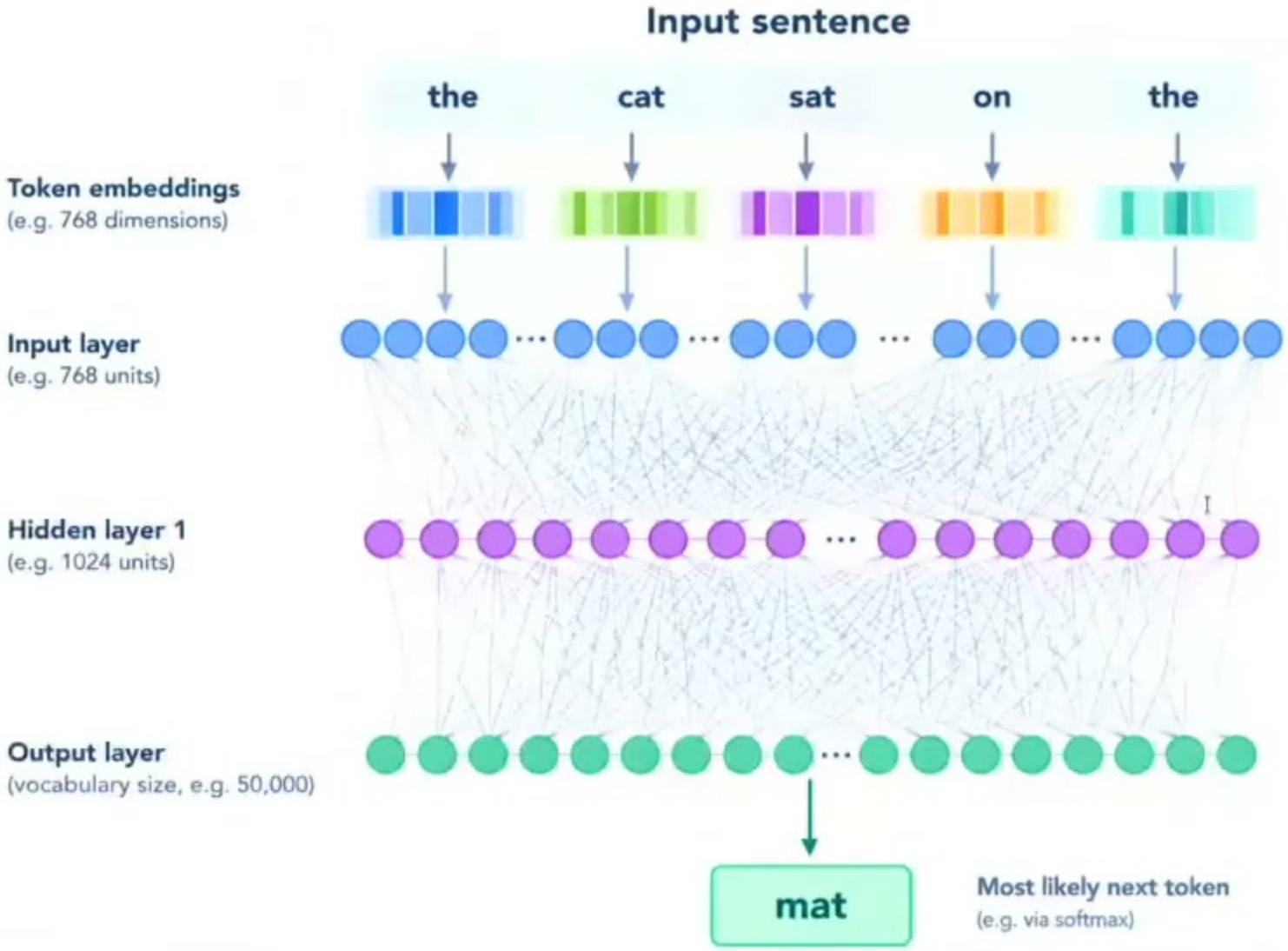
SOFAIR.UK

Science of Fundamental AI : A UK AI Research Lab

David Barber <david.barber@ucl.ac.uk>

Manufacturing
Smarter production.
Higher quality.
Less waste.
Healthcare
Earlier detection.
Personalised treatment.
Healthier lives.
Finance
Smarter decisions.
Greater inclusion.
More prosperity.
Environment
Cleaner planet.
Sustainable resources.
A healthier future.
AI
A smarter, healthier, more sustainable future for all
Entertainment
More creativity.
More connection.
More immersive experiences.
Drug Discovery
Faster breakthroughs.
Better treatments.
More lives saved.
Transport
Safer journeys.
Less congestion.
Lower emissions.
Robotics
Safer workplaces.
Greater productivity.
New possibilities.


Input sentence
the
cat
sat
on
the
Token embeddings
(e.g. 768 dimensions)
Input layer
(e.g. 768 units)
Hidden layer 1
(e.g. 1024 units)
Output layer
(vocabulary size, e.g. 50,000)
mat
Most likely next token
(e.g. via softmax)

Best Frontier Models openllm leaderboard

Sounds great, but . . .

- What if government X blocks access to frontier models?
- Is access to openweight models enough?
- Can the UK train its own foundation models?
- Should we partner with other countries and organisations to train AI?
- Are there better technologies available than current systems?
- Should we just try to get more data and more compute to be competitive?
- How to lessen environmental impact to train and serve AI systems?

- Universities are a good place to look for new ideas in AI.

UK AI Research Labs : £60M investment

UCL

THE UNIVERSITY of EDINBURGH

UNIVERSITY OF CAMBRIDGE

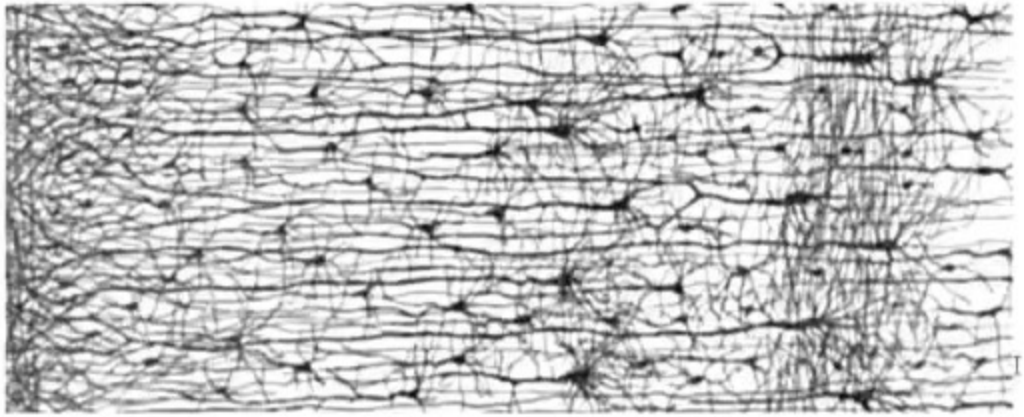
UNIVERSITY OF OXFORD

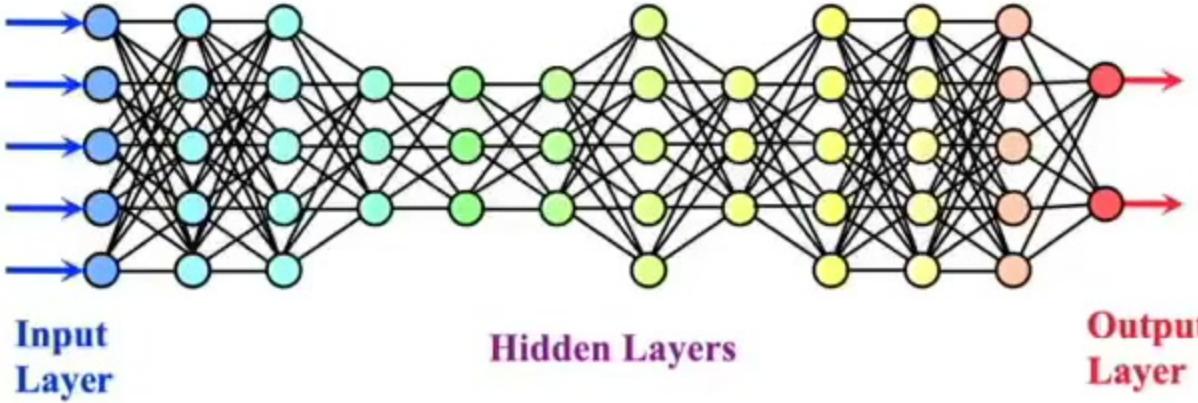
Reliable Access to Reliable Models

- Foundation models are the building blocks of modern AI.
- Models used for government services, the NHS and defence require full transparency over training and data sources.
- Industry also needs reliable, trusted models.
- Need guaranteed access to high-quality models.

- Develop ecosystem to support frontier training and model deployment.
- Harness UK talent and compute to become "AI Makers".
- Help UK lead global opensource efforts.

Deep Learning and Transformers





mriquestions.com

The models are all very similar (MoE Transformers)

gpt-oss-120b	Aug 2025, 117B, 5.1B, Undisclosed, 4.4%, MoE
Llama 4 Maverick	Apr 2025, 400B, 17B, 128 + 1 shared, 4.3%, MoE
Llama 4 Scout	Apr 2025, 109B, 17B, 16, 15.6%, MoE
DeepSeek R1	Jan 2025, 671B, 37B, 256 + shared, 5.5%, MoE
GLM-5	Feb 2026, 744B, 44B, 256, 5.9%, MoE
Mistral Small 4	Mar 2026, 119B, 6B, 128, 5.0%, MoE
Qwen 3.5	Feb 2026, 397B, 17B, 512, 4.3%, MoE
Kimi K2	Jul 2025, 1,040B, 32B, 384, 3.1%, MoE

Eight models. Zero dense. The last major dense frontier model was Qwen 2.5-72B,

birjob.com

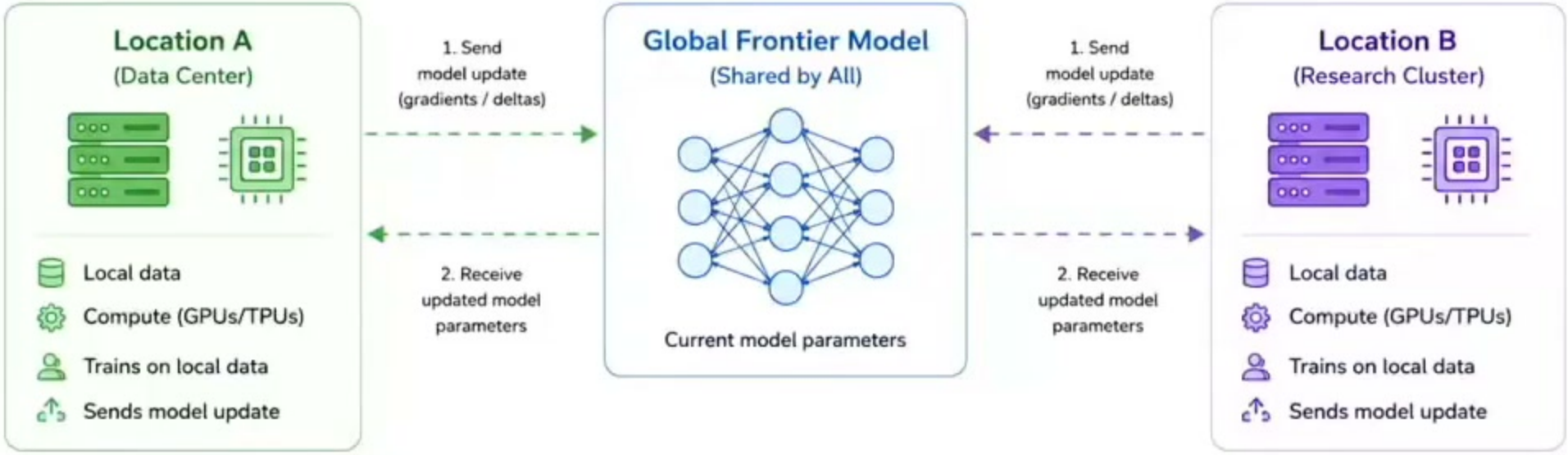
Replacing the Transformer

- Transformer Model is powerful, but very expensive (10 times longer context requires 100 times more compute)
- Mixture of Experts Transformer is hard to train (unstable)
- Recurrent attention is faster, but not as accurate – is there a happy middle ground?
- Agentic AI is successful, but how to make it more effective and train end-to-end?
- Hallucinations are still an issue – how to improve confidence and reliability?
- How to learn rapidly from domain experts
- How to build a system that is much more energy efficient?

Science for AI

Centralised Training

Decentralised Training



SOFAIR Vision

Novel Architectures

- Statistics
- Mathematics
- Neuroscience
- Computer Science

Distributed Training

- Train across the world
- Unite 1000s of GPUs
- Train large-scale models

Opensource

- Bring the global opensource community together
- Build the best opensource models

Supported by 40 Postdoc Researchers and 10 PhD students.

SOFAIR : Not academic business as usual

- A 6 year focussed project
- SOFAIR will bring together the best AI researchers in the UK
- Researchers outside of Edinburgh, Cambridge, Oxford, UCL will be invited to join us
- We will work with leading opensource researchers globally
- PhD students will train from the best talents in the UK
- Work is goal oriented to deliver impact at scale (not just about publishing papers)
- Lab will host leading talents from academia and industry
- We will create an ecosystem of basic AI science to entrepreneurship

Global Opensource AI Summit (30th November)

- Openweight is not enough.
- Most opensource efforts replicate existing methods.
- UK is well positioned to help with global opensource efforts.

Ai2 EURO LLM Masakhane
A grassroots NLP community for Africa, by Africans

APERTUS Open Athena

Supported by Jump TRADING

Basic Science → Entrepreneurship Success

- UK has great talents in AI
- Fundamental algorithmic improvements → commercial opportunity
- Eg Our PhD programme in Foundational AI → 8 spinouts
- Collectively, Edinburgh, UCL, Oxford, Cambridge have generated over $500M spinout acquisitions, with current valuations exceeding $10B
- SOFAIR will continue this, generating spinouts
- Our **Venturer Programme** works with corporates and investors – please get in touch if interested
- ConceptionX will help with PhD student spinout creation

Supporters

Jump
TRADING

FACULTY

e l l i s

NVIDIA

Open
Athena

NPL
National Physical Laboratory

conceptionx

GRAPHCORE

The
Alan Turing
Institute

Distributed Training (AI Grid)

UK AI Grid*

Zenith

+

Dawn

+

Isambard

* there are some existing noble related efforts around for inference, scientific simulations etc

SOFAIR
Science of Fundamental AI Research Lab

SOFAIR
Science of Fundamental AI Research Lab

Systems: High-level Overview
Influence on model training

CaMLSys mlsys.cst.cam.ac.uk

Flower flower.ai

Homogeneous compute: identical GPUs with balanced workloads.

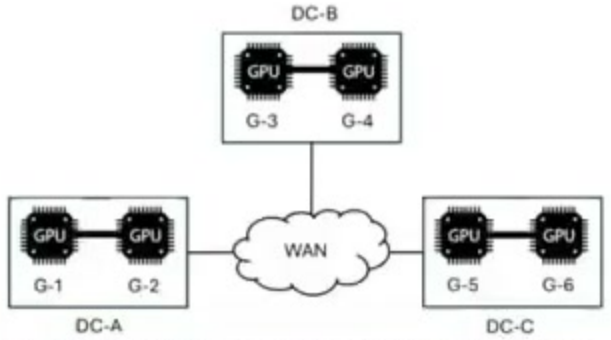
Full local replicas: the model, optimiser state and training working memory fit on each GPU.

Local data: each worker has its own batch available for training.

BUT!

Slow inter-site links: degrades the efficacy of traditional parallel training methods.

Fast local compute can be held back by slow global communication.



Source: Palak et al. (2024), Fig. 1, arXiv:2411.14458v1. Controlled WAN experiment.

Source: Sani et al., Photon: Federated LLM Pre-Training, Fig. 3

Vanilla distributed data parallelism (DDP)

Require: N: N. of devices (workers), θ: Model parameters , T: N. of epochs
Require: $\mathcal{D}$: Dataset partitioned across devices $\mathcal{D}_i$ where $i \in \{1, 2, \ldots, N\}$
Require: `RingAllReduce`: aggregation operation across devices on a ring
Require: `Opt`: Optimizer for updating θ with gradients

Initialize: Randomly initialize model parameters θ_0 on each device
for $t = 1$ to T **do**
  Step 1: Parallel Local Training
  for each device $i \in \{1, 2, \ldots, N\}$ **in parallel do**
    Compute local mini-batch loss $\mathcal{L}_i(\theta_{t-1}, \mathcal{D}_i)$
    Compute local gradients $\nabla_{\theta_{t-1}} \mathcal{L}_i(\theta_{t-1})$
  Step 2: Distributed `RingAllReduce` Gradient Aggregation
  $\nabla_{\theta_{t-1}} \mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta_{t-1}} \mathcal{L}_i$
  Each device now possesses the global gradient $\nabla_{\theta_{t-1}} \mathcal{L}$
  Step 3: Parallel Model Update
  for each device $i \in \{1, 2, \ldots, N\}$ **in parallel do**
    $\theta_t = \texttt{Opt}(\theta_{t-1}, \nabla_{\theta_{t-1}} \mathcal{L})$
Output: Trained model parameters θ_T

CaMLSys mlsys.cst.cam.ac.uk

Flower flower.ai

Require: N: Number of clients, f_θ: Model with parameters θ
Require: T: Number of federated rounds, K: Number of local steps
Require: $\{\mathcal{D}_i\}$: Federated dataset, i.e., a set of private $\mathcal{D}_i$, $i \in \{1, \ldots, N\}$
Require: `ClientOpt`: local client optimizer, `ServerOpt`: server optimizer

Initialize: Randomly initialize global model parameters θ_0 on the server
for $t = 1$ to T **do**
  Step 1: Broadcast model parameters
  Server sends θ_t to all N clients
  Step 2: Parallel Local Training
  for each client $i \in \{1, 2, \ldots, N\}$ **in parallel do**
    $\omega_{i,0} \leftarrow \theta_{t-1}$
    for each local iteration $k \in \{1, 2, \ldots, K\}$ **do**
      Compute local mini-batch loss $\mathcal{L}_i(\omega_{i,k-1}, \mathcal{D}_i)$
      Compute local gradients $\nabla_{\omega_{i,k-1}} \mathcal{L}_i(\omega_{i,k-1})$
      $\omega_{i,k} \leftarrow \texttt{ClientOpt}(\omega_{i,k-1}, \nabla_{\omega_{i,k-1}} \mathcal{L}_i(\omega_{i,k-1}))$
    $\Delta\theta_{t-1,i} \leftarrow \omega_{i,K} - \theta_{t-1}$
  Step 3: Global Model Update (on the server)
  $\theta_t = \texttt{ServerOpt}(\theta_{t-1}, \{\Delta\theta_{t-1,i}\})$
Output: Trained model parameters θ_T

K

Vanilla federated alt.

https://arxiv.org/abs/2405.10853; **MLSys '25**

Bi-level Optimisation

Choose the inner optimiser, outer optimiser and local interval K.

The framework leaves many open questions:

- Which optimiser states should be shared, and when? **(DES-LOC)**
- How do we preserve global memory while responding to local updates? **(MT-DAO)**
- How can we compress the state that is stored and communicated? **(LoRDO)**

This motivates a co-design principle: how communication-efficient can we make the training progress whilst maintaining the quality of training?

Algorithm 1 FEDOPT

Input: x_0, CLIENTOPT, SERVEROPT
for $t = 0, \cdots, T-1$ **do**
  Sample a subset $\mathcal{S}$ of clients
  $x_{i,0}^t = x_t$
  for each client $i \in \mathcal{S}$ **in parallel do**
    for $k = 0, \cdots, K-1$ **do**
      Compute an unbiased estimate $g_{i,k}^t$ of $\nabla F_i(x_{i,k}^t)$
      $x_{i,k+1}^t = \text{CLIENTOPT}(x_{i,k}^t, g_{i,k}^t, \eta_l, t)$
    $\Delta_i^t = x_{i,K}^t - x_t$
  $\Delta_t = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \Delta_i^t$
  $x_{t+1} = \text{SERVEROPT}(x_t, -\Delta_t, \eta, t)$

Feasibility Analysis: Training Time (Days)

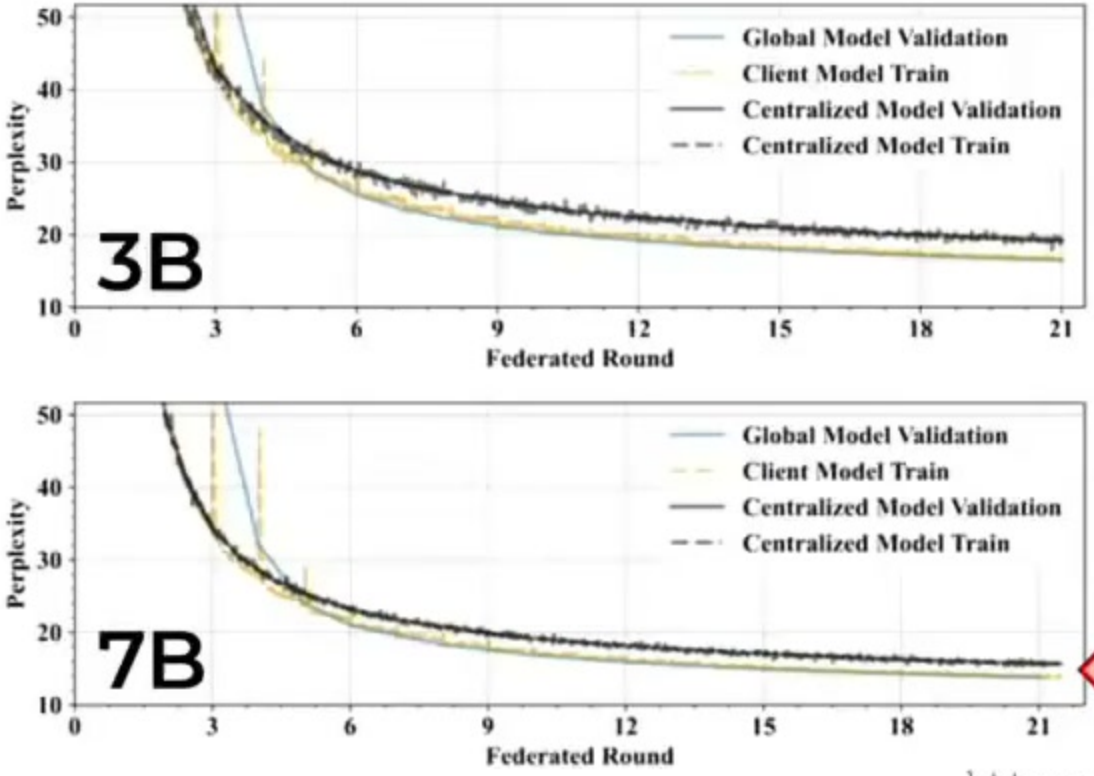
10B

	Speed	Sites	GPUs	Training Days
DDP	100Gb/s	1	384	0.7
Photon	10Gb/s	4	384	0.9

100B

	Speed	Sites	GPUs	Training Days
DDP	100Gb/s	1	384	72.9
Photon	10Gb/s	4	384	78.4

Federated LLMs are at least as good as conventional data-center LLMs

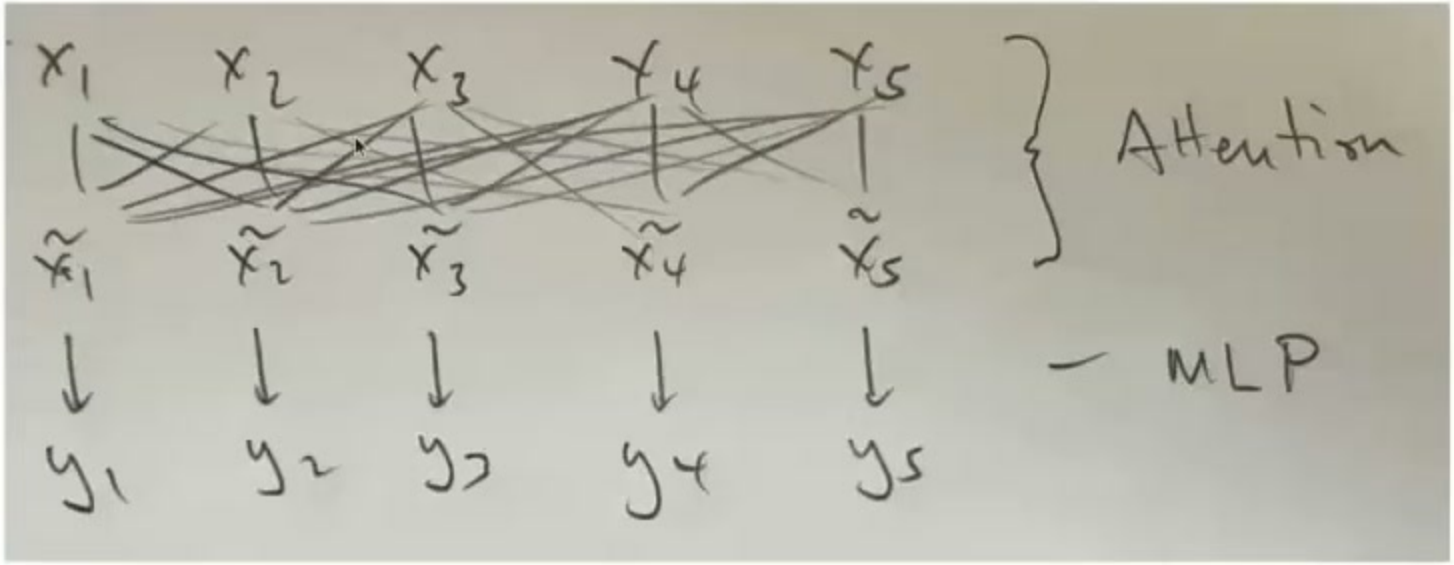


Size	Fed PP	Cent PP
1.3B	**20.1**	23.2
3B	**15.7**	18.2
7B	**13.8**	16.6

https://arxiv.org/abs/2405.10853; **MLSys '25**

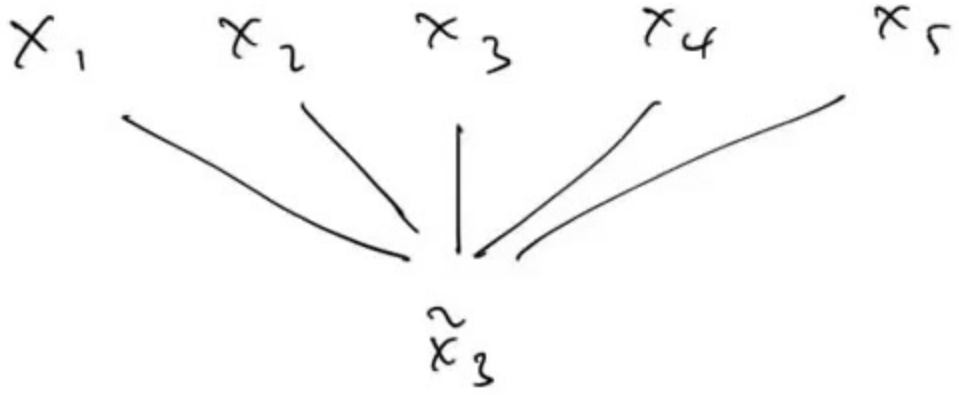
Faster Attention

Transformer



- A sequence of tokens $x_1, \ldots, x_T$ is transformed to a new sequence $y_1, \ldots, y_T$
- Repeated many times, taking the output of one layer and feeding it into the next.

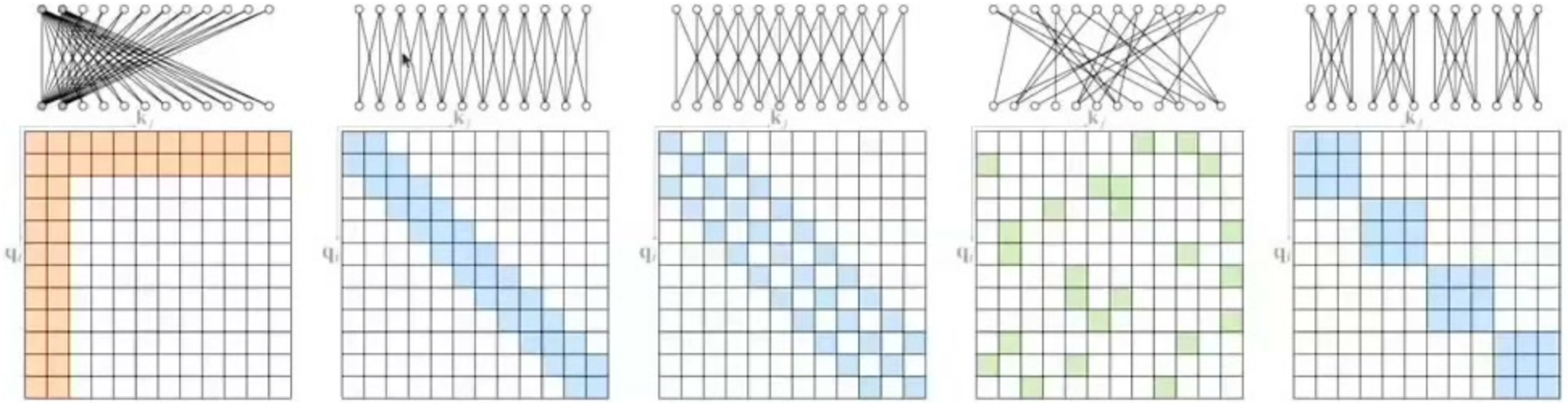
Attention Mechanism



$$\tilde{x}_t = \frac{\sum_s e^{x_s^T x_t} x_s}{\sum_s e^{x_s^T x_t}} = \sum_s p(s|t) x_s$$

- Good : Long range dependencies are accounted for.
- Good : weights are input dependent (not fixed).
- Bad : Permutation Invariant. Need to include positional encoding to counter this.
- Bad : Scaling is poor. Takes $O(T^2)$ to compute all the transformed vectors.

Sparse Attention ("A survey of transformers" Liu et al AI Open 2022)



(a) global (b) band (c) dilated (d) random (e) block local

Bidirectional (non causal) Latte

Note that we can write

$$\tilde{x}_t = \sum_{s=1}^{T} p(s|t) v_s$$

We introduce a hidden variable with L states and model the attention using

$$p(s|t) = \sum_{l=1}^{L} p(s|l)p(l|t), \qquad p(l|t) = \frac{e^{x_t^\top w_l^q}}{\sum_{j=1}^{L} e^{x_t^\top w_j^q}}, \qquad p(s|l) = \frac{e^{x_s^\top w_l^k}}{\sum_{s=1}^{T} e^{x_s^\top w_l^k}}$$

Attention Mechanism

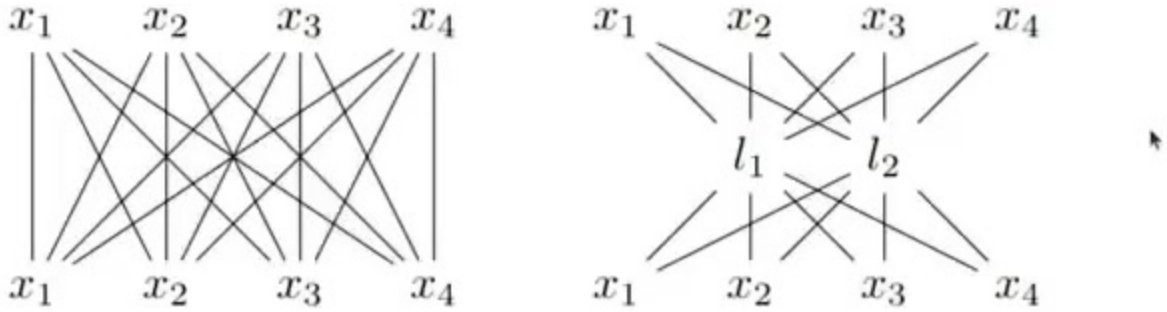
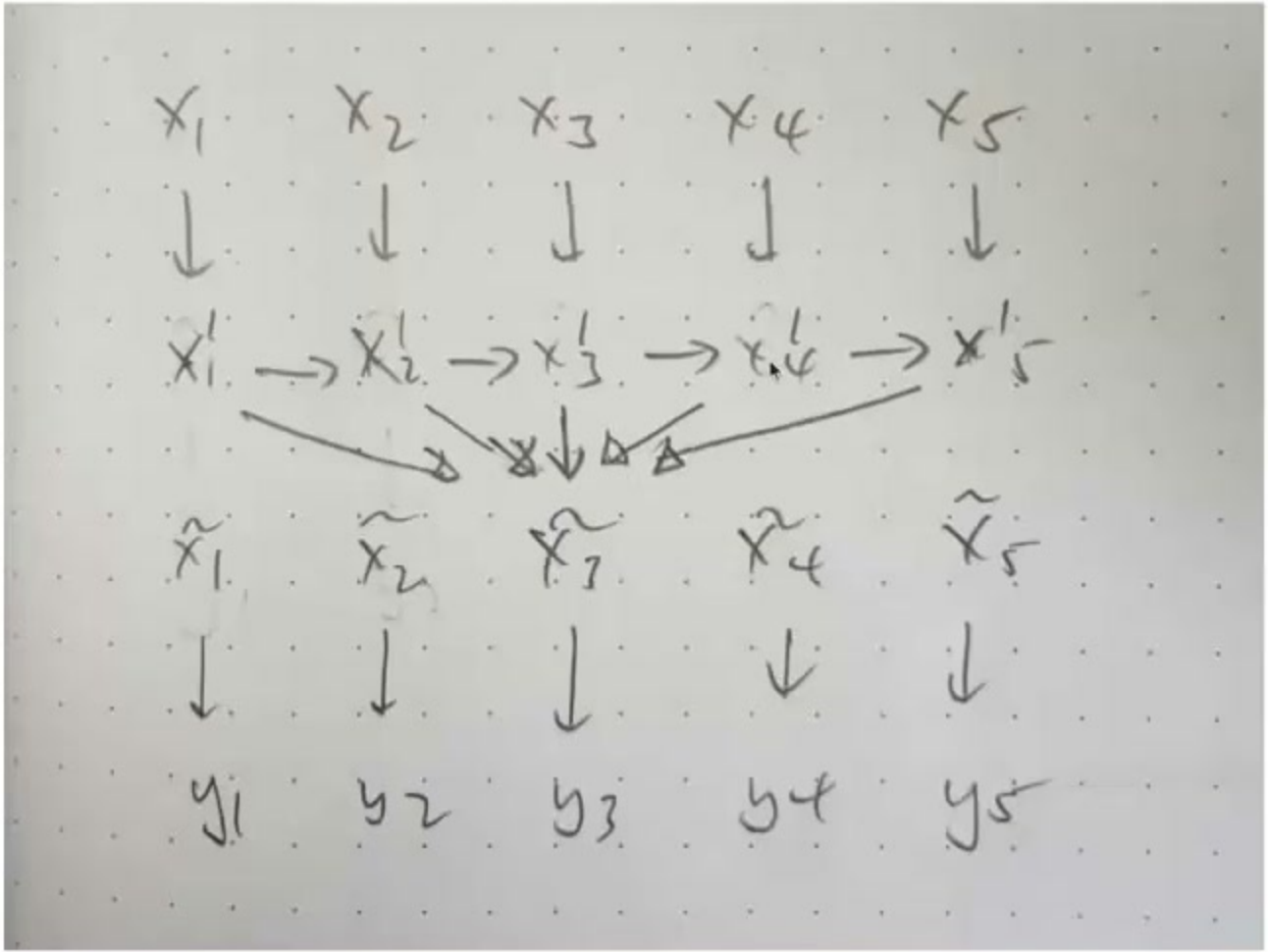


Figure 1: Non causal attention. The time complexity of each algorithm is $O(ED)$ where E is the number of edges in each graph and D is the dimension of each x. (Left) Standard Attention computes all pairwise similarities between the elements of the sequence. (Right) Latte computes only the pairwise similarities between each element of the sequence and each latent embedding.

Recurrent encoding



OpenWebText next token prediction

Model	Params.	PPL ↓
Mega Ma et al. (2023)	153M	23.75
Retnet Sun et al. (2023)	197M	21.59
H3 Fu et al. (2023)	125M	21.0
RWKV Peng et al. (2023a)	153M	18.97
Griffin De et al. (2024)	139M	18.83
R-SWA++ (our)	141M	18.25
Mamba Gu & Dao (2023)	149M	17.70
GLA Yang et al. (2024)	206M	19.10
Ligth.Att Qin et al. (2024a)	166M	23.67
GatedDeltaNet Yang et al. (2025)	190M	21.41
Latte-R-SWA++	153M	**17.64**

Join us!

contact@sofair.uk

We're looking for:

- Compute
- Research Engineers
- Research Scientists
- Philanthropic Donations (PhD studentships, Postdocs)
- Investors
- We will have internships and host researchers within the lab.