{Q}, \mathcal{K}, \mathcal{V})$
    $\mathcal{U}^{k+1}, \mathcal{D}_r^{k+1} \leftarrow \text{PartitionResiduals}(q, \mathcal{K}^{k+1}, V, \mathcal{U}^k, \mathcal{D}_r^k)$ ▷ Based on dissimilarity
return $\mathcal{B}^k, \mathcal{K}^k, \mathcal{U}^k$

Worldwide Federated Training of Language Models

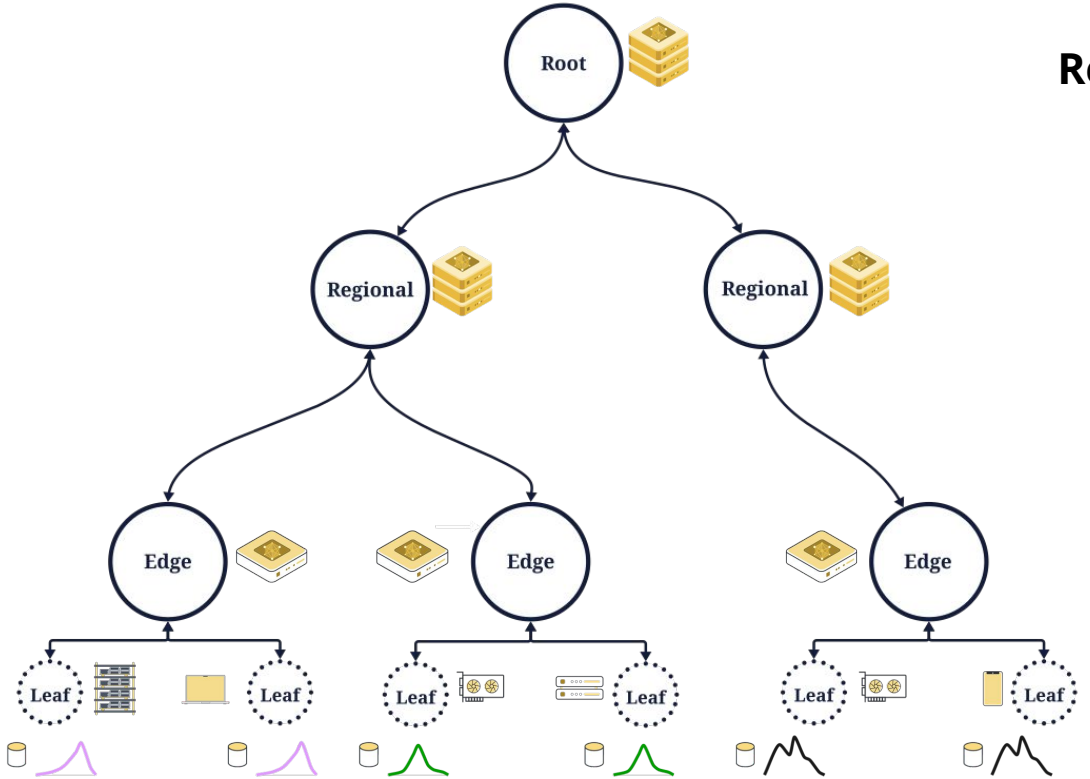
Alex Iacob[1,2] Lorenzo Sani[1,2]
Bill Marino[1] Preslav Aleksandrov[1] William F. Shen[1] Nicholas D. Lane[1,2]

Abstract

The reliance of language model training on massive amounts of computation and vast datasets scraped from potentially low-quality, copyrighted, or sensitive data has come into question practically, legally, and ethically. Federated learning provides a plausible alternative by enabling previously untapped data to be voluntarily gathered from collaborating organizations. However, when scaled globally, federated learning requires collaboration across heterogeneous legal, security, and privacy regimes while accounting for the inherent locality of language data; this further exacerbates the established challenge of federated statistical heterogeneity. We propose a Worldwide Federated Language Model Training (*WorldLM*) system based on **federations of federations**, where each federation has the autonomy to account for factors such as its industry, operating jurisdiction, or competitive environment. *WorldLM* enables such autonomy in the presence of statistical heterogeneity via partial model localization by allowing sub-federations to attentively aggregate key layers from their constituents. Furthermore, it can adaptively share information across federations via residual layer embeddings. Evaluations of language modeling on naturally heterogeneous datasets show that *WorldLM* outperforms standard federations by up to 1.91×, approaches the personalized performance of fully local models, and maintains these advantages under privacy-enhancing techniques.

https://arxiv.org/pdf/2405.14446

Worldwide Federated Language Model (WorldLM) – *extending Photon*

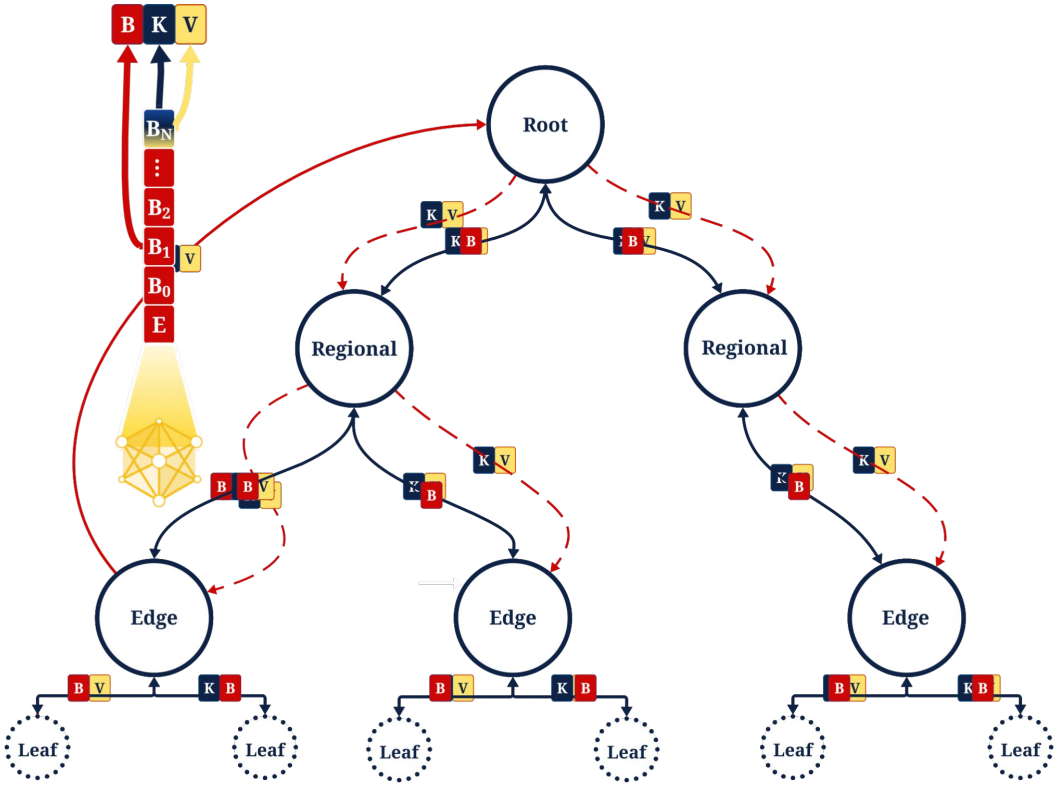


Regional Governance Silos

- Legal
- Privacy
- Security
- Legacy FL or AI systems

https://arxiv.org/pdf/2405.14446

Worldwide Federated Language Model (WorldLM) – *extending Photon*



Key Techniques

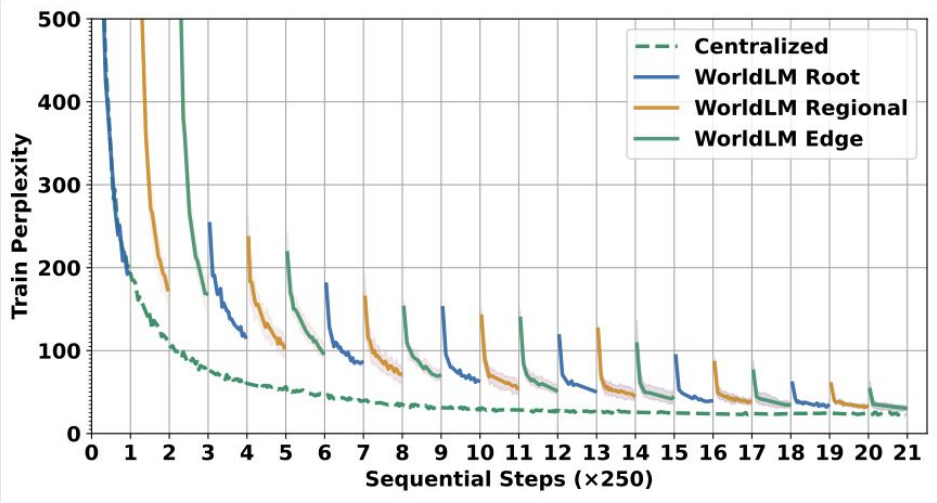
- Federation of Federations
- Hierarchy w/ data driven routing
- Scalable Differential Privacy

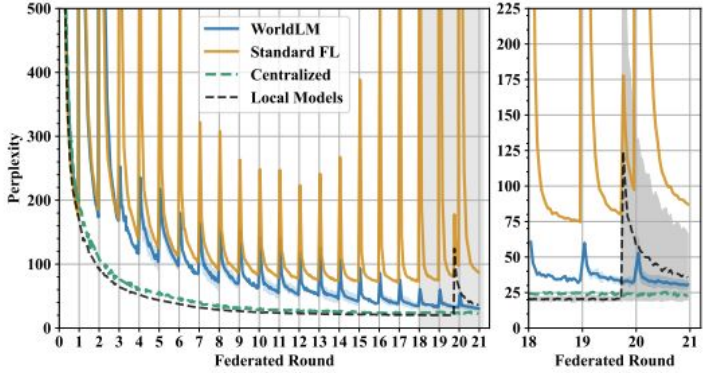
https://arxiv.org/pdf/2405.14446

Comparisons to natural alternatives

Dataset	Model	*WorldLM*	FL	Centralized
Pile	75M	**73.82 ± 44.18**	107.31 ± 52.50	85.81 ± 24.42
Pile	125M	**48.34 ± 32.41**	53.92 ± 24.24	29.61 ± 13.17
MC4	250M	**80.47 ± 68.53**	153.27 ± 95.47	72.21 ± 49.78
C4	75M	167.31 ± 2.92	**145.32 ± 3.53**	67.01 ± 1.67

- Size: 75M to 250M
- Three datasets
  - The Pile
  - MC4 and C4

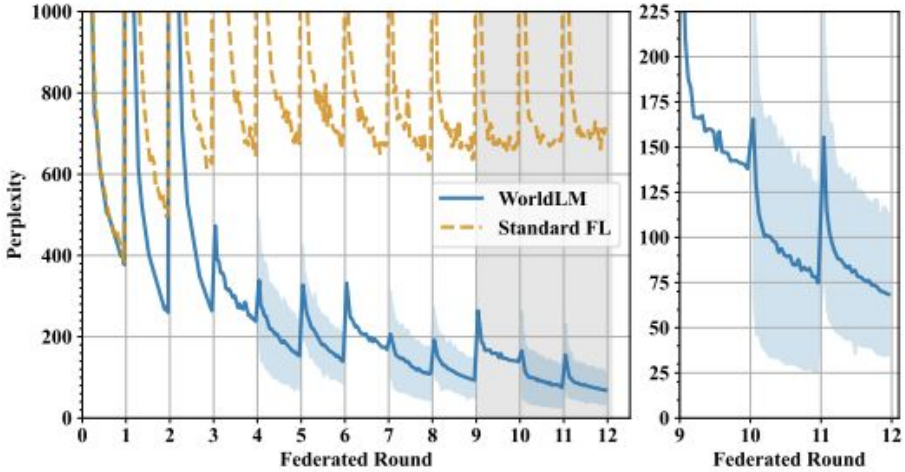




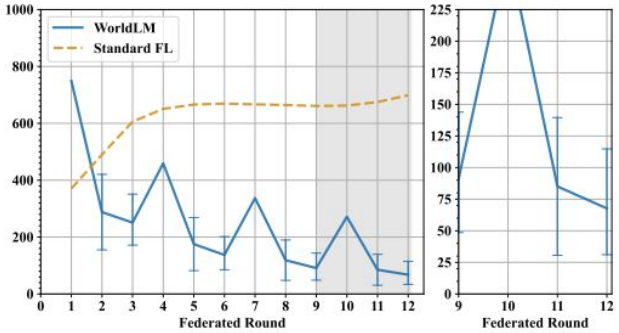
https://arxiv.org/pdf/2405.14446

Enabling Differential Privacy

Method	Pile	$DP_{CC,WK}$	$DP_{PBC,PBA}$
WorldLM	**73.82 ± 44.18**	**101.78 ± 88.48**	**103.68 ± 90.53**
FL	107.31 ± 52.50	724.56 ± 251.89	724.24 ± 250.98

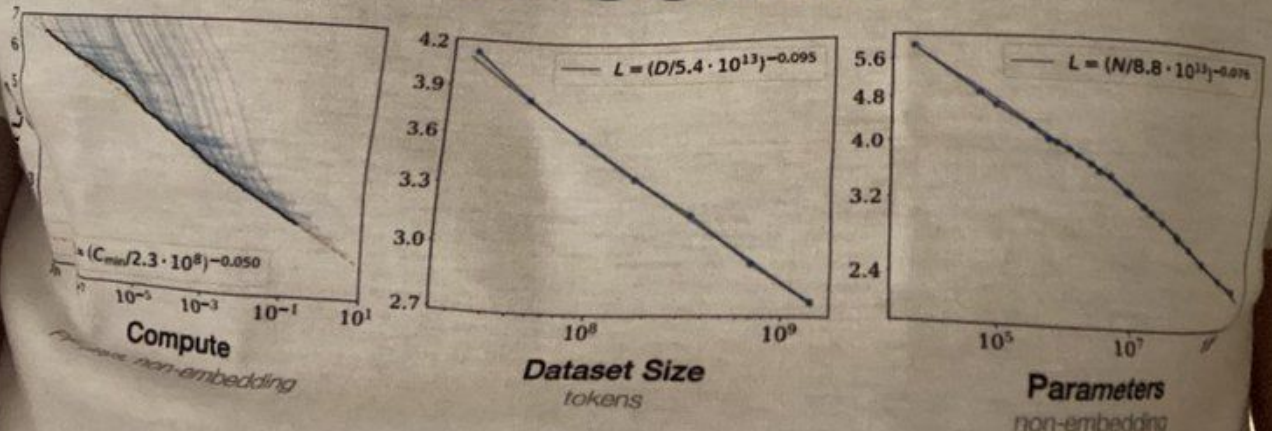


- The Pile variations
  - Common Crawl, Wikipedia, PubMed Central, and PubMed Abstracts
- DP settings local to a region; σ = 0.5

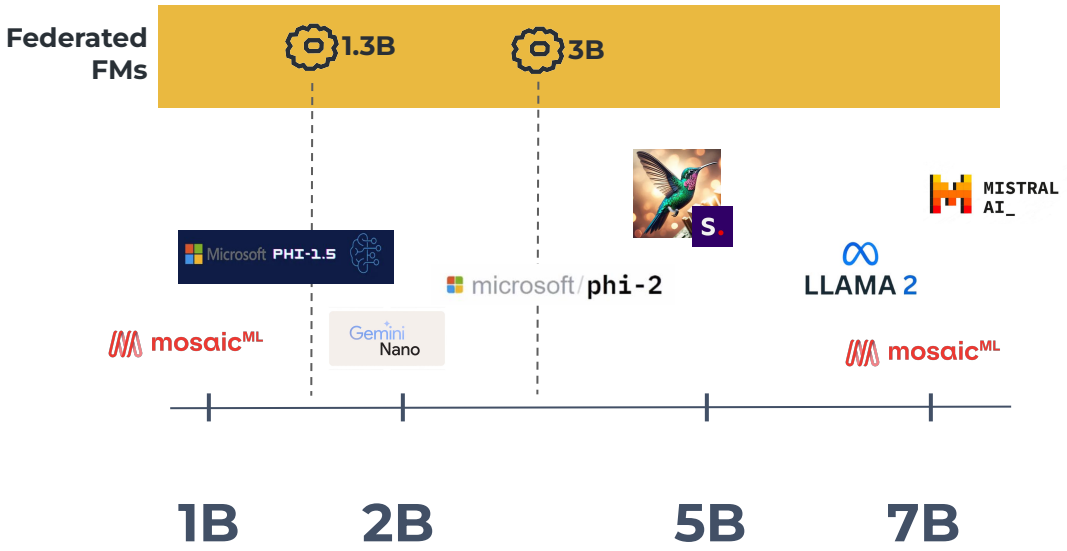


https://arxiv.org/pdf/2405.14446

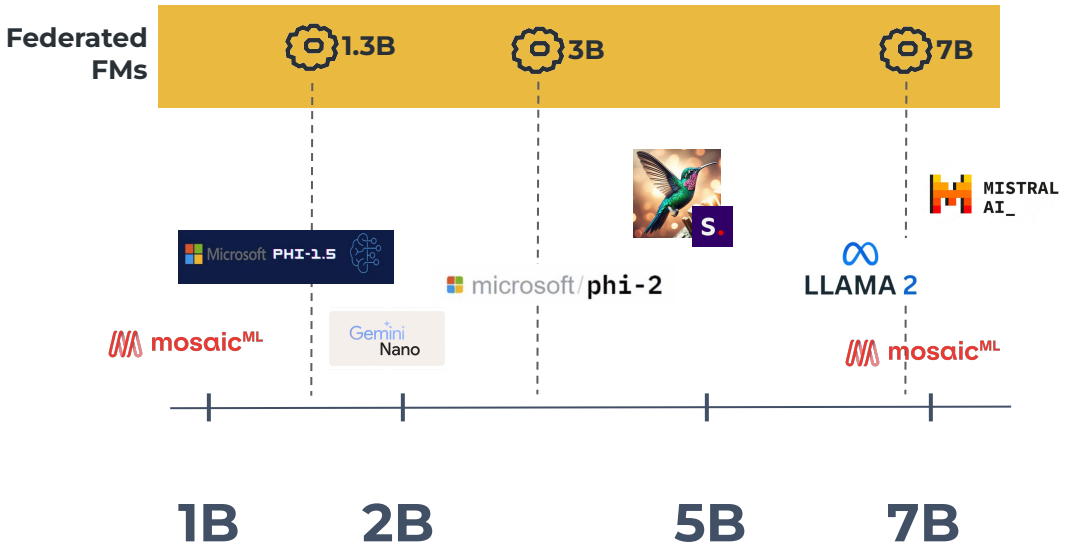
SCALE IS ALL YOU NEED - AGI IS COMING



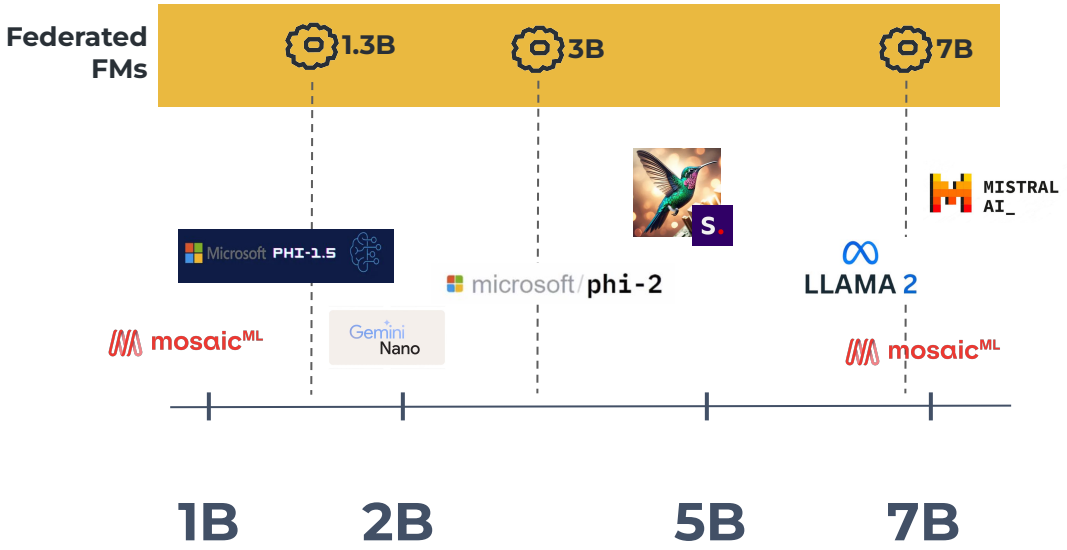
Next steps for FlowerLLM and Photon



Next steps for FlowerLLM and Photon



Next steps for FlowerLLM and Photon

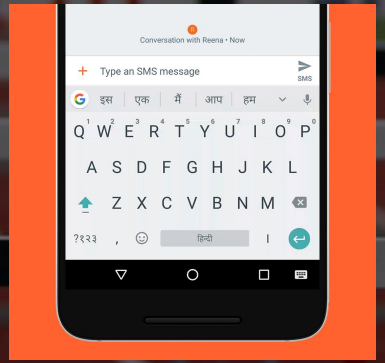


FlowerLLM-30B

GitHub

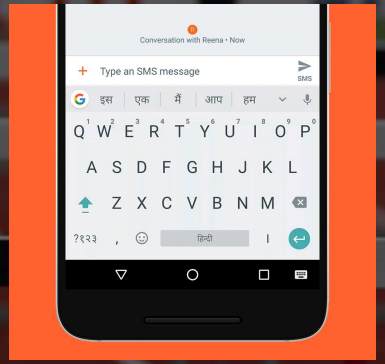
Revisiting the FL research agenda

- Rethink the key important settings and scenarios for FL
- Extreme comms and local compute/memory overhead
- Scaling to large model scales (e.g., FMs, LLMs)
- Coping with heterogeneity (clients/devices)
- Challenges of non-I.I.D data (i.e., data heterogeneity)
- Overly complex and limited tooling
- Primitive MLOps and tuning capabilities
- Gaps in theoretical understanding and/or empirical best-practices
- …



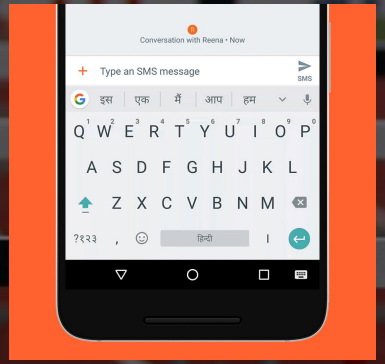
Revisiting the FL research agenda

- Rethink the key important settings and scenarios for FL
- Extreme comms and local compute/memory overhead
- Scaling to large model scales (e.g., FMs, LLMs)
- Coping with heterogeneity (clients/devices)
- Challenges of non-I.I.D data (i.e., data heterogeneity)
- Overly complex and limited tooling
- Primitive MLOps and tuning capabilities
- Gaps in theoretical understanding and/or empirical best-practices
- …



Revisiting the FL research agenda

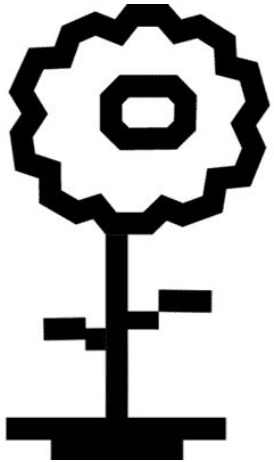
- Rethink the key important settings and scenarios for FL
- Extreme comms and local compute/memory overhead
- Scaling to large model scales (e.g., FMs, LLMs)
- Coping with heterogeneity (clients/devices)
- Challenges of non-I.I.D data (i.e., data heterogeneity)
- Overly complex and limited tooling
- Primitive MLOps and tuning capabilities
- Gaps in theoretical understanding and/or empirical best-practices
- …



Open Questions for HW Design and Design Automation

- Boring but correct answer: Anything for to model training
- Memory and storage hierarchy
  - More of it, energy efficient, and considering training data flow

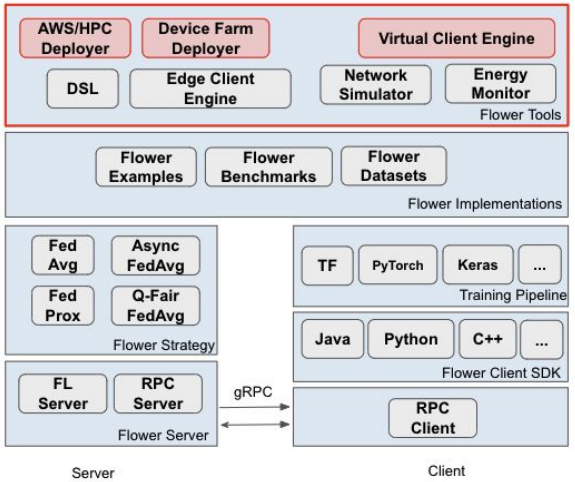
- Hardware accelerated privacy algs (SA, FHE etc)
- Co-processors, bespoke networking/comms, algs in HW, anything goes! The opportunity it indeed that valuable
- ...



Flower

https://flower.ai

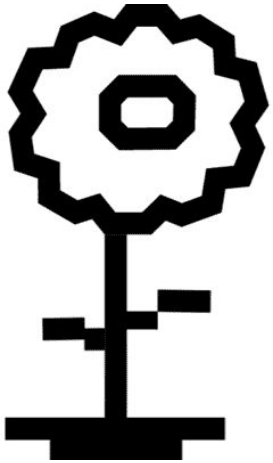
```python
import flwr as fl

class MyClient(fl.KerasClient):
    def __init__(self, model, ds_train, val)
        self.model = model
        self.ds_train = ds_train
        self.ds_val = ds_val

    def get_parameters():
        return model.get_weights()

    def fit(self, weights, config):
        model.set_weights(weights)
        model.fit(ds_train, epochs=config["epochs"])
        return model.get_weights()

    def evaluate(self, weights, config):
        model.set_weights(weights)
        return model.evaluate(ds_test)

server_address, model, ds_train, ds_test = ...
client = MyClient(model, ds_train, ds_test)
fl.app.client.start_client(server_address, client)
```

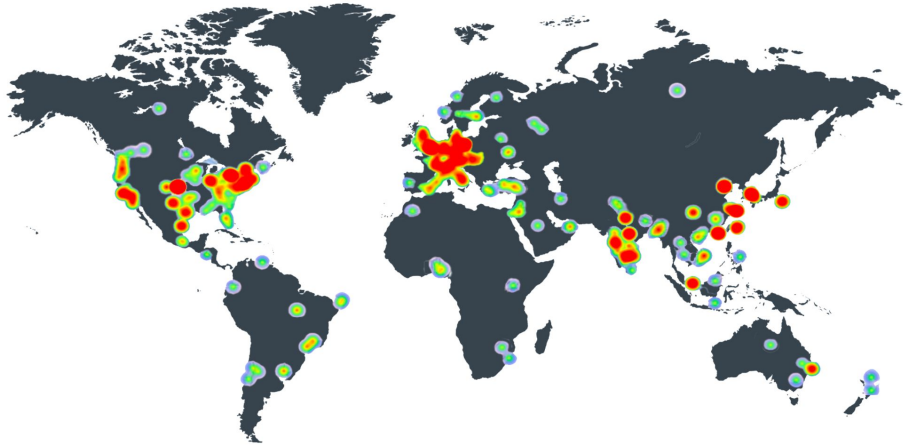


https://arxiv.org/abs/2007.14390
https://arxiv.org/abs/2104.03042
https://arxiv.org/abs/2104.14297
https://arxiv.org/abs/2102.07627
https://arxiv.org/abs/2205.06117
https://arxiv.org/abs/2208.02507
https://arxiv.org/abs/2204.02804
https://arxiv.org/abs/2206.11239
https://arxiv.org/abs/2212.04084

OPEN SOURCE

\+ Community Driven





Scientists + AI Devs ❤️ Flower

4,700+
Stars

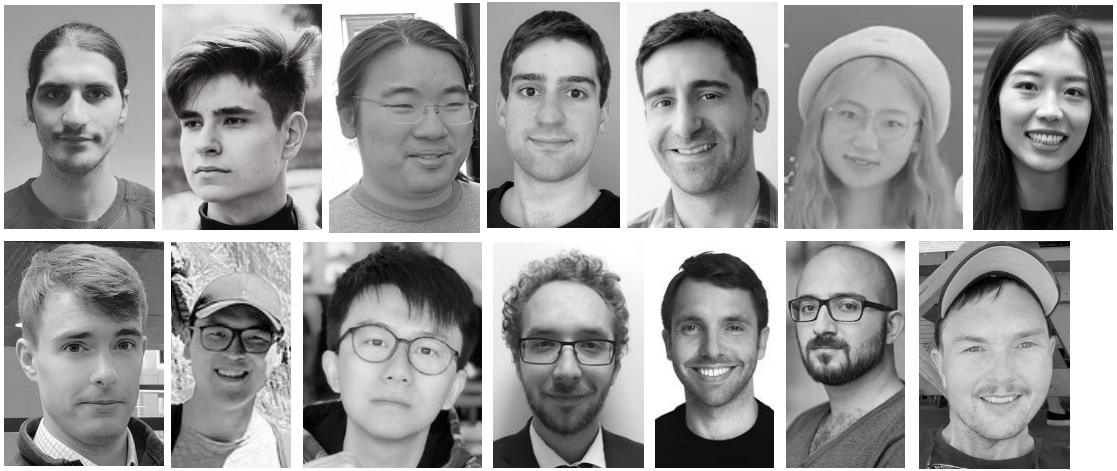
4,200+
Developers

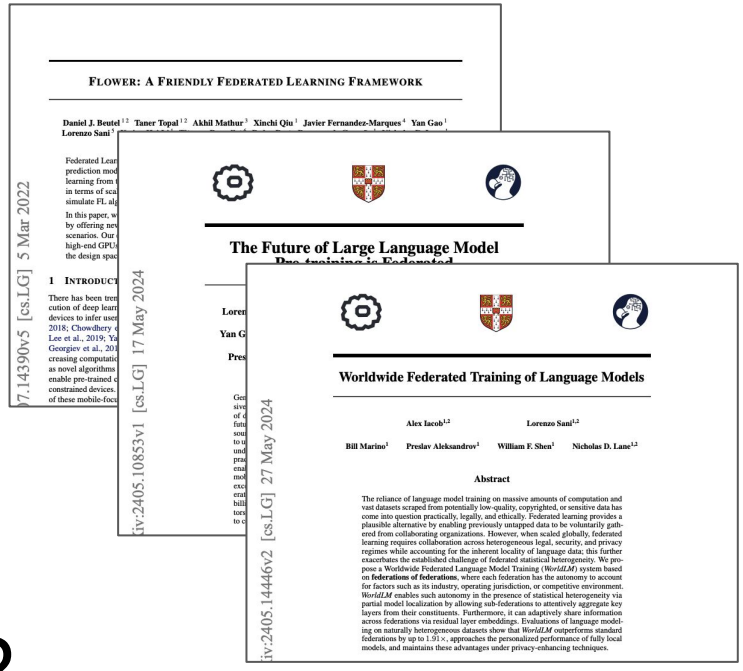
1,500+
Dependents

Flower

https://flower.ai

CaMLSys http://mlsys.cst.cam.ac.uk





Flower

https://flower.ai

Cambridge Machine Learning Systems Lab
mlsys.cst.cam.ac.uk

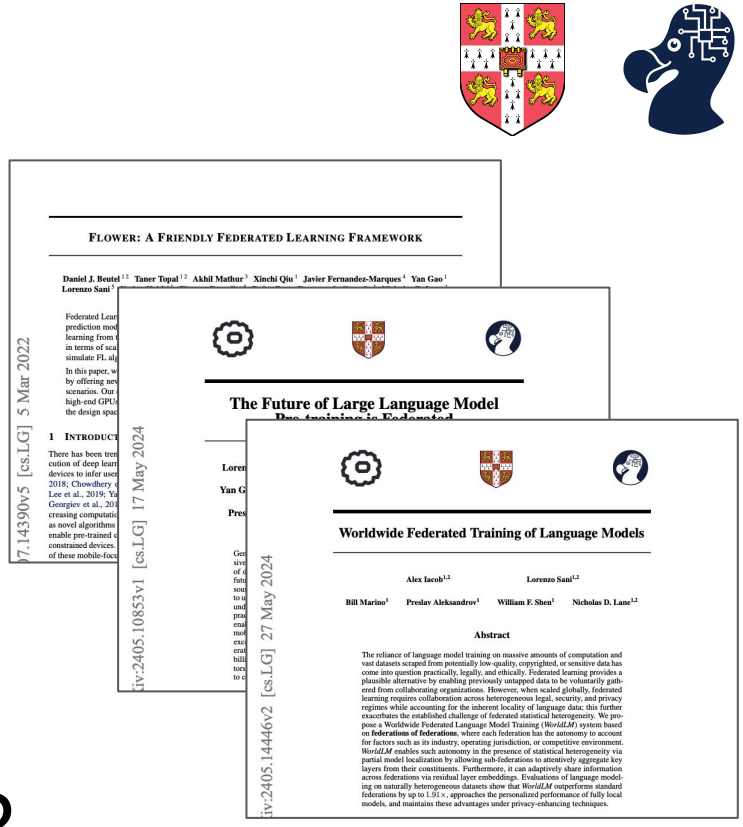
Questions? Comments?

(Would love to hear thoughts on the impact to HW Design or Design Automation Tooling…)

Nicholas D. Lane

University of Cambridge | Flower Labs

@niclane7



FLOWER: A FRIENDLY FEDERATED LEARNING FRAMEWORK

Daniel J. Beutel [1 2] Taner Topal [1 2] Akhil Mathur [3] Xinchi Qiu [1] Javier Fernandez-Marques [4] Yan Gao [1]
Lorenzo Sani [3]

[cs.LG] 5 Mar 2022

The Future of Large Language Model

[cs.LG] 17 May 2024

Worldwide Federated Training of Language Models

Alex Iacob[1,2] Lorenzo Sani[1,2]

Bill Marino[1] Preslav Aleksandrov[1] William F. Shen[1] Nicholas D. Lane[1,2]

Abstract

[cs.LG] 27 May 2024

"In the near future, every SOTA AI model will be trained using federated learning"

Questions? Comments?

(Would love to hear thoughts on the impact to HW Design or Design Automation Tooling…)

Nicholas D. Lane
University of Cambridge | Flower Labs
@niclane7