LightOn
We bring Light to AI

Photonic computing for large-scale AI applications

A (non-linear) entrepreneurial journey from optical computing to AI agents

Laurent DAUDET

LightOn, Co-Founder

Université Paris Cité, Professor

IT4I Workshop on Quantum Computing, 3 December 2025

A spinoff from university research

LightOn

4 Co-founders

Igor Carron

CEO

Management of complex nuclear engineering/ aerospace projects

- Former Assistant Director at Spacecraft Technology Center at Texas A&M Univ.

Florent Krzakala

ML Advisor

- Professor of Physics and Electrical Engineering at EPFL

Laurent Daudet

Expert in Signal Processing

- Physics professor at Université Paris Cité

Sylvain Gigan

Optics Advisor

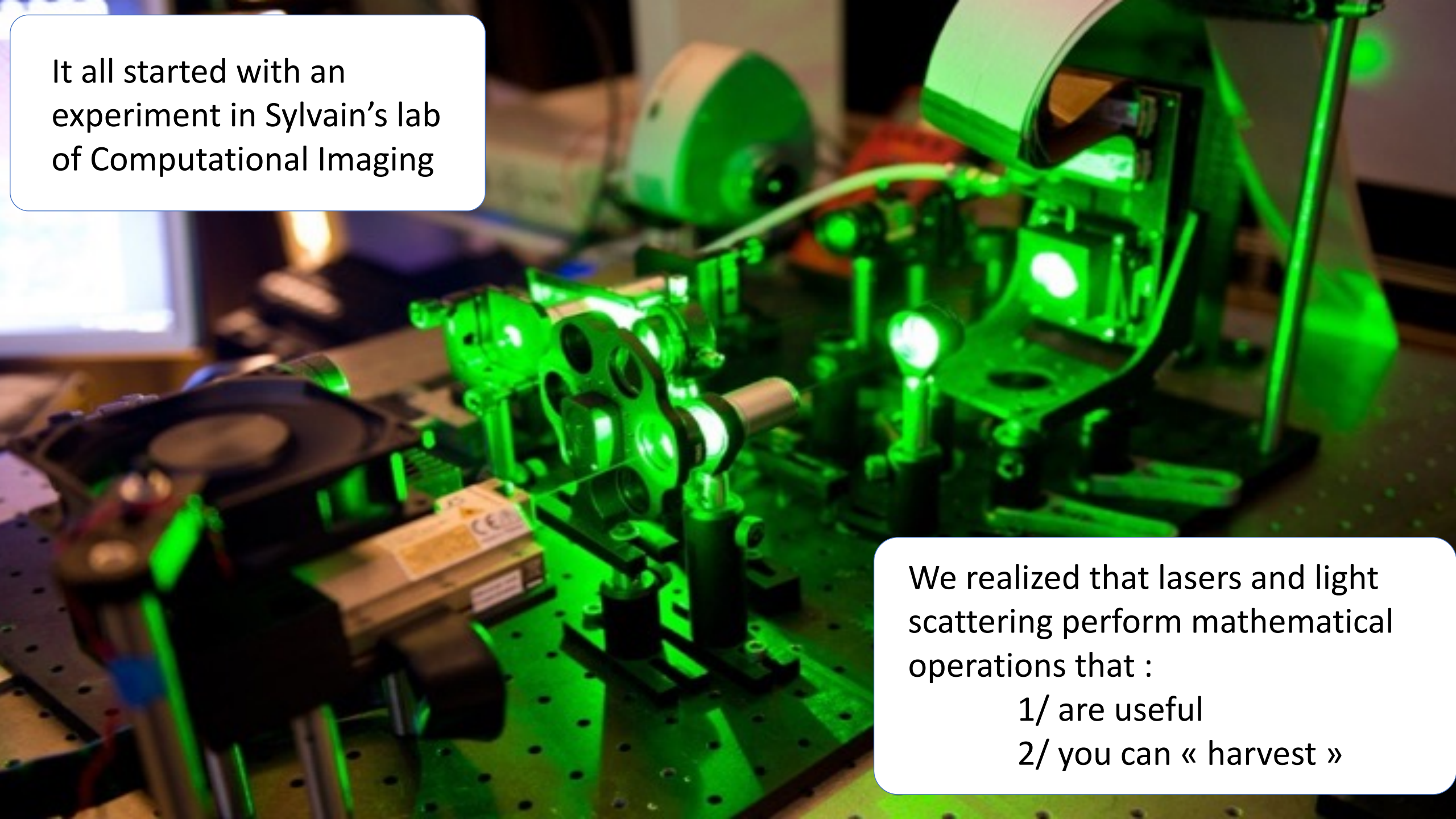
- Physics professor at Sorbonne University

A team of ~45

How LightOn *actually* started

Procrastination

Coffee



It all started with an experiment in Sylvain's lab of Computational Imaging

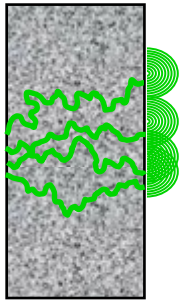
We realized that lasers and light scattering perform mathematical operations that :

1/ are useful

2/ you can « harvest »

Scattering: a coherent process

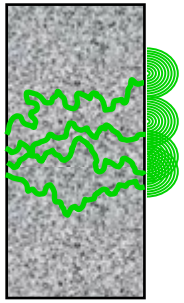
laser



« Speckle »

Scattering: a coherent process

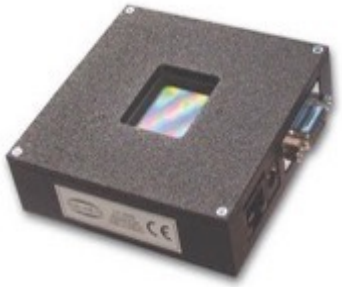
laser



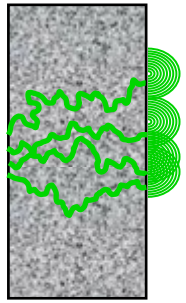
« Speckle »

Modulated beam

laser



Spatial Light Modulator



CMOS camera

Scattering: a coherent process

Discrete input vector

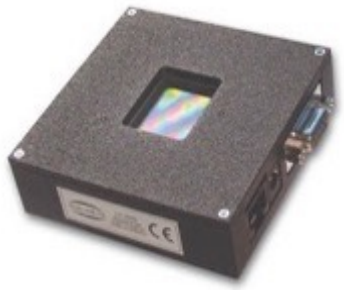
$$\begin{pmatrix} h_{1,1} & h_{1,2} & \dots & h_{1,N} \\ h_{2,1} & h_{2,2} & \dots & h_{2,N} \\ \vdots & & \ddots & \vdots \\ h_{M,1} & h_{M,2} & \dots & h_{M,N} \end{pmatrix}$$

Random transmission matrix

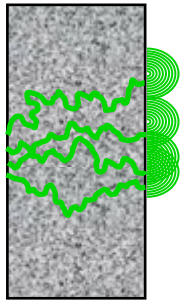
Discrete output vector (speckle intensity)

$$x \qquad H \qquad y = | H x |^2$$

laser



Spatial Light Modulator



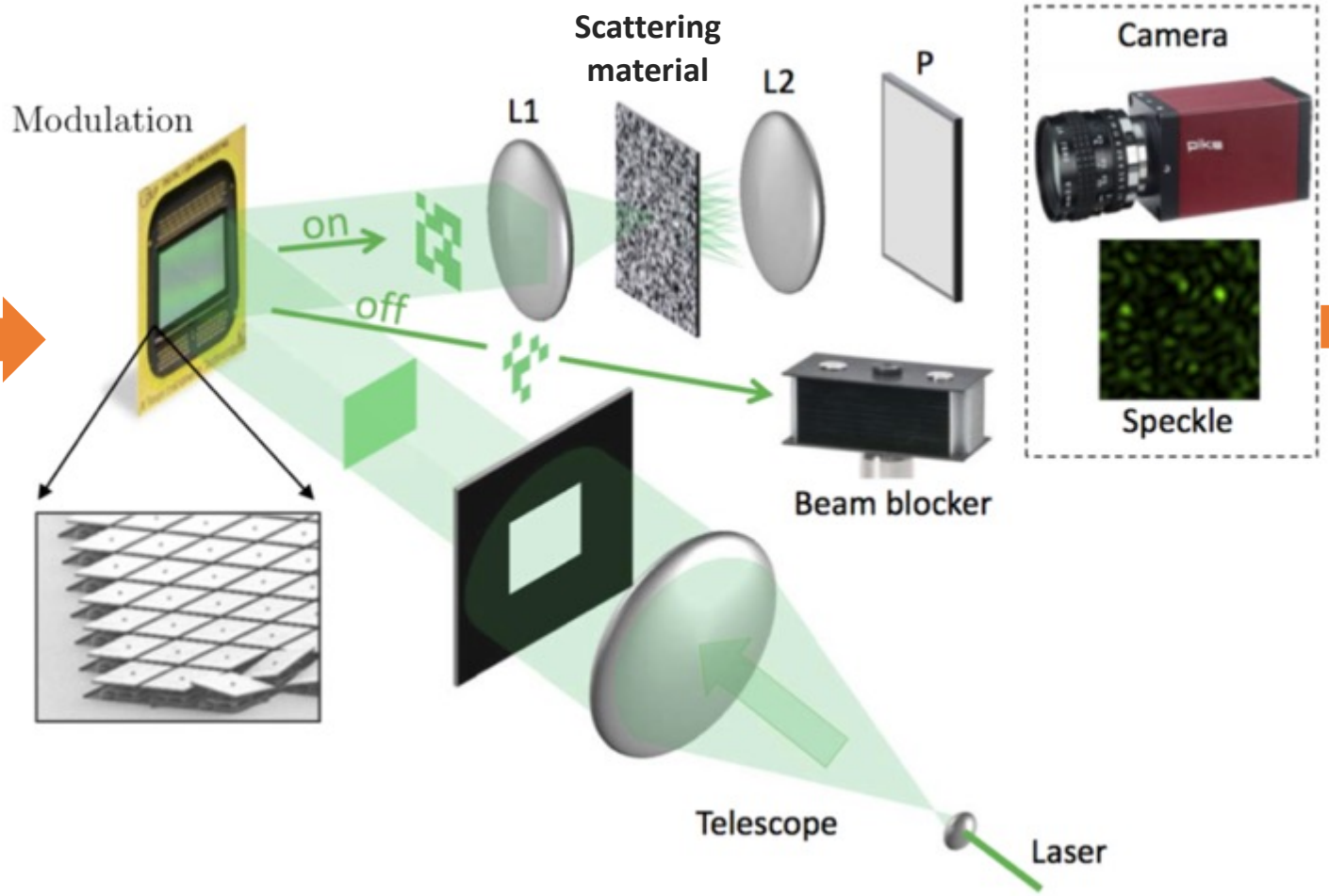
CMOS camera

Popoff et al. *Nat. Commun.* 1:81 doi: 10.1038/ncomms1078 (2010)

Optical Processing Unit technology

Data in

Vector x

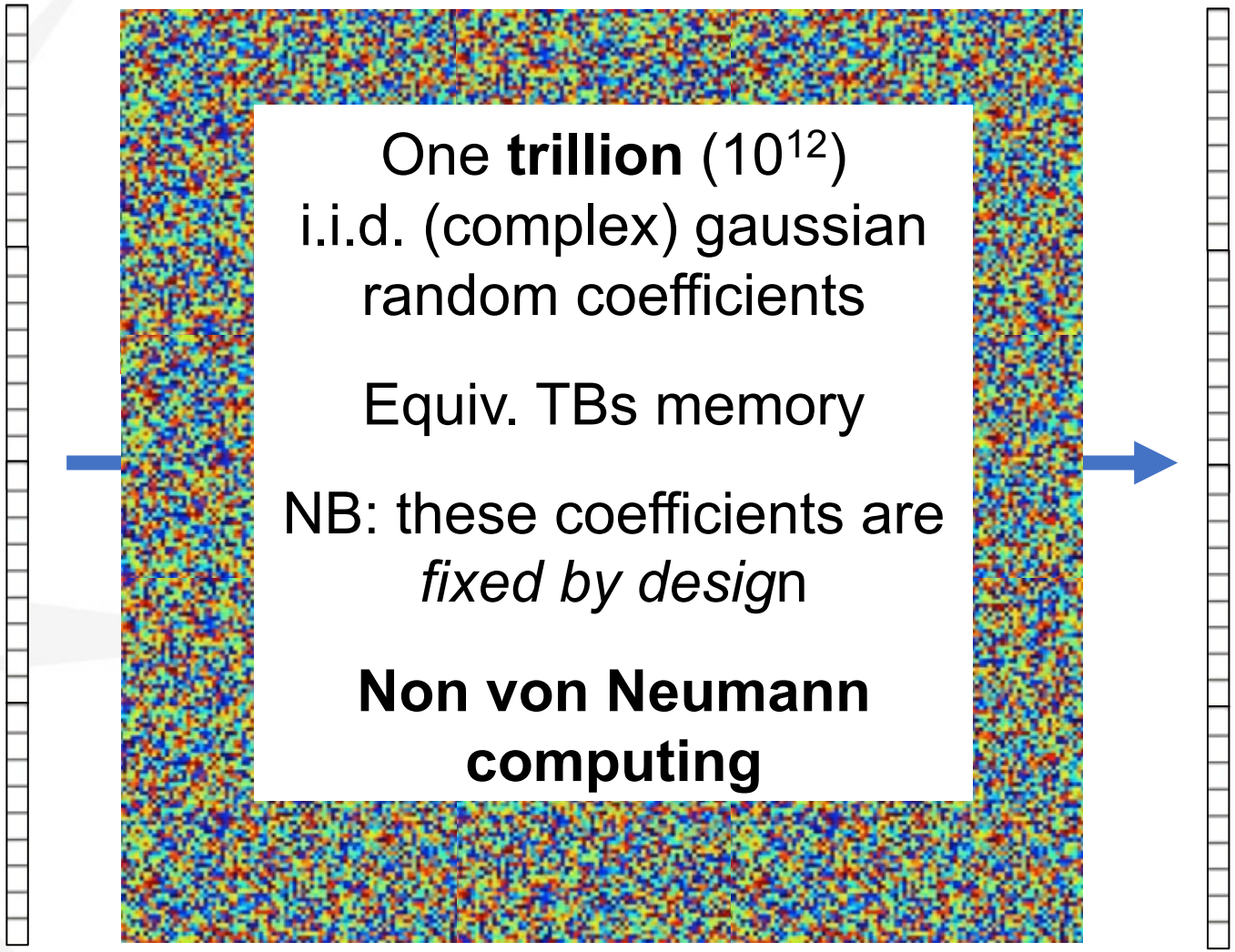


Data out

Vector y = $|M x|^2$

Matrix-vector multiplication through light scattering

1 million independent input pixels



1 million independent output pixels

Random Projections in Machine Learning

- Random Projections act as distance-preserving point cloud embeddings

Johnson-Lindenstrauss Lemma (1984)

Lemma For any $0 < \epsilon < 1$ and any interger n let k be a po sitive interger such that

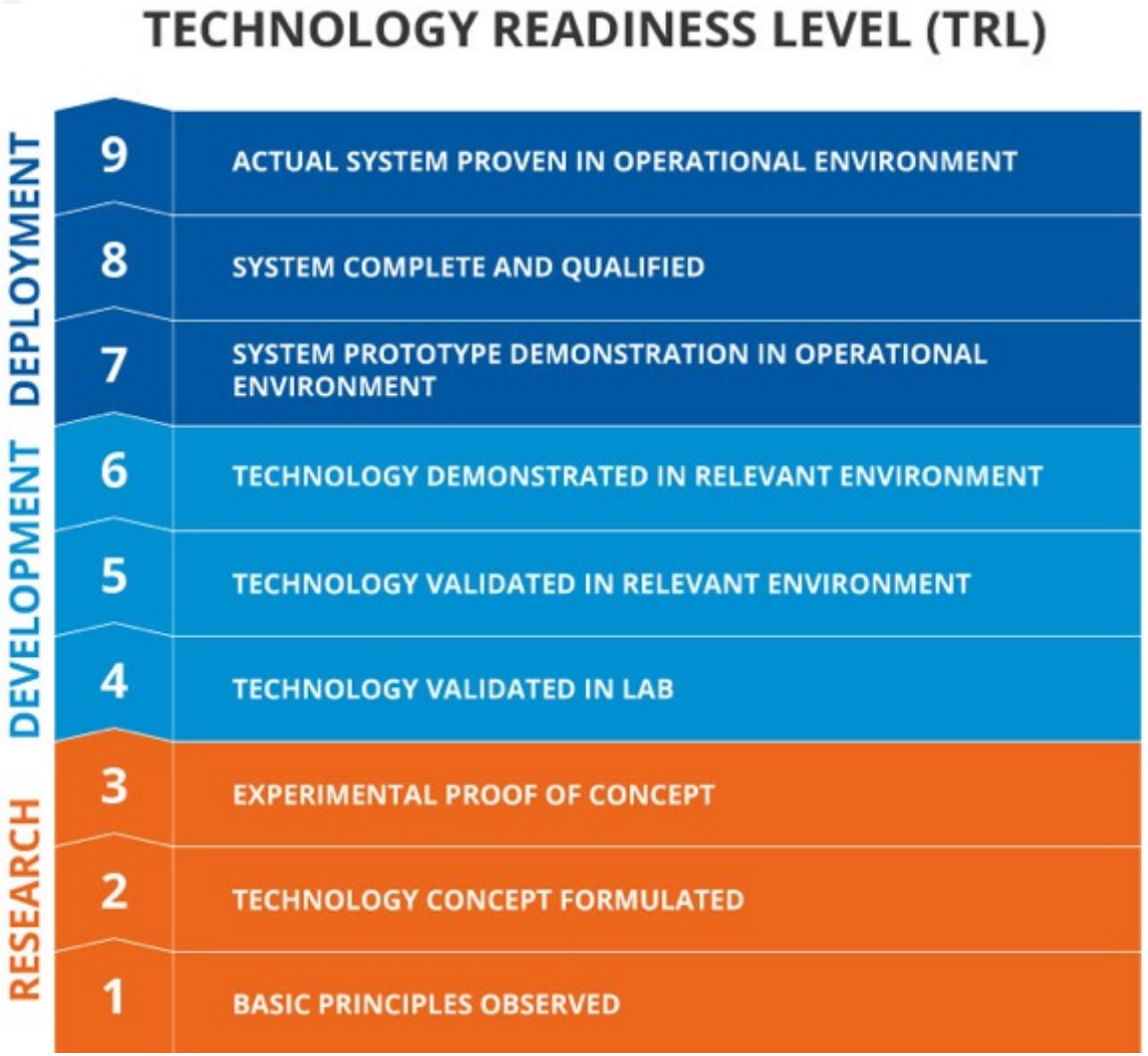
$$k \geq \frac{24}{3\epsilon^2 - 2\epsilon^3} \log n$$

then for any set A of n points $\in \Re^d$ there exists a map $f : \Re^d \rightarrow \Re^k$ such that for all $x_i, x_j \in A$

$$(1-\epsilon)||x_i - x_j||^2 \leq ||f(x_i) - f(x_j)||^2 \leq (1+\epsilon)||x_i - x_j||^2$$

- NeurIPS 2017 Test of Time Award
  "Random Features for Large-scale Kernel Machines", Rahimi, Recht, 2008

The TRL scale



The first 4 years 2016-2020

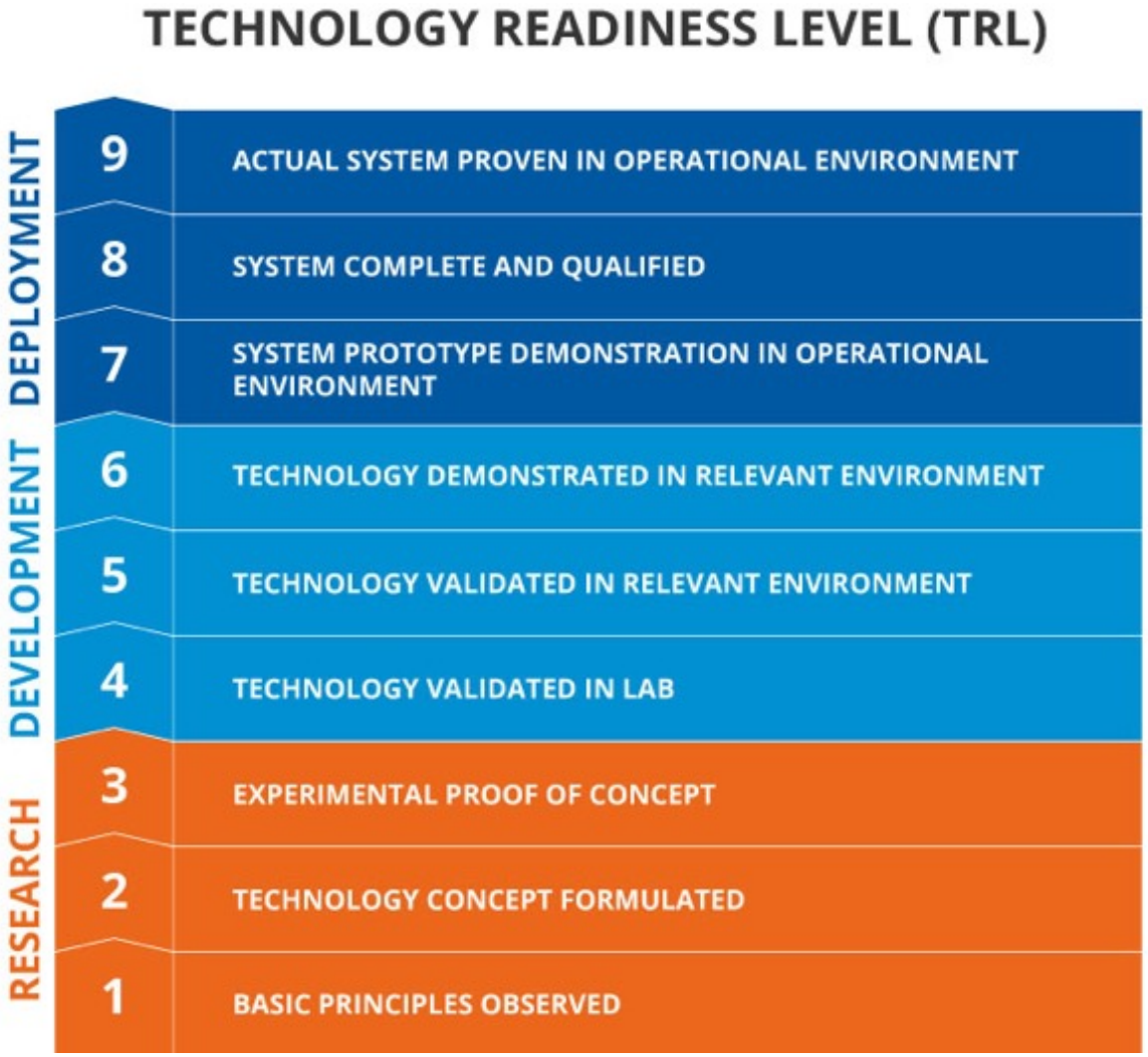
LightOn

Climbing up the TRL ladder

Lab experiment

Component

Photonic hardware



Optical Processing Unit technology

LightOn Optical Processing Unit (OPU)
the world's first photonic AI co-processor publicly available

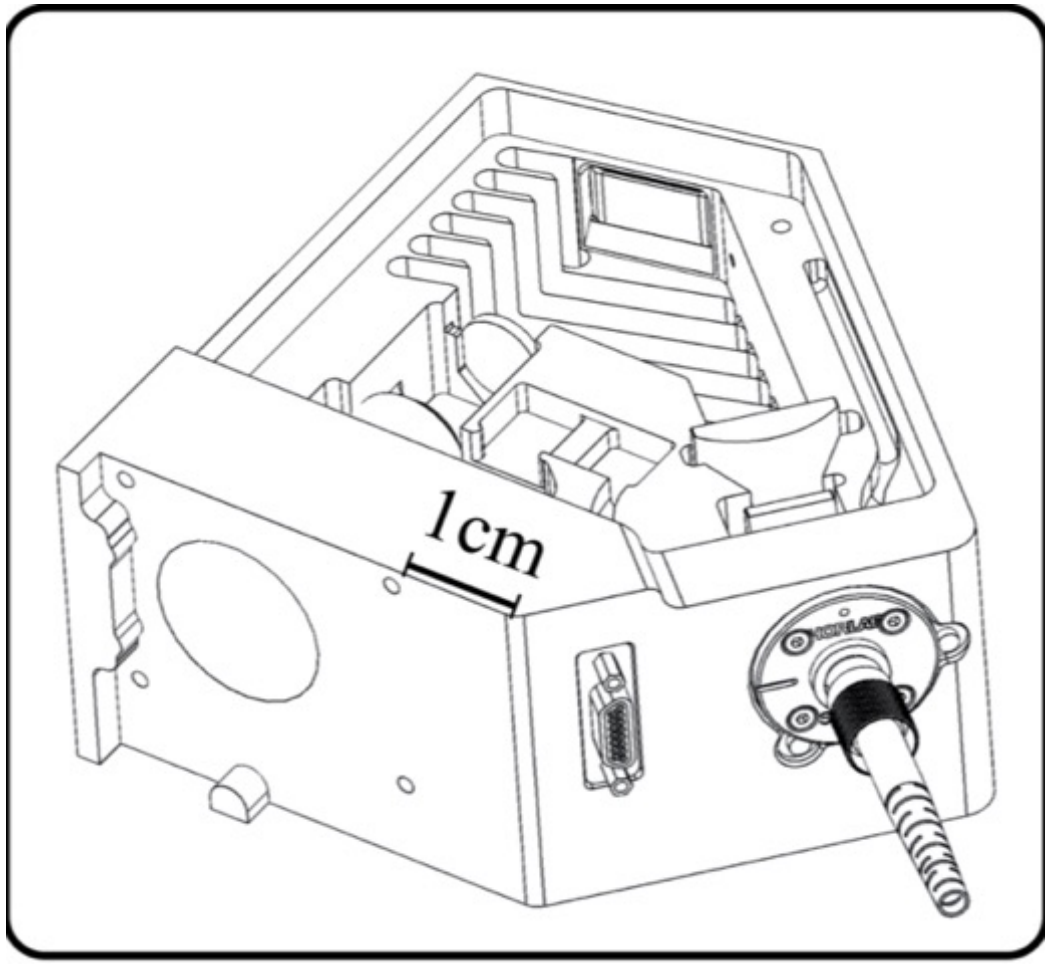
2200 TOPS
In a single photonic core

With only 30 W TDP
200 times better in #OPS/W than NVIDIA V100 GPU boards

Used by AI researchers through LightOn's cloud platform → build a community

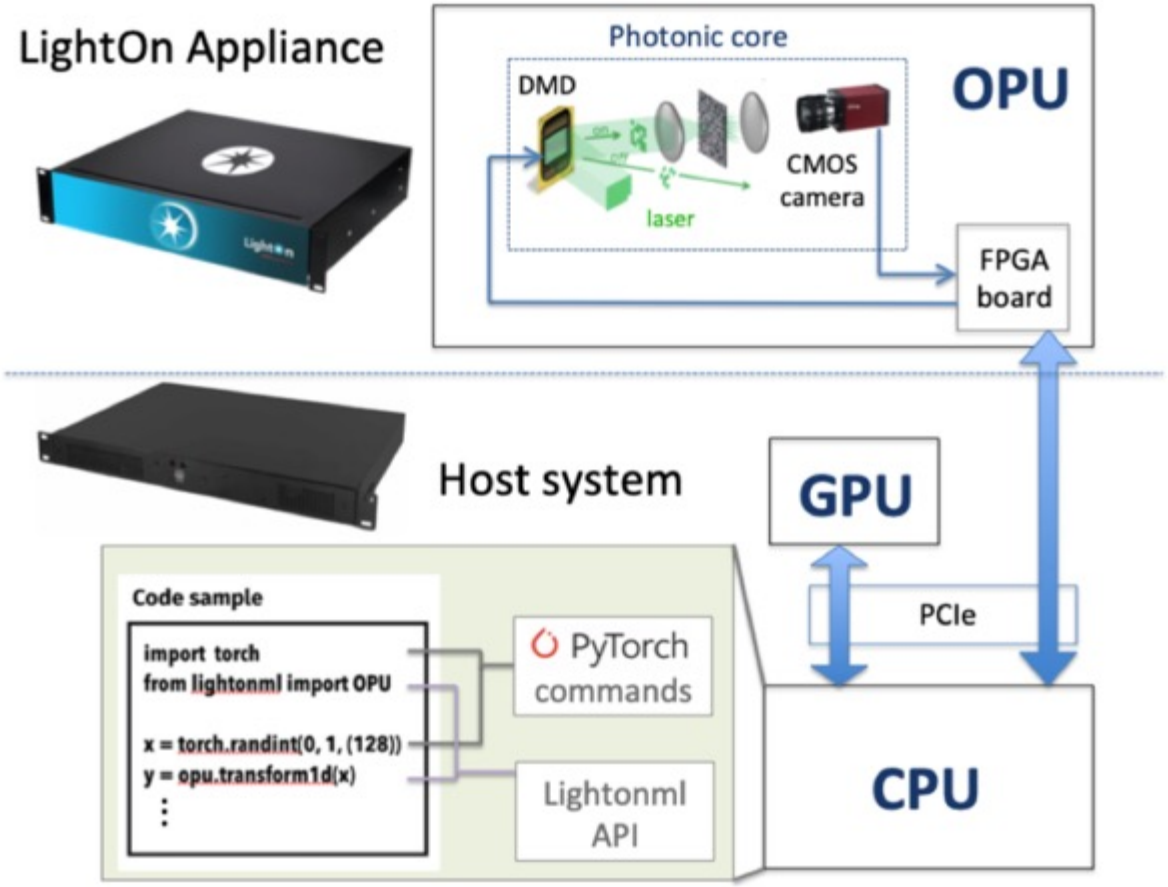
Optical Processing Unit technology

Optics integration in aluminum monobloc



Optical Processing Unit technology

Software integration for hybrid data processing architecture



Developing applications

Lab experiment

Component

Photonic hardware

Application layer

System

AI software

+

Photonic hardware

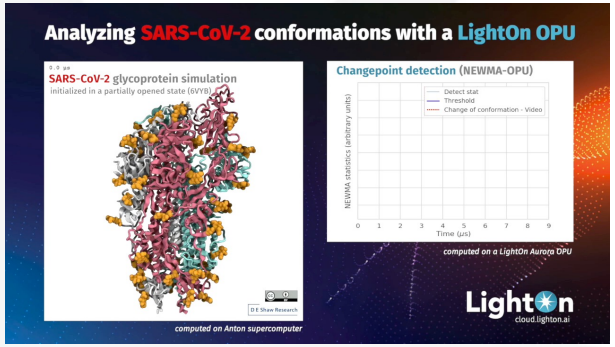
Hybrid computing in AI pipelines

Ivan Dokmanic
Associate Prof.
UNIVERSITY OF ILLINOIS

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2019

David Rousseau @dhpmrou · 4 avr.
Our talk on analysing #hep data with random matrices in @LightOnIO Optical Processor Unit accepted at @ichep2020 conference ! (Remote in July or Prague beg 2021)#hepml @Laurent_Daudet @IgorCarron

6 15



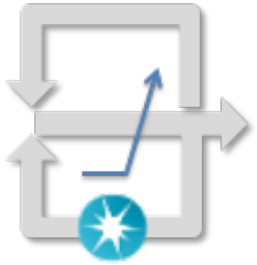
Accelerating SARS-COv2 Molecular Dynamics Studies with Optical Random Features

Amélie Chatelain
LightOn ML R&D engineer

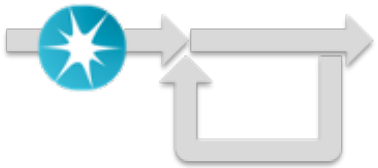
Pre-conditioning
for RandSVD

Transfer learning
in Convolutional Deep NN

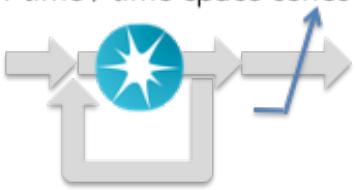
Reinforcement Learning



Sketching
for changepoint detection



Reservoir computing
for time / time-space series



Adversarial robustness / differential privacy
by design

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2020

Collaboration with Criteo

criteo

Collaboration with FAIR

Facebook AI Research

Accelerated scientific computing
with
Randomized numerical linear algebra

HPC Use case: Accelerated Scientific Computing

Randomized Numerical Linear Algebra

DOE RASC report (Jan 2021):
randomized algorithms are "essential to the future of computational science and AI for Science."

- Approximate matrix multiplications
- Randomized SVD
- ... And much more

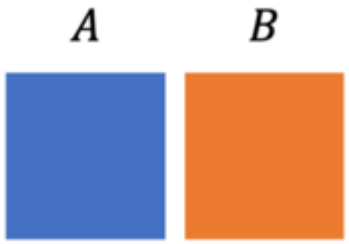
Photonic co-processors in HPC: using LightOn OPUs for Randomized Numerical Linear Algebra

Daniel Hesslow, Alessandro Cappelli, Igor Carron, Laurent Daudet, Raphaël Lafargue, Kilian Müller, Ruben Ohana, Gustave Pariente, and Iacopo Poli
LightOn, Paris, France

See Hesslow D. *et al*, HotChips2021 proceedings, https://arxiv.org/abs/2104.14429

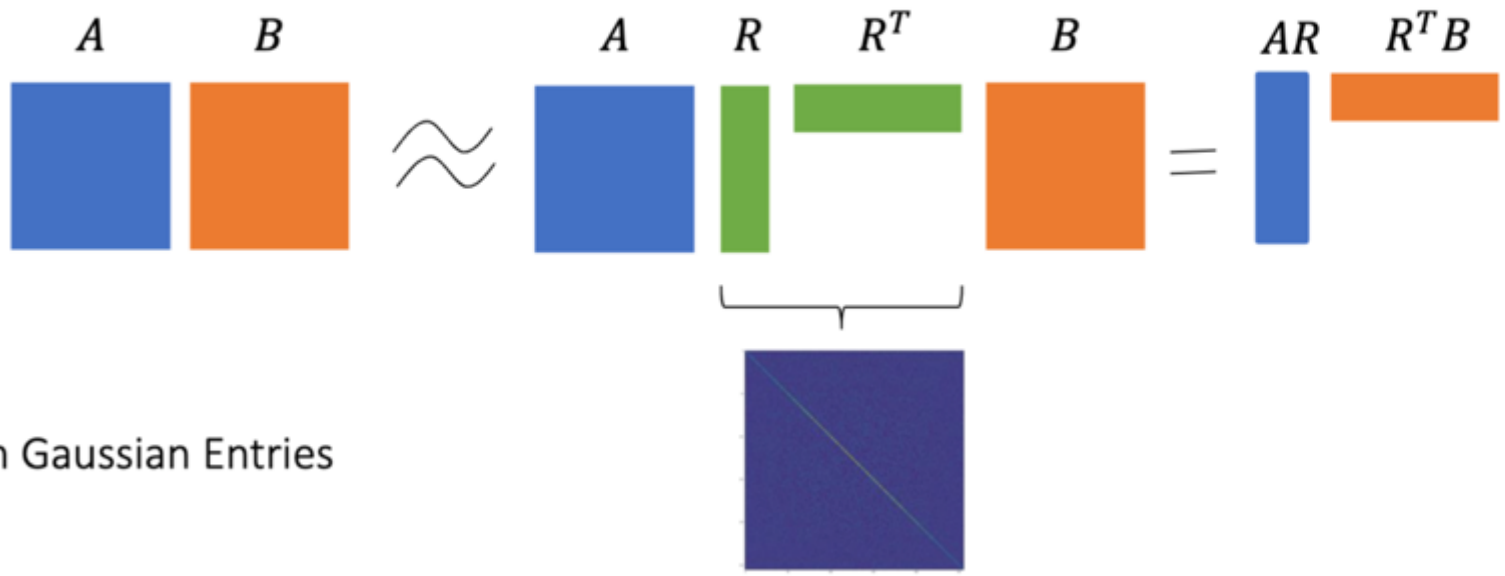
HPC Use case: Accelerated Scientific Computing

TOY EXAMPLE
Multiplying 2 matrices



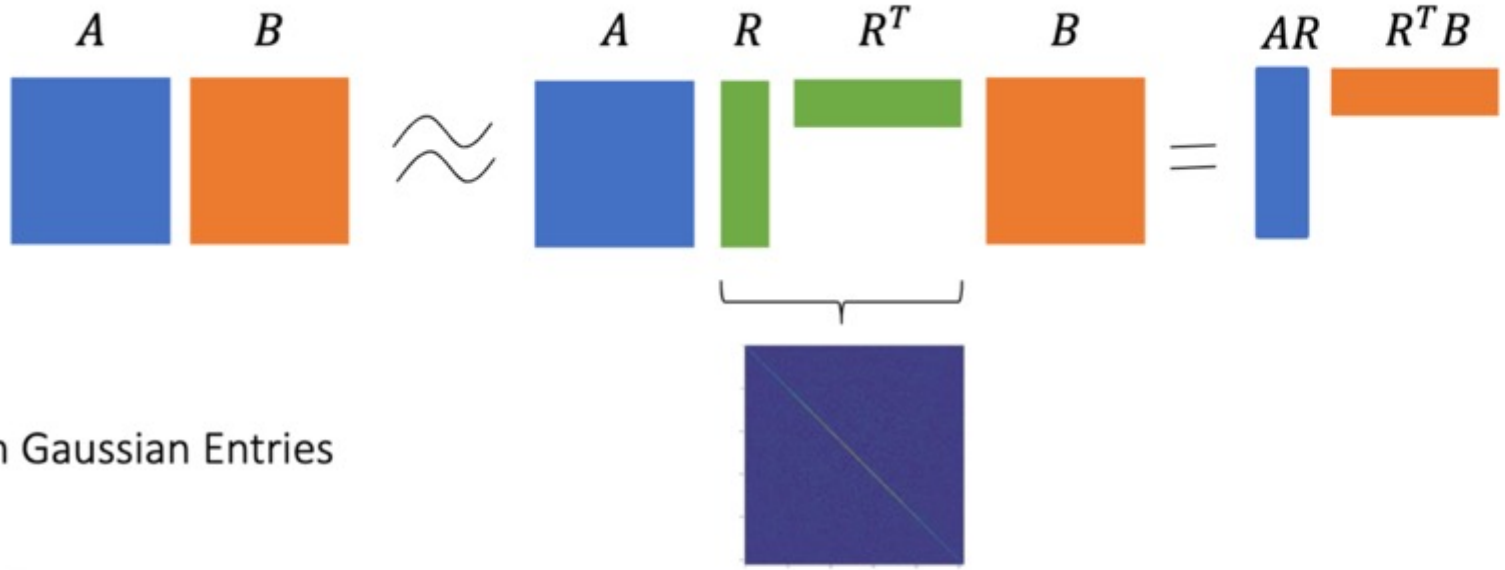
NB: DON’T DO THIS IN PRACTICE !

HPC Use case: Accelerated Scientific Computing

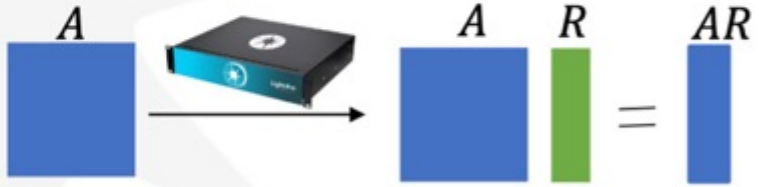


Matrix with Random Gaussian Entries

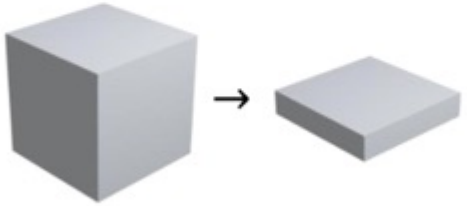
HPC Use case: Accelerated Scientific Computing



Matrix with Random Gaussian Entries



Computational Complexity: $O(n^3) \rightarrow O(n^2 m)$

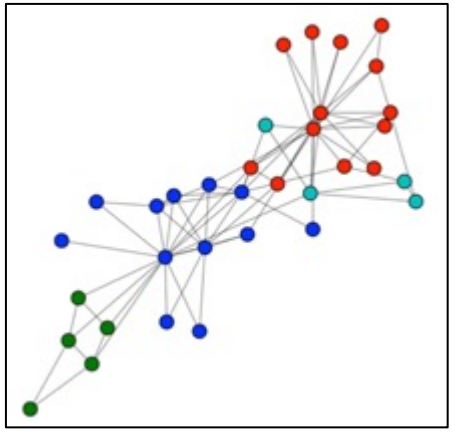


Randomized Trace estimators – application to Graph Neural Networks

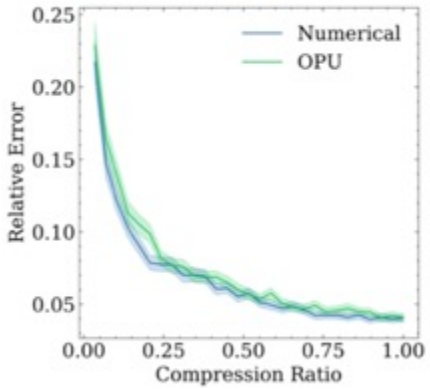
Hutchinson's trace estimator

$$\mathrm{Tr}(A) \approx \mathrm{Tr}(RAR^\top)$$

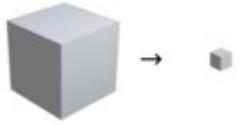
$$\mathrm{Tr}(A^3) \approx \mathrm{Tr}(RA^3R^\top) \approx \mathrm{Tr}((RAR^\top)^3)$$



A³ TRACE ESTIMATION



$O(n^3) \rightarrow O(m^3 + n)$

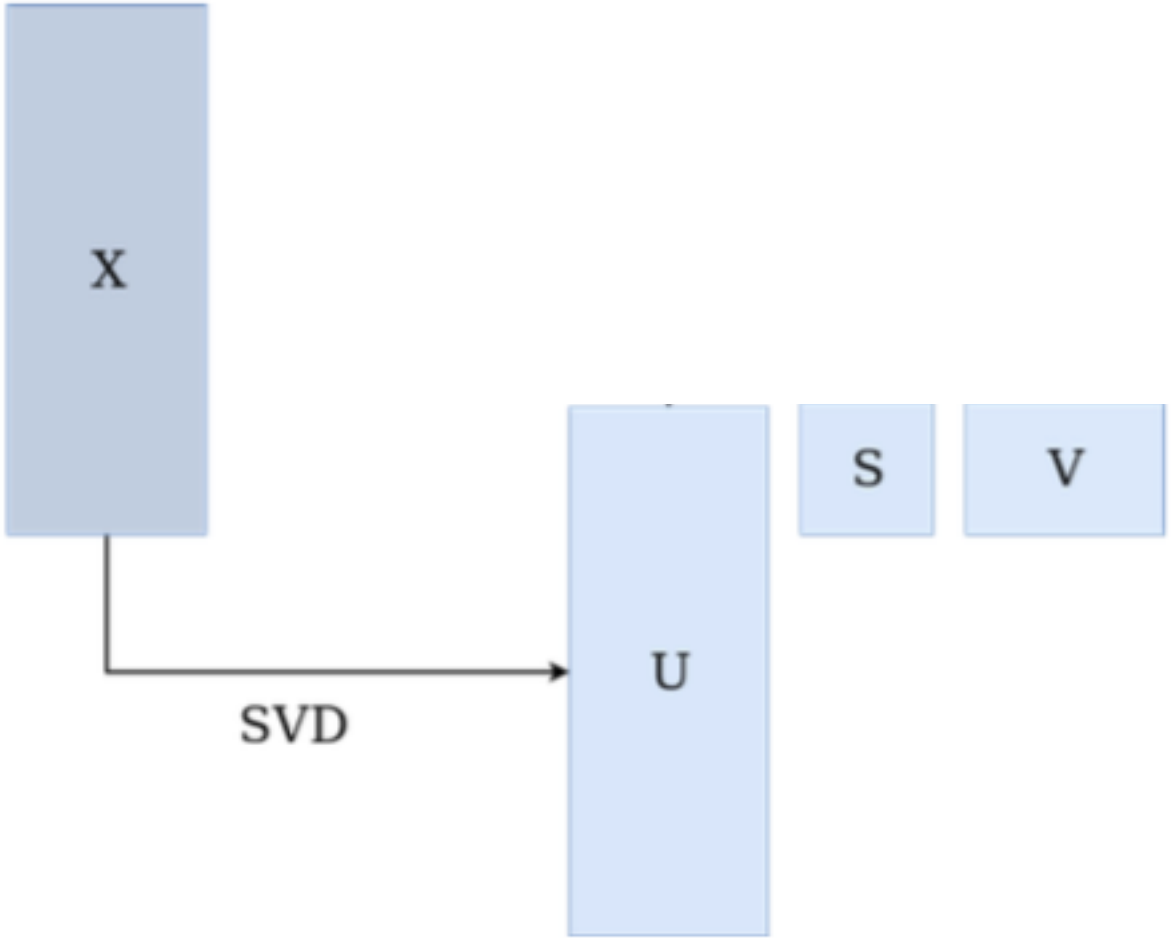


Community detection in networks
→ Triangle counting on graphs

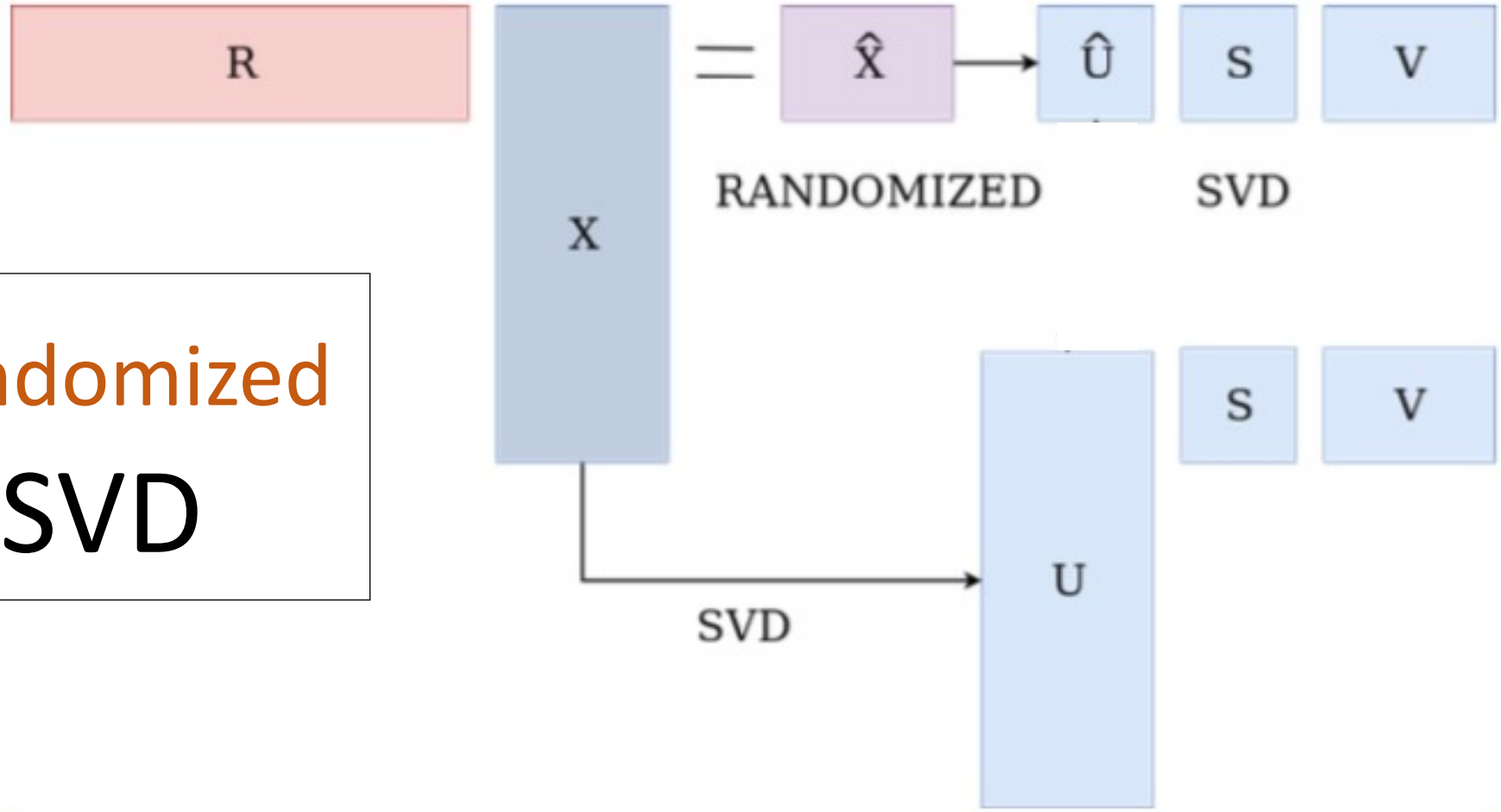
(figure from Rossetti et al. Applied Network Science (2019) 4:52)

HPC Use case: Accelerated Scientific Computing

SVD



HPC Use case: Accelerated Scientific Computing



Randomized

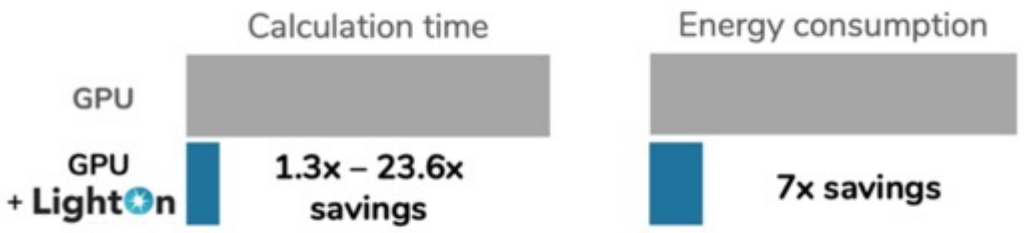
SVD

Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, Halko, N., Martinsson, P., Tropp, J., 2009, arXiv:0909.4061

Some benchmarks

LightOn Appliance for AI acceleration

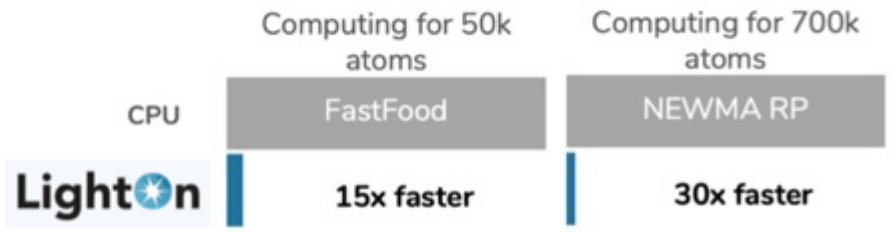
Approximating Kernel Ridge Regression for classification tasks



Performance gains on Kernel ridge approximation for classification tasks. Dataset qm7 (quantum chemistry), high energy physics, and others. The OPU is compared to an NVIDIA P100 GPU (250 W). GPU RAM limit was hit at 32GB. Results acquired extrapolating to 1M features. OPU: Aurora 1.5 (30 W).

Real-time AI analysis of large-scale HPC results

Change detection in Molecular Dynamics



15x faster than FastFood on CPU at 50k atoms. For 700k + atoms, NEWMA RP on OPU is expected to be 30x faster than NEWMA FF on CPU. Library: LightOnML, Dataset: Molecular Dynamics simulations (HPC, Anton), OPU: Aurora 1.5

The pivot

May 2020: supercomputer-sized AI

Training *a single* GPT-3 model :

- 3 Million GPU–hours (on NVIDIA V100s)
- 550 T CO_2 equiv.
- Estimated price 5-10 M $ for training only

Could OPU technology speed this up ?

Deploying in operational environments

Lab experiment

Component

Photonic hardware

System

AI software

+

Photonic hardware

we put an OPU in a #top500 supercomputer !

HPC wire

LightOn Photonic Co-processor Integrated Into France's Jean Zay Supercomputer

December 23, 2021

PARIS, Dec. 23, 2021

Jean Zay supercomputer


LightOn

As we first learned how to train LLMs on GPUs …

we realized there was an immediate market for it !

2021 pivot: decoupling hardware and software

Lab experiment

Component

Photonic hardware

System

AI software

+

Photonic hardware

Software

Large Language Models

Hardware as internal R&D

Contributions to open-source LLMs

Lyra

PAGnol

Rita

Mambaoutai

Alfred

ModernBERT

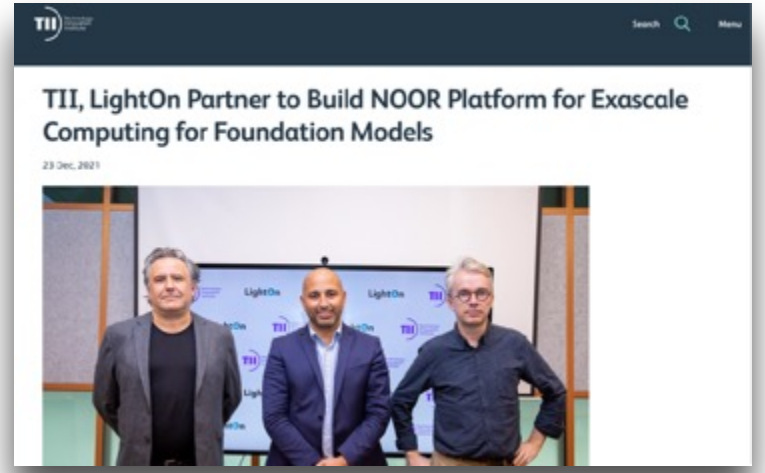
with Answer.AI

NOOR: world's first LLM in Arabic

LightOn partnered with TII Technology Innovation Institute Abu Dhabi to build

noor

The world's first Large Language Model in Modern Standard Arabic - released 2022



Noor powered the first AI-written journal article in MSA
AlEtihad News; Nov 19th, 2022

LightOn in 2025

Our mission :
turn GenAI into an enterprise-ready solution

Finally reaching the top !

Lab experiment

Component

Photonic hardware

System

AI software

+

Photonic hardware

Software

Large Language Models

Hardware as internal R&D

Platform

PARADIGM

The LLM platform for the Enterprise

Turn GenAI into an enterprise-ready solution:

- Safe
- Reliable
- Compliant
- Customizable
- Sustainable

State-of-the-art LLM,
customized to your needs

Lighton

Turn GenAI into an enterprise-ready solution

Multimodal document parsing

On-premises / private cloud with scalable inference

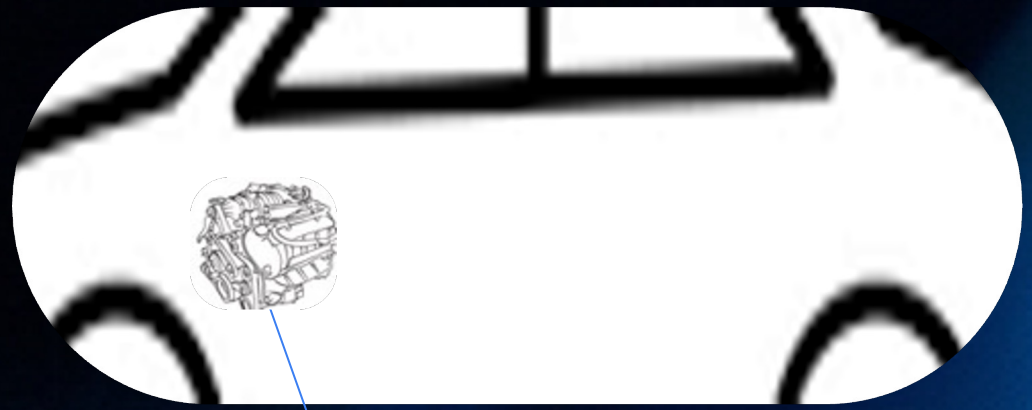
User interface / API

Enhanced RAG pipeline

PARADIGM

Updates management

SSO / role-based access



Workflows / agents

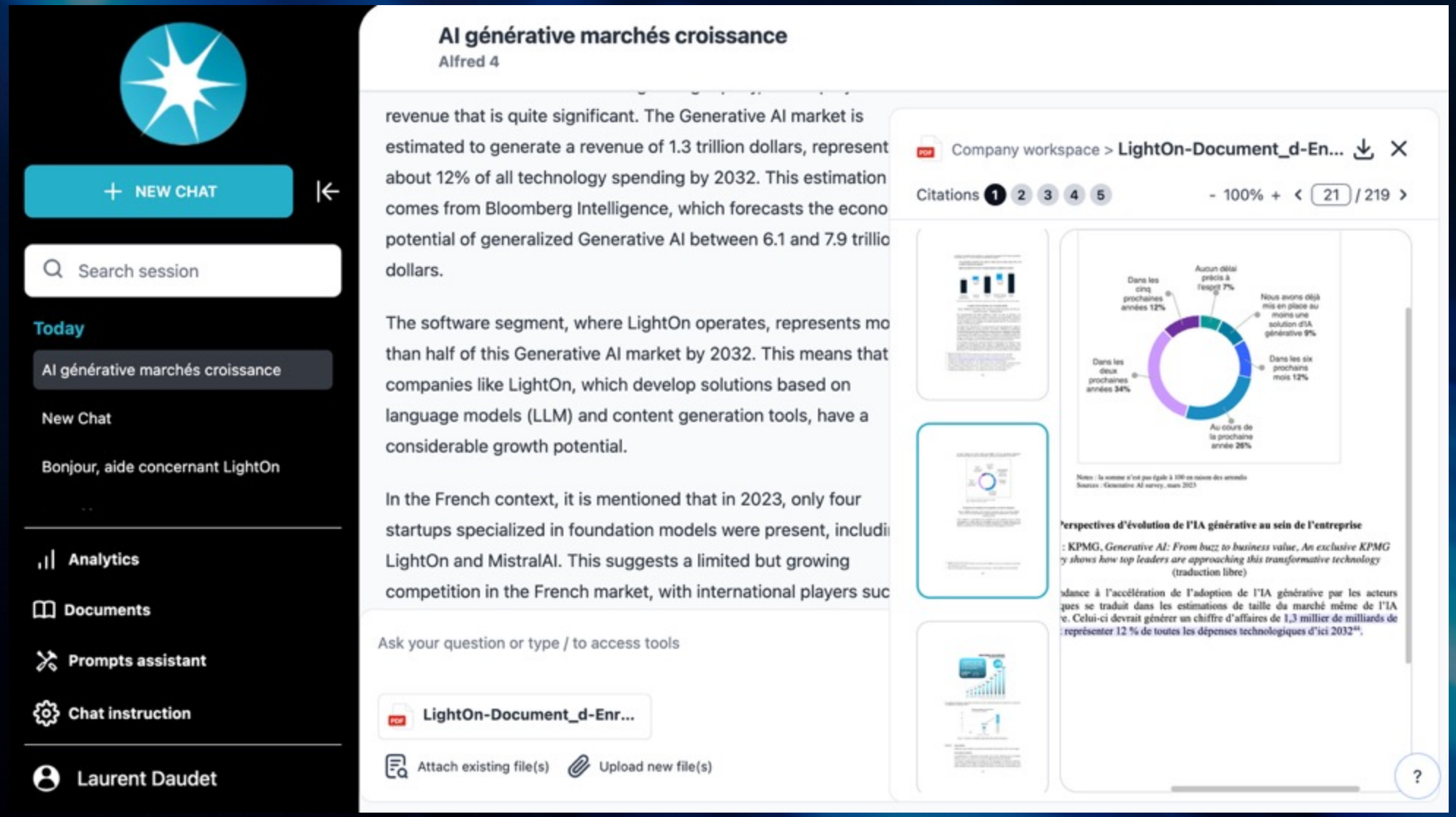
Compliance Audit logs

State-of-the-art LLM, customized to your needs

. . .

Lighton

From documents to insights



More complex workflows

You asked (using OCR)

make a graph illustrating the total vehicle production capacity for each site

Current Installed Annual Vehicle Capacity

Region	Model	Capacity	Status
California	Model S / Model X	100,000	Production
	Model 3 / Model Y	>550,000	Production
Shanghai	Model 3 / Model Y	>950,000	Production
Berlin	Model Y	>375,000	Production
Texas	Model Y	>250,000	Production
	Cybertruck	>125,000	Production
Nevada	Tesla Semi	-	Pilot production
Various	Next Gen Platform	-	In development
TBD	Roadster	-	In development

7BIXJ.png
75.0 KB

OCR executed

Details

Agent answered

Based on the text extracted from the image using OCR, I can provide a graph illustrating the total production capacity of vehicles for each site.

execute_python executed Success

Details

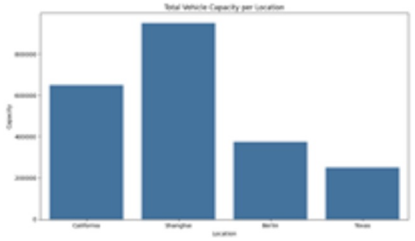


chart.png
21.0 KB

This workflow calls on a range of tools

Optical Character Recognition

Python execution

You asked (using OCR)

make a graph illustrating the total vehicle production capacity for each site

Current Installed Annual Vehicle Capacity

Region	Model	Capacity	Status
California	Model S / Model X	100,000	Production
	Model 3 / Model Y	>550,000	Production
Shanghai	Model 3 / Model Y	>950,000	Production
Berlin	Model Y	>375,000	Production
Texas	Model Y	>250,000	Production
	Cybertruck	>125,000	Production
Nevada	Tesla Semi	-	Pilot production
Various	Next Gen Platform	-	In development
TBD	Roadster	-	In development

7BIXJ.png
75.0 KB

OCR executed

Details

Agent answered

Based on the text extracted from the image using OCR, I can provide a graph illustrating the total production capacity of vehicles for each site.

execute_python executed Success

Details

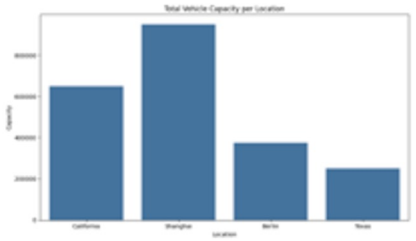


chart.png
21.0 KB

One more thing...

Can you use optics to train a (very) large digital model ?

Wang, Muller, et al., arXiv:2409.12965(2024)

Still plenty of interesting research questions

Lighton

Lab experiment

Component

Photonic hardware

System

AI software

+

Photonic hardware

Software

Large Language Models

Hardware as internal R&D

Platform

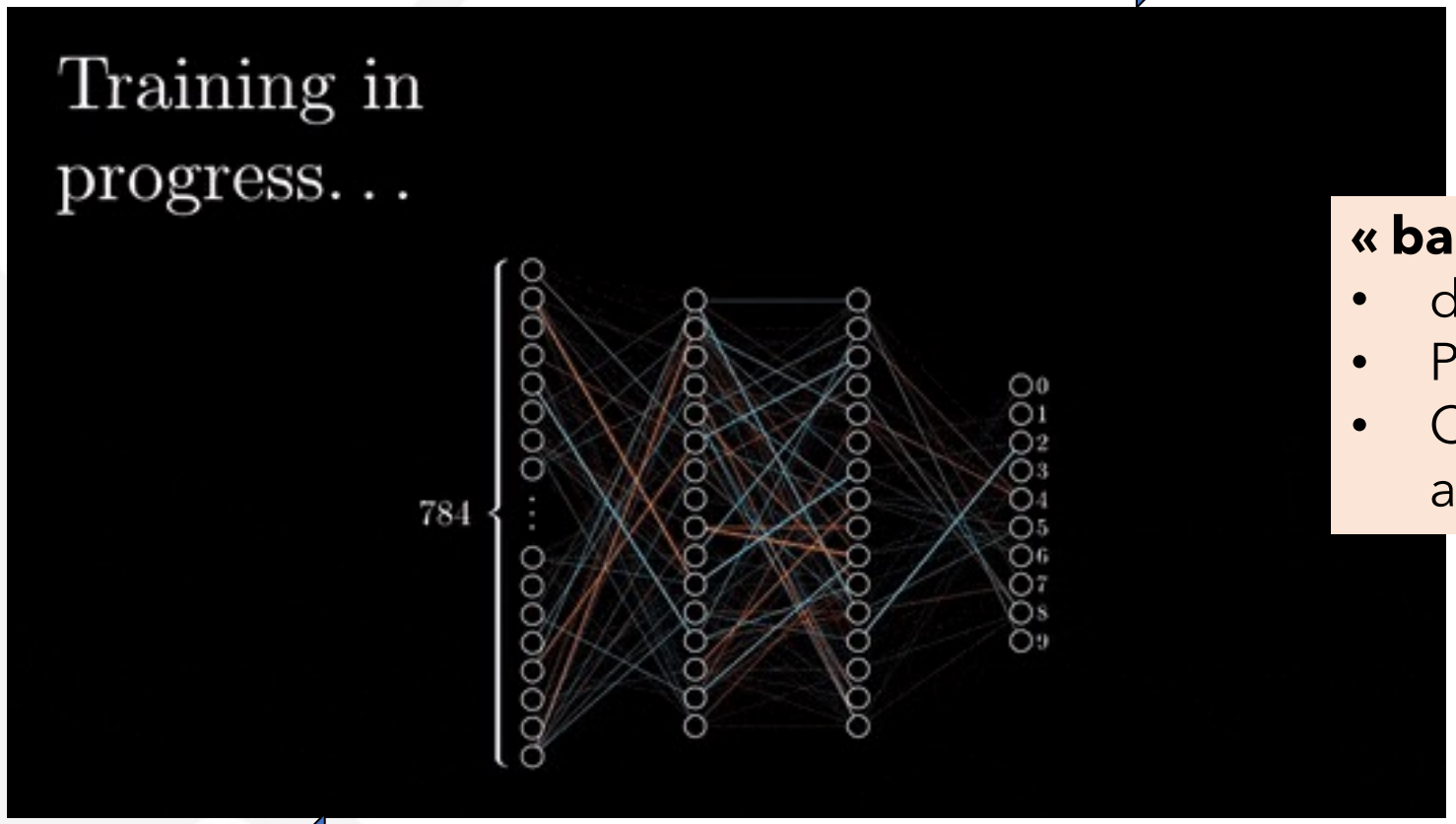
PARADIGM

The LLM platform for the Enterprise

Back to the lab

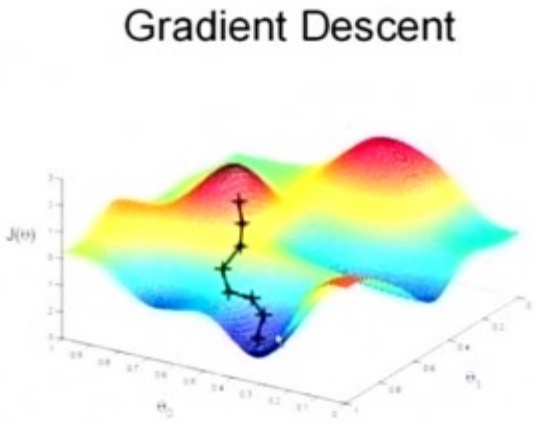
Training with backpropagation

FORWARD : inference



« backprop' »

- de-facto standard for training
- Power-intensive
- Constraints to a layered NN architecture

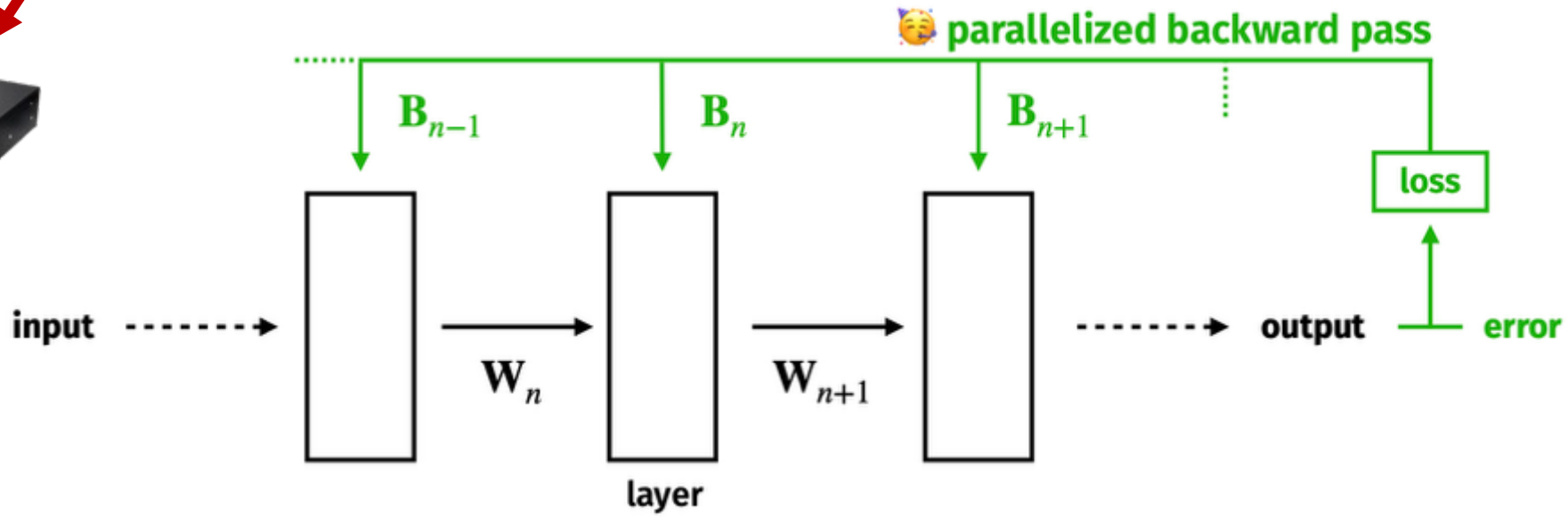


BACKWARD : Error propagation and weight-tuning based on local gradients

Training with Direct Feedback Alignment (DFA)

Lighton

🔁 Use random weights to directly deliver feedbacks from the global loss.



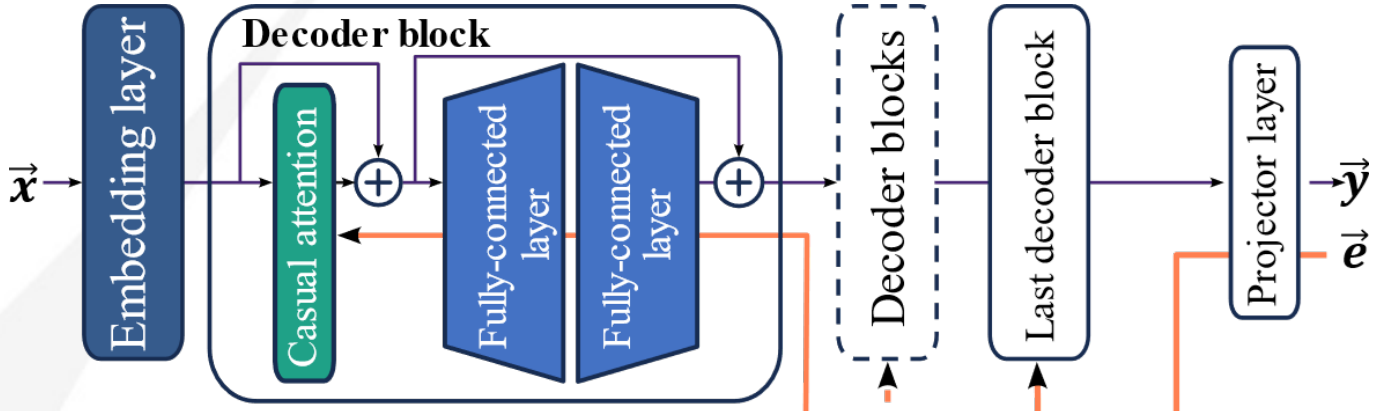
gradient signal from layers above — random projection of the loss gradient

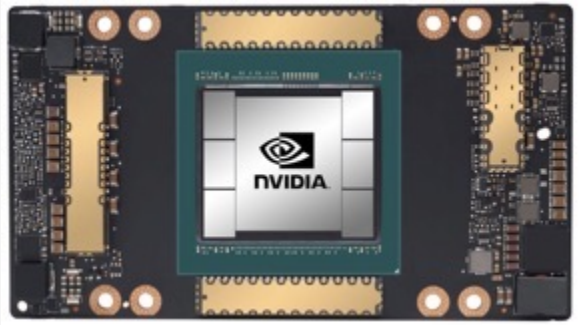
$$\delta \mathbf{W}_i^{BP} = -\left[(\mathbf{W}_{i+1}^T \delta \mathbf{a}_{i+1}) \odot f_i'(\mathbf{a}_i)\right]\mathbf{h}_{\mathbf{i-1}}^{\mathbf{T}} \longrightarrow \delta \mathbf{W}_i^{DFA} = -\left[(\mathbf{B}_i^T \mathbf{e}) \odot f_i'(\mathbf{a}_i)\right]\mathbf{h}_{\mathbf{i-1}}^{\mathbf{T}}$$

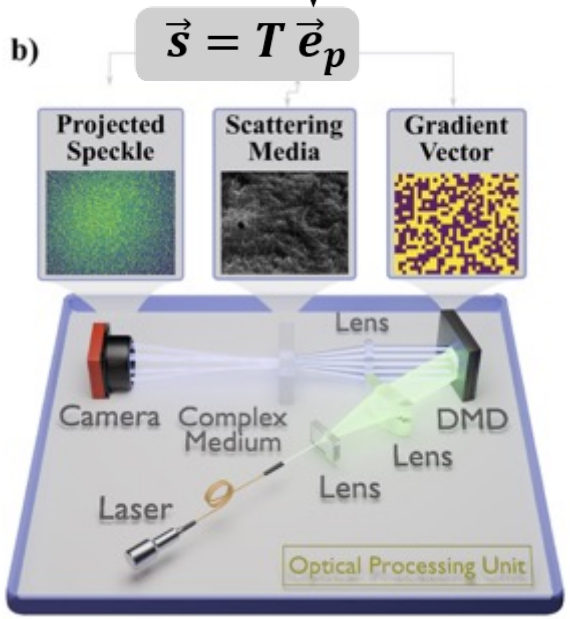
General-purpose: agnostic to model architecture & scale well

Original paper : Arild Nøkland, NeurIPS 2016 arXiv:1609.01596

ODFA on Text Transformer







- 1B parameters GPT-like transformer
- 400M parameters directly receive optical signal

Cornell Movie-Dialogs Corpus
Example in the Dataset

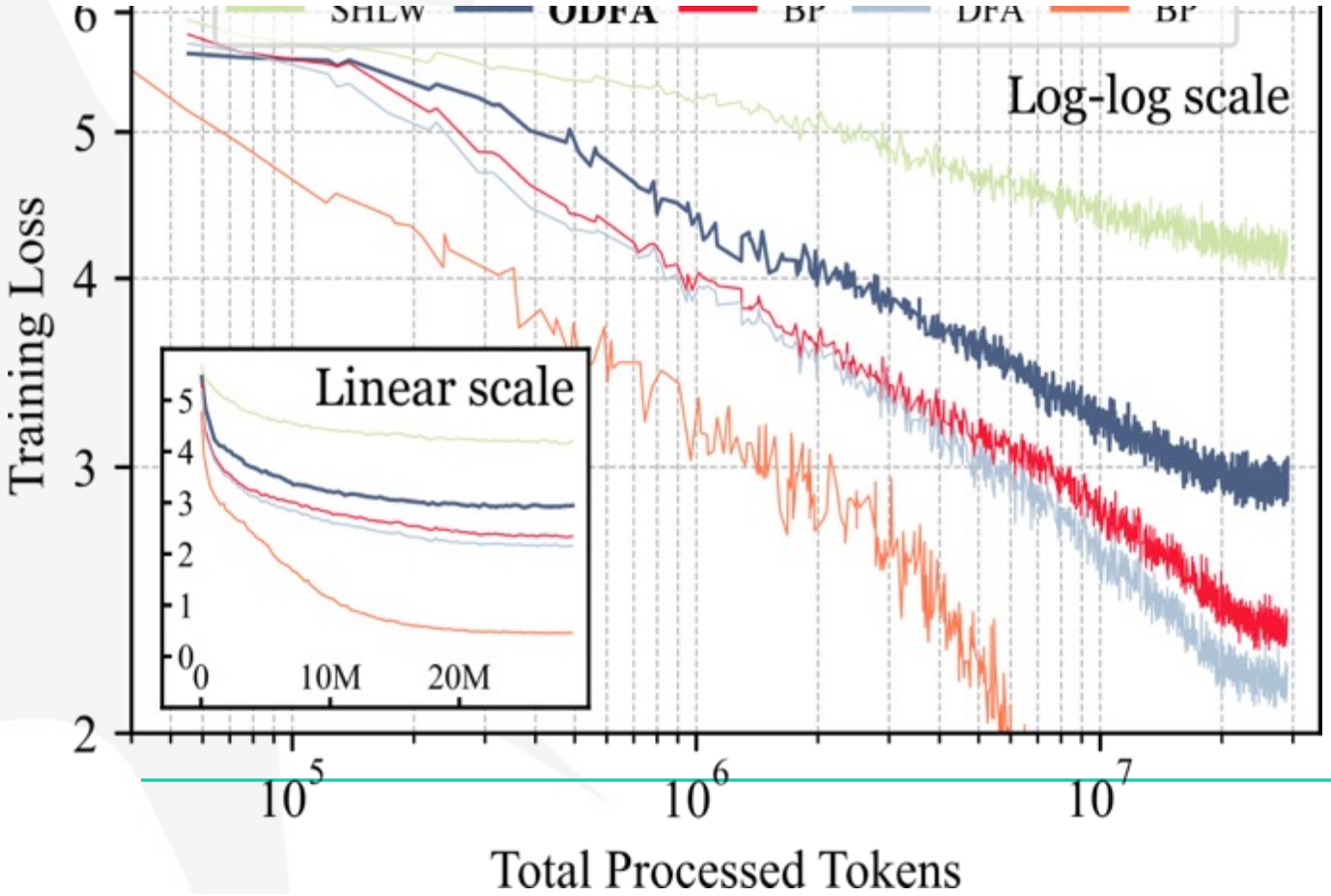
- MILES: Back at you.
- JACK: Love you, man.
- MILES: Yeah.
- JACK: So I'll see you at the rehearsal.

Ziao Wang
(LKB)

Kilian Müller
(LightOn)

Wang, Muller, et al., arXiv:2409.12965(2024)

ODFA on Text Transformer: Performance

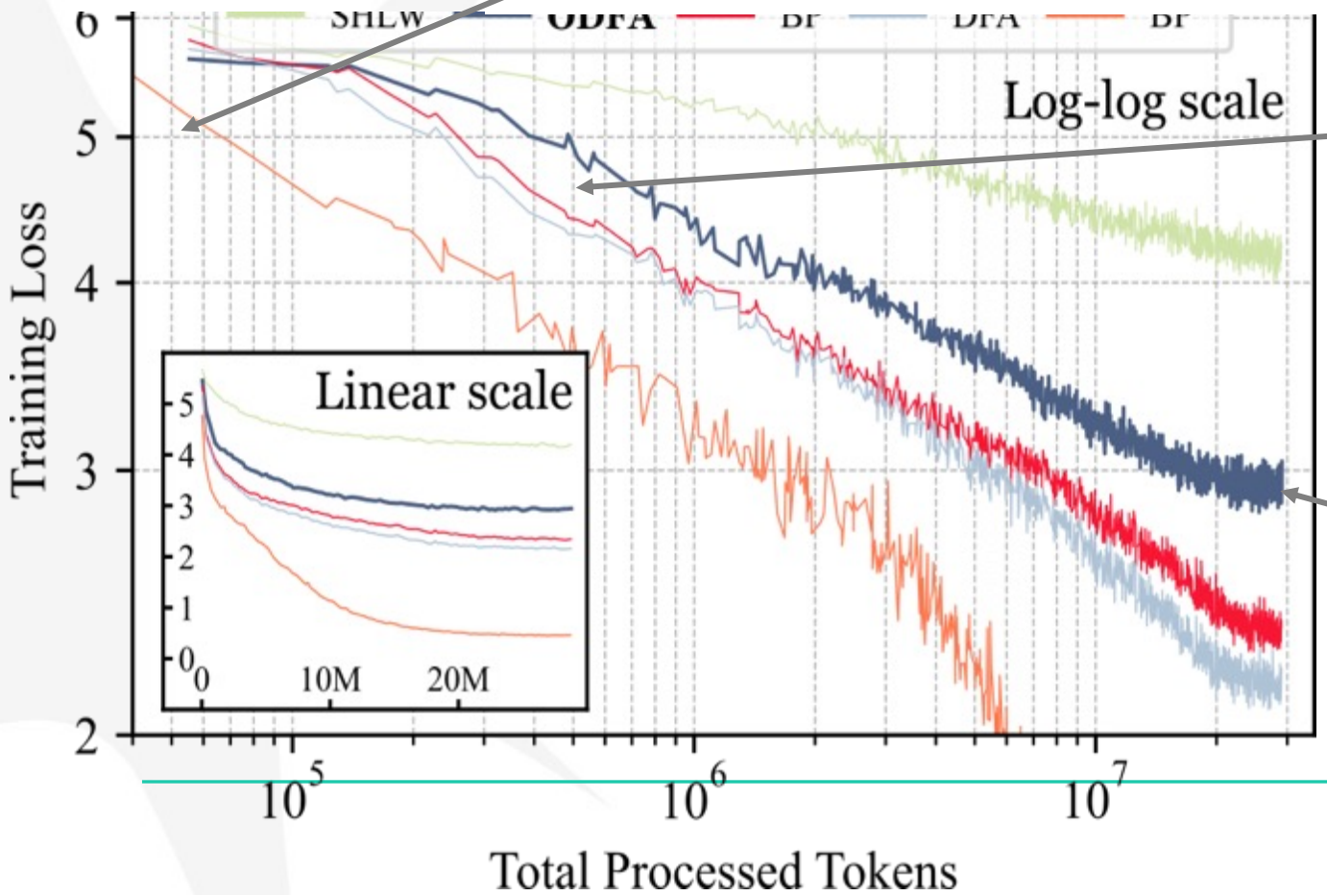


Wang, Muller, et al., arXiv:2409.12965(2024)

ODFA on Text Transformer: Performance

0% Training

aQ o do wantQain do do twoain o
re[through show" happenQ start. 2lt
JULIANNE JULIANNE



20% Training

JACKIE : Snkups to the brags, but I know that. ↲
CRAYC : It ' s just a concer. ↲

100% Training

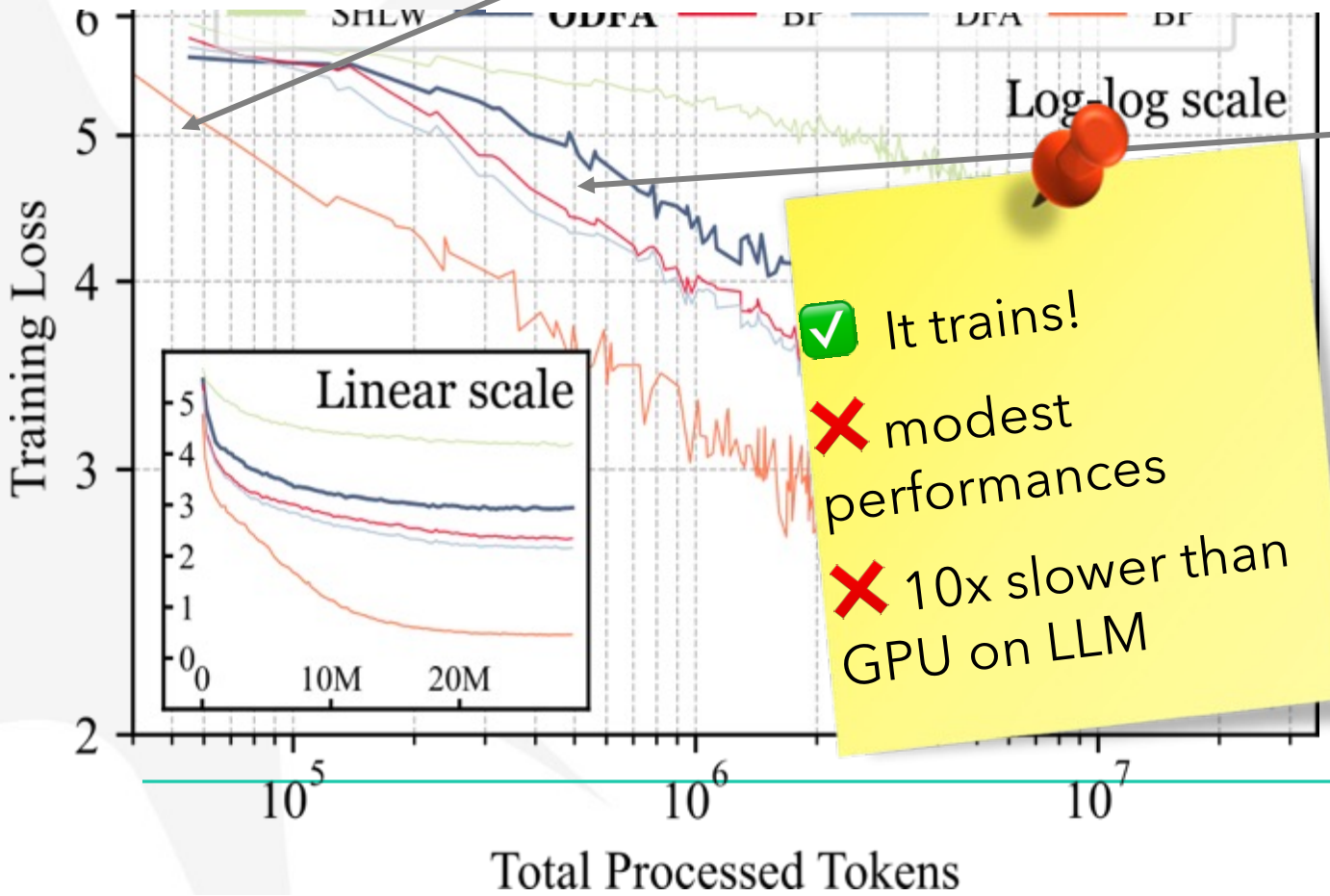
JOE : So what are you talking about? ↲
ED : I mean I'm sorry, you know how to know you are. ↲

Wang, Muller, et al., arXiv:2409.12965(2024)

ODFA on Text Transformer: Performance

0% Training

aQ o do wantQain do do twoain o
re[through show" happenQ start. 2lt
JULIANNE JULIANNE



20% Training

JACKIE : Snkups to the brags, but I know that. ↲
CRAYC : It ' s just a concer. ↲

✅ It trains!

❌ modest performances

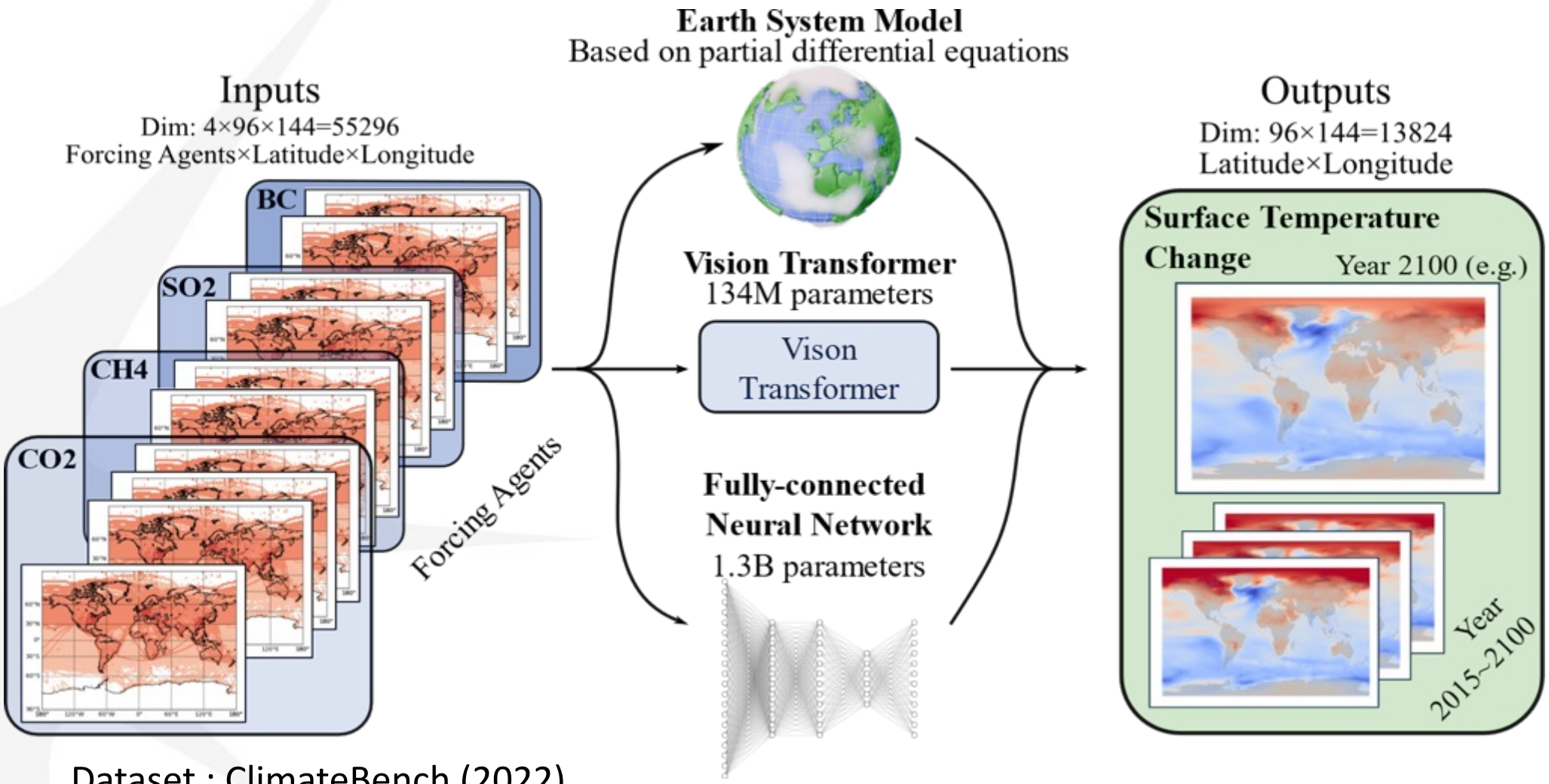
❌ 10x slower than GPU on LLM

100% Training

JOE : So what are you talking about? ↲
ED : I mean I'm sorry, you know how to know you are. ↲

Wang, Muller, et al., arXiv:2409.12965(2024)

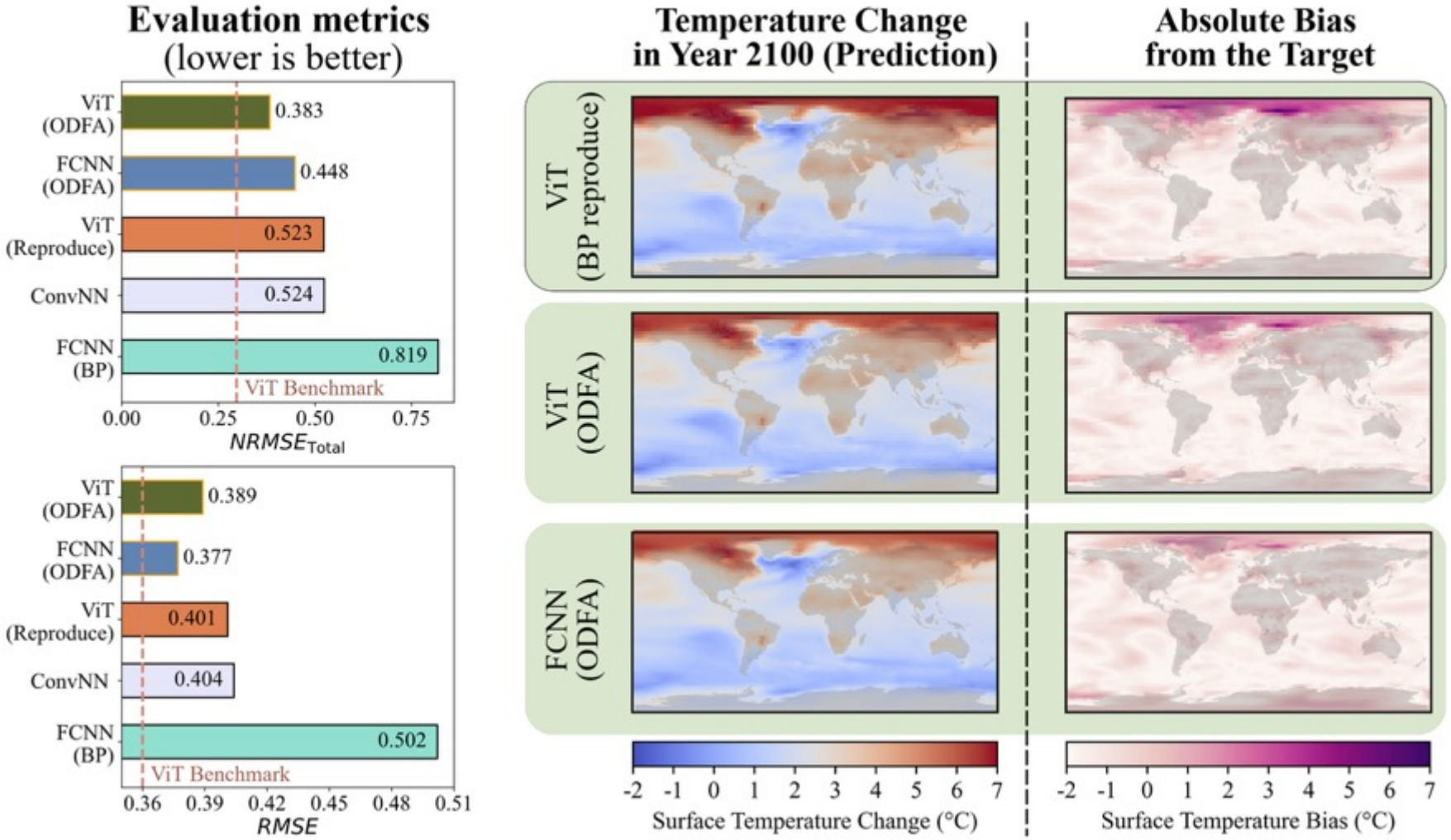
Vision Task: More Suitable for ODFA



Dataset : ClimateBench (2022)

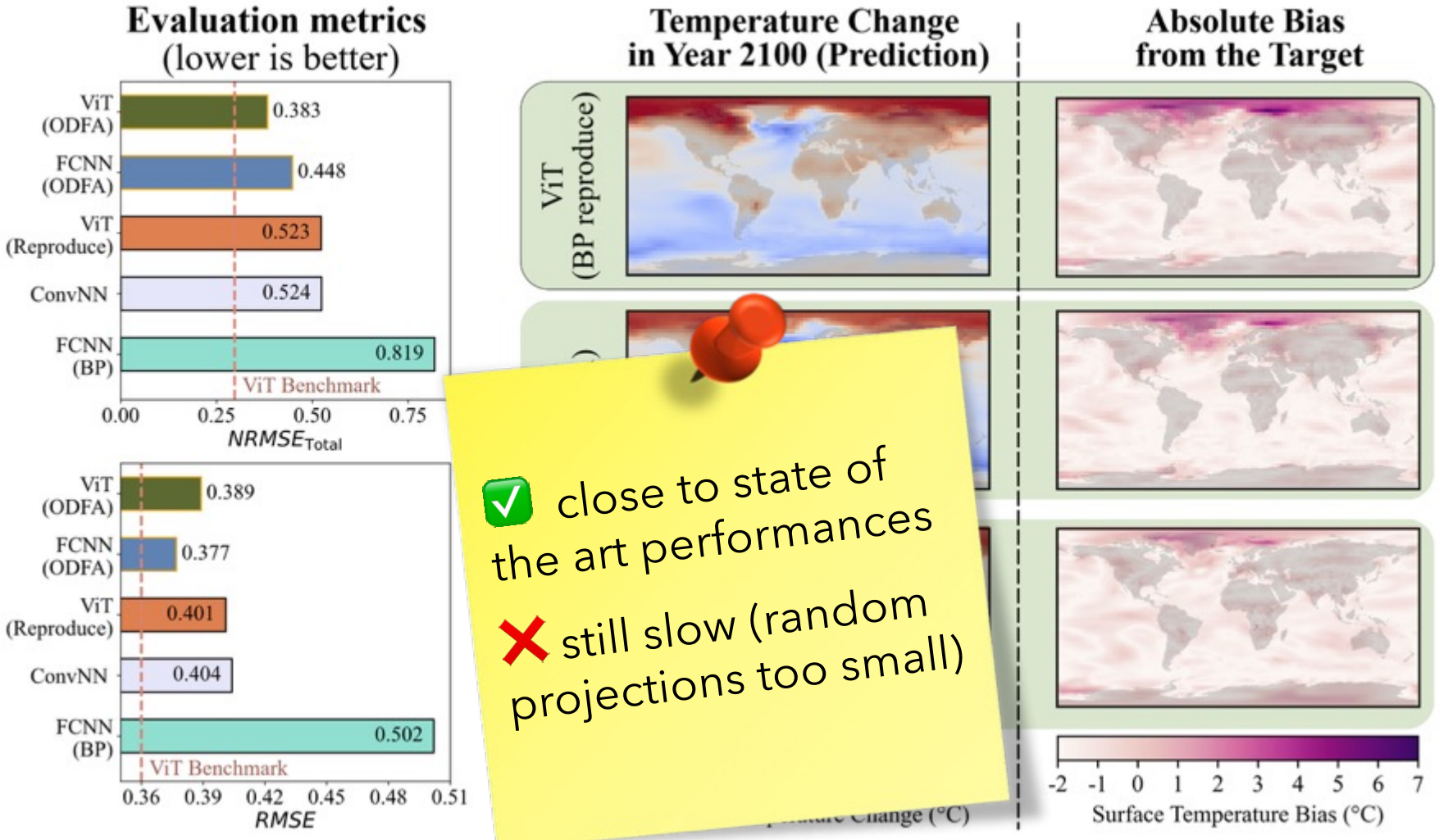
Wang, Muller, et al., arXiv:2409.12965(2024)

FCNN and ViT with ODFA on Climate Projection



Wang, Muller, et al., arXiv:2409.12965(2024)

FCNN and ViT with ODFA on Climate Projection

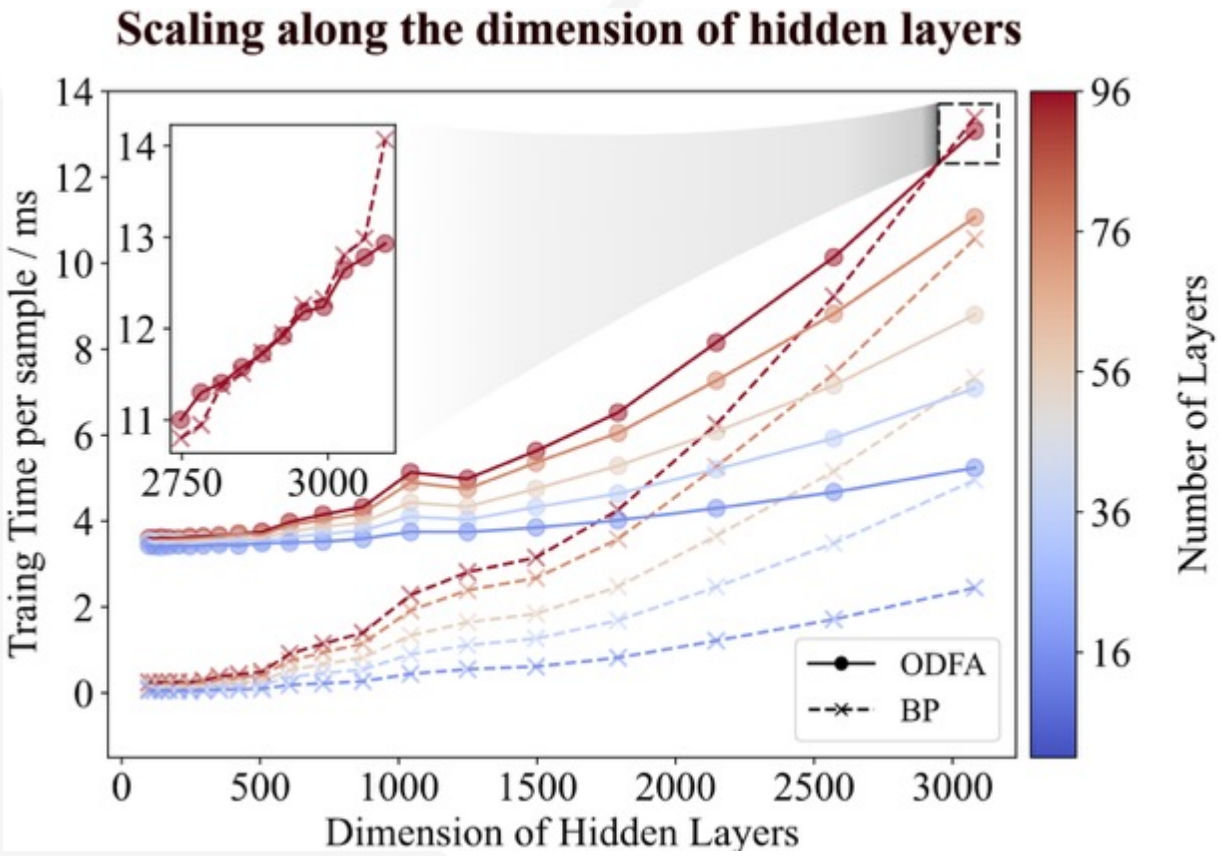


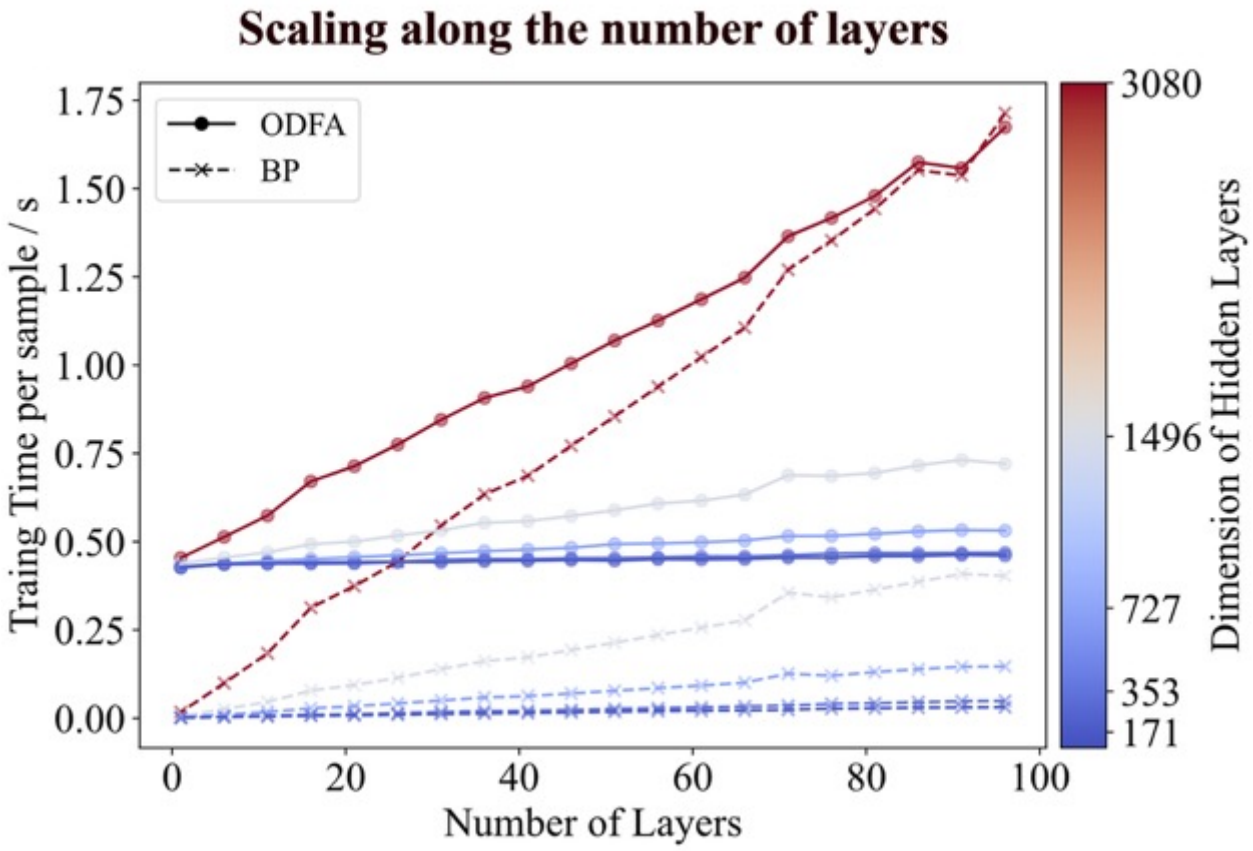
✅ close to state of the art performances

❌ still slow (random projections too small)

Wang, Muller, et al., arXiv:2409.12965(2024)

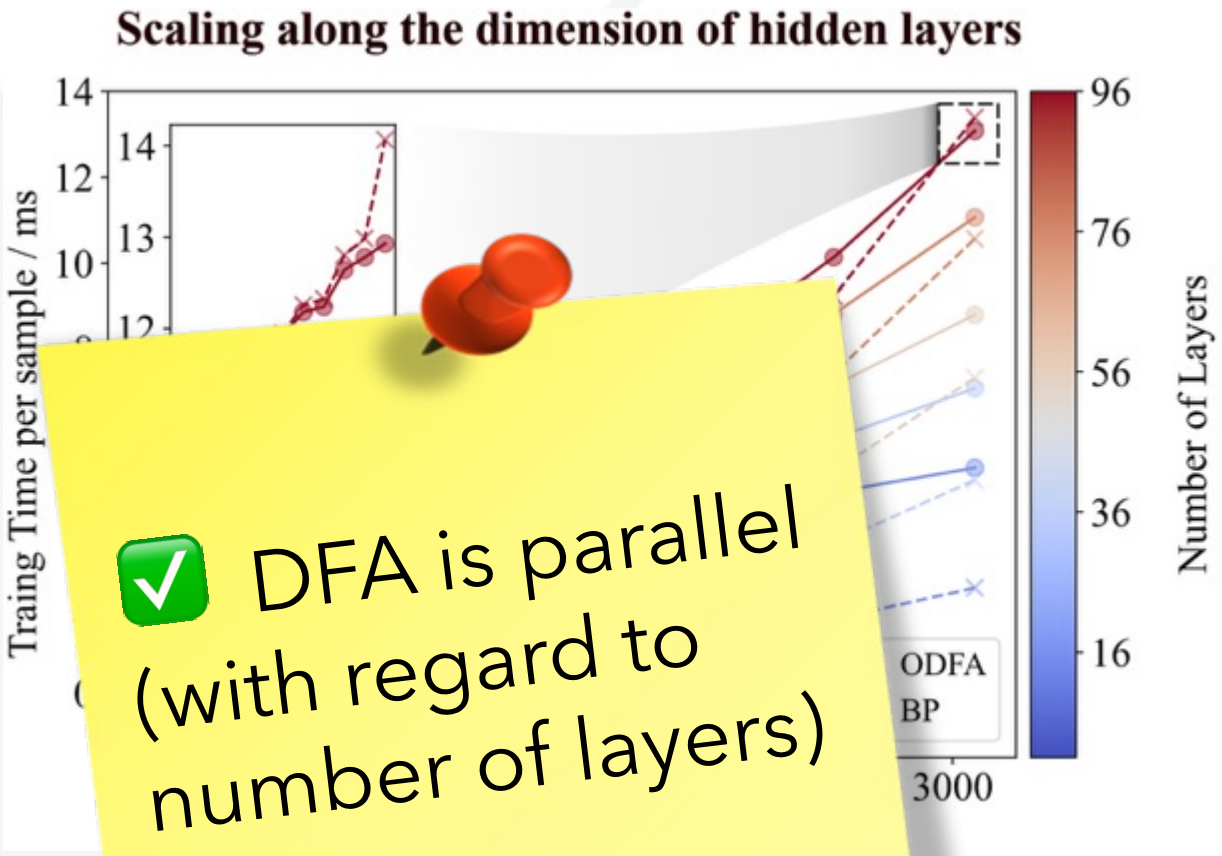
Towards Extreme Scale: When will ODFA be faster than BP?



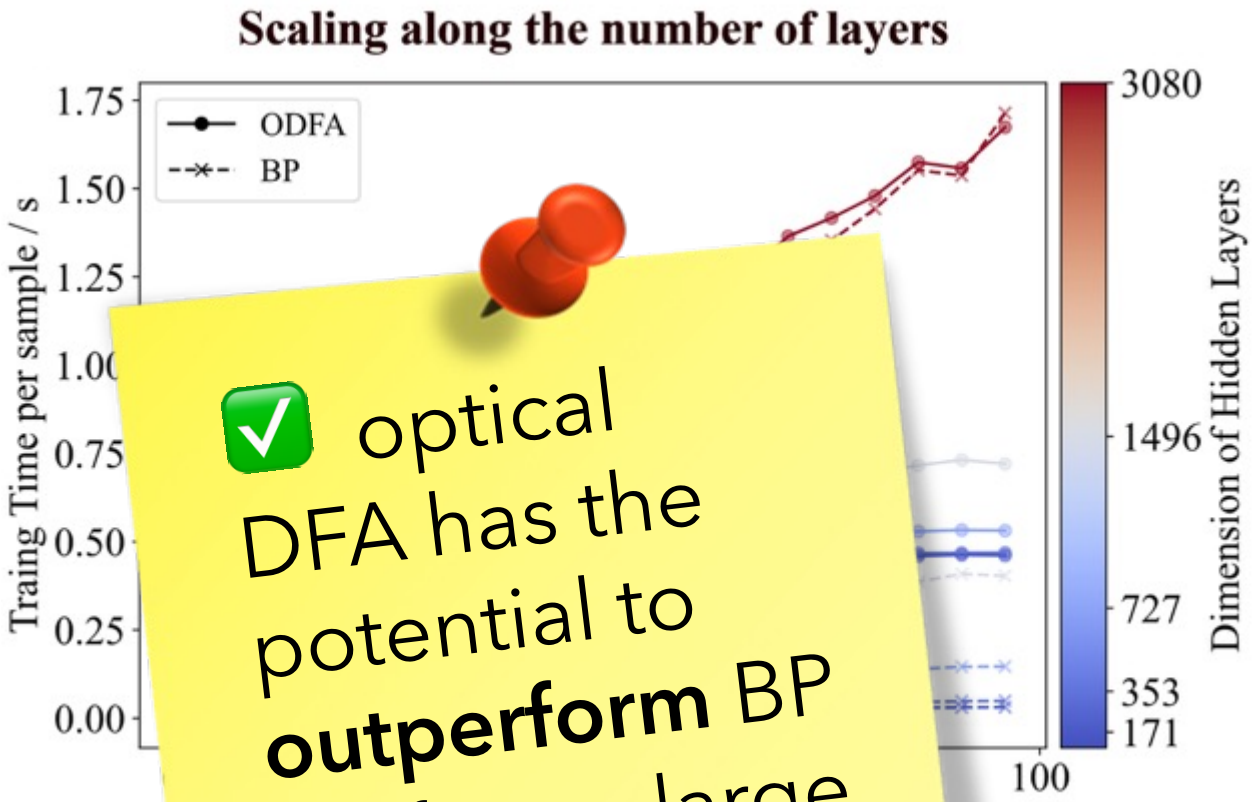


Wang, Muller, et al., arXiv:2409.12965(2024)

Towards Extreme Scale: When will ODFA be faster than BP?



✅ DFA is parallel (with regard to number of layers)



✅ optical DFA has the potential to **outperform** BP for future large models !

Wang, Muller, et al., arXiv:2409.12965(2024)

By way of conclusion

The first publicly traded GenAI company in Europe !

Nov 26th, 2024

Lighton

The entrepreneurial journey