

A large, light grey graphic of the letter 'C' composed of several concentric, slightly offset curved segments, positioned on the left side of the slide.

# Training larger, more accurate models with less compute

*Cerebras Systems for The AI Conference 2023*

27 September 2023

A. Hock, PhD

VP Product, Cerebras Systems

Large language models (LLM) and generative AI (genAI)  
have **transformative potential for enterprise**

Large language models (LLM) and generative AI (genAI)  
have **transformative potential for enterprise**

So...if you're an enterprise data scientist, ML/AI researcher, or CxO,  
figuring out **how to build the right model for your business can be daunting**

Large language models (LLM) and generative AI (genAI)  
have **transformative potential for enterprise**

So...if you're an enterprise data scientist, ML/AI researcher, or CxO,  
figuring out **how to build the right model for your business can be daunting**

Going it alone **using traditional infrastructure is complex and costly;**  
we know **there's a better way...**



cerebras



# The Wafer-Scale Engine

- The **largest chip ever made**, the heart of Cerebras' AI computing solution
- Built for this work: outperforms traditional, general-purpose processors by **orders of magnitude** on every dimension relevant to AI computation

**Cluster-scale AI in a single chip**



# The Cerebras CS-2 System

The world's fastest  
AI accelerator

- ✓ Deploy easily into existing racks
- ✓ Cluster-scale in a single system
- ✓ Datacenter-scale in a cluster
- ✓ Available on-prem or remote / in cloud





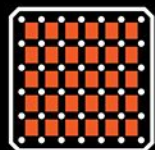


# Condor Galaxy 1 AI Supercomputer



**64**

CS-2 nodes



**54 million**

AI cores



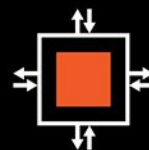
**4 exaFLOPS**

AI compute  
at FP16



**82 TB**

parameter  
memory



**388 Tbps**

internal  
bandwidth



**72,704**

AMD EPYC™  
cores



**10 days**

to first  
training run







# The right platform for enterprise LLM / genAI

At Cerebras, we provide **end-to-end support to accelerate model training** and fine-tuning:

- **Hardware:** We build the fastest AI accelerators that train 10B+ models in days not months.
- **Software:** Our platform scales models from 1B to 1T parameters with no code changes.
- **Expertise:** We help our customers build state-of-the-art models optimized for their domains.

# Cerebras-GPT

## A family of state-of-the-art GPTs

- 111M-13B parameters
- The first family of GPT models using Chinchilla's compute-efficient recipe
- First scaling laws developed on public data
- Permissive Apache 2.0 open-source license

## Built entirely on Cerebras systems

- With a few people over a few weeks (not hundreds over months)
- The only large language models trained using exclusively data-parallel execution
- Largest models trained from scratch on non-GPU/TPU systems

...illustrative as a case study on **how to build the right model for your work**

## Cerebras-GPT: Open Compute-Optimal Language Models Trained on the Cerebras Wafer-Scale Cluster

Nolan Dey, Gurpreet Gosal, Zhiming (Charles) Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, Joel Hestness

Cerebras Systems {nolan,joel}@cerebras.net

### Abstract

We study recent research advances that improve large language models through efficient pre-training and scaling, and open datasets and tools. We combine these advances to introduce Cerebras-GPT, a family of open compute-optimal language models scaled from 111M to 13B parameters. We train Cerebras-GPT models on the Eleuther Pile dataset following DeepMind Chinchilla scaling rules for efficient pre-training (highest accuracy for a given compute budget). We characterize the predictable power-law scaling and compare Cerebras-GPT with other publicly-available models to show all Cerebras-GPT models have state-of-the-art training efficiency on both pre-training and downstream objectives. We describe our learnings including how Maximal Update Parameterization ( $\mu P$ ) can further improve large model scaling, improving accuracy and hyperparameter predictability at scale. We release our pre-trained models and code, making this paper the first open and reproducible work comparing compute-optimal model scaling to models trained on fixed dataset sizes. Cerebras-GPT models are available on HuggingFace: <https://huggingface.co/cerebras>.

# SlimPajama

The **right data** matters...

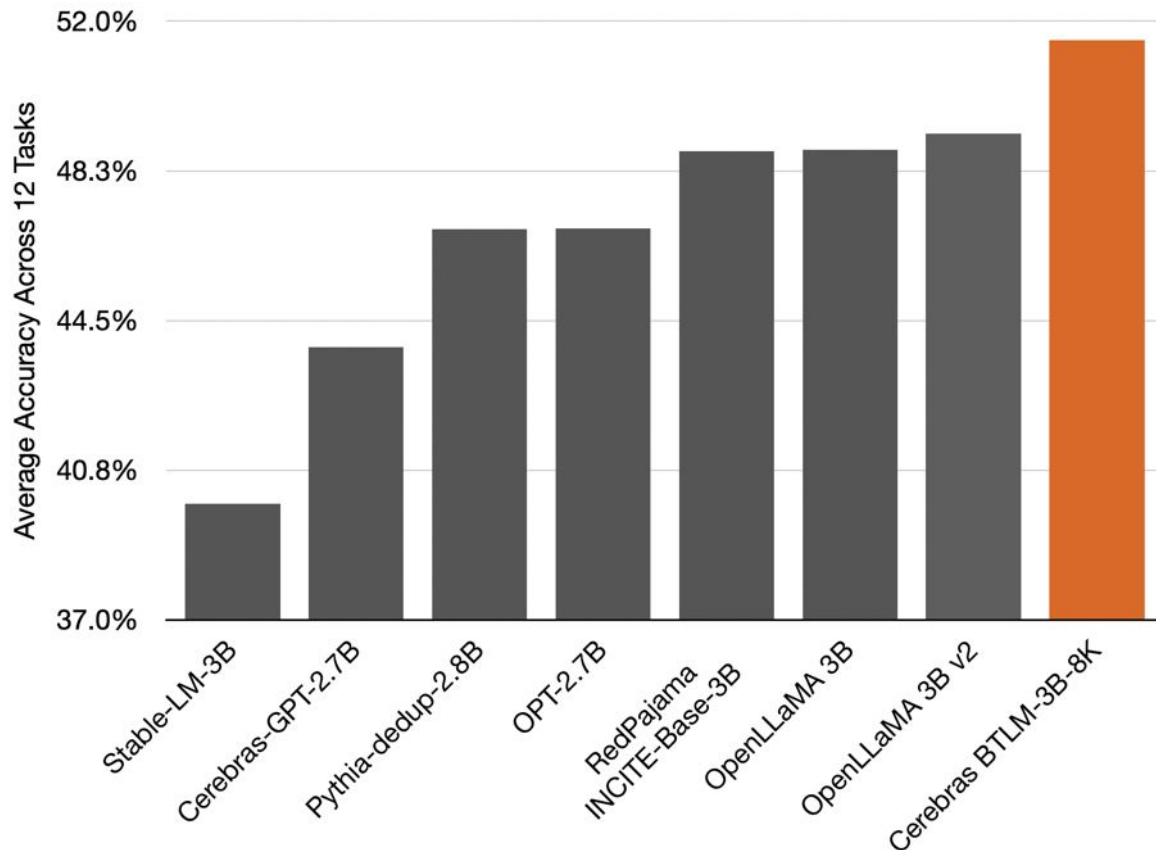
- Largest fully deduplicated, multi-corpora, open-source dataset: 627 Billion tokens
  - The highest quality open-source English dataset available
- Enables model training to the same accuracy in half the time and half the power of the parent dataset RedPajama 1.2T
  - Data pre-processing scripts open-sourced on [Cerebras Github](#)
- Done in **partnership** with OpenTensor and TOGETHER





# BTLM-3B-8K: 7B performance in a 3B model

BTLM-3B-8K: The New State-of-the-Art 3B Model



Bringing the **right compute, right data, and latest methods** to bear

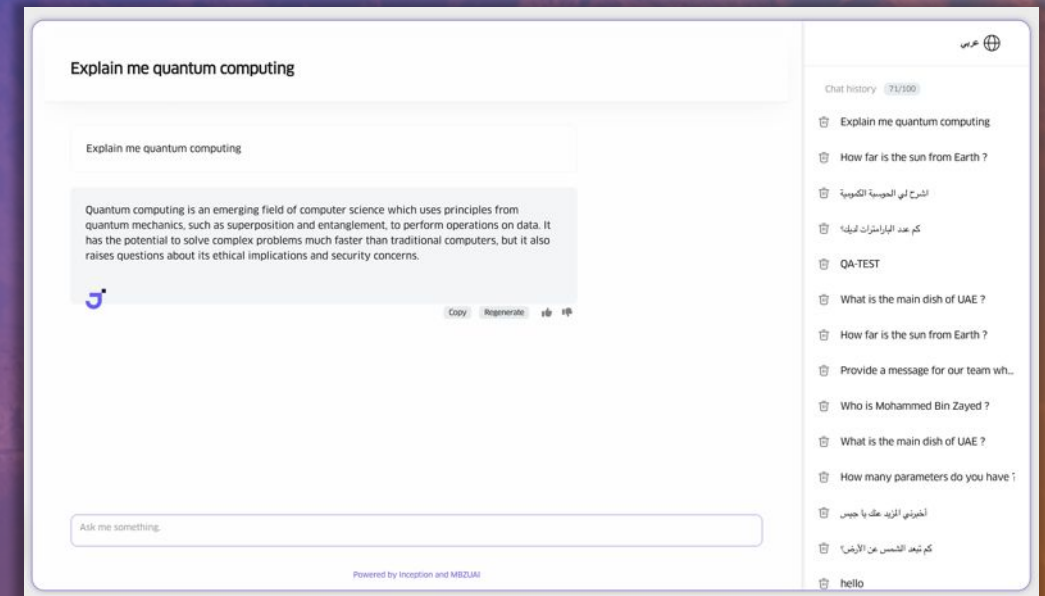
- Model by Cerebras for OpenTensor
  - Trained on Condor Galaxy 1
  - Using SlimPajama dataset
- Highest accuracy 3B model
  - Sets a new benchmark across 12 English language tasks
  - Supports 8K sequence length inference
  - Designed for extremely efficient low cost inference
  - Inference runs natively on MacBook, iPhone, Raspberry Pi thanks to 3GB memory footprint
- Competitive even with leading 7B models

# Introducing Jais

Bringing it all together for **global enterprise**...

- The world's best-performing Arabic Large Language Model
- A partnership between Inception, MBZUAI and Cerebras Systems
- 13-billion-parameter, open-source, bilingual Arabic-English model
  - Trained on 116 billion Arabic tokens and 279 billion English tokens of data
  - Trained on Condor Galaxy 1 AI Supercomputer
  - Homegrown in Abu Dhabi to bring the power of Generative AI to 400m Arabic speakers
  - Available for download on HuggingFace

**From kickoff to SoTA model in just a few weeks.**



# Language modeling best practices

- Invest in the **right data**: cleaned, de-duplicated
- Define relevant **evaluation criteria** and representative tests, harnesses in advance
- Run many experimental **sub-scale trainings** to define the “ML recipe” for your use case:
  - Data tokenization, mix(es), MSL(s)
  - Training parameters (e.g. using muP)
  - Scaling laws
- To iterate and converge on this ^ quickly, you need the **right hardware platform**...
  - Simple programming – minimize setup time
  - Large model and methods support “out of the box” – minimize idiosyncratic hardware optimizations
  - Seek out simple and ~linear performance scaling – when you go big, you want to go fast ;)
- ...with the **right partner(s)**
  - The full solution often involves not just compute
  - Expertise: data, model/ML, business domain



# Wrapping up

Cerebras was built for this: we are **the world's premier LLM / genAI model factory**

- The right hardware platform: fastest AI compute; built for this work
- World-class ML researchers and engineers that work directly with our customers and partners
- From zero to industry-leading models tailored to your business in weeks-months, not years

What we see next / **over the horizon?**

- More, larger, multi-modal models – deployed into end-user production applications
- More high quality, multi-modal, and non-English datasets

Curious how generative AI can help and how to **build the right models for your business?**

Standing up a new foundation model project or **large-scale AI initiative?**

We can't wait to talk to you.

**Thank you!**

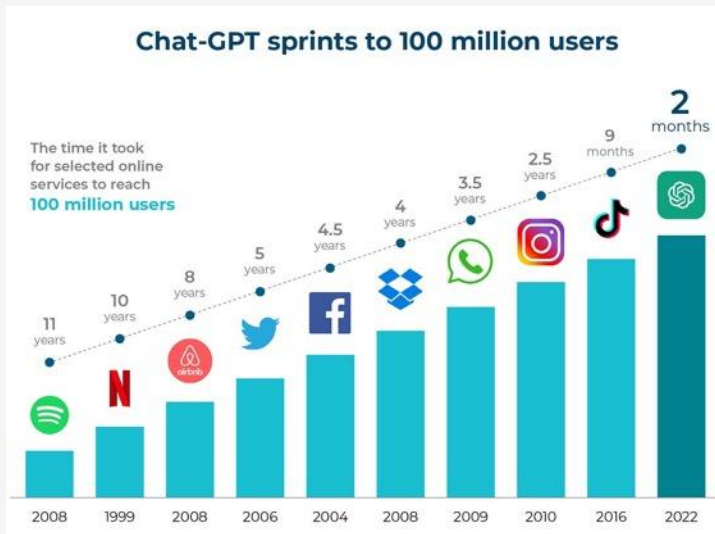


# Backup

# Generative AI Is Big News

## User Demand Exploding

ChatGPT among the fastest growing apps in history



## Investment Soaring

Investor interest in generative AI at an all-time high: Billions invested

### Microsoft Invests \$10 Billion in ChatGPT Maker OpenAI

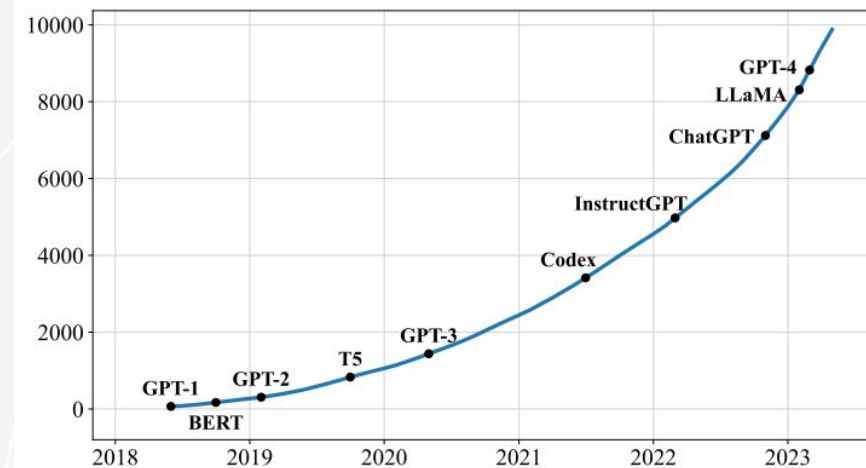
Microsoft-backed AI startup Inflection raises \$1.3 billion from Nvidia and others

AI startup Cohere, now valued at over \$2.1B, raises \$270M

2. [Runway](#), \$141M, artificial intelligence: New York-based [Runway](#) made this list not that long ago after [Business Insider](#) reported on this

## Generative AI Is Top Of Mind

Articles that include “language model” are everywhere in the press

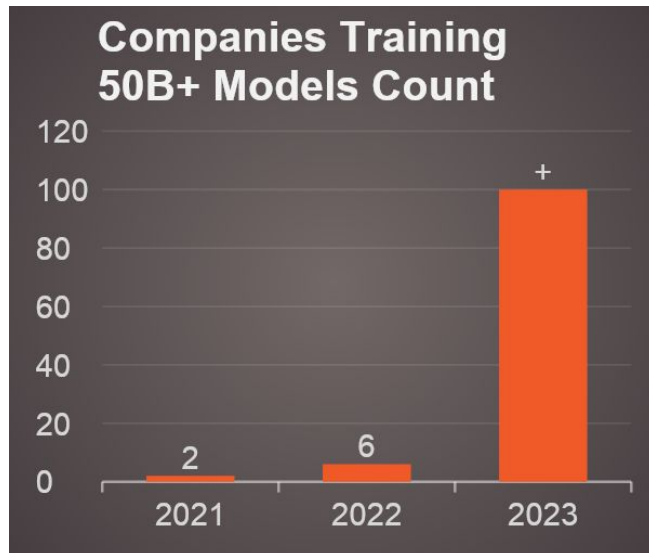




# The Foundation of Generative AI: Compute, Models, and Data

## Compute

Generative AI has a vast and insatiable demand for compute



## Models

New models, new applications are proliferating and are now incorporated across industries

A table titled "Generative AI use cases across industries" showing various use cases and their application across different industries. The industries are eCommerce, Healthcare, Travel and transportation, Manufacturing and supply chain, Utilities, Entertainment, and Education. The use cases are Chatbots and virtual assistants, Design and development, Content creation and repurposing, Data analytics, Risk mitigation, and Predictive maintenance. Checkmarks indicate the presence of each use case in each industry.

	eCommerce	Healthcare	Travel and transportation	Manufacturing and supply chain	Utilities	Entertainment	Education
Chatbots and virtual assistants	✓	✓	✓	✓	✓	✓	✓
Design and development	✓	✓	✓	✓	✓	✓	✓
Content creation and repurposing	✓	✓	✓			✓	✓
Data analytics	✓	✓	✓	✓	✓	✓	✓
Risk mitigation	✓	✓	✓	✓	✓	✓	✓
Predictive maintenance		✓	✓	✓	✓		

## Data

Data is the “new gold”... Datasets are huge, ranging from 600 Billion – 5 Trillion tokens

Dataset Name	Tokens
RefinedWeb-600B	600B
SlimPajama	627B
MPT	1T
RedPajama	1.21T
LLaMa	1.4T
MassiveText	1.4T
RefinedWeb-5T	5T



# Cerebras Wafer-Scale Engine

Outperforms state-of-the-art chips across key dimensions

	Cerebras WSE-2	Nvidia H100*	Cerebras vs. H100
<b>Release date</b>	04/2021	09/2022	
<b>Chip size</b>	46,225 mm <sup>2</sup>	814 mm <sup>2</sup>	<b>56 X</b>
<b>Cores</b>	850,000	16,896 + 528	<b>50 X</b>
<b>On-chip memory</b>	40 GB	50 MB	<b>800 X</b>
<b>Memory bandwidth</b>	20 PB/sec	3.0 TB/sec	<b>6,667 X</b>
<b>Fabric bandwidth</b>	220 Pb/sec	7.2 Tb/sec	<b>30,555 X</b>

# Cerebras Wafer-Scale Cluster

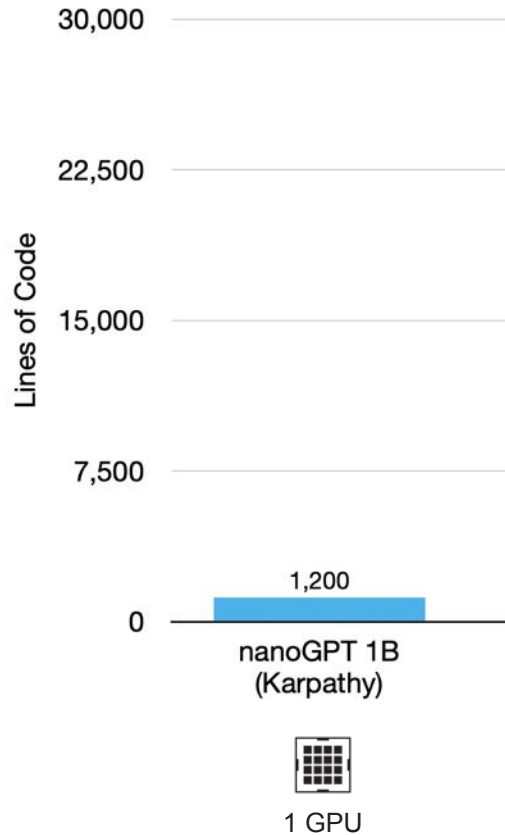
Purpose-built high performance, scalable appliance

- Designed to scale from one to thousands of CS-2s and achieve linear performance improvements
- Can support models with trillions of parameters
- Purposefully built to remove any bottlenecks during training





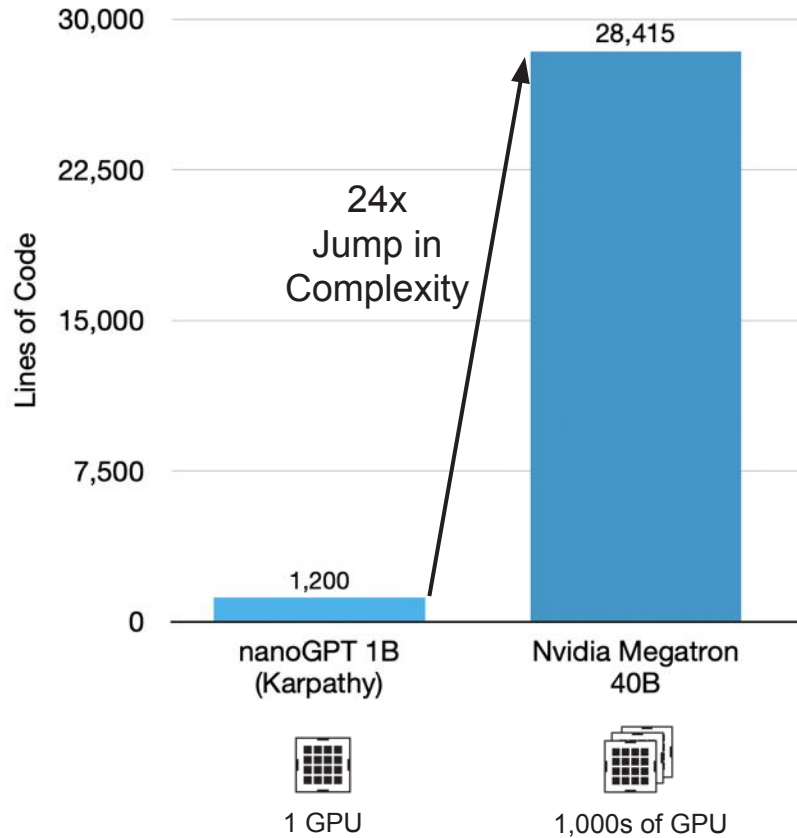
# Cerebras Eliminates The Complexity In Training Large Models



A 1B parameter is simple to write and train on one GPU.

But it takes an army of engineers and 40,000 lines to train a 100B parameter across thousands of GPUs.

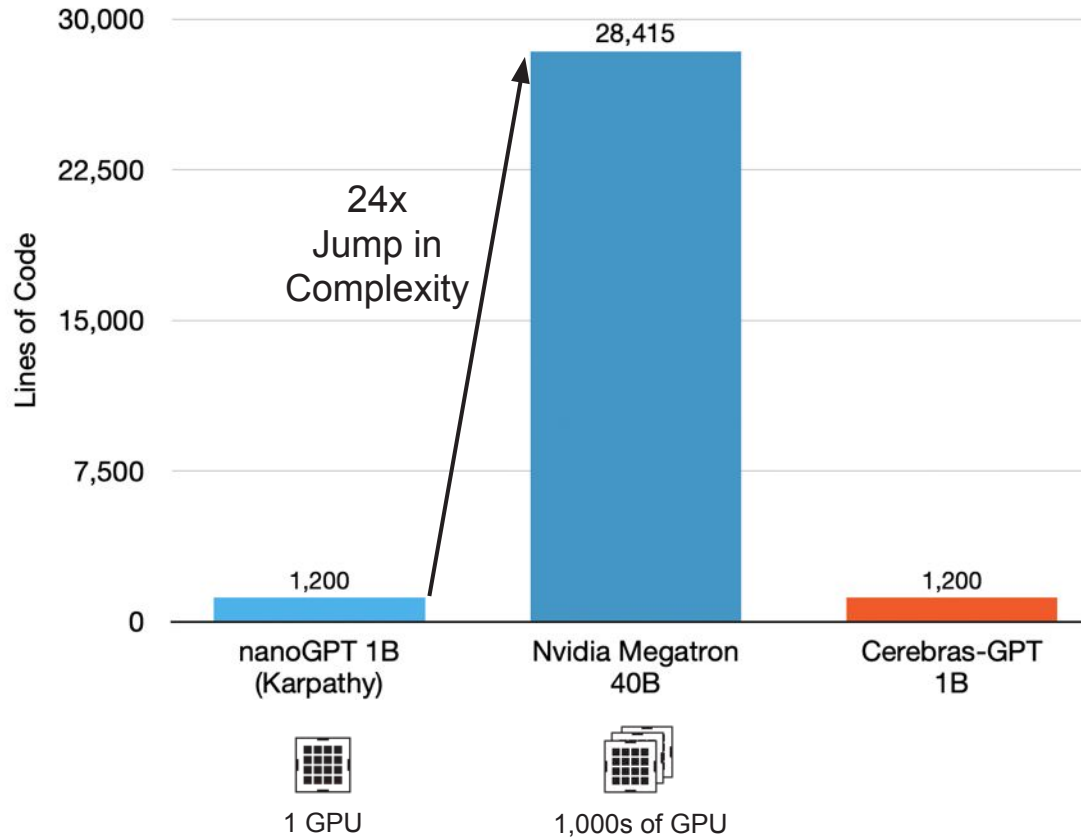
# Cerebras Eliminates The Complexity In Training Large Models



A 1B parameter is simple to write and train on one GPU.

But it takes an army of engineers and 28,415 lines of code to train a 40 Billion parameter model across thousands of GPUs.

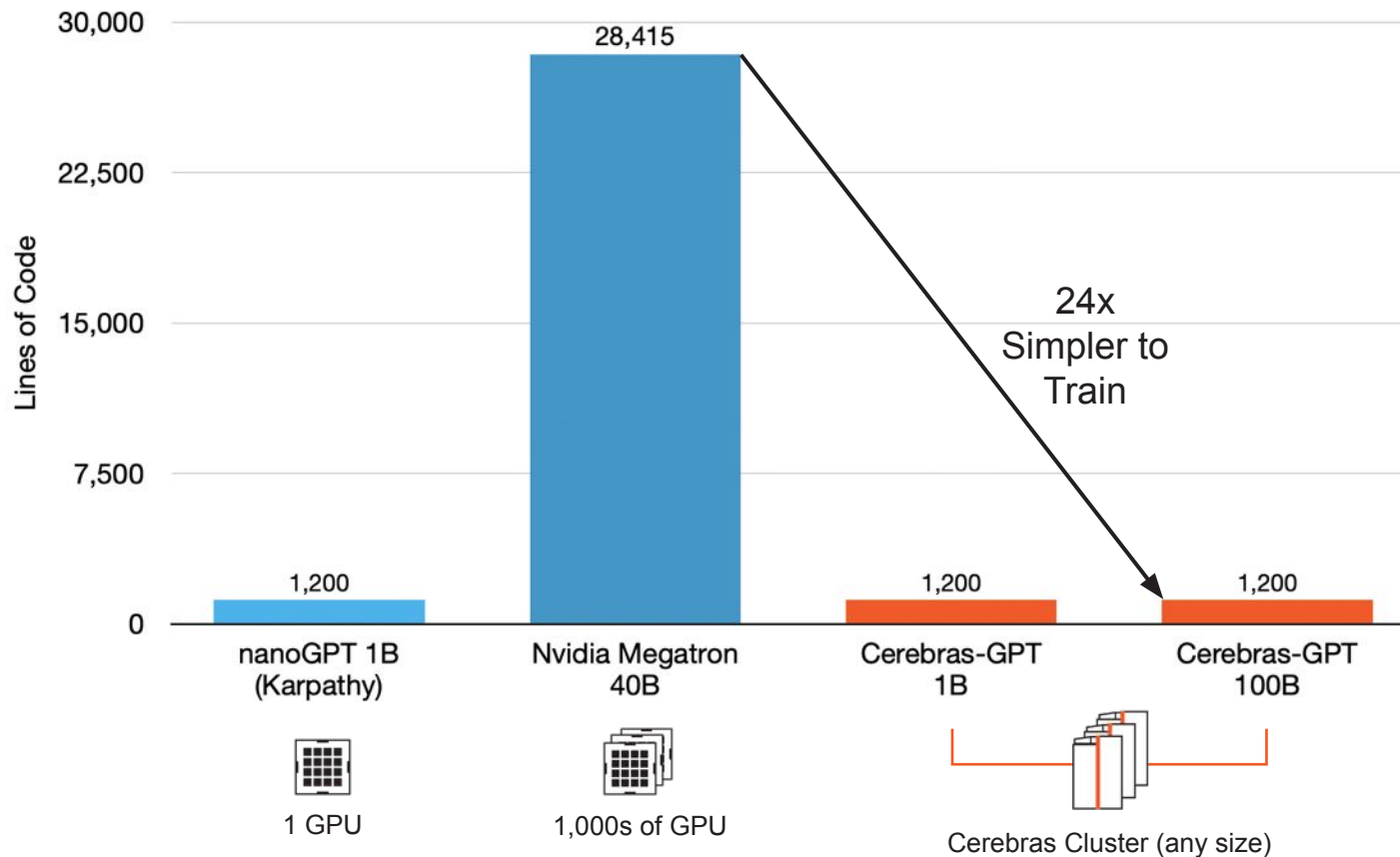
# Cerebras Eliminates The Complexity In Training Large Models



A 1B parameter is simple to write and train on one Cerebras CS-2.



# Cerebras Eliminates The Complexity In Training Large Models, Requires 27,215 Fewer Lines of code



A 1B parameter is simple to write and train on one GPU. It requires 1200 lines of code.

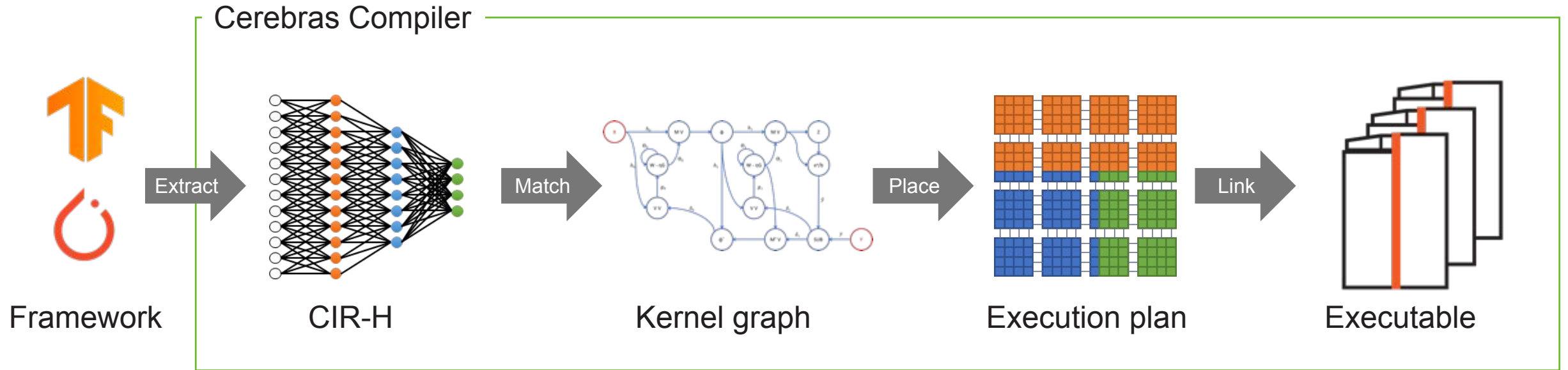
A 40B parameter model on 1,000 GPUs requires 28,415 lines of code to run.

A 1B parameter is simple to write and train on one CS-2. It requires 1200 lines of code.

A 100B parameter model across 64 Cerebras CS-2s is also easy to train. It requires the same 1200 lines of code as a 1B parameter model on 1 CS-2

**Avoid writing 27,215 lines of code with Cerebras Clusters.**

# The Cerebras software platform



Program a cluster-scale resource with the ease of a single node

# Programming / training with the cluster is simple

## Define the model

- Write in TensorFlow or PyTorch
- Parameterize based on yaml file
- Write *logical* model for *single* device

## Train the model

- Point to the model parameters
- Specify the number of CS-2s
- Specify the number of steps
- Run!

params\_gpt3xl.yaml

```
### GPT-3 XL 1.3B

hidden_size: 2048
num_hidden_layers: 24
num_heads: 16
```

training:

```
python run.py \  
--params params_gpt3xl.yaml \  
--num_csx 1 \  
--num_steps 100 \  
--model_dir model_dir \  
--mode train
```



# Scaling compute to more CS-2s is simple

## Scaling compute

- Change the number of CS-2s
  - Let's run GPT-3 XL 1.3B on 4x CS-2s
- Fully data-parallel training
- Run!

```
python run.py
--params params.yaml ← Where's your dataset?
--num_csx = 1 ← How many nodes?
--model_dir = model_dir ← Where to store weights?
--num_steps = 1000 ← How many training steps?
--mode=train ← Train, evaluate or infer?
```



# Scaling to larger models is simple

## Scaling the model

- Change the model parameters in yaml
  - Let's run GPT-NeoX 20B on 4x CS-2s
- Fully data-parallel training
- Run!

params\_gptneox.yaml

```
### GPT-NeoX 20B

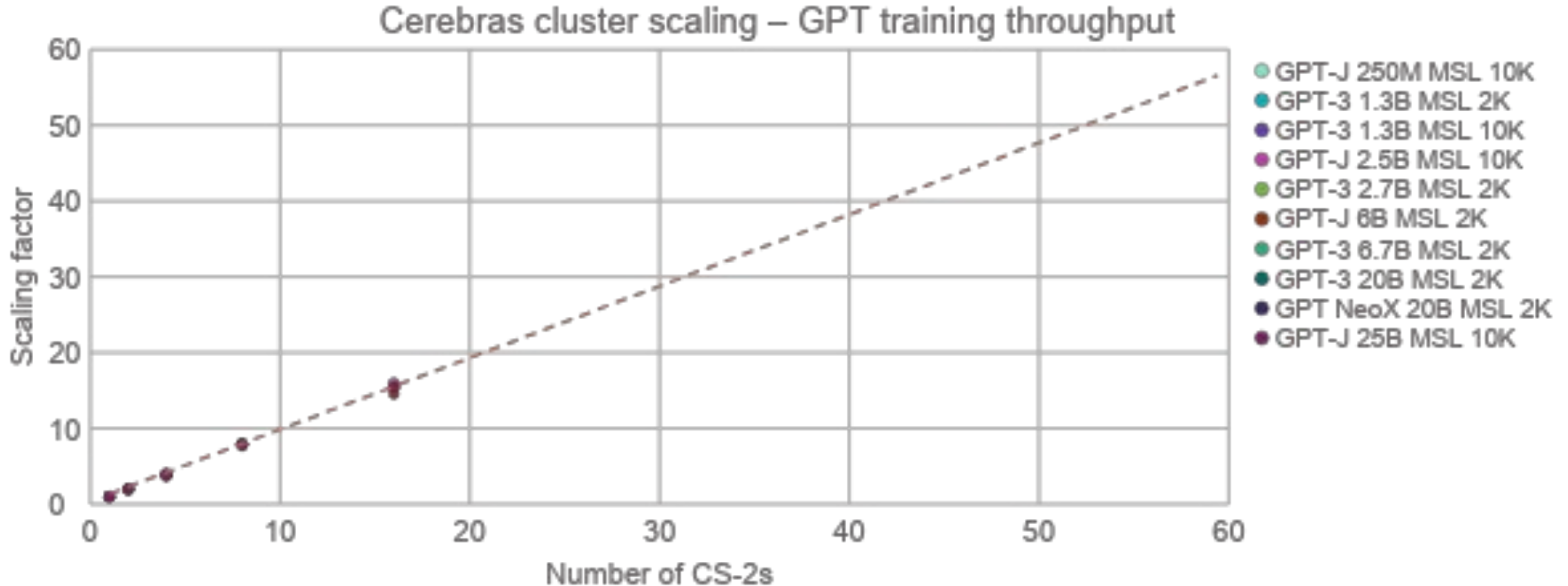
hidden_size: 6144
num_hidden_layers: 44
num_heads: 64
```

training:

```
python run.py \  
--params params_gptneox.yaml \  
--num_csx 4 \  
--num_steps 100 \  
--model_dir model_dir \  
--mode train
```

# And the results are extraordinary...

- Even the largest state-of-the-art models can train on a single CS-2
- Near-linear time to solution scaling across multiple CS-2s in a wafer-scale cluster




**Figure.** Measured training throughput scaling for 250M-20B GPT models over 1-16 CS-2 systems; projected scaling to 64 systems.


# Cerebras Open-sources Seven GPT-3 Models


A Family of Open, Compute-efficient, Large Language Models


- Open, compute-optimal GPT models up to 13B trained on Cerebras Wafer-Scale Cluster
- Trained on public Eleuther Pile dataset
- Fully open, Apache 2.0
- Compute optimal scaling law model family 111M, 256M, 590M, 1.3B, 2.7B, 6.7B, 13B
- Trained in just weeks on the CS-2 platform!
- Checkpoints: <https://huggingface.co/cerebras>
- Source code: <https://github.com/Cerebras/modelzoo>
- Paper: <https://arxiv.org/abs/2304.03208>

 Models 7  Hugging Face

 cerebras/Cerebras-GPT-13B  
📄 • Updated Apr 7 • ↓ 20.1k • ❤️ 601

 cerebras/Cerebras-GPT-6.7B  
📄 • Updated Apr 7 • ↓ 6.6k • ❤️ 59

 cerebras/Cerebras-GPT-2.7B  
📄 • Updated Apr 7 • ↓ 10.4k • ❤️ 33

 cerebras/Cerebras-GPT-1.3B  
📄 • Updated Apr 7 • ↓ 10.3k • ❤️ 39

 cerebras/Cerebras-GPT-590M  
📄 • Updated Apr 7 • ↓ 3.44k • ❤️ 16

 cerebras/Cerebras-GPT-256M  
📄 • Updated Apr 7 • ↓ 4k • ❤️ 19

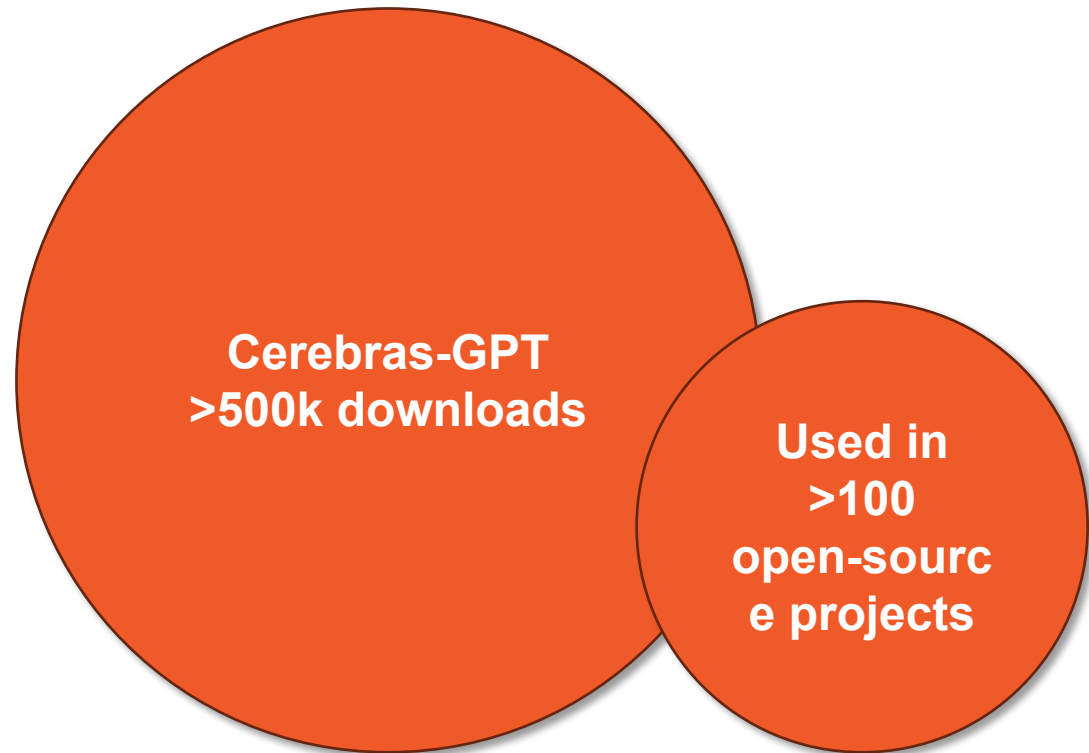
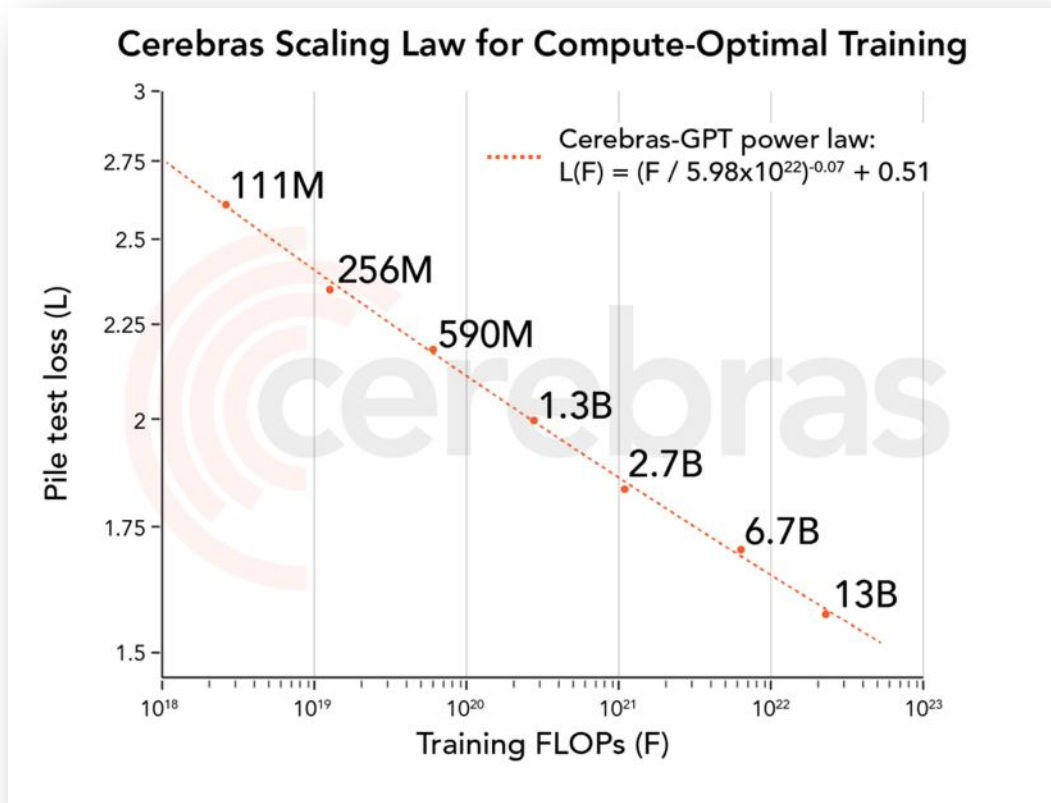
 cerebras/Cerebras-GPT-111M  
📄 • Updated Apr 7 • ↓ 22.5k • ❤️ 52



In March, Cerebras introduced **Cerebras-GPT**: the first open-source family of Chinchilla-optimal large language models. 111M-13B parameters trained on The Pile data, exclusively on Cerebras CS-2 machines.

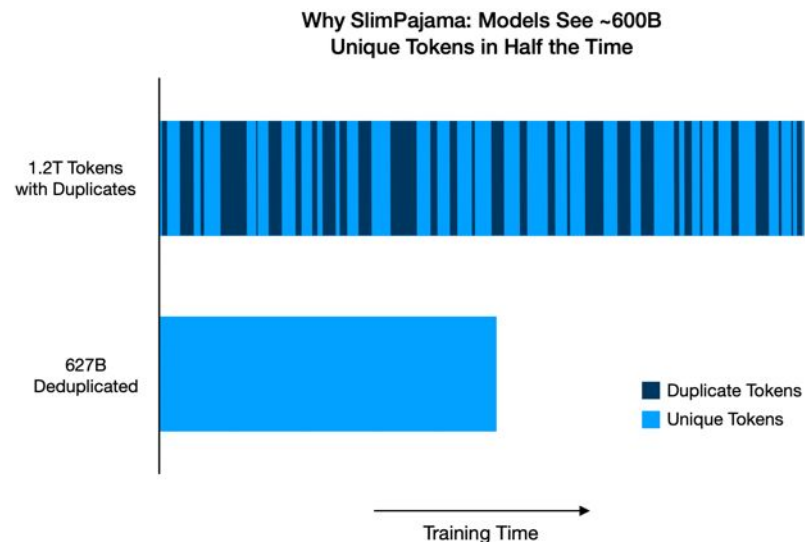
These models were trained on Cerebras in weeks with a handful of people, not months with hundreds. They are available under a permissive Apache open-source license.

Since launch, **Cerebras-GPT models have been downloaded more than 500,000 times and used in more than 50 other open-source projects.**



# SlimPajama: a better LLM dataset from CS experts

- Largest fully deduplicated, multi-corpora, open dataset
- Twice the training speed & compute efficiency of RedPajama 1.2T
- Data pre-processing scripts released open source on [Cerebras Github](#)




June 9, 2023

f t p in x ↗

In Machine Learning, Software, Cloud, Blog, Developer Blog, Large Language Model, NLP, Deep Learning

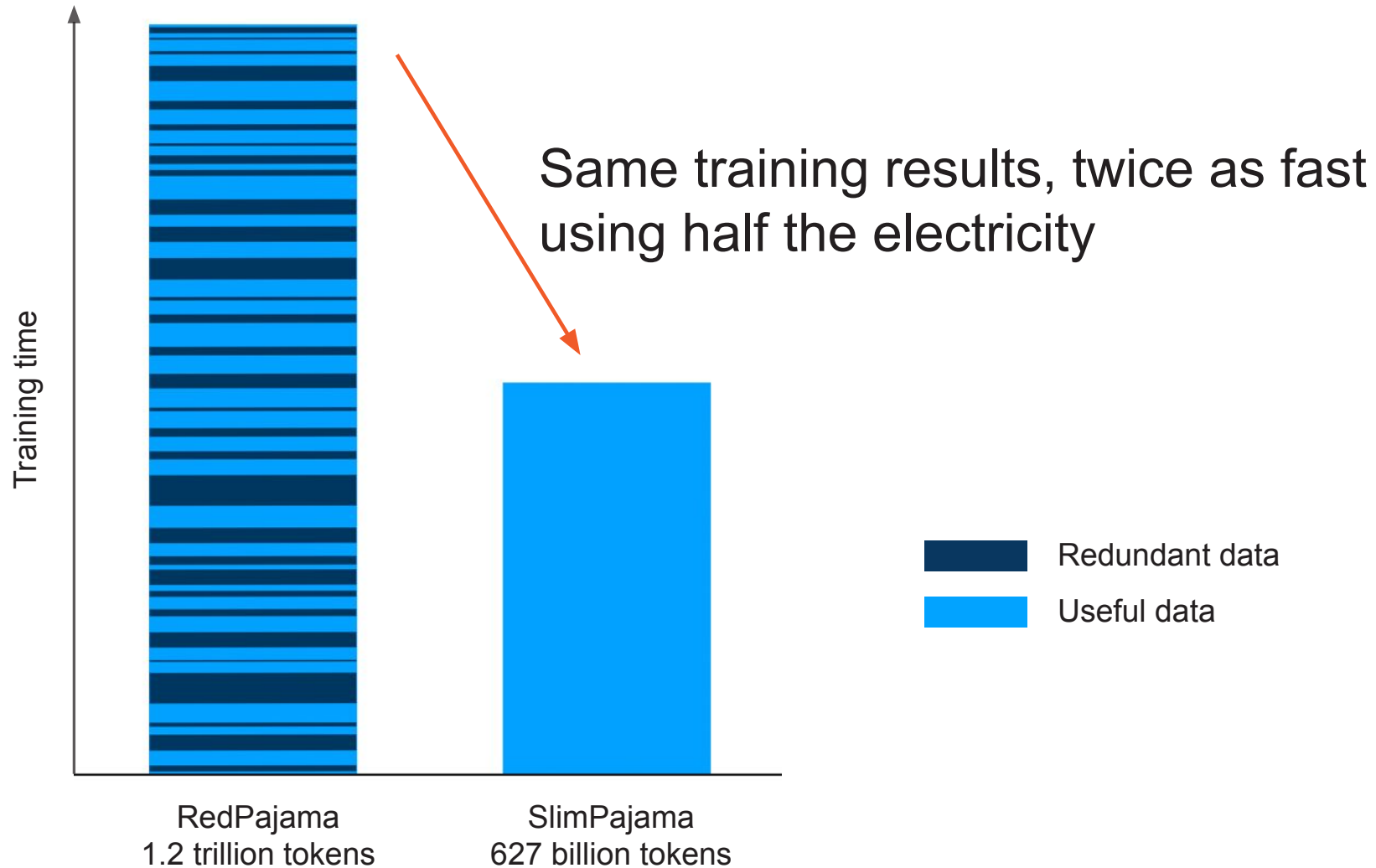
## SlimPajama: A 627B token cleaned and deduplicated version of RedPajama

Today we are releasing SlimPajama - the largest deduplicated, multi-corpora, open-source, dataset for training large language models.

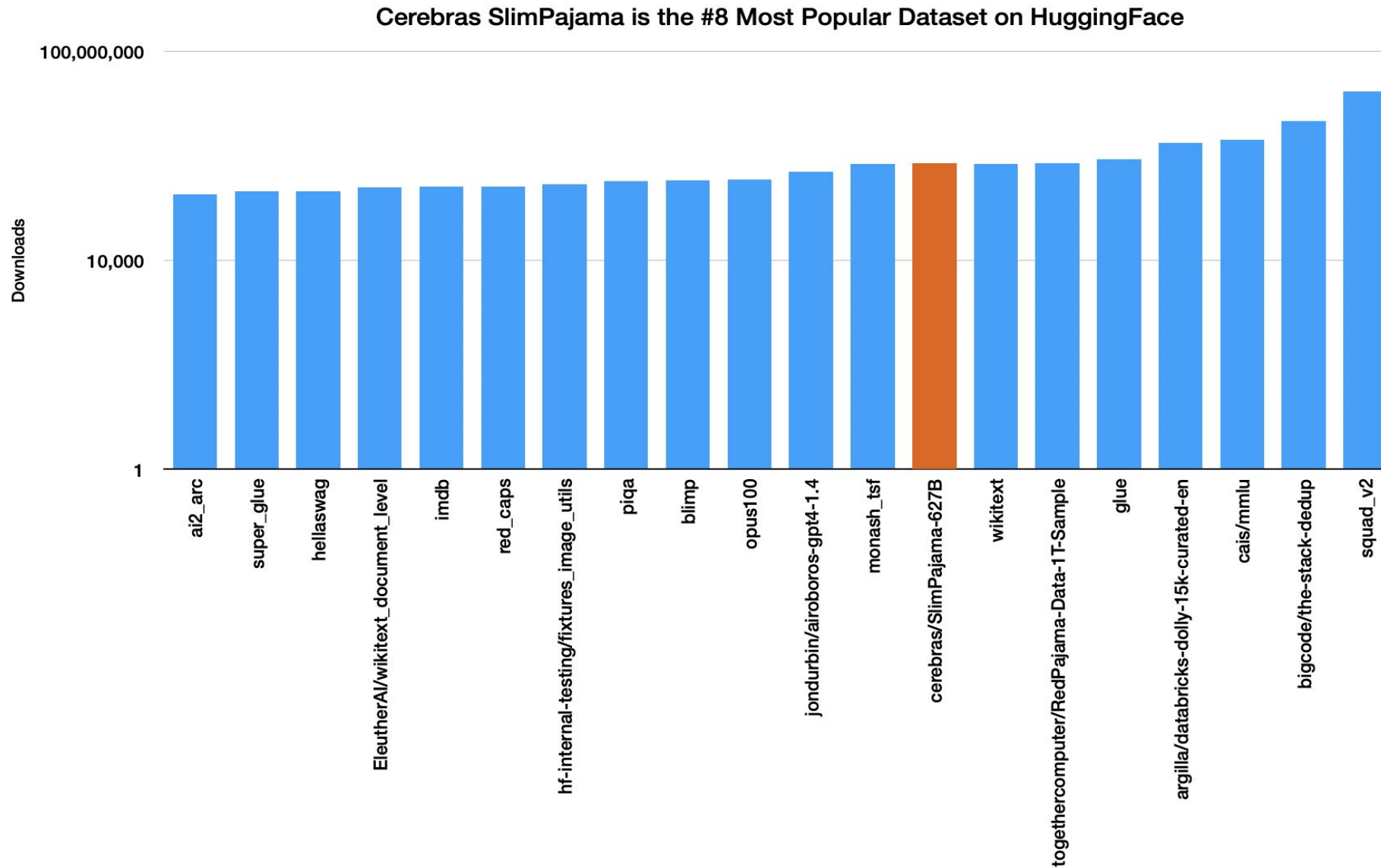


	Tokens	Open source	Curated data sources	Deduplication level
SlimPajama	627B	Yes	Yes	Extensive
RedPajama	1.21T	Yes	Yes	Partial
RefinedWeb-600B	600B	Yes	No	Extensive
RefinedWeb-5T	5T	No	No	Extensive
LLaMA	1.4T	No	Yes	Partial
MPT	1T	No	Yes	Partial
MassiveText	1.4T	No	Yes	Extensive

# Clean Data is Better Data



# World-Class Datasets from Cerebras

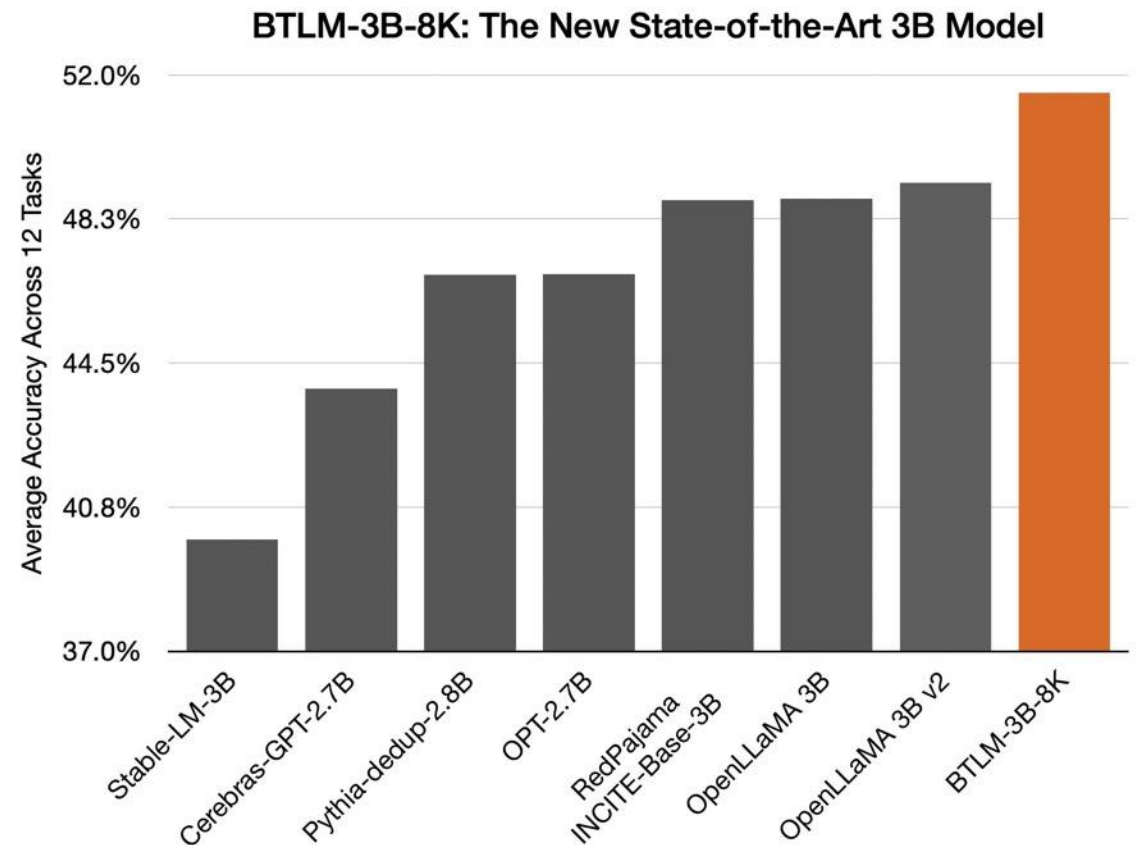




# Cerebras and OpenTensor Announce BTLM-3B-8K

The state-of-the-art 3 billion parameter open-source language model

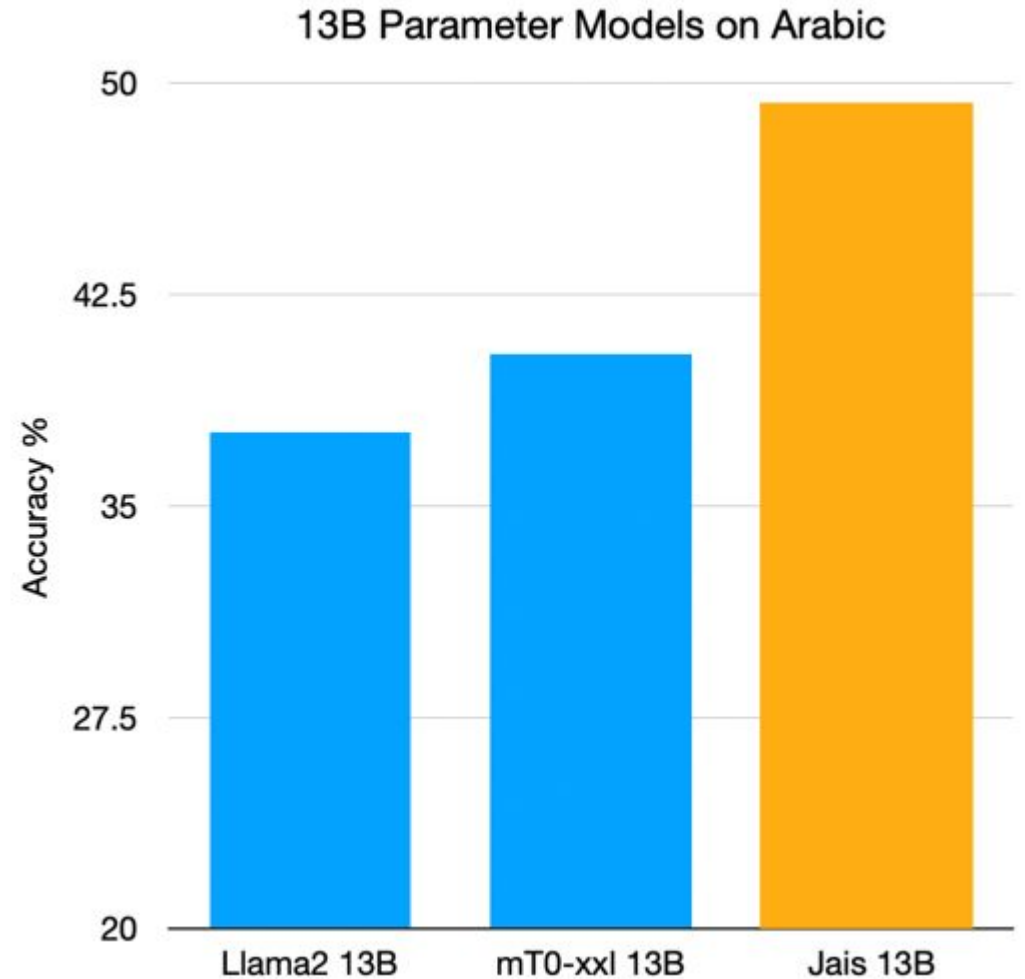
- **The most popular 3B parameter model on Hugging Face with >1 million downloads**
- Trained using the CG-1 AI supercomputer
- Performance of a 7B parameter model in a 3B parameter model
- Optimized for long sequence length inference 8K or more
- First model trained on the SlimPajama, the largest fully deduplicated open dataset
- Runs on devices with as little as 3GB of memory when quantized to 4-bit (e.g., mobile devices, Raspberry Pi)
- Apache 2.0 license for commercial use



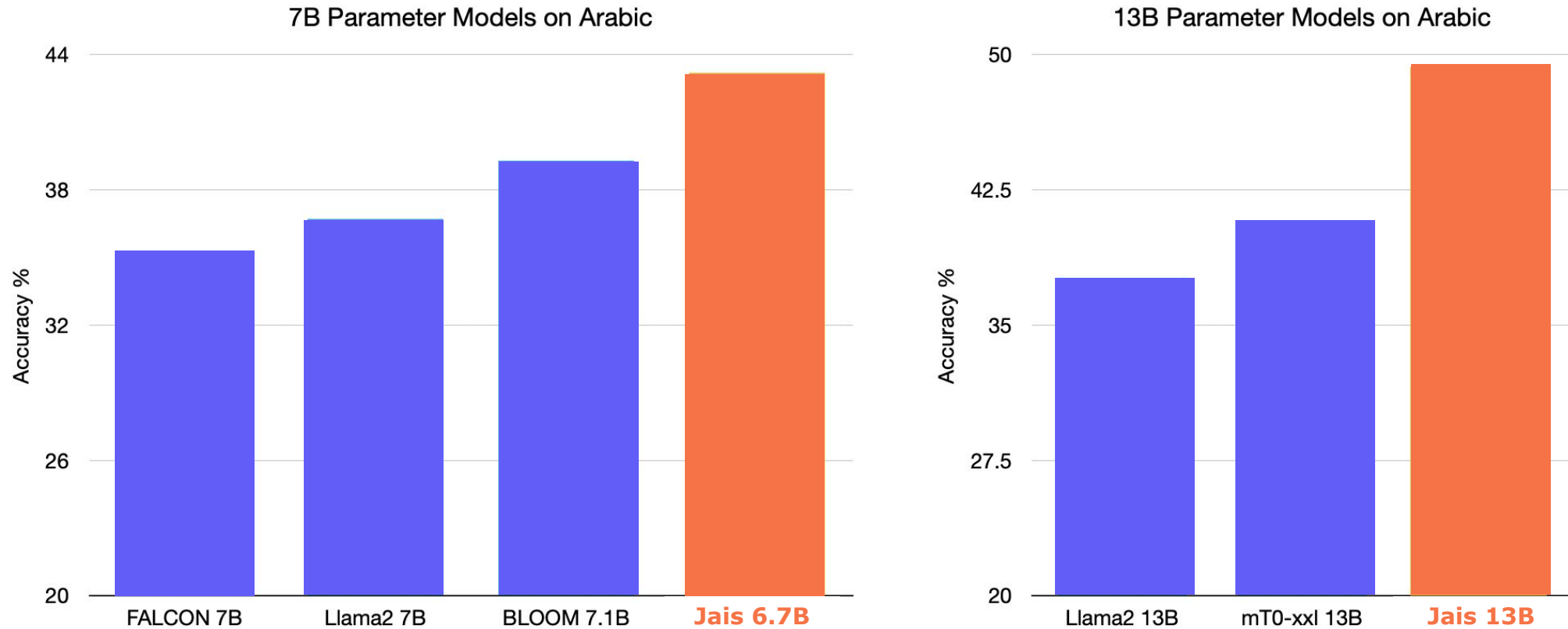
# Cerebras Introduces Jais, a Bilingual Model

The World's Best Arabic-English Language Model, in partnership with G42

- 13B Bilingual Model Trained on 400B Tokens
- Outperforms all Arabic models, sets new benchmark
- Achieves Llama Level Performance in English while using ~10% the training data
- Developed by G42's Inception and MBZU AI
- Trained on Condor Galaxy 1
- Fully open, Apache 2.0
  
- Checkpoints: <https://huggingface.co/inception-mbzuai>
- Paper: <https://arxiv.org/abs/2308.16149>



# Jais: the new standard for open Arabic LLM

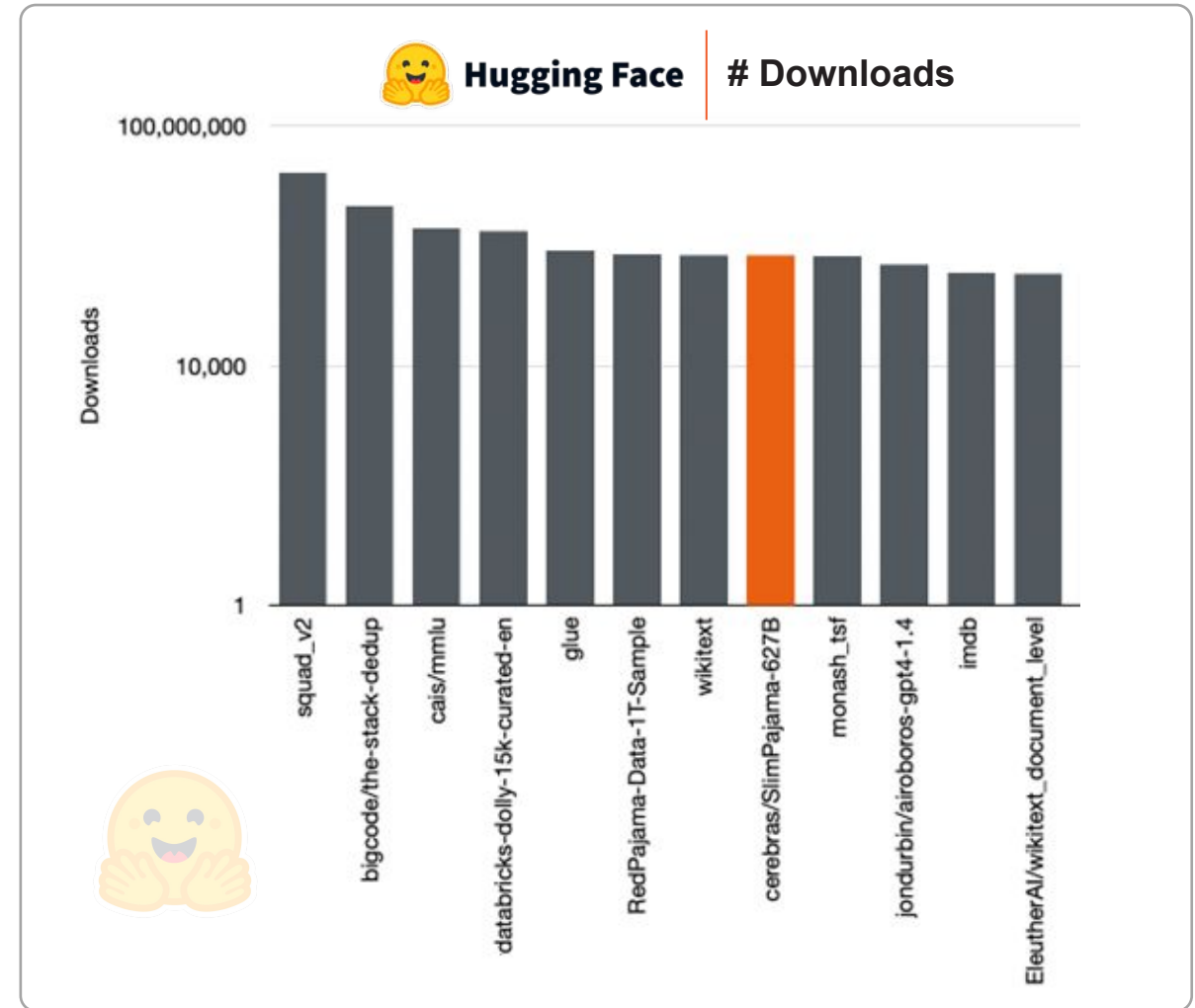


**Figure.** State of the art Arabic with respect to recent standards like Bloom, Llama, Falcon

# Cerebras Announces SlimPajama

The #8 Most Popular Dataset on HuggingFace

- Largest fully deduplicated, multi-corpora, open-source dataset
- 627 Billion tokens
- The highest quality English language dataset in existence
- Twice the training speed and uses half the power of RedPajama 1.2T
- Data pre-processing scripts released open source on [Cerebras Github](#)
- Done in partnership with OpenTensor





# Case Study – Global Technology Conglomerate Achieves State-of-The-Art on Multilingual Model



**Objective:** Train large language model from scratch using state-of-the-art NLP technologies, multi-lingual vocabularies and datasets to transform machine translation and search.

Establish new scaling laws to inform future optimal choices for dataset and model sizes.

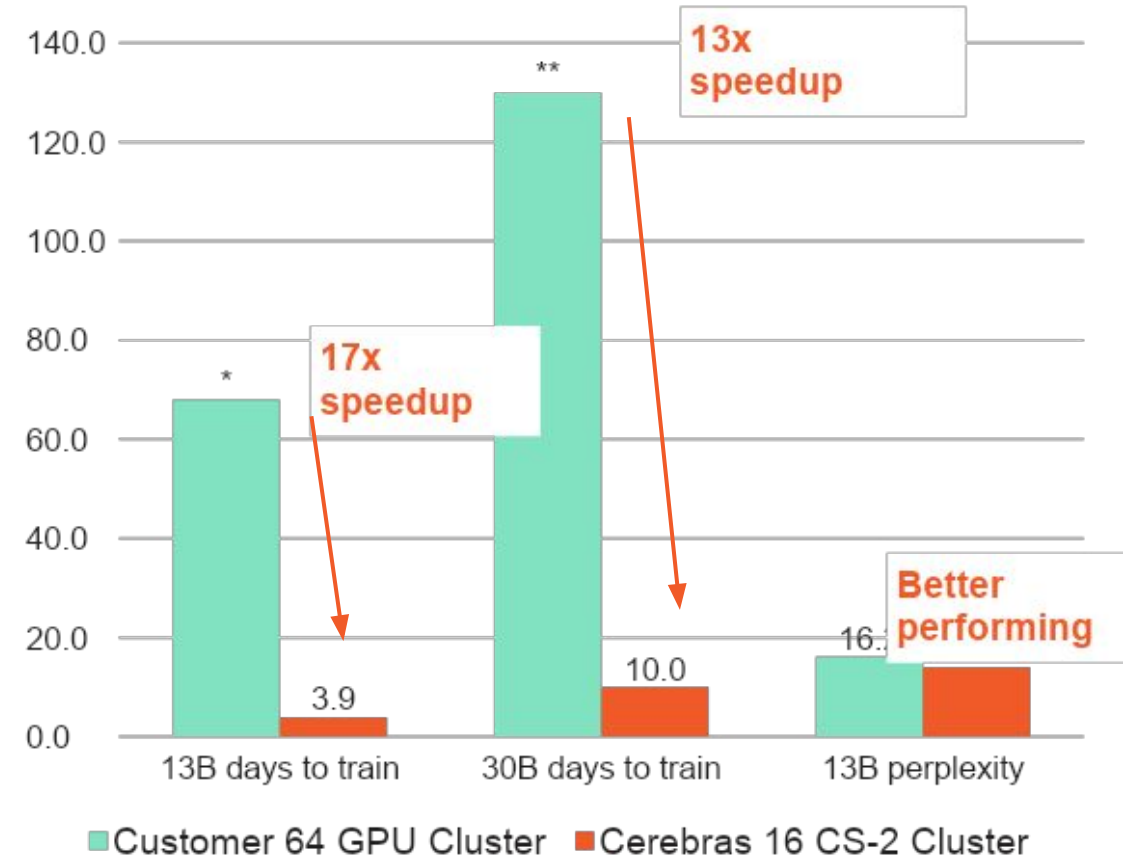


**Challenge:** Training from scratch was intractable using conventional hardware, making experimentation impractical due to prohibitively-long training time.



**Outcome:** 13B parameter GPT model trained from scratch in 3.9 days on 16-node Cerebras Cluster, compared to 68 days to train smaller, 13B model on 64-GPU baseline.

Measured near-linear speedup from one to 16xCS-2 with no code changes. Total engagement time <3 weeks



\* To 95% completion  
\*\* Customer estimate

# Case Study - Top 5 US Bank Accelerates NLP Model Training



**Objective:** Improve large language model accuracy for financial services applications by training the model from scratch using domain-specific data.



**Problem:** Rapid experimentation with model parameters is difficult on traditional infrastructure



**Outcome:** The performance and ease of use of the Cerebras CS-2 system enabled fast iteration to a better-performing model

**Time-to-Train**

**15x**

The CS-2 system reduced training time from 10.5 days to just 17 hours compared to a leading 8-GPU server

**Predictive Power**

**2.6x**

Perplexity of more than 2x lower compared to the baseline

**Energy Consumption**

**45%**

Energy consumption nearly halved compared to baseline

[Full Case Study Here](#)

# Case Study – Award-Winning COVID Research Accomplished Only on Cerebras 16-node Cluster



ACM Gordon Bell Special Prize  
for High Performance Computing-Based  
COVID-19 Research



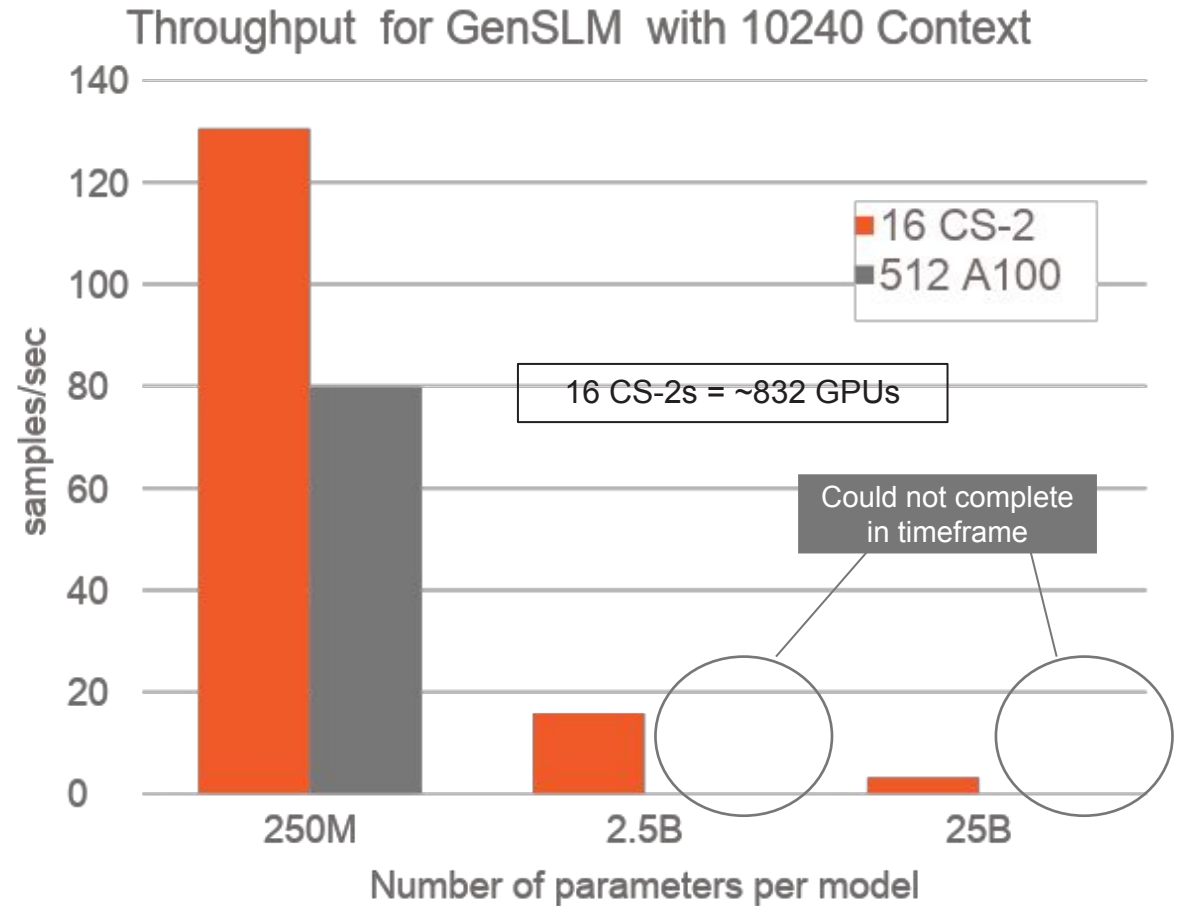
**Objective:** Accelerate COVID-19 understanding by training 250M-25B GPT models on full-length SARS-CoV-2 genome data, using 10,240 token sequence length



**Challenge:** Team of ML PhDs found training from scratch to be intractable using conventional hardware, making experimentation impractical due to prohibitively-long training time.

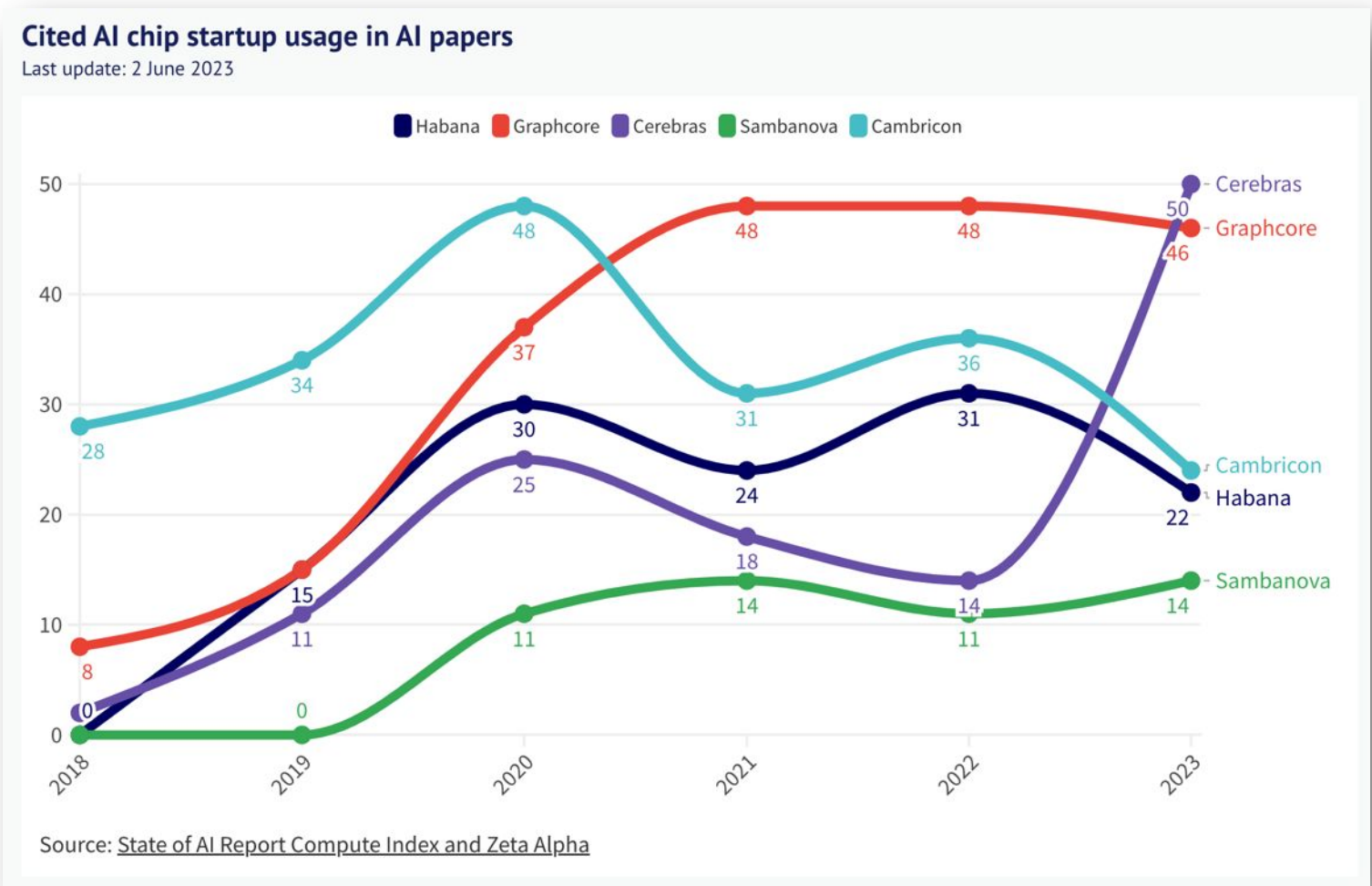


**Outcome:** GPUs suffered out-of-memory errors on sequences of 10,240 tokens. CS-2 supported 10k context length out-of-the-box. No code changes or specialized expertise required.



Developer adoption. Emerging AI computing platforms are powering new AI research, model and application development. Cerebras Systems, with its focus on performance and ease of use at scale, is emerging as the leading non-GPU AI platform for the latest research.

This chart shows recent **trends in AI hardware system use** as measured by AI paper citation.





# Cerebras: AI Solutions at the Intersection of Compute, Models, Data

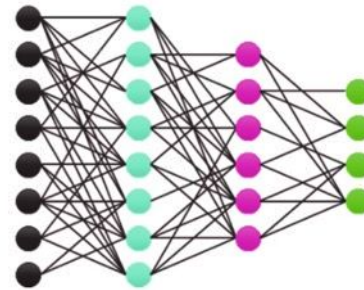
## Compute



**Wafer Scale Engine**  
Fastest AI Processor in  
the World

**Condor Galaxy 1**  
4 Exaflop AI  
Supercomputer

## Models



**Cerebras-GPT**  
First family of open source  
GPT models

**BTLM-3B-8K**  
#1 3B parameter model  
1M downloads on HuggingFace

## Data



**SlimPajama**

The **highest quality** open  
source, multi-copra English  
dataset.

**Ranked #8** among 52,241  
datasets