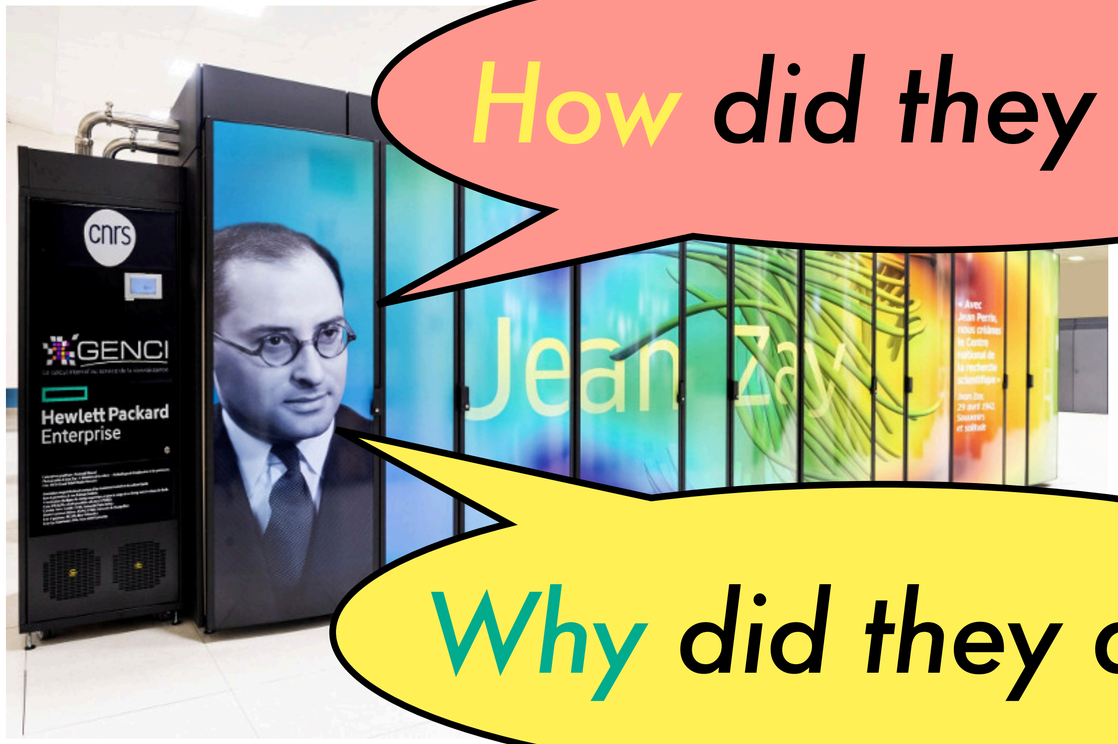


How did they do it?

Why did they do it?

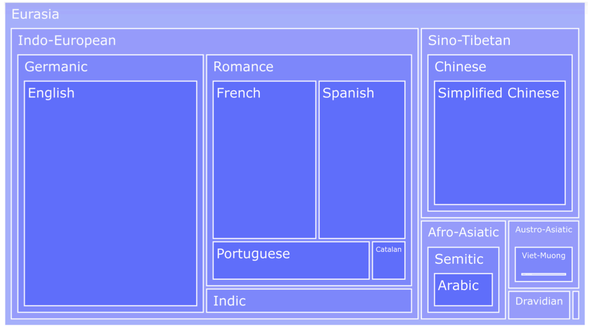
a BigScience initiative

BLOOM

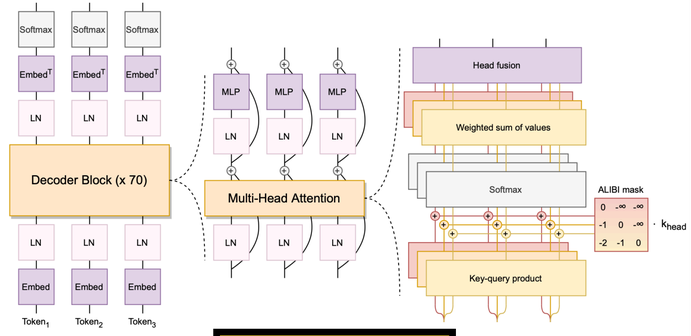
176B params · 59 languages · Open-access

BigScience

Initiative



Training Data



Model

BLOOM: 176B Open-Access Multilingual LLM

BigScience Workshop

Paper: T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (Nov. 2022)

Motivation

Pretrained Large Language Models (LLMs) now play key NLP role - they can adapt with less labelled data

Recipe 1: pretrain & finetune ELMo ULMFiT GPT BERT

Recipe 2: pretrain & zero-shot GPT-2 GPT-3

Empirical observation: model scale helps Scaling laws

Training LLMs is unaffordable for most organisations

Until recently, most LLMs were not publicly released

LLMs mostly English PanGu-α (CH) HyperClova (KO)

BLOOM aims to address these challenges

(**B**igScience **L**arge **O**pen-science **O**pen-access **M**ultilingual Language Model)

BLOOM is a 176B parameter language model

Trained on 46 natural languages 13 coding languages

Compute funded by French public grant Jean Zay

Goals:

- Release a strong multilingual LLM
- Document coordinated process of development

References:
(ELMo) M. Peters et al., "Deep Contextualized Word Representations", *NAACL* (2018)
(ULMFiT) J. Howard et al., "Universal Language Model Fine-tuning for Text Classification", ACL (2018)
(GPT) A. Radford et al., "Improving language understanding by generative pre-training" (2018)
(BERT) J. Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding", NAACL-HLT (2019)
(GPT-2) A. Radford et al., "Language Models are Unsupervised Multitask Learners" (2019)
(GPT-3) T. Brown et al., "Language models are few-shot learners", NeurIPS (2020)
(Scaling laws) J. Hestness et al., "Deep learning scaling is predictable, empirically", arxiv (2017)
(Scaling laws) J. Kaplan et al., "Scaling laws for neural language models", arxiv (2020)
(PanGu-α) W. Zeng, et al. "PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation", arxiv (2021)
(HyperClova) B. Kim et al., "What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers", arxiv (2021)

Background

Language modelling

Model probability of token sequences

Token are units of text [word] [subword] [char] [byte]

This work (& many others) models autoregressively:

$$p(x) = p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t \mid x_{<t})$$

Early lang. models used n-grams [Shannon (1948)]

n-gram model issues [scaling] [unseen n-grams]

NN [Miikkulainen ('91)] [Schmidhuber ('96)] [Bengio ('00)]

Architecture evolution: [FFN] → [RNN] → [Transformer]

Transfer/Few/Zero-shot learning

Transfer learning: pretrain then finetune on task

Word vectors: pretrain embedding [word2vec ('13)]

Pretrain & finetune network [(almost) from scratch ('11)]

Benefits of pretrained transformers [GPT] [BERT]

Zero/few-shot abilities come with scale [GPT-2] [GPT-3]

Social limitations of LLM development 🦜

Environmental impact of training [Strubell et al. (2019)]

Tech. shapes lives (like regulation) [Winner ('77,'80)]

Concentrated in corps. (EleutherAI rare exception)

References:

T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)

(Tokenisation) S. Mielke et al., "Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP", arxiv (2021)

C. Shannon, "A mathematical theory of communication", Bell System Technical Journal (1948)

(FGREP) R. Miikkulainen et al., "Natural language processing with modular PDP networks and distributed lexicon", Cognitive Science (1991)

J. Schmidhuber et al., "Sequential neural text compression", IEEE Transactions on Neural Networks (1996)

Y. Bengio et al., "A neural probabilistic language model", NeurIPS (2000)

(RNN LM) T. Mikolov et al., "Recurrent neural network based language model", Interspeech (2010)

(Transformer) A. Vaswani et al., "Attention is all you need", NeurIPS (2017)

(word2vec) T. Mikolov et al., "Distributed representations of words and phrases and their compositionality", NeurIPS (2013)

(NLP (almost) from scratch) R. Collobert et al., "Natural language processing (almost) from scratch", JMLR (2011)

(GPT) A. Radford et al., "Improving language understanding by generative pre-training" (2018)

(BERT) J. Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding", NAACL-HLT (2019)

(GPT-2) A. Radford et al., "Language Models are Unsupervised Multitask Learners" (2019)

(GPT-3) T. Brown et al., "Language models are few-shot learners", NeurIPS (2020)

(🦜) E. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜", FAccT (2021)

E. Strubell et al., "Energy and Policy Considerations for Deep Learning in NLP", ACL (2019)

L. Winner, "Autonomous technology: Technics-out-of-control as a theme in political thought", MIT Press (1978)

L. Winner, "Do artifacts have politics?", Computer Ethics (1980)

BigScience

Participants

BLOOM was developed by BigScience

Open research collaboration for an LLM

Initially: Hugging Face & French NLP folks

Funding support for compute from GENCI

Eventually, 1200+ registered participants

Backgrounds: ML | Computer Science | Stats | Law | Linguistics | Philosophy | Socio-cultural anthropology

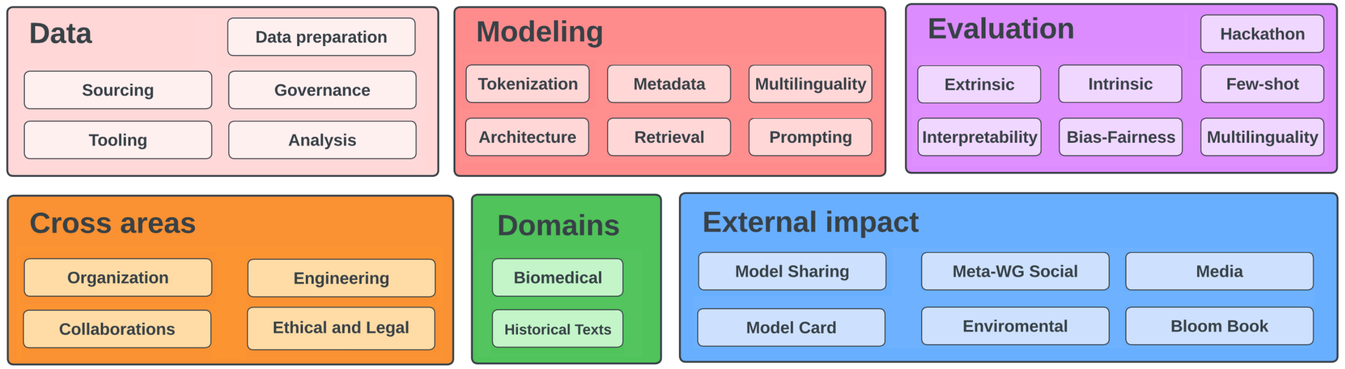
Hundreds contributed to project artifacts

38 countries were represented 

Organisation

Organised around working groups (each with a chair)

Participants encouraged to join multiple working groups



Ethical considerations

Collaboratively designed Ethical Charter (to start addressing LLM social limitations within BigScience)

Research on non-US regulation ("NLP Legal Playbook")

Ethical Charter values:

Inclusivity & diversity | openness & reproducibility | responsibility

References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
C. Akiki et al., "BigScience: A Case Study in the Social Construction of a Multilingual LLM", WBRC (2022)
(Ethical Charter) https://bigscience.huggingface.co/blog/bigscience-ethical-charter
(NLP Playbook) https://bigscience.huggingface.co/blog/legal-playbook-for-natural-language-processing-researchers

Training Dataset: Motivation

Motivation

Data work is undervalued "Everyone wants to do the model work, not the data work" (Sambasivan et al., 2021)

Focus in LLM data: use heuristics to get as much "high-quality" data as possible

High-quality - max performance on downstream tasks (minus content judged offensive)

This yields TBs of data, but compounds biases in source data

Dodge et al. (2021) Block lists for "pornographic" terms can skew data LGBTQ+ AAE Hisp

Johnson et al. (2022) GPT-3 aligns w. rep. US dominant values (curation/filtering affects values)

Birhane et al. (2021) CLIP filtering can exacerbate bias in multimodal dataset creation

Abstractive approaches to data curation - hard to document & govern (loss of provenance)

Efforts to prioritise compilations of documented sources over crawling can help The Pile

BigScience workshop: aim to prioritise human involvement local expertise language expertise

References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
N. Sambasivan et al., ""Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI", CHI (2021)
J. Dodge et al., "Documenting large webtext corpora: A case study on the colossal clean crawled corpus", arxiv (2021)
R. Johnson et al., "The Ghost in the Machine has an American accent: value conflict in GPT-3", arxiv (2022)
A. Birhane et al., "Multimodal datasets: misogyny, pornography, and malignant stereotypes", arxiv (2021)
(The Pile) L. Gao et al., "The pile: An 800gb dataset of diverse text for language modeling", arxiv (2020)
(The Pile) S. Biderman et al., "Datasheet for the Pile", arxiv (2022)

Data Governance

As datasets grow, need systems to account for interests of developers | data subjects | rights holders

Aim of BigScience: address needs by combining expertise technical | legal | sociological

Two goals of governance effort (different time scales):

- Long-term international governance Data Stewardship Org + data/rights holders (Jernite et al., 2022)
- Concrete recommendations for BigScience project data

1. Explicit permission from specific data providers within BigScience context where possible
2. Keep data sources separate until final preprocessing to assist traceability
3. Composite release: support reproducibility | respect source-dependent needs

Provide tools to inspect/visualise data (demos to analyse full corpus)

223/498 datasets made available after accounting for licensing | privacy risk | custodian agreements

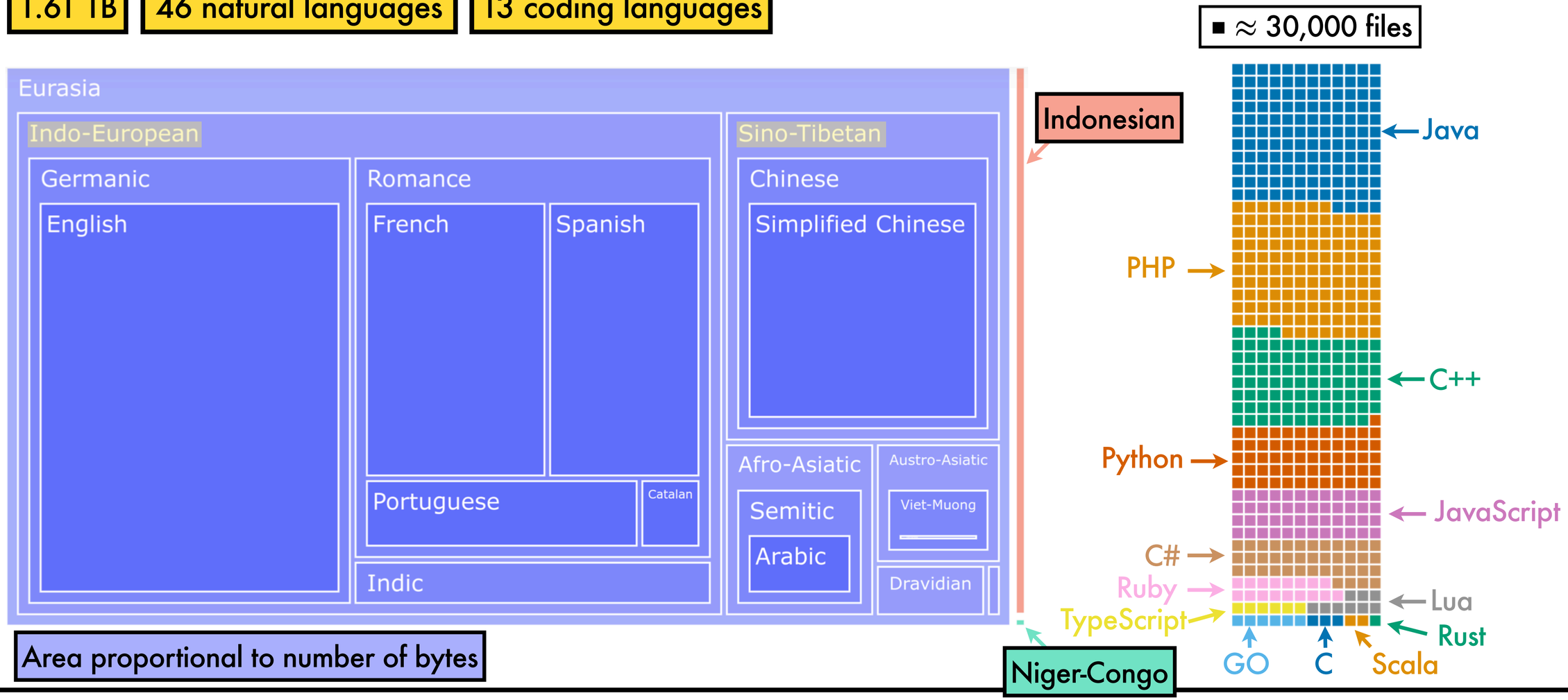
For research to analyse BLOOM that requires full dataset access, a signup form is provided

References:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
Y. Jernite et al., "Data governance in the age of large-scale data-driven language technology", FAccT (2022)

Training Dataset: ROOTS

BLOOM trained on **R**esponsible **O**pen-science **O**pen-collaboration **T**ext **S**ources (ROOTS)

1.61 TB | 46 natural languages | 13 coding languages



References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(ROOTS) H. Laurençon et al., "The BigScience ROOTS Corpus: A 1.6 TB Composite Multilingual Dataset", NeurIPS Datasets Track (2022)

Data Sources

Dataset composition - driven by goals in natural tension:

- Build LLM accessible to as many people as possible globally
- Limit to languages where project has sufficient expertise to curate and document data

Languages: start from 8 largest languages (by number of speakers) invite fluent speakers

Expanded to include: Niger-Congo languages | Indic languages

Include languages with 3+ fluent participants (avoid issues from lack of expertise) Kreutzer et al. (2022)

Source selection: BigScience Catalogue built via community hackathons McMillan-Major et al. (2022)

Extend with compilation of language-specific sources and additional websites

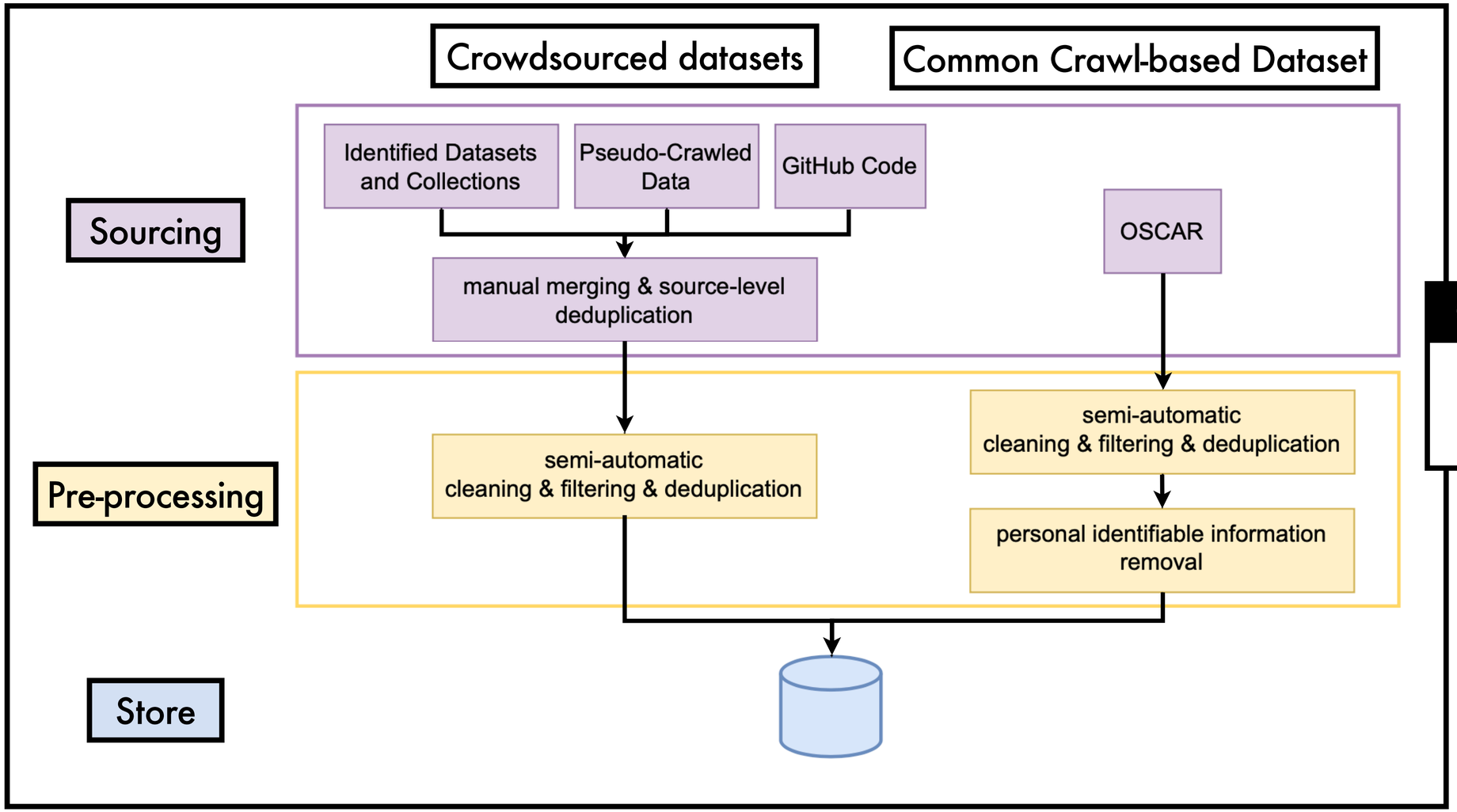
Code: Collected from GitHub via BigQuery, mimicking the languages used by AlphaCode

OSCAR: to get enough data for BLOOM (& match prior work), OSCAR (Feb 2021) was used

OSCAR provides 38% of the ROOTS corpus

References:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
W. Nekoto et al., "Participatory research for low-resourced machine translation: A case study in african languages", arxiv (2020)
A. Kunchukuttan et al., "Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages", arxiv (2020)
J. Kreutzer et al., "Quality at a glance: An audit of web-crawled multilingual datasets", ACL (2022)
A. McMillan-Major et al., "Documenting geographically and contextually diverse data sources: The bigscience catalogue of language data and resources", arxiv (2022)
Y. Li et al., "Competition-level code generation with AlphaCode", arxiv (2022)
(BigQuery) https://cloud.google.com/blog/topics/public-datasets/github-on-bigquery-analyze-all-the-open-source-code
(OSCAR) P. Ortiz Suárez et al., "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures", CMLC-7 (2019)
(OSCAR) J. Abadji et al., "Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus", CMLC-9 (2021)

Data Pipeline Overview



References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)

Prompted Datasets

Multitask prompted finetuning (aka instruction finetuning) finetunes on tasks specified via prompts

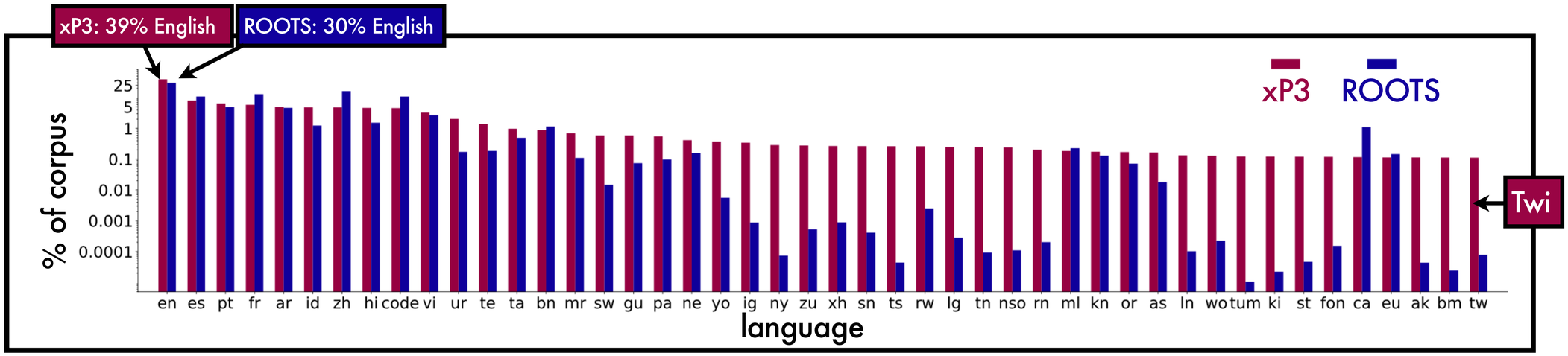
T0 trained on Public Pool of Prompts (P3) - 2000+ prompts curated through hackathons

P3 was collected using PromptSource (open source tool created as part of BigScience)

T0 showed LLMs finetuned on multitask mixtures of prompted datasets generalise to zero-shot tasks

To apply this to BLOOM, P3 extended to non-English data and extra tasks (e.g. translation)

Result: **xP3** - prompts 86 datasets 46 languages 16 tasks Prompts also machine-translated (xP3mt)



References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(T0) V. Sanh et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization", ICLR (2022)
(xP3) N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(ROOTS) H. Laurençon et al., "The BigScience ROOTS Corpus: A 1.6 TB Composite Multilingual Dataset", NeurIPS Datasets Track (2022)

Model Architecture: Design Methodology

Exhaustive exploration of architectures is infeasible

One option: simply adopt existing LLM architecture

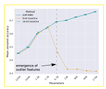
Narang et al. (2021) "improved" transformers have seen little adoption but still potentially useful

Adopt middle ground: use models known to scale well with ablations aim for best use of budget

Ablations: study zero-shot generalisation (more practical than finetuning)

Evaluation data: 29 tasks EleutherAI LM Eval Harness 9 tasks T0 eval set

Models: 6.7B pretrain objective Wang et al. (2022) 1.3B pos. embeds, activations, LNorm Le Sco et al. (2022)

Dettmers et al. (2022) identify "outlier features" above 6.7B (1.3B may not generalise) 

Out of scope: mixture-of-experts LLMs Shazeer et al. (2017) recent Switch Transformers (2022)

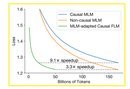
State-space models Gu et al. (2020) recent H3 (2022)

References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
S. Narang et al., "Do transformer modifications transfer across implementations and applications?", arxiv (2021)
(EleutherAI Eval) L. Gao et al., "A framework for few-shot language model evaluation", https://doi.org/10. 5281/zenodo.5371628 (2021)
(T0) V. Sanh et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization", ICLR (2022)
T. Wang et al., "What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?", arxiv (2022)
T. Le Scao et al., "What Language Model to Train if You Have One Million GPU Hours?." arxiv (2022)
T. Dettmers et al., "Llm. int8 (): 8-bit matrix multiplication for transformers at scale", arxiv (2022)
(MoE) N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer", ICLR (2017)
(State Space models) A. Gu et al., "Hippo: Recurrent memory with optimal polynomial projections", NeurIPS (2020)
(Switch Tx) W. Fedus et al. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity", arxiv (2021)
(H3) Anon, "Hungry hungry hippos: Towards language modeling with state space models", ICLR 2023 submission (2022)

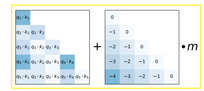
Architecture, Objective & Model Details

Study architecture: encoder-decoder | decoder-only loss: causal | prefix | masked

Wang et al. (2022) causal decoder-only models best after pretraining (approach of SotA LLMs)

Can be efficiently adapted to non-causal architecture/objective  Tay et al. (2022)

Two architectural changes to a standard causal decoder-only model are adopted

ALiBi Position Embeddings attenuate attention based on query-key distance 

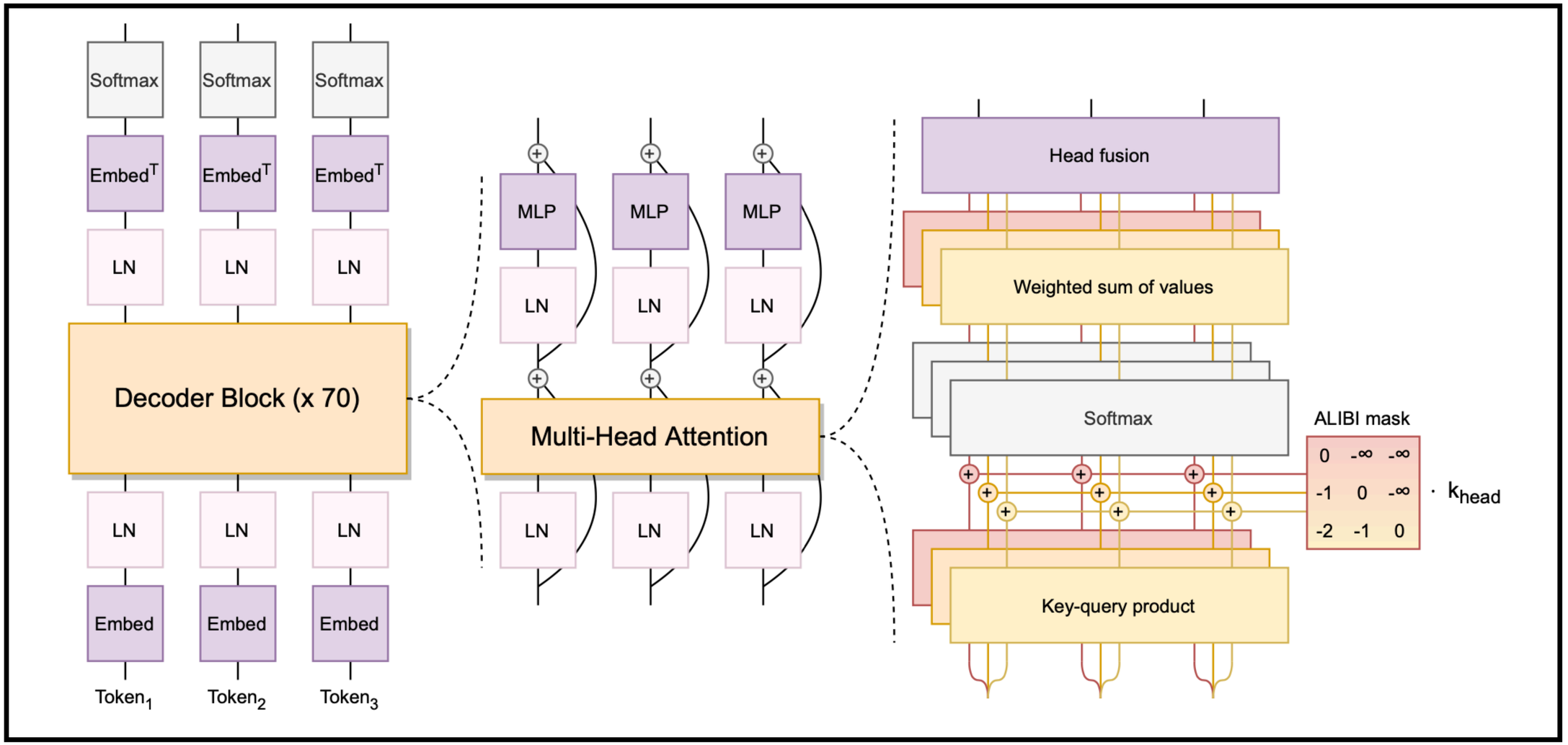
Embedding LayerNorm | bitsandbytes improves training stability, harms zero-shot generalisation

Note: later studies highlight instability from `float16` GLM-130B

BLOOM ultimately used `bfloat16` (embedding LayerNorm may not be needed)

References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
T. Wang et al., "What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?", arxiv (2022)
Y. Tay et al., "Transcending scaling laws with 0.1% extra compute", arxiv (2022)
(ALiBi) O. Press, "Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation", ICLR (2021)
(bitsandbytes) T. Dettmers et al., "Llm. int8 (): 8-bit matrix multiplication for transformers at scale", arxiv (2022)
A. Zeng et al., "GLM-130B: An Open Bilingual Pre-trained Model", arxiv (2022)

BLOOM: Decoder-only Architecture



References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(ALiBi) O. Press, "Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation", ICLR (2021)

BLOOM: Tokenisation

Mielke et al. (2022) note it is common to re-use tokenisers GPT-3 OPT re-use GPT-2 tokeniser

Given the diverse training data for BLOOM, this may be suboptimal

Tokeniser trained on ROOTS, removing duplicates (e.g. URLs) preserving language ratios

Vocab size of 250K tokens was chosen (250,680 for efficiency)

Tokeniser uses learned Byte-level BPE tokenisation (bytes rather than characters) GPT-2 tokeniser

No normalisation was applied to keep the model as general as possible

Pre-tokenisation: `?[^(\s|[.,!?,])]+` - split on words, keep space/line break sequences

The tokeniser was assessed with fertility (# subwords per word) Ács (2019)

Goal: avoid increasing fertility per language >10% vs monolingual tokenisers

Fertility assessed on subsets of UD 2.9 OSCAR goal mostly achieved (but +13% for Arabic)

References:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
S. Mielke et al., "Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP", arxiv (2021)
(GPT-3) T. Brown et al., "Language models are few-shot learners", NeurIPS (2020)
(OPT) S. Zhang et al., "OPT: Open pre-trained transformer language models", arxiv (2022)
(GPT-2) A. Radford et al., "Language Models are Unsupervised Multitask Learners" (2019)
J. Ács, "Exploring bert's vocabulary", http://juditacs.github.io/2019/02/19/bert-tokenization-stats.html (2019)
(UD) J. Nivre et al., "Universal dependencies v1: A multilingual treebank collection", LREC (2016)
(OSCAR) P. Ortiz Suárez et al., "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures", CMLC-7 (2019)

Engineering: Hardware

Training BLOOM: 3.5 months | 384 (80GB) A100s | 1,082,990 compute hours | Jean Zay cluster

48 nodes for training (4 spare for hardware failures) | 8 GPUs | 512GB RAM

Parallel file system: GPFS (mix of flash, HDDs) | **Intra-node:** 4 NVLink GPU-to-GPU interconnects

Inter-node: 4 Omni-Path 100 Gbps links per node | 8D hypercube global topology

References/Image credit:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(Jean Zay) http://www.idris.fr/eng/jean-zay/jean-zay-presentation-eng.html

Engineering: Framework

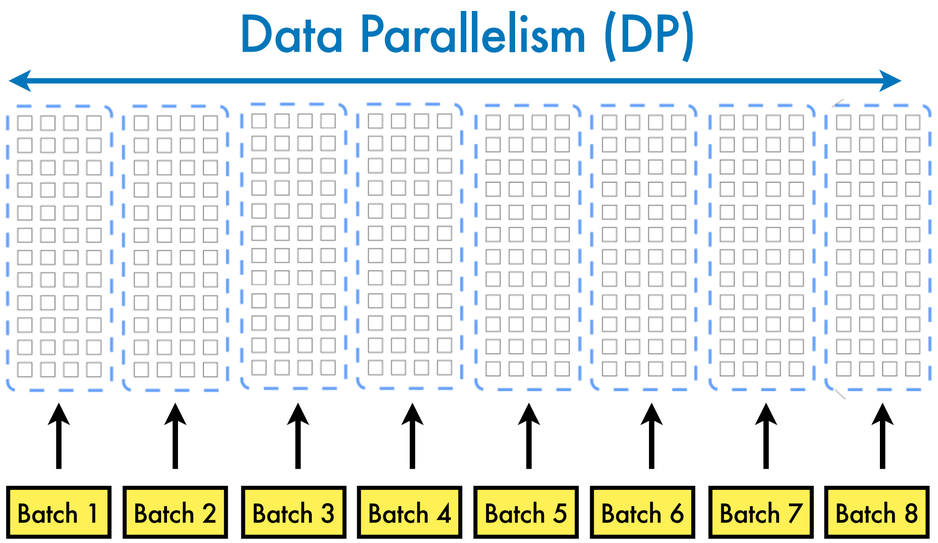
Training framework

Megatron-DeepSpeed

Megatron-LM Transformer, tensor parallelism, data loading

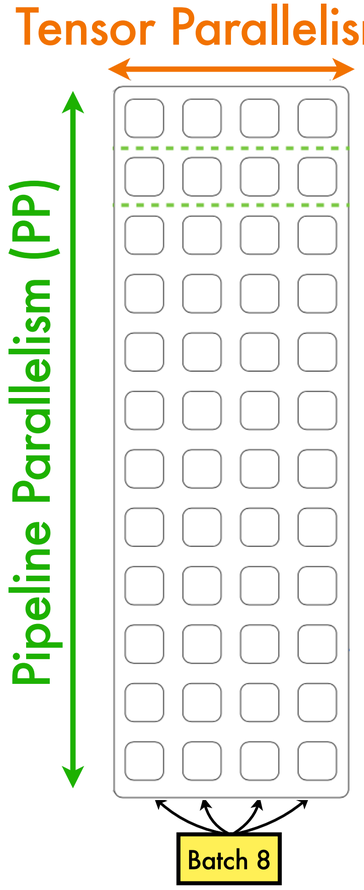
DeepSpeed ZeRO, model pipelining, distributed training

8 copies of model trained in parallel



NVIDIA 80GB A100

ZeRO stage 1 (shard optimiser states)



Data parallelism (DP)

replicates model across devices

each replica sees different data

Tensor parallelism (TP)

splits within individual layers across multiple GPUs

Pipeline parallelism (PP)

splits different layers across GPUs

3D parallelism: DP + PP + PP

156/312 TFLOPs (FP32, BFP16)

References/Image credit:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(Megatron-DeepSpeed) S. Smith et al., "Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model", arxiv (2022)
(Megatron-LM) M. Shoeybi et al., "Megatron-lm: Training multi-billion parameter language models using model parallelism", arxiv (2019)
(DeepSpeed) J. Rasley et al., "Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters", SIGKDD (2020)
S. Rajbhandari et al., "Zero: Memory optimizations toward training trillion parameter models", SC (2020)

Engineering Details

Floating point

Issues training 104B model on V100s

Possibly due to `IEEE float16`

A100s support `bfloat16` format:

- same dynamic range as `float32`
- much lower precision

Used mixed-precision training:

- `float32` for sensitive operations
- `bfloat16` for others

Resolved instabilities MT-NLG-530B

CUDA kernel fusion

Kernel fusion - multiple operations in 1 kernel call Wu et al. (2012)

Custom fused CUDA kernels used from Megatron-LM:

LayerNorm | Combinations of scaling, masking, softmax

Fuse bias term with GeLU activation via PyTorch JIT

Other challenges

To scale up to 384 GPUs required:

Disabling asynchronous CUDA kernel launches - avoid deadlocks

Splitting parameter groups - avoid excessive CPU mem allocations

Hardware failures: 1 - 2 GPUs per week

Not too disruptive (checkpoints saved every 3 hours)

References:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(bfloat16) https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus
(Mixed Precision Training) P. Micikevicius et al., "Mixed Precision Training", ICLR (2018)
(Megatron-DeepSpeed) S. Smith et al., "Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model", arxiv (2022)
H. Wu et al., "Optimizing data warehousing applications for GPUs using kernel fusion/fission", IPDPS (2012)

Training

Pretraining

Architectures 559M 1.1B 1.7B 3B 7.1B 176B

Model depth/width for <176B approximately follow prior work GPT-3 OPT

ROOTS contains 341B tokens - planned to train for equiv. number of tokens

Chinchilla revised scaling laws, trained for extra 25B tokens on repeated data

Multitask finetuning

BLOOMZ finetuned from BLOOM

Approx. hyperparams T0 FLAN

Plateaus after ≤6B tokens finetuning

Contrastive finetuning

SGPT Bi-Encoder recipe (1.3B, 7.1B)

Strong text embeddings IR STS

bitext mining reranking feat. extraction

References:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(GPT-3) T. Brown et al., "Language models are few-shot learners", NeurIPS (2020)
(OPT) S. Zhang et al., "OPT: Open pre-trained transformer language models", arxiv (2022)
(Chinchilla) J. Hoffmann et al., "Training Compute-Optimal Large Language Models", arxiv (2022)
(BLOOMZ) N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(SGPT) N. Muennighoff, "SGPT: GPT Sentence Embeddings for Semantic Search", arxiv (2022)
(MTEB) N. Muennighoff et al., "MTEB: Massive Text Embedding Benchmark", arxiv (2022)

Carbon Footprint

Prior studies of carbon footprint focus on model training — Strubell et al. (2019) — Patterson et al. (2021)

Approach with BLOOM - inspired by Life Cycle Assessment (LCA) — Klöpffer (1997)

Account for manufacturing, training, deployment

Emissions associated with training: 81 tons CO_2eq

manufacturing: 11 tons/14%

training power: 25 tons/30%

idle consumption of equipment: 45 tons/55%

BLOOM training: 37% of emissions (OPT≈50%)

Model name	Number of parameters	Power consumption	CO_2eq emissions
GPT-3	175B	1,287 MWh	*502 tons*
Gopher	280B	*1,066 MWh*	*352 tons*
OPT	175B	*324 MWh*	70 tons
BLOOM	176B	433 MWh	25 tons

Accounting not standardised — *italics*: estimated — nuclear

BLOOM API deployment estimate:

Hosted on GCP instance (16 GPUs in `us-central1`): 20 kg CO_2eq

not representative of all uses

References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
A. Luccioni et al., "Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model", arxiv (2022)
E. Strubell et al., "Energy and Policy Considerations for Deep Learning in NLP", ACL (2019)
D. Patterson et al., "Carbon emissions and large neural network training", arxiv (2021)
W. Klöpffer, "Life cycle assessment", Environmental Science and Pollution Research (1997)
(OPT) S. Zhang et al., "OPT: Open pre-trained transformer language models", arxiv (2022)

Release

Model Card

BLOOM Model Card: Mitchell et al. (2019)

- technical specs
- training details
- intended uses
- out-of-scope uses
- limitations

Licensing

Aimed for balance between:

unrestricted open-access

behavioural-use clauses | Contractor et al. (2022)

The latter are used in "Responsible AI Licenses (RAILs)"

RAIL for BLOOM separates "source code" & "model"

Source code: Apache 2.0 license

Model: 13 behavioural-use restrictions

Based on Model Card and BigScience ethical charter

Definitions for "use"/"derived works" include:

prompting | finetuning | distillation | logits | prob. distributions

Extra | Rationale: blog (CarlosMF) | RAIL critique (Yannic Kilcher)

References:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
M. Mitchell et al., "Model cards for model reporting", FAccT (2019)
D. Contractor et al., "Behavioral use licensing for responsible AI", FAccT (2022)
(blog post on Open RAIL rationale) https://huggingface.co/blog/open_rail
(video critique by Yannic Kilcher) https://www.youtube.com/watch?v=W5M-dvzpzSQ

Evaluation Overview

Overview

BLOOM is compared to LLMs on:

- Zero-shot tasks
- One-shot tasks
- Multitask finetuning

HELM

Multilingual probing used for analysis

Prompts

Prompts crowd-sourced promptsource

Multiple prompts per task (peer reviewed)

Infrastructure

Extend EleutherAI LM Eval harness (add promptsource)

Prompted LM evaluation harness (open-sourced)

Datasets

SuperGLUE subset of tasks (selected for low compute)

Machine Translation (MT) WMT14 Flores-101 DiaBLa

Summarisation WikiLingua (9 languages)

Baselines

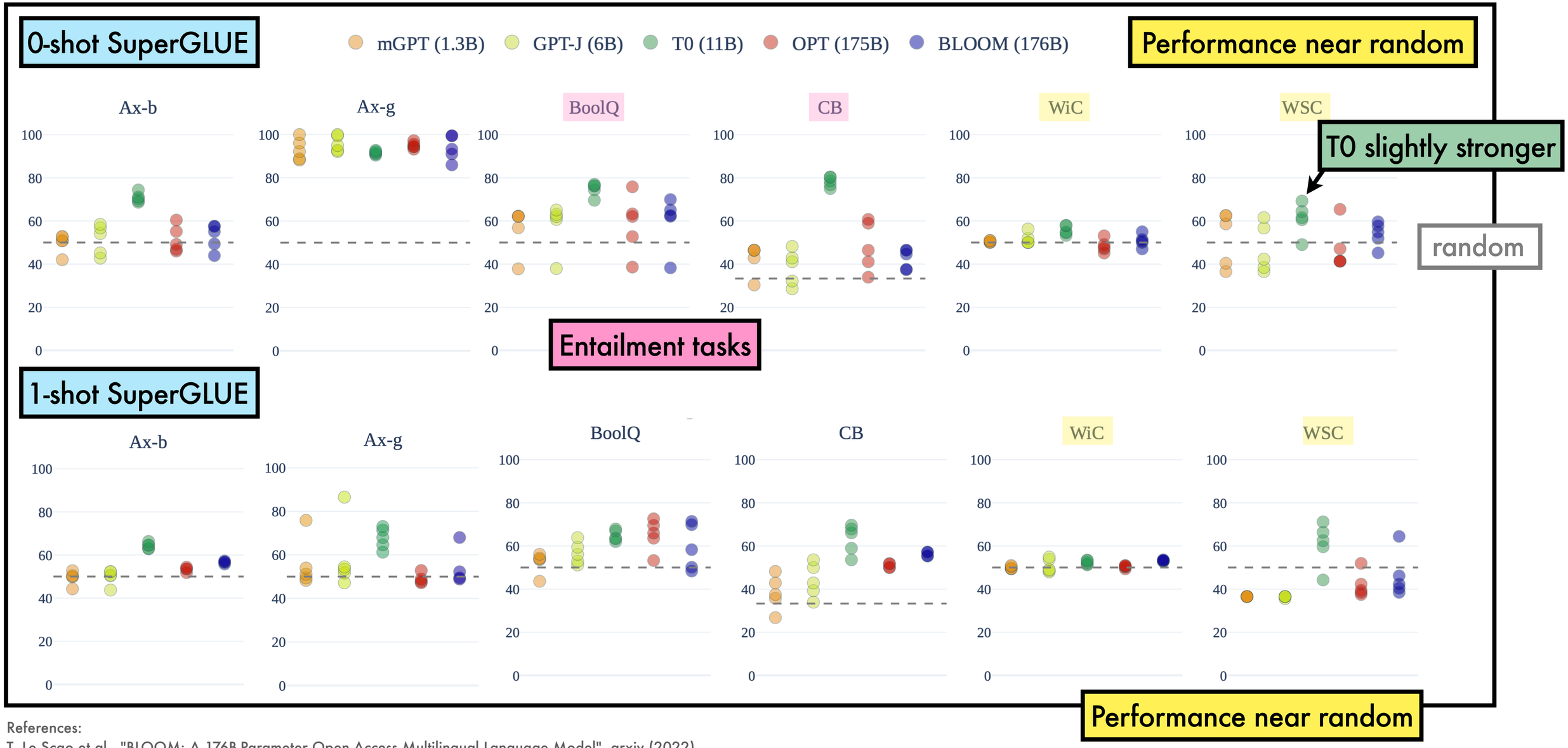
mGPT GPT-Neo GPT-J-6B GPT-NeoX T0 OPT XGLM

M2M AlexaTM mTk-Instruct Codex GPT-fr

References/image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(Helm image and benchmark) https://crfm.stanford.edu/helm/latest/
S. Bach et al., "Promptsource: An integrated development environment and repository for natural language prompts", arxiv (2022)
(LM-Evaluatio-harness) L. Gao et al., "A framework for few-shot language model evaluation", Version v0.0.1 (2021)
(SuperGLUE) A. Wang et al., "Superglue: A stickier benchmark for general-purpose language understanding systems", NeurIPS (2019)
(WMT14) O. Bojar et al., "Findings of the 2014 workshop on statistical machine translation", WMT (2014)
N. Goyal et al., "The flores-101 evaluation benchmark for low-resource and multilingual machine translation", TACL (2022)
R. Bawden et al., "DiaBLa: a corpus of bilingual spontaneous written dialogues for machine translation", LREC (2021)
F. Ladhak et al., "WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization", arxiv (2020)

(mGPT) O. Shliazhko et al., "mGPT: Few-shot learners go multilingual", arxiv (2022)
(GPT-Neo) S. Black et al., "GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow", (2021)
(GPT-J-6B) B. Wang et al., "GPT-J-6B: A 6 billion parameter autoregressive language model" (2021)
(GPT-NeoX) S. Black et al., "GPT-NeoX-20B: An open-source autoregressive language model", arxiv (2022)
(T0) V. Sanh et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization", ICLR (2022)
(OPT) S. Zhang et al., "OPT: Open pre-trained transformer language models", arxiv (2022)
(XGLM) X. Lin et al. "Few-shot learning with multilingual language models", arxiv (2021)
(M2M) A. Fan et al., "Beyond English-Centric multilingual machine translation", JMLR (2021)
(AlexaTM) S. Soltan et al., "AlexaTM 20B: Few-shot learning using a large-scale multilingual seq2seq model, arxiv (2022)
(mTk-Instruct) Y. Wang et al., "Benchmarking generalization via in-context instructions on 1,600+ language tasks", (2022)
(Codex) M. Chen et al. "Evaluating large language models trained on code", arxiv (2021)
(GPT-Fr) A. Simoulin et al., "Un modèle Transformer Génératif Pré-entrainé pour le _____ francais", ATALA (2021)

SuperGLUE Experiments



References:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)

Machine Translation Benchmarks

WMT'14 (zero-shot with BLOOM)

Prompt	en → fr		fr → en		en → hi		hi → en	
Shots	0	1	0	1	0	1	0	1
a_good_translation	15.38	36.39	16.81	36.56	1.90	14.49	13.04	24.60
gpt3	7.90	32.55	12.73	33.14	0.26	6.51	0.66	9.99
version	21.96	34.22	26.79	35.42	1.96	13.95	11.48	25.81
xglm	4.16	28.99	11.23	33.30	0.63	14.55	4.10	13.22

Poor performance with "gpt3" and "xglm"

Significant variation across prompts

Reasonable 1-shot performance with the right prompt

DiaBLa (zero-shot)

informal dialogue dataset

Prompt	en → fr			fr → en		
	T0	mGPT	BLOOM	T0	mGPT	BLOOM
MT sent-level	0.33	0.09	0.05	12.53	0.27	0.11
MT complete (1-orig-context)	0.87	0.13	1.08	13.77	0.59	1.31

Performance not great

Problems: Over-generation | Not producing correct language

References:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(WMT14) O. Bojar et al., "Findings of the 2014 workshop on statistical machine translation", WMT (2014)
(DiaBLa) R. Bawden et al., "DiaBLa: a corpus of bilingual spontaneous written dialogues for machine translation", LREC (2021)

Machine Translation Benchmarks

Flores-101 1-shot (devtest)

Src↓	Trg→	eng	ben	hin	swh	yor
eng	M2M	–	23.04	28.15	29.65	2.17
	BLOOM	–	25.52	27.57	21.7	2.8
ben	M2M	22.86	–	21.76	14.88	0.54
	BLOOM	30.23	–	16.4	–	–
hin	M2M	27.89	21.77	–	16.8	0.61
	BLOOM	35.40	23.0	–	–	–
swh	M2M	30.43	16.43	19.19	–	1.29
	BLOOM	37.9	–	–	–	1.43
yor	M2M	4.18	1.27	1.94	1.93	–
	BLOOM	3.8	–	–	0.84	–

Low resource languages

SWH/YOR (<50K tokens)

Src ↓	Trg →	eng	fre	hin	ind	vie
eng	M2M	–	41.99	28.15	37.26	35.1
	BLOOM	–	44.4	27.57	38.75	28.83
fre	M2M	37.17	–	22.91	29.14	30.26
	BLOOM	45.11	–	17.04	29.50	31.66
hin	M2M	27.89	25.88	–	21.03	23.85
	BLOOM	35.40	27.83	–	–	–
ind	M2M	33.74	30.81	22.18	–	31.4
	BLOOM	44.59	29.75	–	–	–
vie	M2M	29.51	28.52	20.35	27.1	–
	BLOOM	38.77	28.57	–	–	–

High → mid-resource language pairs

solid results

References:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
N. Goyal et al., "The flores-101 evaluation benchmark for low-resource and multilingual machine translation", TACL (2022)
(M2M) A. Fan et al., "Beyond English-Centric multilingual machine translation", JMLR (2021)

Machine Translation Cont.

Flores-101 1-shot (devtest)

Src↓	Trg→	cat	spa	fre	por
cat	M2M	–	25.17	35.08	35.15
	BLOOM	–	29.12	34.89	36.11
spa	M2M	23.12	–	29.33	28.1
	BLOOM	31.82	–	24.48	28.0
Galician glg	M2M	30.07	27.65	37.06	34.81
	BLOOM	38.21	27.24	36.21	34.59
fre	M2M	28.74	25.6	–	37.84
	BLOOM	38.13	27.40	–	39.60
por	M2M	30.68	25.88	40.17	–
	BLOOM	40.02	28.1	40.55	–

Romance languages

solid results

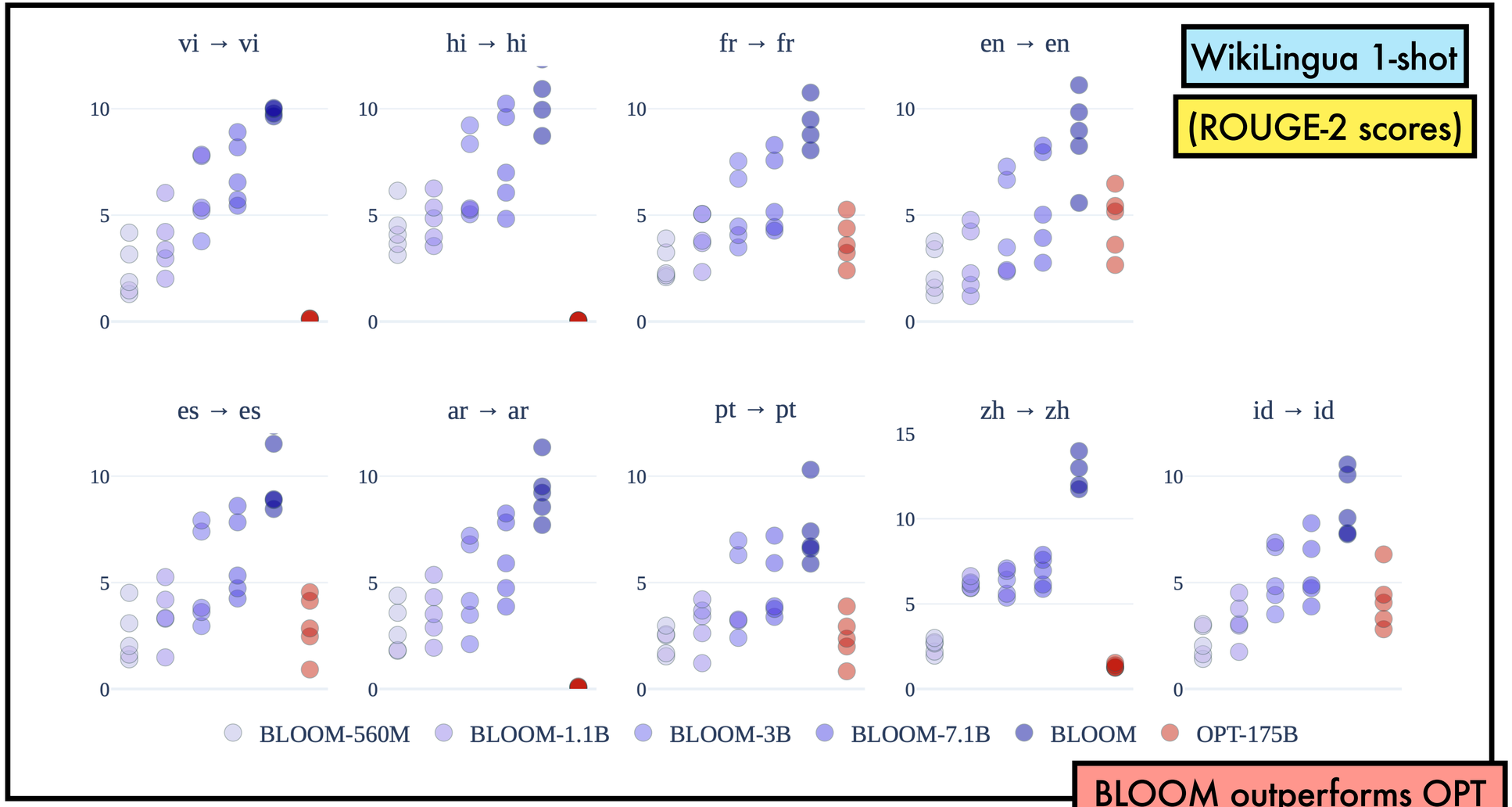
Flores-101 1-shot (devtest)

Src ↓	Trg →	ara	fre	eng	chi	spa
ara	M2M	–	25.7	25.5	13.1	16.74
	XGLM	–	17.9	27.7	–	–
	AlexaTM	–	35.5	41.8	–	23.2
	BLOOM	–	33.26	40.59	18.88	23.33
fre	M2M	15.4	–	37.2	17.61	25.6
	XGLM	5.9	–	40.4	–	–
	AlexaTM	24.7	–	47.1	–	26.3
	BLOOM	23.30	–	45.11	22.8	27.4
eng	M2M	17.9	42.0	–	19.33	25.6
	XGLM	11.5	36.0	–	–	–
	AlexaTM	32.0	50.7	–	–	31.0
	BLOOM	28.54	44.4	–	27.29	30.1
chi	M2M	11.55	24.32	20.91	–	15.92
	XGLM	–	–	–	–	–
	AlexaTM	–	–	–	–	–
	BLOOM	15.58	25.9	30.60	–	20.78
spa	M2M	12.1	29.3	25.1	14.86	–
	XGLM	–	–	–	–	–
	AlexaTM	20.8	33.4	34.6	??	–
	BLOOM	18.69	24.48	33.63	20.06	–

High resource languages

References:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
N. Goyal et al., "The flores-101 evaluation benchmark for low-resource and multilingual machine translation", TACL (2022)
(M2M) A. Fan et al., "Beyond English-Centric multilingual machine translation", JMLR (2021)
(XGLM) X. Lin et al., "Few-shot learning with multilingual language models", arxiv (2021)
(AlexaTM) S. Soltan et al., "AlexaTM 20B: Few-shot learning using a large-scale multilingual seq2seq model, arxiv (2022)

Summarisation



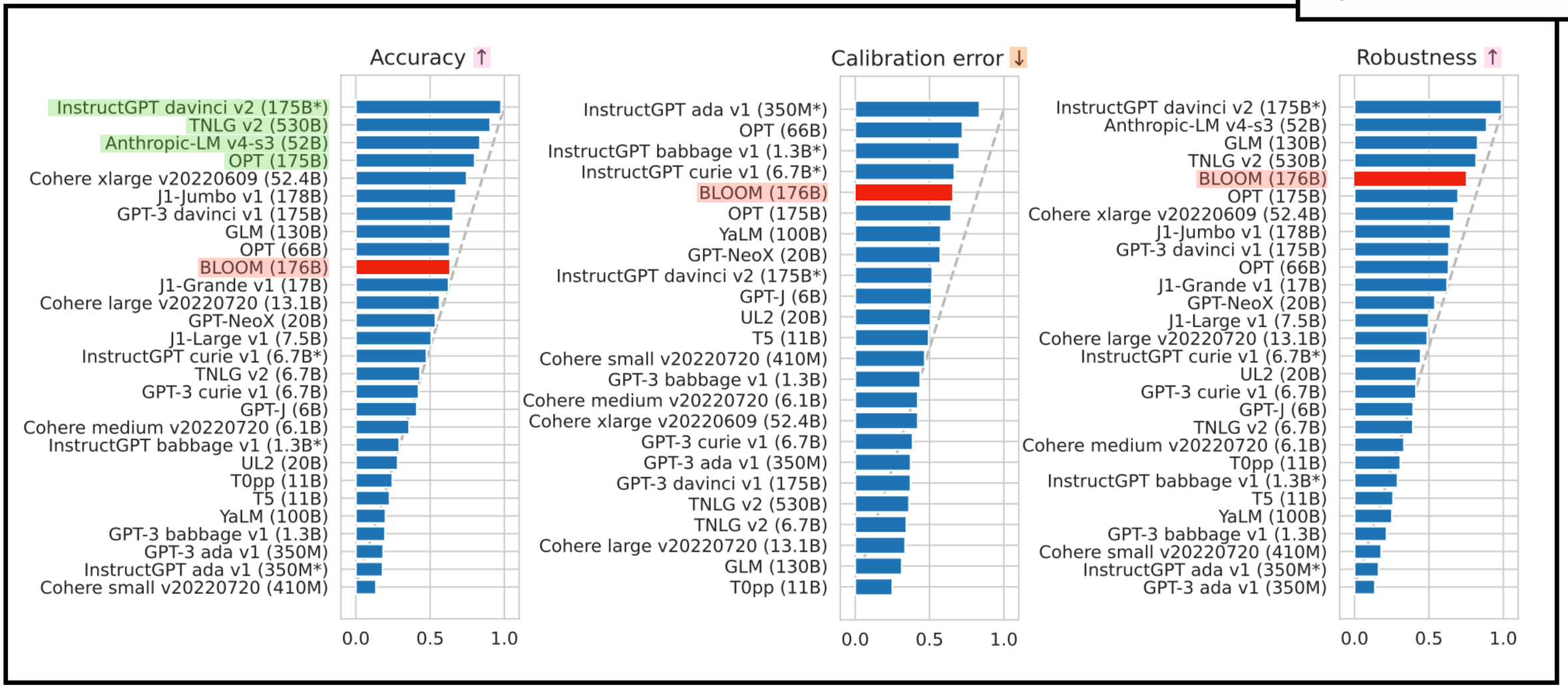
WikiLingua 1-shot

(ROUGE-2 scores)

BLOOM outperforms OPT

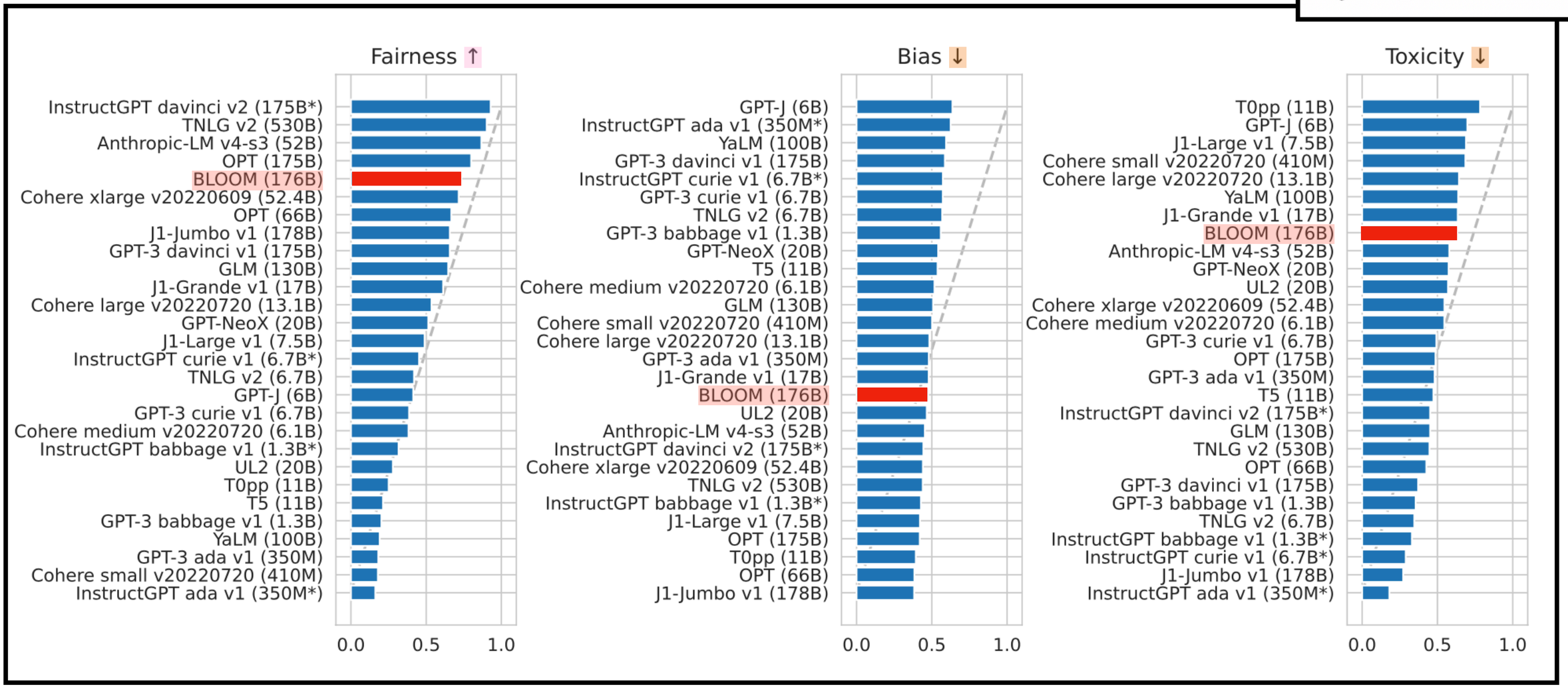
References:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
F. Ladhak et al., "WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization", arxiv (2020)
(OPT) S. Zhang et al., "OPT: Open pre-trained transformer language models", arxiv (2022)

HELM Benchmark



References:
P. Liang et al., "Holistic evaluation of language models", arxiv (2022)

HELM Benchmark Cont.

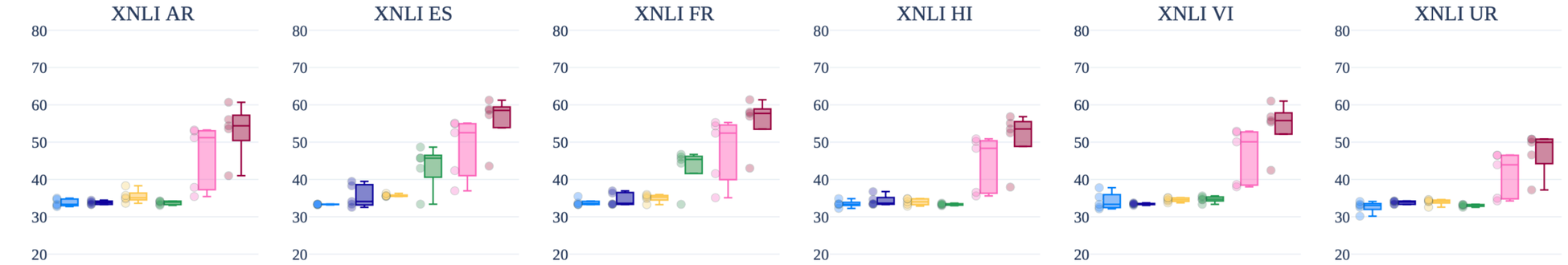


References:
P. Liang et al., "Holistic evaluation of language models", arxiv (2022)

Multilingual Multitask Finetuning

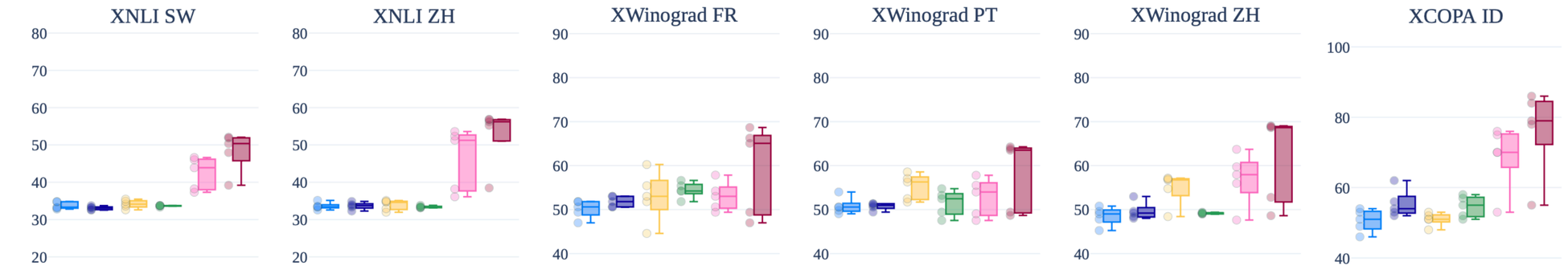
BLOOMZ: BLOOM with multitask finetuning on xP3

Zero-shot Natural Language Inference



Multilingual multitask finetuning brings boosts zero-shot performance

Zero-shot Coreference Resolution



XGLM-7.5B BLOOM mTk-13B T0-11B BLOOMZ-7.1B BLOOMZ

Note: T0 also benefits from xP3 finetuning

References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(xP3) N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)
(XNLI) A. Conneau et al., "XNLI: Evaluating Cross-lingual Sentence Representations", EMNLP (2018)
(XGLM) X. Lin et al., "Few-shot learning with multilingual language models", arxiv (2021)
(mTk-Instruct) Y. Wang et al., "Benchmarking generalization via in-context instructions on 1,600+ language tasks", (2022)
(T0) V. Sanh et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization", ICLR (2022)
(XWinograd) A. Tikhonov et al., "It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning", ACL/IJCNLP (2021)

Code Generation

HumanEval (function completion)

	Pass@k $k=1$	$k=10$	$k=100$
GPT-Neo 1.3B	4.79%	7.47%	16.30%
GPT-Neo 2.7B	6.41%	11.27%	21.37%
GPT-J 6B	11.62%	15.74%	27.74%
GPT-NeoX 20B	15.4%	25.6%	41.2%
Codex-300M	13.17%	20.37%	36.27%
Codex-679M	16.22%	25.7%	40.95%
Codex-2.5B	21.36%	35.42%	59.5%
Codex-12B	28.81%	46.81%	72.31%
BLOOM-560M	0.82%	3.02%	5.91%
BLOOM-1.1B	2.48%	5.93%	9.62%
BLOOM-1.7B	4.03%	7.45%	12.75%
BLOOM-3B	6.48%	11.35%	20.43%
BLOOM-7.1B	7.73%	17.38%	29.47%
BLOOM	15.52%	32.20%	55.45%
BLOOMZ-560M	2.18 %	4.11%	9.00%
BLOOMZ-1.1B	2.63%	6.22%	11.68%
BLOOMZ-1.7B	4.38%	8.73%	16.09%
BLOOMZ-3B	6.29%	11.94%	19.06%
BLOOMZ-7.1B	8.06%	15.03%	27.49%
BLOOMZ	12.06%	26.53%	48.44%

higher is better

trained on "The Pile" (11% code)

trained on ROOTS (13% code)

limited code completion in xP3

−3.5 %

Performance drops for BLOOMZ

References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(HumanEval) M. Chen et al., "Evaluating large language models trained on code", arxiv (2021)
(BLOOMZ) N. Muennighoff et al., "Crosslingual Generalization through Multitask Finetuning", arxiv (2022)

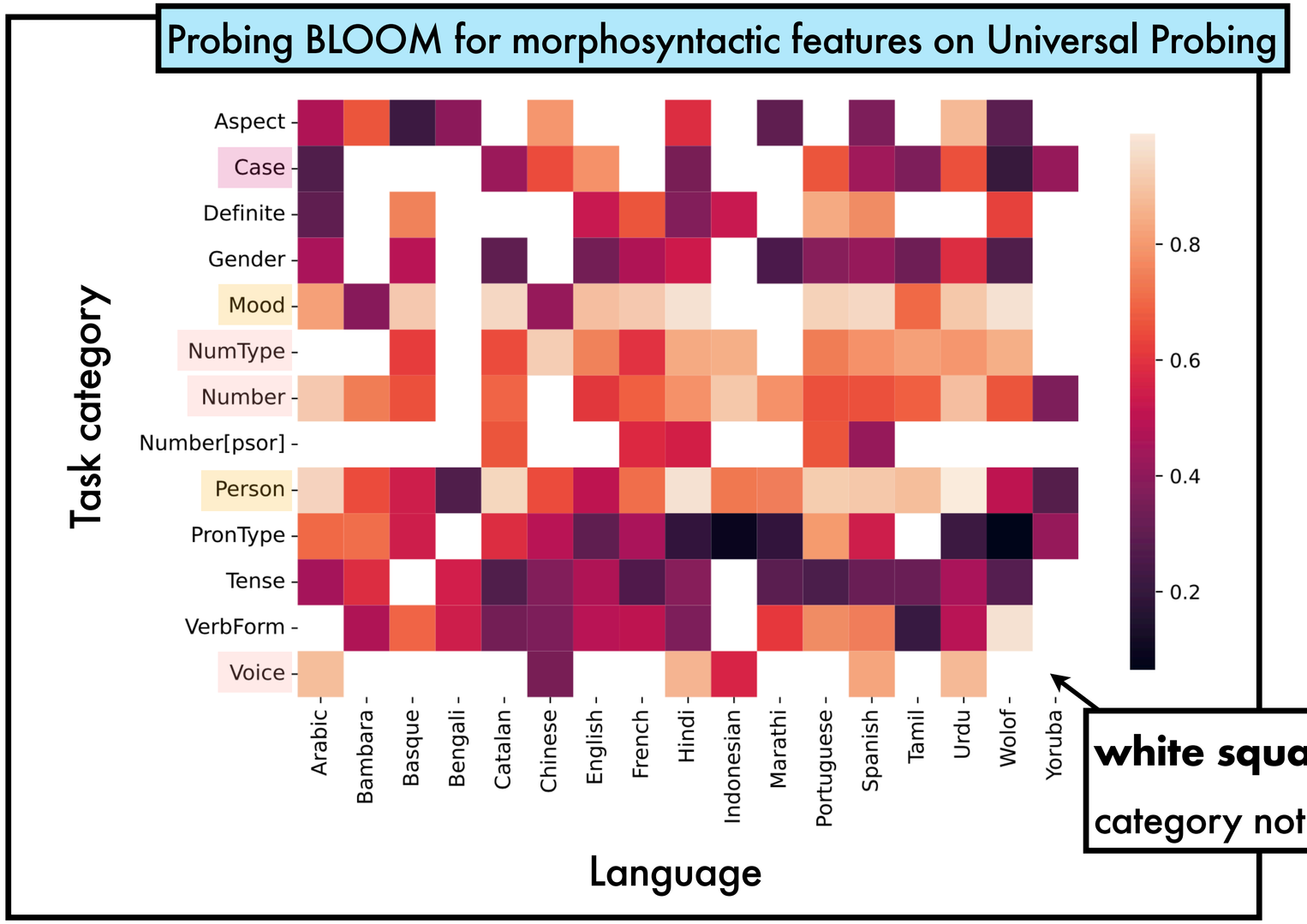
Embedding Evaluations

	ST5-XL	LASER2	MiniLM-L12 [33]	MPNet[34]	LaBSE	SGPT-BLOOM-1.7B	SGPT-BLOOM-7.1B
Embedding classification performance on MASSIVE (FitzGerald et al., 2022) scored using accuracy							
Arabic (ar)	4.18	37.16	51.43	45.14	50.86	54.59	**59.25**
Bengali (bn)	2.60	42.51	48.79	35.34	58.22	57.76	**61.59**
English (en)	**72.09**	47.91	69.32	66.84	61.46	66.69	69.67
Spanish (es)	57.97	45.44	64.43	59.66	58.32	61.77	**66.35**
French (fr)	60.99	46.13	64.82	60.25	60.47	64.58	**66.95**
Hindi (hi)	3.02	40.20	62.77	58.37	59.40	60.74	**63.54**
Indonesian (id)	41.53	45.81	**65.43**	59.85	61.12	60.07	64.06
Kannada (kn)	2.79	4.32	50.63	40.98	**56.24**	48.56	53.54
Malayalam (ml)	2.98	41.33	54.34	42.41	57.91	55.10	**58.27**
Portuguese (pt)	57.95	48.55	64.89	61.27	60.16	62.52	**66.69**
Swahili (sw)	30.60	31.89	31.95	29.57	**51.62**	43.90	49.81
Tamil (ta)	1.79	29.63	50.17	36.77	55.04	52.66	**56.40**
Telugu (te)	2.26	36.03	52.82	40.72	**58.32**	49.32	54.71
Urdu (ur)	2.70	26.11	56.37	52.80	56.70	51.00	**56.75**
Vietnamese (vi)	21.47	44.33	59.68	56.61	56.67	59.85	**64.53**
Semantic textual similarity on STS22 (Madabushi et al., 2022) scored using spearman correlation of cosine similarities							
Arabic (ar)	29.60	42.57	52.19	46.20	57.67	48.64	**58.67**
English (en)	64.32	39.76	63.06	61.72	60.97	61.45	**66.13**
Spanish (es)	58.16	54.92	59.91	56.56	63.18	61.81	**65.41**
French (fr)	77.49	58.61	74.30	70.55	77.95	73.18	**80.38**
Chinese (zh)	33.55	49.41	61.75	58.75	63.02	58.53	**66.78**

SGPT-BLOOM-7.1B is a larger model than baselines

References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
N. Muennighoff et al., "MTEB: Massive text embedding benchmark", arxiv (2022)
(ST5) J. Ni et al., "Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models", ACL (2022)
(LASER2) K. Heffernan et al., "Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages", arxiv (2022)
(MiniLM-L12) W. Wang et al., "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers", NeurIPS (2020)
(MPNet) K. Song et al., "MPNet: Masked and permuted pre-training for language understanding", NeurIPS (2020)
(finetuning for MiniLM, MPNet) N. Reimers et al., "Sentence-bert: Sentence embeddings using siamese bert-networks", EMNLP (2019)
(LaBSE) F. Feng et al., "Language-agnostic BERT Sentence Embedding", ACL (2022)
(SGPT) N. Muennighoff, "SGPT: GPT Sentence Embeddings for Semantic Search", arxiv (2022)
(MASSIVE) J. FitzGerald et al., "MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages", arxiv (2022)
(STS22) H. Madabushi et al., "SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding", arxiv (2022)

Multilingual Probing



white square → morphosyntactic category not represented in language

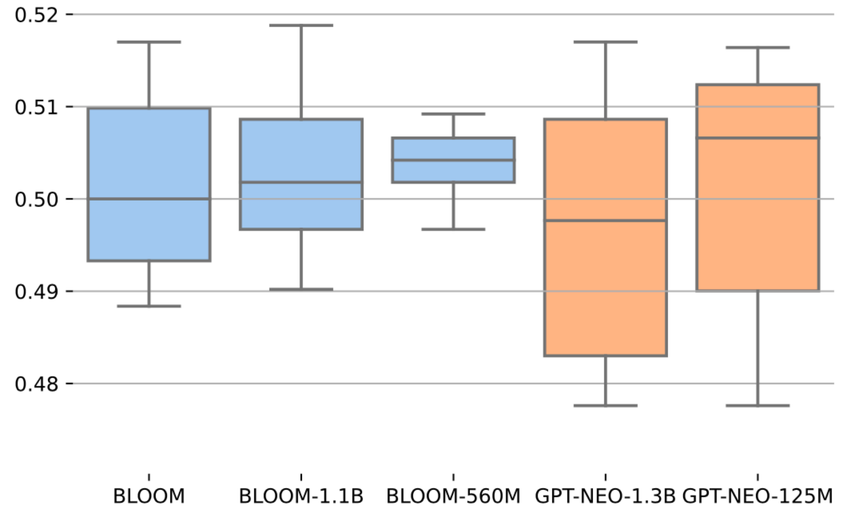
References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(Universal Probing) O. Serikov et al., "Universal and Independent: Multilingual Probing Framework for Exhaustive Model Interpretation and Evaluation", arxiv (2022)

Bias

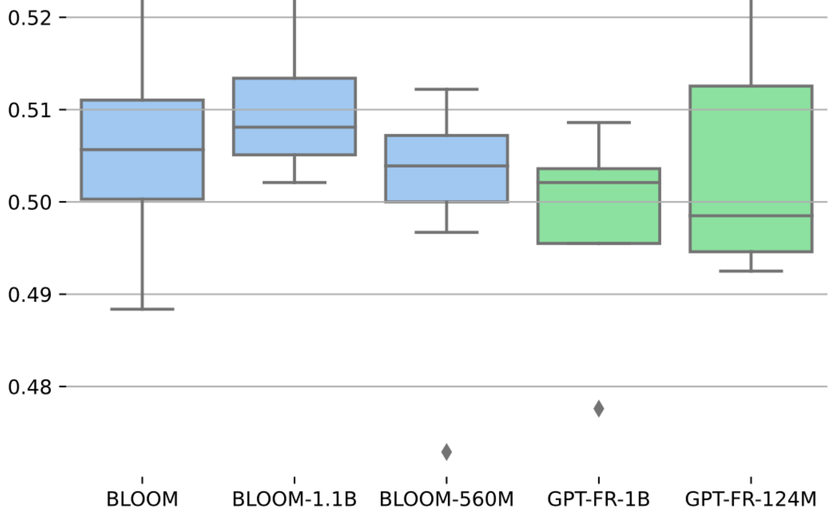
Preliminary study of BLOOM biases | CrowS-Pairs (2020) | French CrowS-Pairs (2022)

Compare stereotyped/non-stereotyped statements: **Women** can't drive | **Men** can't drive

English



French



Accuracy close to 0.5 (suggesting limited bias)

Limitations: Blodgett et al. discuss validity issues with CrowS-Pairs

While promising, examinations do not cover the breadth of possible usage scenarios

References/Image credits:
T. Le Scao et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model", arxiv (2022)
(CrowS-Pairs) N. Nangia et al., "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models", EMNLP (2020)
(French CrowS-Pairs) A. Névéol et al. "French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English", ACL (2022)
(Limitations of CrowS-Pairs) S. L. Blodgett et al., "Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets", ACL (2021)