

**BigScience**

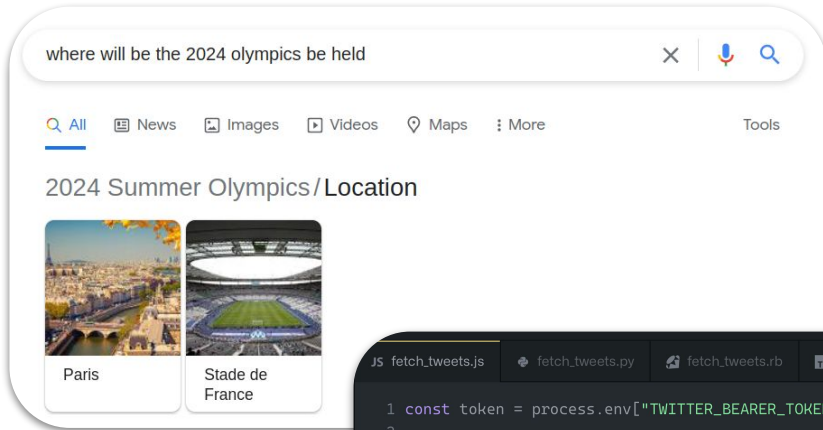


# Collaboratively training a large multilingual language model

Lucile Saulnier, Thomas Wang

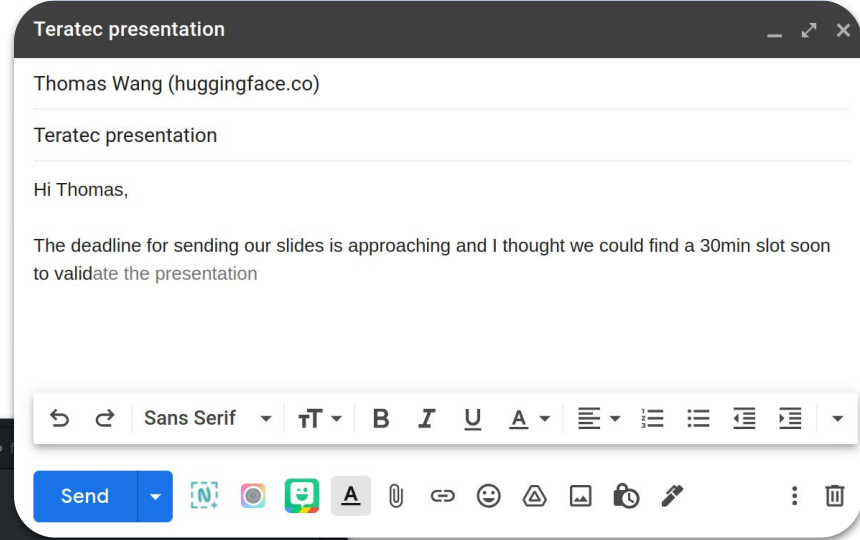
What motivated us to do BigScience?

# Context: why language models are useful?

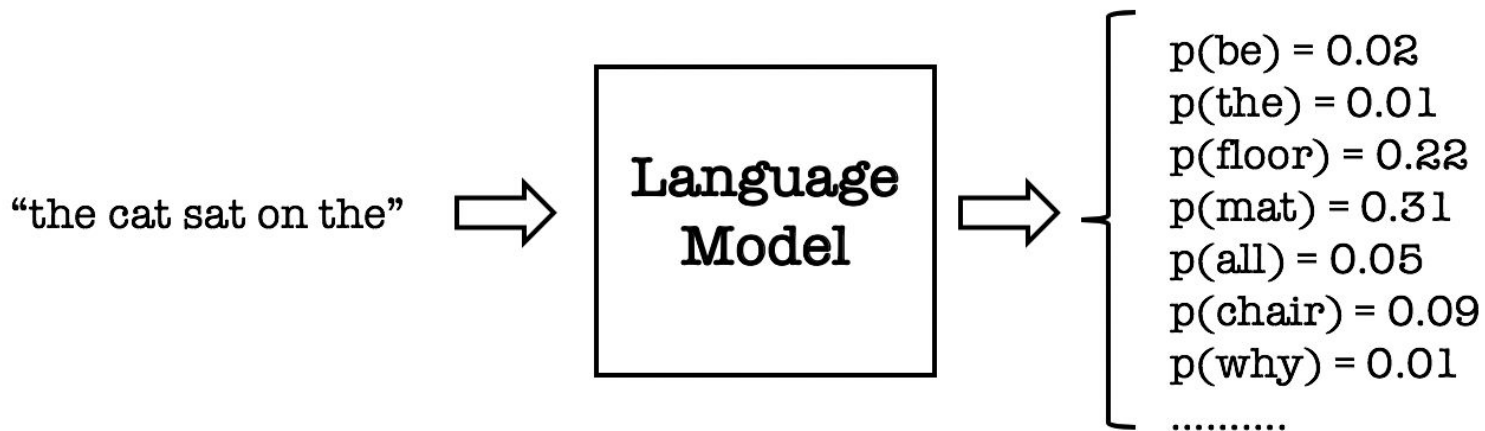


```
1 const token = process.env["TWITTER_BEARER_TOKEN"]
2
3 const fetchTweetsFromUser = async (screenName, count) => {
4   const response = await fetch(
5     `https://api.twitter.com/1.1/statuses/user_timeline.json?screen_name=${screenName}&count=${count}`
6   )
7   const headers = {
8     Authorization: `Bearer ${token}`,
9   }
10  }
11 }
12 const json = await response.json()
13 return json
14 }
```

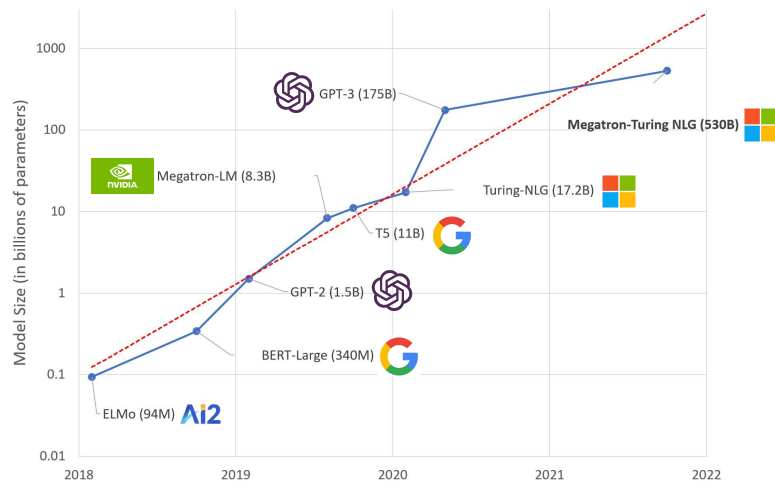
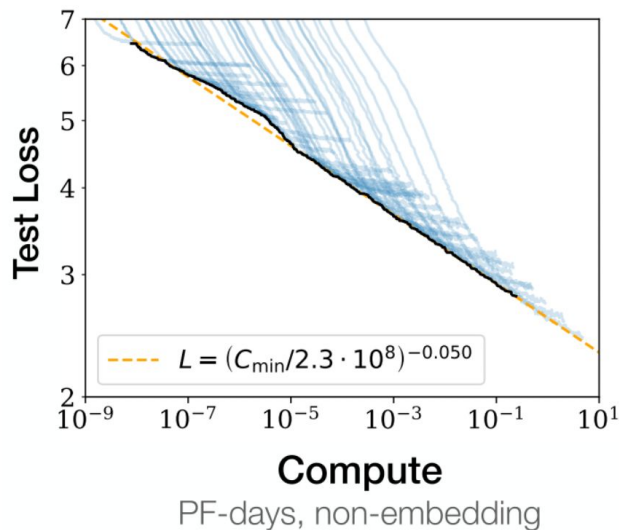
The code editor shows a Python function named `fetchTweetsFromUser` that uses `fetch` to retrieve tweets from a specific user. The code is written in a dark-themed editor with syntax highlighting. A Copilot logo is visible in the bottom left corner of the editor.



# Context: what is a language model?



# Large Language model: why train one?



## GPT-3's generation example:

[...]

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

# Large Language model: why research on LLMs is hard today?

## Closed access for most of them

### Training cost

- typically \$2-5M
- million of gpu hours

**VB** VentureBeat

Naver trained a 'GPT-3-like' Korean language model

Naver claims the system learned 6,500 times more Korean data than OpenAI's ... Some experts believe that while HyperCLOVA, GPT-3, PanGu-a, ...

1 Jun 2021



**TC** TechCrunch

Anthropic is the new AI research outfit from OpenAI's Dario Amodei, and it has \$124M to burn

Anthropic, as it's called, was founded with his sister Daniela and its goal is to create "large-scale AI systems that are steerable, ...

28 May 2021



**VB** VentureBeat

AI21 Labs trains a massive language model to rival OpenAI's GPT-3

"AI21 Labs was founded to fundamentally change and improve the way people read and write. Pushing the frontier of language-based AI requires ...

1 month ago



**FC** Fast Company

Ex-Googleers raise \$40 million to democratize language AI

This story has been updated with more information about Cohere's approach to responsible AI. About the author. Fast Company Senior Writer Mark ...

2 days ago



# Large Language model: why opacity in LLMs is an issue today?

## Research

- Difficult to do real research: no access to data, training artifacts, checkpoints
- Academic researchers: not involved
- Lack of fields diversity: English/Chinese only, ML-only teams

## Environmental

- Training similar models: Duplication of energy requirements

## Ethical and societal around datasets/design

- Shortcomings in the text corpora used to train these models: Representativeness, stereotypes, PII
- Ethical/bias/usage questions: Only asked a-posteriori

# BigScience: what is it?

“During **one-year**, from May 2021 to May 2022, 1000+ researchers from 60 countries and more than 250 institutions are **creating together a very large multilingual neural network language model** and a **very large multilingual text dataset** on the 28 petaflops Jean Zay (IDRIS) supercomputer located near Paris, France.

During the workshop, the participants plan to investigate the dataset and the model from all angles: bias, social impact, capabilities, limitations, ethics, potential improvements, specific domain performances, carbon impact, general AI/cognitive research landscape.”



# BigScience: what is it?

## Endeavour to generate momentum on research over LLMs

- Open-sourcing
- Collaborative / transparent
- Create working groups over scientific questions

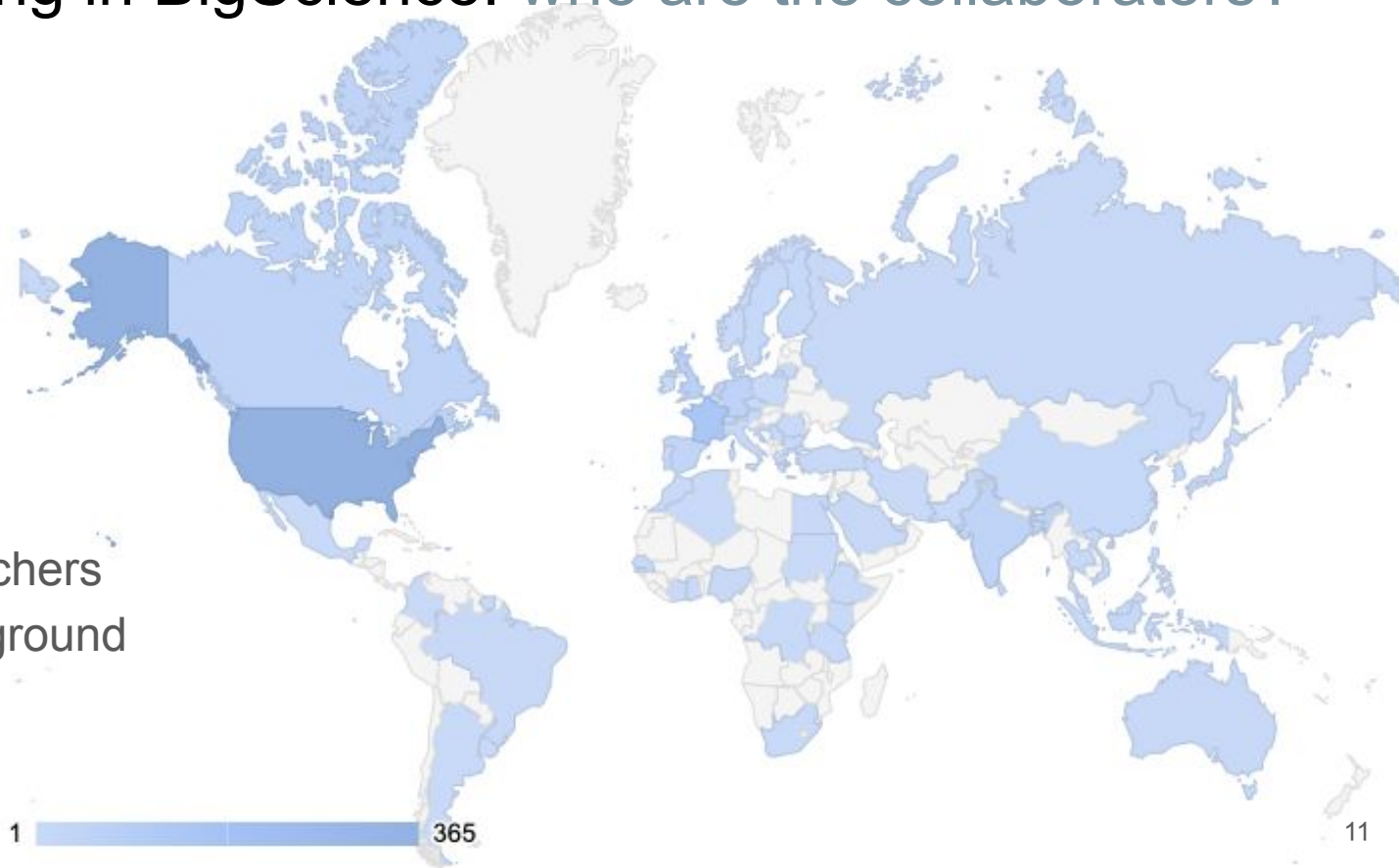
## Training big models is hard from an engineering perspective

- Train a 176B multilingual model
- Openly discuss engineering problems and solutions throughout the project

**BigScience is a collaborative effort**

# People working in BigScience: who are the collaborators?

- 1000+ researchers
- Diverse background




# BigScience's collaborators: what are they working on?

## WORKING GROUPS

PROJECT	DATA	TOKENIZATION	MODELING	EVALUATION	DOMAINS
Organization	Sourcing	<b>INTERPRETABILITY</b>	Architecture and Scaling	Intrinsic	Biomedical
Ethical and Legal Scholarship	Governance	<b>ENGINEERING</b>	Multilinguality	Extrinsic	Historical Texts
Accessibility	Tooling	<b>CARBON FOOTPRINT</b>	Prompt Engineering	Multilinguality	Math
Collaborations and Education	Privacy		Retrieval	Bias, Fairness, Social Impact	
	Analysis and Visualization		Metadata	Few-shot	



# Artifacts: what came out of BigScience? (1/4)




## BigScience Workshop

Research workshop on large language models - The Summer of Language Models 21

<https://bigscience.huggingface.co> [@BigScienceW](#) [bigscience-contact@googlegroups.com](mailto:bigscience-contact@googlegroups.com)

[Overview](#) [Repositories 29](#) [Projects 4](#) [Packages](#) [People 22](#)



### Popular repositories

#### [promptsource](#) Public

Toolkit for creating, sharing and using natural language prompts.

Python 660 144

#### [bigscience](#) Public

Central place for the engineering/scaling WG: documentation, SLURM scripts and logs, compute environment and data.

Shell 277 26

#### [Megatron-DeepSpeed](#) Public

Ongoing research training transformer language models at scale, including: BERT & GPT-2

Python 145 34

#### [t-zero](#) Public

Reproduce results and replicate training for T0 (Multitask Prompted Training Enables Zero-Shot Task Generalization)

Python 143 26

#### [biomedical](#) Public

Tools for curating biomedical training data for large-scale language modeling


Python 125 69

#### [evaluation](#) Public

Code and Data for Evaluation WG

Python 36 23

### People



[View all](#)

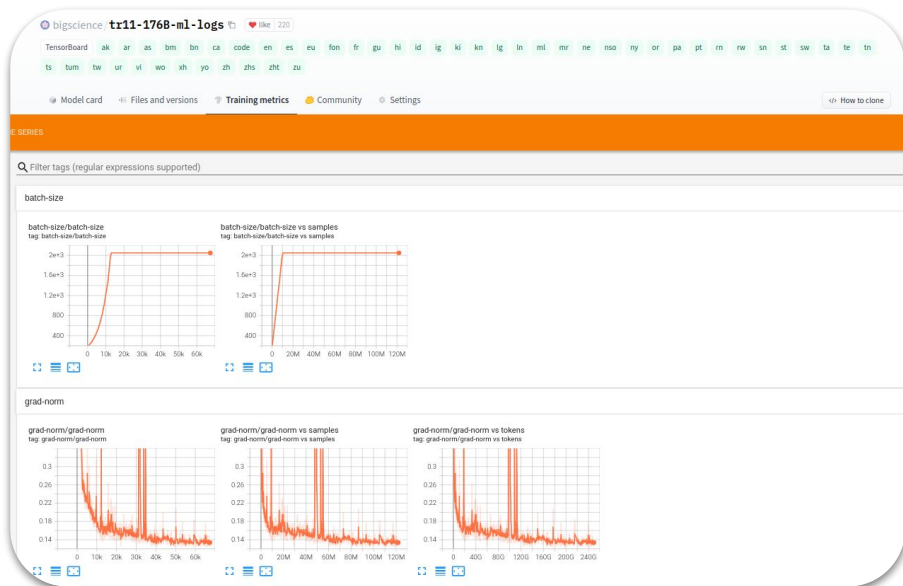
### Top languages

Python Jupyter Notebook HTML  
Shell Makefile

### Most used topics

[machine-learning](#) [nlp](#)

# Artifacts: what came out of BigScience? (1/4)



```
from transformers import AutoModel, AutoTokenizer

model_name = "bigscience/bloom"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModel.from_pretrained(model_name)
```

<https://huggingface.co/bigscience/bloom>

<https://huggingface.co/bigscience/tr11-176B-ml-logs>

## Artifacts: what came out of BigScience? (2/4)



### **BigScience**

#### **BigScience RAIL License v1.0**

This is the home of the BigScience RAIL License v1.0. If you would like to download the license you can get it as [.txt](#), [.docx](#), or [.html](#) file.

<https://huggingface.co/spaces/bigscience/license>

# Artifacts: what came out of BigScience? (3/4)

## Masader: Metadata Sourcing for Arabic Text and Speech Data Resources

Zaid Alyafei<sup>1</sup>, Maraim Masoud<sup>2</sup>, Mustafa Ghaleb<sup>1</sup>, and Maged S. Al-shaibani<sup>1</sup>

<sup>1</sup> King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia  
<sup>2</sup> Independent Researcher

### Abstract

The NLP pipeline has evolved dramatically in the last few years. The first step in the pipeline is to find suitable annotated datasets to evaluate the tasks we are trying to solve. Unfortunately, most of the published datasets lack metadata annotations that describe their attributes. Not to mention, the absence of a public catalogue that indexes all the publicly available datasets related to specific regions or languages. When we consider low-resource dialectal languages, for example, this issue becomes more prominent. In this paper we create *Masader*, the largest public catalogue for Arabic NLP datasets, which consists of 200 datasets annotated with 25 attributes. Further-

and so on. This study attempts to identify the publicly available Arabic NLP datasets and to provide a catalogue of Arabic datasets to researchers. The catalogue will increase the discoverability and provide some key metadata that will help researchers identify the most suitable dataset for their research questions.

We

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

**Abstract**

What are the units of text that we want to model? From bytes to multi-word expressions, text can be analyzed and generated at many granularities. Until recently, most natural language processing (NLP) models operated over words, treating those as discrete and atomic tokens, but starting with byte-pair encoding (BPE), subword-based approaches have become dominant in many areas, enabling small vocabularies while still allowing for fast inference. Is the end of the road character-level model or byte-level processing? In this survey, we connect several lines of work from the pre-neural and neural era, by showing how hybrid approaches

## MULTITASK PROMPTED TRAINING ENABLES ZERO-SHOT TASK GENERALIZATION

Victor Sanh<sup>\*</sup> Hugging Face    Albert Webson<sup>\*</sup> Brown University    Colin Raffel<sup>\*</sup> Hugging Face    Stephen H. Bach<sup>\*</sup> Brown & Snorkel AI

Lintang Sutawika<sup>\*</sup> BigScience    Zaid Alyafei<sup>\*</sup> KFUPM    Antoine Chaffin<sup>\*</sup> IRISA & IMATAG    Arnaud Stiegler<sup>\*</sup> Hyperscience    Teven Le Scao<sup>\*</sup> Hugging Face

Arun Raja<sup>\*</sup> I<sup>2</sup>R, Singapore    Manan Dey<sup>\*</sup> SAP    M Saiful Bari<sup>\*</sup> NTU, Singapore    Chen Xu<sup>\*</sup> UCSD & Hugging Face    Urmish Thakker<sup>\*</sup> SambaNova Systems

Shanya Sharma<sup>\*</sup> Walmart Labs    Eliza Szczecchia<sup>\*</sup> BigScience    Taewoon Kim<sup>\*</sup> VU Amsterdam    Gunjan Chhabhani<sup>\*</sup> BigScience    Nihal V. Nayak<sup>\*</sup> Brown University

Debajyoti Datta<sup>\*</sup> University of Virginia    Jonathan Chang<sup>\*</sup> ASUS    Mike Tian-Jian Jiang<sup>\*</sup> ZEALS, Japan    Han Wang<sup>\*</sup> NYU    Matteo Manica<sup>\*</sup> IBM Research

Sheng Shen<sup>\*</sup> UC Berkeley    Zheng-Xin Yong<sup>\*</sup> Brown University    Harshit Pandey<sup>\*</sup> BigScience    Michael McKenna<sup>\*</sup> Parity    Rachel Bawden<sup>\*</sup> Inria, France

Thomas Wang<sup>\*</sup> Inria, France    Trishala Neeraj<sup>\*</sup> BigScience    Jos Rozen<sup>\*</sup> Naver Labs Europe    Abheesh Sharma<sup>\*</sup> BITS Pilani, India    Andrea Santilli<sup>\*</sup> University of Rome

### Between words and characters:

### A Brief History of Open-Vocabulary Modeling and Tokenization in NLP

Sabrina J. Mielke<sup>1,2</sup>    Zaid Alyafei<sup>3</sup>    Elizabeth Salesky<sup>1</sup>  
 Colin Raffel<sup>2</sup>    Manan Dey<sup>4</sup>    Matthias Gallé<sup>5</sup>    Arun Raja<sup>6</sup>  
 Chenglei Si<sup>7</sup>    Wilson Y. Lee<sup>8</sup>    Benoît Sagot<sup>9\*</sup>    Samson Tan<sup>10\*</sup>

BigScience Workshop Tokenization Working Group

<sup>1</sup>Johns Hopkins University    <sup>2</sup>HuggingFace    <sup>3</sup>King Fahd University of Petroleum and Minerals    <sup>4</sup>SAP  
<sup>5</sup>Naver Labs Europe    <sup>6</sup>Institute for Infocomm Research, A\*STAR Singapore    <sup>7</sup>University of Maryland  
<sup>8</sup>BigScience Workshop    <sup>9</sup>Inria Paris    <sup>10</sup>Salesforce Research Asia & National University of Singapore  
 s.jm@s.jm.i.e.k.e. com

### Abstract

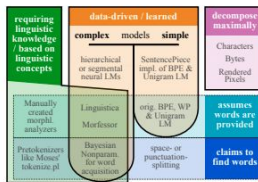


Figure 1: A taxonomy of segmentation and tokenization algorithms and research directions

### ABSTRACT

It has recently been shown to attain reasonable zero-shot performance set of tasks (Brown et al., 2020). It has been hypothesized that implicit multitask learning in language models (Sanh et al., 2019). Can zero-shot generalization instead be directly attributed to multitask learning? To test this question at scale, we develop a large-scale evaluation of modeling choices and their impact on zero-shot generalization. In particular, we focus on text-to-text models and experiment with three model architectures (causal/non-causal decoder-only and encoder-decoder), trained with two different pretraining objectives (autoregressive and masked language modeling), and evaluated with and without multitask prompted finetuning. We train

And many mores..

## What Language Model to Train if You Have One Million GPU Hours?

### The BigScience Architecture & Scaling Group

Teven Le Scao<sup>1\*</sup>    Thomas Wang<sup>1\*</sup>    Daniel Hesse<sup>2\*</sup>    Lucile Saulnier<sup>1\*</sup>    Stas Bekman<sup>1\*</sup>  
 M Saiful Bari<sup>3</sup>    Stella Biderman<sup>4,5</sup>    Hady Elsahar<sup>6</sup>    Jason Phang<sup>7</sup>    Ofir Press<sup>7</sup>    Colin Raffel<sup>1</sup>  
 Victor Sanh<sup>1</sup>    Sheng Shen<sup>9</sup>    Lintang Sutawika<sup>10</sup>    Jaesung Tae<sup>1</sup>    Zheng Xin Yong<sup>11</sup>

Julien Launay<sup>2,12†</sup>    Iz Beltagy<sup>13†</sup>

<sup>1</sup>Hugging Face    <sup>2</sup>LightOn    <sup>3</sup>NTU, Singapore    <sup>4</sup>Booz Allen    <sup>5</sup>EleutherAI    <sup>6</sup>Naver Labs Europe    <sup>7</sup>New York University  
<sup>8</sup>University of Washington    <sup>9</sup>Berkeley University    <sup>10</sup>Big Science    <sup>11</sup>Brown University    <sup>12</sup>LPENS    <sup>13</sup>Allen Institute for AI

Modeling methods have been a well-motivated transfer across impact of modeling the emergence of parameters models, reusing pre-training. Notably, how modeling capabilities, use mainly from linguistic language scale, our goal training setup

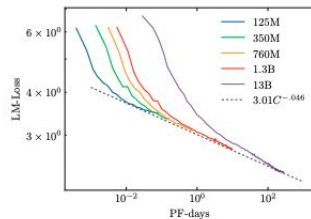


Figure 1: Smooth scaling of language modeling loss as compute budget and model size increase. We observe a power-law coefficient  $\alpha_C \sim 0.046$ , in-line with pre-training setup

## What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?

### The BigScience Architecture & Scaling Group

Thomas Wang<sup>1\*</sup>    Adam Roberts<sup>2\*</sup>  
 Daniel Hesse<sup>3</sup>    Teven Le Scao<sup>1</sup>    Hyung Won Chung<sup>2</sup>  
 Iz Beltagy<sup>4</sup>    Julien Launay<sup>3,5†</sup>    Colin Raffel<sup>1†</sup>

<sup>1</sup>Hugging Face    <sup>2</sup>Google    <sup>3</sup>LightOn


<sup>4</sup>Allen Institute for AI    <sup>5</sup>LPENS, École Normale Supérieure

### Abstract

Large pretrained Transformer language models have been shown to exhibit zero-shot generalization, i.e. they can perform a wide variety of tasks that they were not explicitly trained on. However, the architectures and pretraining objectives used across state-of-the-art models differ significantly, and there has been limited systematic comparison of these factors. In this work, we present a large-scale evaluation of modeling choices and their impact on zero-shot generalization. In particular, we focus on text-to-text models and experiment with three model architectures (causal/non-causal decoder-only and encoder-decoder), trained with two different pretraining objectives (autoregressive and masked language modeling), and evaluated with and without multitask prompted finetuning. We train



# Artifacts: what came out of BigScience? (4/4)


 **Stas Bekman**  
@StasBekman

The embed matrix with 250k multi-lingual vocab is on par in size with the transformer block, so rebalancing the pipeline to count embedding matrices as transformer blocks leads to even faster throughput and less memory usage on ranks 0 and -1

Benchmarks:

**bigscience-workshop/**  
**bigscience**


Central place for the engineering/scaling WG: documentation, SLURM scripts and logs, compute environment and data.




13 Contributors   2 Issues   272 Stars   24 Forks

github.com  
bigscience/chronicles-prequel.md at master · bigscience-workshop/bigscience  
Central place for the engineering/scaling WG: documentation, SLURM scripts and logs, compute environment and data. - bigscience/chronicles-prequel.md a...

6:34 AM · Mar 4, 2022 · Twitter Web App

 **BigScience Large Model Training**  
@BigScienceLLM

 Currently doing live surgery on the 176B model during training 🙏

While testing the checkpoint weights for integration in transformers we discovered that layer norms were not in sync across TP ranks contrary to what expected - trying to fix this while training 🙏

Wish us luck!

12:38 PM · Mar 25, 2022 · Twitter Web App

<https://github.com/bigscience-workshop/bigscience/blob/master/train/lessons-learned.md>

# Training a 176B model

# Jean Zay: accessing the french public supercomputer compute

✿ This work was granted access to the HPC resources of *Institut du développement et des ressources en informatique scientifique* (IDRIS) du *Centre national de la recherche scientifique* (CNRS) under the allocation 2021-A0101012475 made by *Grand équipement national de calcul intensif* (GENCI).

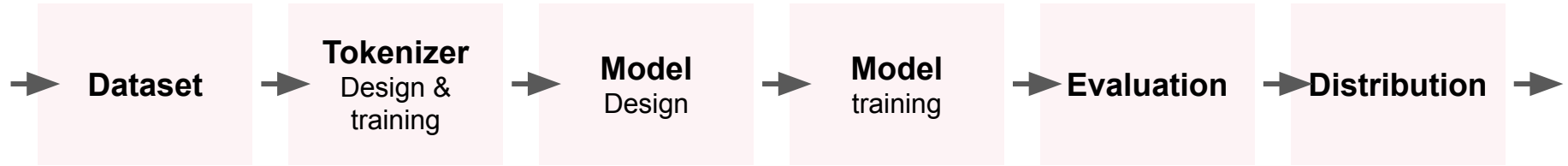
✿ Compute grant:

- 2.5M V100 hours
- 1.25M A100 hours: a reserved allocation of 416 A100 (80GB)
- and a ton of CPU

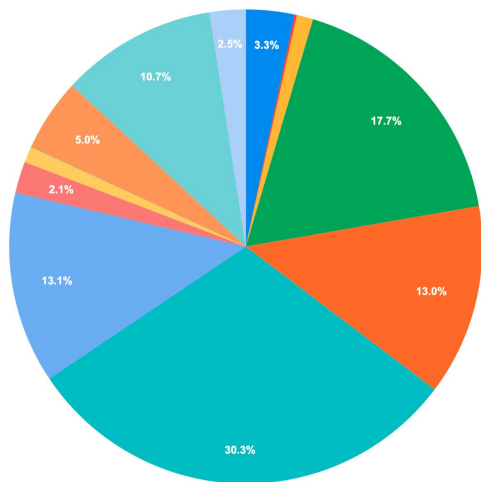
✿ Technical support - Thanks Rémi Lacroix!



# DL pipeline: what do we need to train a language model?



# Training dataset: building a 1.5T multilingual dataset



- Arabic (3,3%)
- Basque (0,2%)
- Catalan (1,1%)
- Chinese (17,7%)
- Code (13%)
- English (30,3%)
- French (13,1%)
- Indic (2,1%)
- Indonesian (1,1%)
- Niger Congo (0,03%)
- Portuguese (5%)
- Spanish (10,7%)
- Vietnamese (2,5%)

## BigScience How did we create a 1,5 TB multilingual dataset?

INPUT FROM WGS

### A/ Sourcing high-quality multilingual data

One of the training components is text extracted from the Catalog (output of the BigScience data Sourcing Hackathon that resulted in 246 data resources).

#### 1 RETRIEVING

- We had a Hackathon on December 2021, during which contributors sourced the data from the Catalog and added it to the Hugging Face Hub.
- We extracted text via a loading script from the downloads and new datasets were loaded to the Hub. Many of the initial datasets were split into many datasets, since we created datasets per languages.

#### 2 PREPROCESSING

- We examined individual sources to remove remaining pre-processing artifacts.
- We performed source-level line deduplication on selected datasets.
- We filtered items by length on higher-resourced languages.

#### 3 LOADING DATA @HF

- ✓ Training data loadable as HuggingFace datasets.

We collected text data from a human-curated catalog of data sources in all BigScience languages to best leverage our language expertise for dataset quality. This step still needed to be complemented with other approaches to meet our scale and diversity requirement, so we also used pseudo-crawled (column B) and crawled (column C) data.

### B/ Identifying seeds from a web crawl

One of the training components is text extracted from a pseudo-crawl. We initially identified seeds (603) from a web crawl to do so. We effectively retrieved text from 535 sources.

#### 1 RETRIEVING

- We created an index using the identified seeds in Common Crawl.
- We queried the index to retrieve WARC files and extracted the web pages (HTML format) from the WARC files.
- We extracted the text content from HTML web pages.

#### 2 PREPROCESSING

- We performed URL-based deduplication.
- We performed a seed level line deduplication.
- We selected high priority filters (cf step C) to remove some pages:
  - Length,
  - Character repetition,
  - Language ID confidence,
  - Common token ratios.

#### 3 LOADING DATA @HF

- ✓ Training data loadable as HuggingFace datasets.

Our pseudo-crawl data was made up of specific websites selected by participants to maximize geographical diversity, especially for English and Spanish-language data. This data required additional filters to handle the noise and artifacts of web content.

### C/ Defining filters to apply to a web crawl

One of the components of the training set is text extracted from web crawl (OSCAR v2)

#### 1 RETRIEVING

- We downloaded OSCAR v2.
- To ensure that humans wrote the retrieved text for humans, it required the creation of filters to exclude the "spam" pages from OSCAR v2. We collected inputs from native speakers for non-language agnostics filters (flagged word ratio, closed class words ratio). We created a tool to manage filtering thresholds for all the other filters.

#### 2 PREPROCESSING

- We were able to retrieve text from 13 languages.
- We performed deduplication.
- We used 13 filters to remove "spam" pages: Length, character repetition, Language ID, Stopwords, flagged word ratio...
- We removed several categories of PII (Personally identifiable information): email, username, IP address...

#### 3 LOADING DATA @HF

- ✓ Training data loadable as HuggingFace datasets.

Catalog (column A) and Pseudo-Crawl (column B) data together accounted for 65% of our target corpus size while still over-representing English. We complemented it with data obtained from a pre-existing web crawl (OSCAR v2) to improve the diversity and balance of the final dataset.

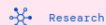
DATASET

350B tokens (1.5 TB) multilingual dataset

# Modeling: why are we training a GPT-3 like model?



The latest from Google Research



Language modelling at scale:  
Gopher, ethical considerations,  
and retrieval

December 8, 2021

RESEARCH

## Democratizing access to large-scale language models with OPT-175B

May 3, 2022

## Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance

Monday, April 4, 2022

---

## YUAN 1.0: LARGE-SCALE PRE-TRAINED LANGUAGE MODEL IN ZERO-SHOT AND FEW-SHOT LEARNING

---

Shaohua Wu*	Xudong Zhao	Tong Yu	
Rongguo Zhang	Chong Shen	Hongli Liu	Feng Li
Hong Zhu	Jiangang Luo	Liang Xu	Xuanwei Zhang



An empirical analysis of compute-optimal large language model training

Download

View publication

## Announcing AI21 Studio and Jurassic-1 Language Models

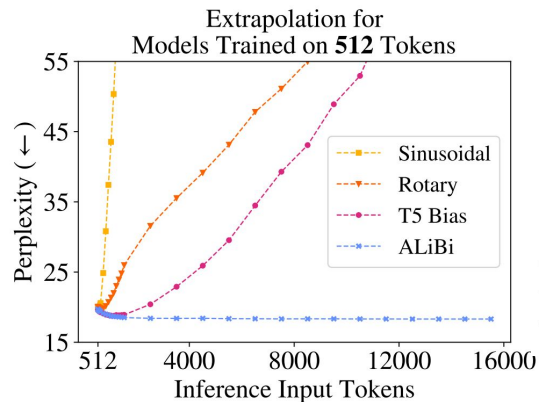
AI21 Labs' new developer platform offers instant access to our 178B-parameter language model, to help you build sophisticated text-based AI applications at scale

## Announcing GPT-NeoX-20B

Announcing GPT-NeoX-20B, a 20 billion parameter model trained in collaboration with CoreWeave.  
February 2, 2022 - Connor Leahy

# Modeling: what did we choose?

- **ALiBi** positional embeddings which allows to do good extrapolations



Positional Embedding	Average EAI Results
None	41.23
Learned	41.71
Rotary	41.46
ALiBi	<b>43.70</b>

Table 2: **ALiBi significantly outperforms other embeddings for zero-shot generalization.** All models are trained on the OSCAR dataset for 112 billion tokens.

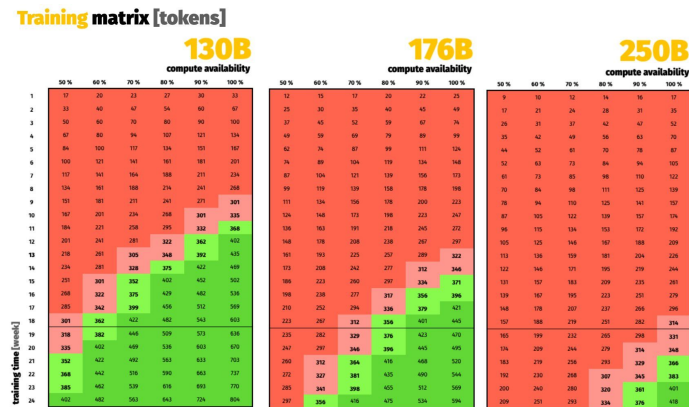
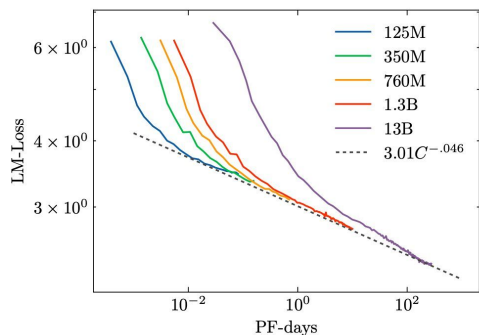
- **GELU** activation which are 30% faster than SwiGLU

Activation function	Average EAI Results
GELU	42.79
SwiGLU	<b>42.95</b>

Table 3: **SwiGLU slightly outperforms GELU for zero-shot generalization.** Models trained on The Pile for 112 billion tokens.

# Modeling: how did we decide on the final dimensions of the model?

Perform scaling law on English...



... Create fixed budget scenarios for different model sizes ...

Model	Size [Bparams.]	Pretraining [Btokens]	Budget [PF-days]	Layers	Hidden dim.	Attention heads num.	Attention heads dim.
LaMDA (Thoppilan et al., 2022)	137	432	4,106	64	8,192	128	64
GPT-3 (Brown et al., 2020)	175	300	3,646	96	12,288	96	128
J1-Jumbo (Lieber et al., 2021)	178	300	3,708	76	13,824	96	144
PanGu- $\alpha$ (Zeng et al., 2021)	207	42	604	64	16,384	128	128
Yuan (Wu et al., 2021)	245	180	3,063	76	16,384		
Gopher (Rae et al., 2021)	280	300	4,313	80	16,384	128	128
MT-530B (Smith et al., 2022)	530	270	9,938	105	20,480	128	160

Model	Size [params.]	Layers	Hidden dim.	Attention heads num.	Attention heads dim.	Memory [GB]	Performance [sec/iter.]	Performance [TFLOPs]
(1)	178	82	13,312	64	208	63	104	152
(2)	178	82	13,312	128	104	60	109	146
(3)	176	70	14,336	112	128	59	105	150

[arxiv.org/abs/2001.08361](https://arxiv.org/abs/2001.08361)

[arxiv.org/abs/2006.12467](https://arxiv.org/abs/2006.12467)


... Read the scientific literature...

... Take into account engineering constraint



# Training a 176B Model from an engineering perspective: how do we train effectively models at this scale?

```
~checkpoints/tr11-176B-ml/checkpoints/main> du -h global_step63600/  
2.3T  global_step63600/
```

 Including optimizer states and checkpoints

After preliminary studies, we selected:



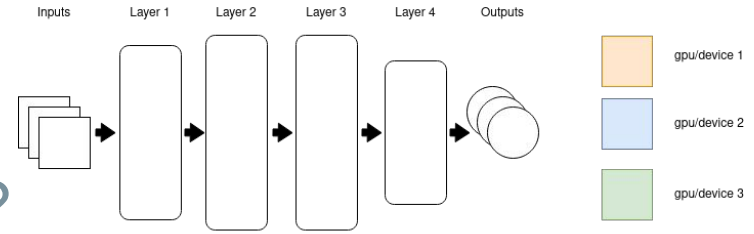
 [microsoft](#) / [Megatron-DeepSpeed](#) Public

forked from [NVIDIA/Megatron-LM](#)



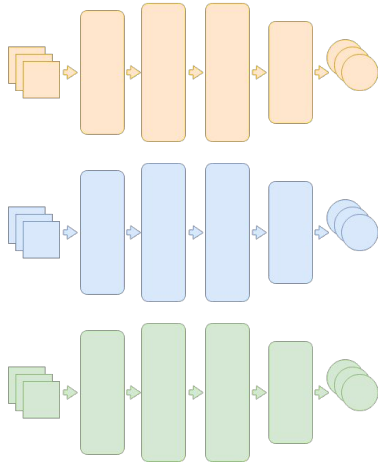
# Parallelism:

## how to use a cluster wisely for DL?



### Data parallelism

to accelerate the training speed

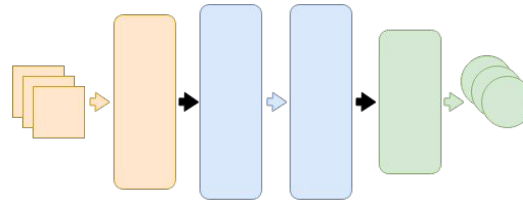


Each device has a replica of the model and receives a different batch of training data on which it performs a forward and backward pass

### Model parallelism

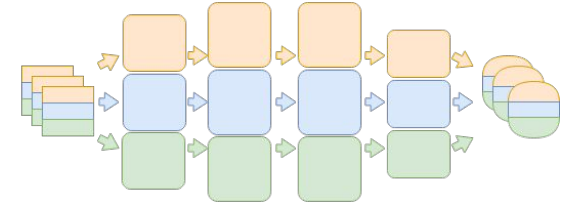
to train models that don't fit in the memory of one device

#### Pipeline parallelism



Only one or several consecutive layers of the model are placed on a single GPU

#### Tensor parallelism

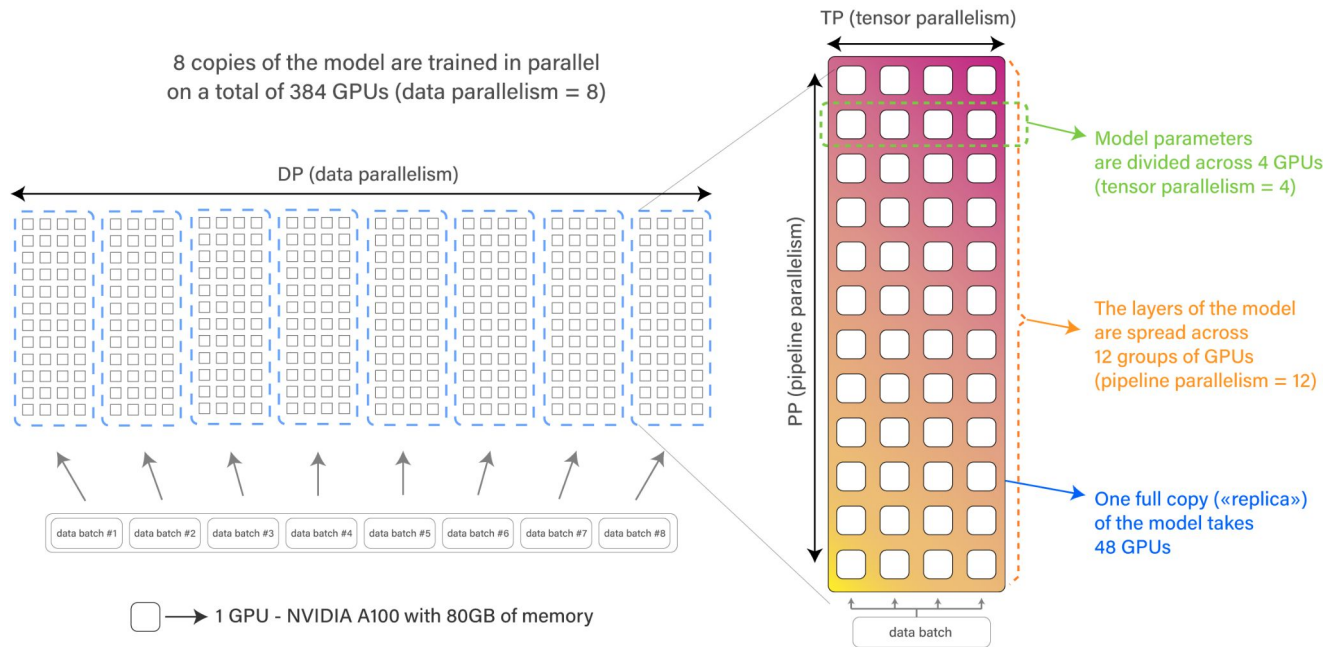


Each tensor is divided into several pieces so that instead of having the whole tensor residing on a single GPU each piece of the tensor resides on a different GPU

# Parallelism: how to use a cluster wisely for DL?

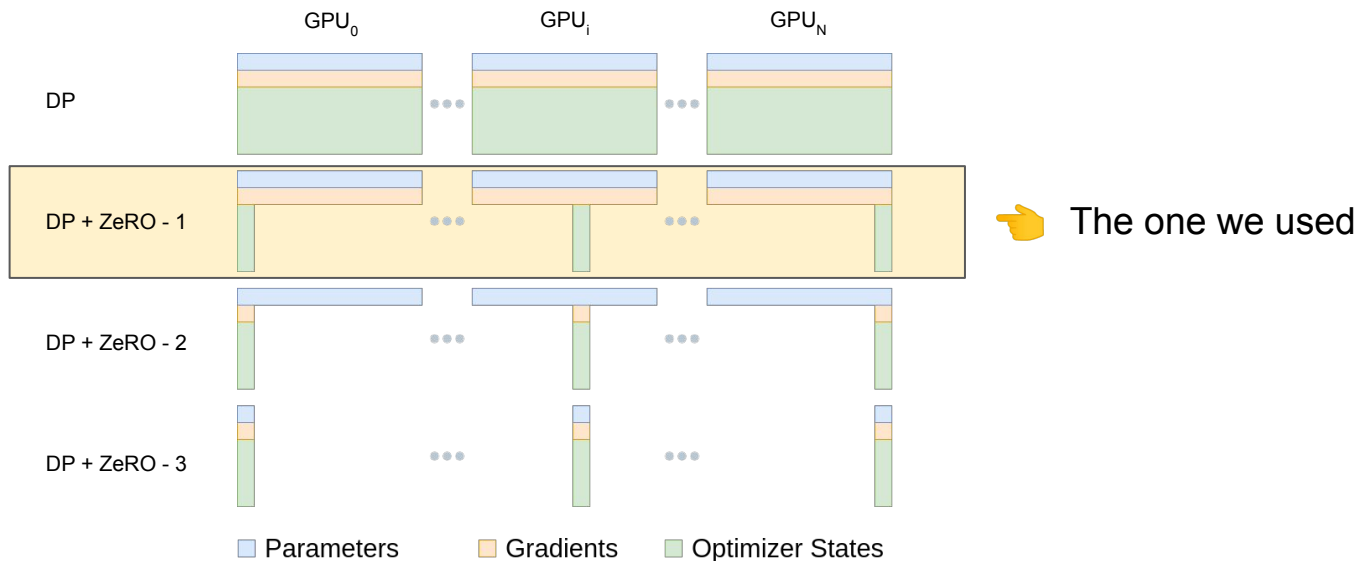
DP?  
TP?  
PP?

All 3 techniques were used!



# ZeRO data parallelism: to make the most of data parallelism

- instead of replicating everything each GPU stores only a slice of it
- free the gpus for larger batch sizes or more layers



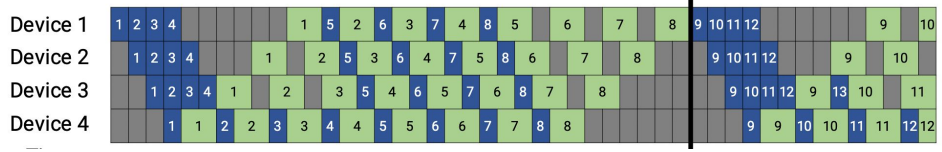
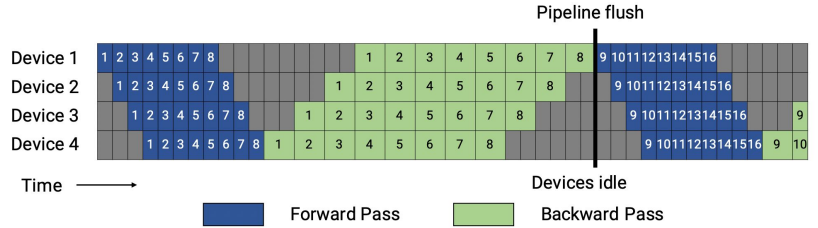
# Pipeline scheduling: improving memory footprint

👉 The one we used

All forward, all backward

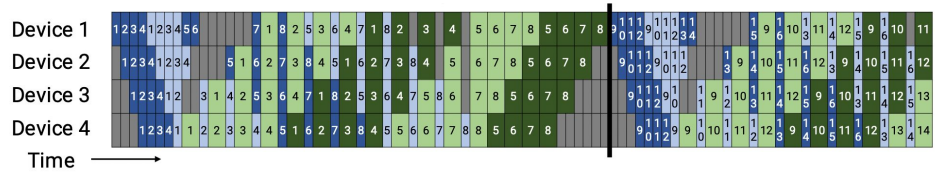
Reduce memory →

One forward, one backward (1f1b)

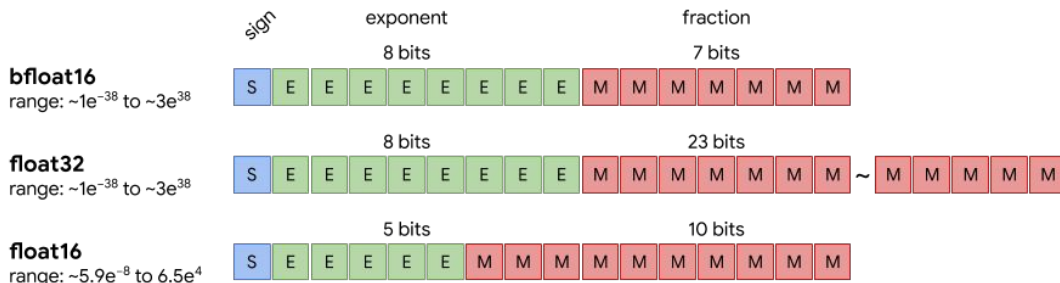


↓ Reduce bubble at the cost of communication

Interleaved 1f1b



# BF16 (+ clean data): a mixed precision enabling stable training



Source: <https://moocholic.medium.com/f684-fb32-fb16-bfloat16-f32-and-other-members-of-the-zoo-a1ca7897d407>

Trade-off between memory footprint, dynamic range and precision



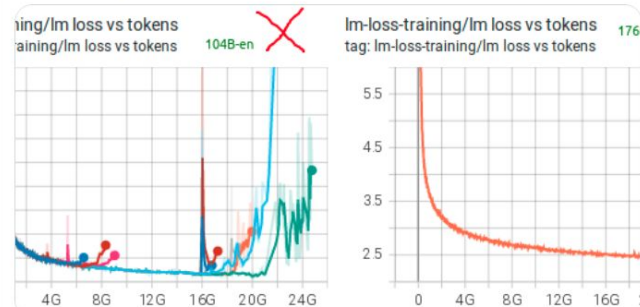
Stas Bekman  
@StasBekman

[1/2] What makes the @BigScienceLLM 176B-ml training so stable?

The 176B-ml succeeded to cross the 24B-token barrier whereas 104B-en failed.

We would love to hear your speculative and experiential reasoning for why this is so!


Following are the main candidates:




6:53 PM · Mar 20, 2022 · Twitter Web App


# Evaluation: A new benchmark to measure the performance of the model - WIP

 **Extrinsic Evaluation:** Focus on downstream, user-facing tasks

 **Intrinsic Evaluation:** Focus on encoding of linguistic and world knowledge

 **Bias/Social Impact:** Quantify encoding of stereotypes and risk of user harm

 **Multilingualism:** Ensure coverage of training and unseen language in all evaluations

 **Few-Shot Generalization:** Focus on evaluation on distributions not seen in pretraining

## Code Bases

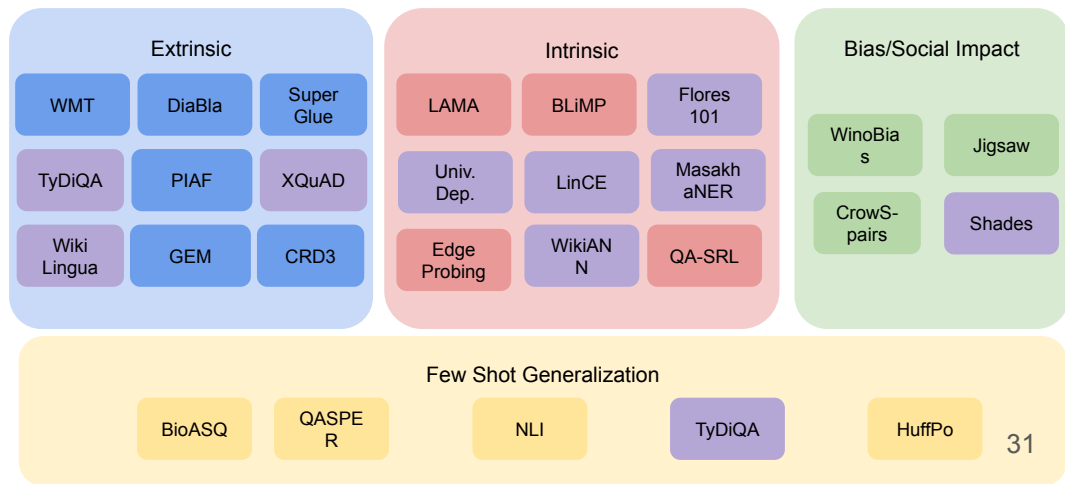


[bigscience-workshop / promptsource](#) Public



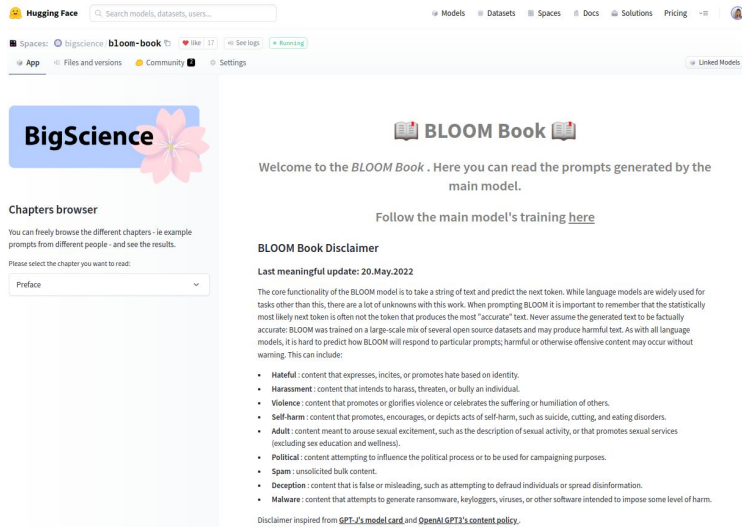
[EleutherAI / lm-evaluation-harness](#) Public

## Benchmark



# Evaluation: possibility to prompt the model

## BLOOM Book



**Chapters browser**

You can freely browse the different chapters - ie example prompts from different people - and see the results.

Please select the chapter you want to read:

Preface

### BLOOM Book

Welcome to the *BLOOM Book*. Here you can read the prompts generated by the main model.

Follow the main model's training [here](#)

#### BLOOM Book Disclaimer

Last meaningful update: 20.May.2022

The core functionality of the BLOOM model is to take a string of text and predict the next token. While language models are widely used for tasks other than this, there are a lot of unknowns with this work. When prompting BLOOM it is important to remember that the statistically most likely next token is often not the token that produces the most "accurate" text. Never assume the generated text to be factually accurate: BLOOM was trained on a large-scale mix of several open source datasets and may produce harmful text. As with all language models, it is hard to predict how BLOOM will respond to particular prompts; harmful or otherwise offensive content may occur without warning. This can include:

- **Hateful** : content that expresses, incites, or promotes hate based on identity.
- **Harassment** : content that intends to harass, threaten, or bully an individual.
- **Violence** : content that promotes or glorifies violence or celebrates the suffering or humiliation of others.
- **Self-harm** : content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders.
- **Adult** : content meant to arouse sexual excitement, such as the description of sexual activity, or that promotes sexual services (excluding sex education and wellness).
- **Political** : content attempting to influence the political process or to be used for campaigning purposes.
- **Spam** : unsolicited bulk content.
- **Deception** : content that is false or misleading, such as attempting to defraud individuals or spread disinformation.
- **Malware** : content that attempts to generate ransomware, keyloggers, viruses, or other software intended to impose some level of harm.

Disclaimer inspired from [GPT-3's model card](#) and [OpenAI GPT-3's content policy](#).

<https://hf.co/spaces/bigscience/bloom-book>

## Checkpoint: 65k steps (240B tokens)

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

We were on a Kindrex toyshop hop and we got to see if Kindrex toys were built to withstand constant jumping, lots of jumping! To do this part, I had to farduddle to simulate jumping.

Prompt  
Generated




# What's next?

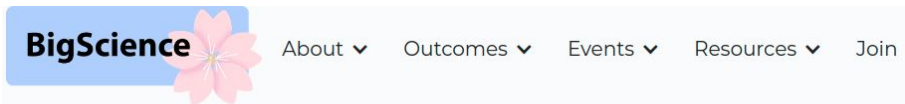
 Finish training



 Finish evaluation benchmark and perform evaluation

 Make the model more accessible - ZeRO off-loading, Distillation, ...

# Follow us!



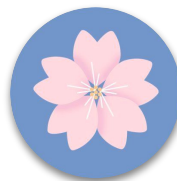
A one-year long  
research workshop  
on large multilingual  
models and datasets

**Update: Big Science model training has launched!** 

You can follow its progress [here](#) and learn more by reading our [blog post](#).

 Website

<https://bigscience.huggingface.co>



BigScience Research Workshop  
[@BigScienceW](#)



BigScience Large Model Training  
[@BigScienceLLM](#)