BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

BigScience Workshop

https://arxiv.org/abs/2211.05100

Presenter: Amirkeivan Mohtashami

All figures in this presentation are adapted from the original paper.

EPFL

Motivation

- Until recently, most LLMs were not publicly released.
- Only a single non-corporate entity outside of China developing large language models.
- Majority of the research community has been excluded from the development of LLMs which has various consequences
  - Hindered prospects for an inclusive, collaborative, and reliable governance of the technology
  - Inflated expectations about its suitability for use
  - Misaligned research and policy priorities with potentially dire consequences
  - Values of the developers emphasized over those of the direct and indirect users

Goal

- Address the above problems and facilitate access to LLMs for research community.
- Train an open-access multilingual LLM with comparable performance to recently developed systems.
- Ensure reproducibility of the training procedure.
- Emphasize inclusivity, diversity, and responsibility.
- Carefully document the whole coordinated process used for development.

What is the problem?

- LLMs are extremely costly to develop and train.
- Many details or essential components are not released or disclosed.
- Computation budget should cover both hyperparameter tuning and the main training.
- Gap between developers and users of the technology particularly apparent in dataset curation (e.g. permission to use data, inclusivity of marginalized population, etc.)

What are the insights?

- Breakdown of the training into various components and forming workgroups for each component:

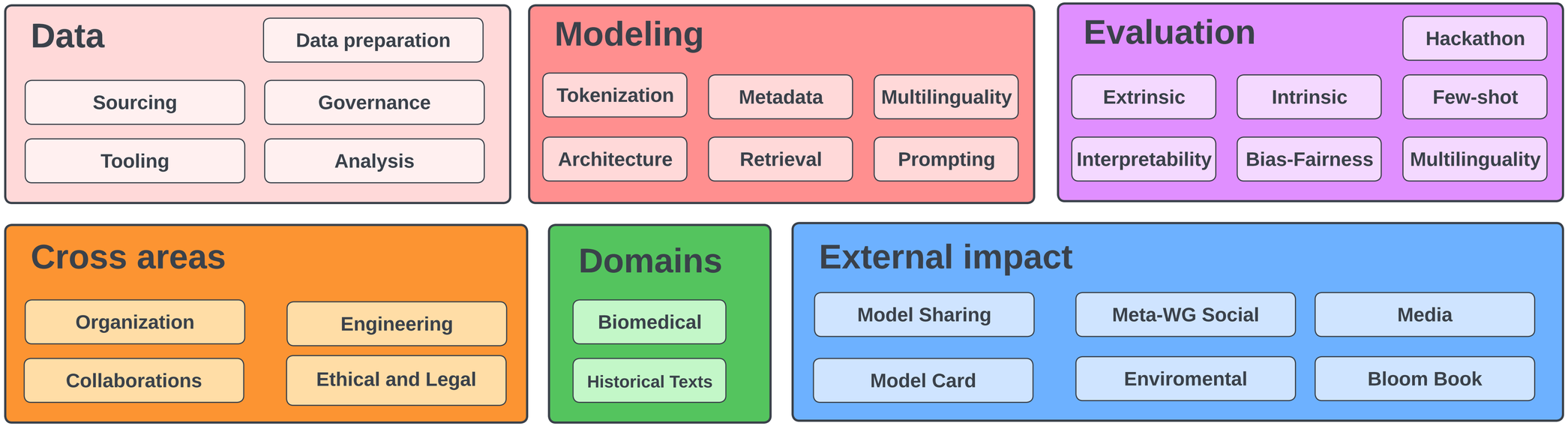

Figure 1: Organization of BigScience working groups.

What are the insights? (Cont.)

- Dataset
  - Abstract filtering of datasets can adversely affect marginalized populations.
  - Early dataset accumulation leads to provenance and authorship of individual items is usually lost.
  - Building on the diversity of the development team, it is possible to find fluent speakers for many languages who volunteer to selecting sources and guiding processing.

What are the insights? (Cont.)

- Training
  - Improving existing architectures has seen relatively little adoption. adopting some of these recommended practices could yield a significantly better model.
  - Better to focus on model families that have been shown to scale well, and that have reasonable support in publicly available tools and codebases.
  - Tokenizer are often neglected in favour of “default” settings but for a multilingual training data, careful design choices are required to ensure that the tokenizer encodes sentences in a lossless manner.

What are the insights? (Cont.)

- Evaluation
  - One of the main draws of LLMs has been their ability to perform tasks in a “zero/few-shot” way.
- Distributed learning
  - Using wrong floating point format can cause numerical instabilities that are known to cause irreversible training divergences.
  - Failures in hardware or even widely used software happen and should be dealt with.

What is the solution?

- Dataset
  - Active outreach in the early stages of the project to invite fluent speakers.
  - Designing a structure for long-term international data governance through various measures such as asking for explicit permission whenever possible and keeping individual sources separate until the final mixing.
  - Gathering and selecting sources by crowd-sourcing the task in the community through hackathons and working groups.
  - Multilingual ROOTS datasets released in two parts: a public part and a sign-up-required part.

What is the solution? (Cont.)

- Training and Evaluation
  - Use smaller models to evaluate architectural decisions
  - Evaluate based on zero-shot generalization.
  - Use decoder-only architecture with causal objective as they performed best.
  - Use ALiBi Positional Embeddings and Embedding LayerNorm.

What is the solution? (Cont.)

- Distributed learning
  - Use mixed-precision training and adapt bfloat16 instead of float16.
  - Allocate spare nodes and checkpoint frequently.

What is the takeaway message?

- Different steps and complexities of training an LLM are described thoroughly.
- Various challenges in different steps are explained in detail.
- Overall can act as a detailed roadmap to train a new LLM.

Will this paper win the test of time award?

- The model is most likely to become obsolete.
- However, the developed and released tools as well as the dataset will remain useful.
- Therefore, in my opinion, yes!

One reason why this paper should have not appeared in top conferences

- The detailed report is interesting for the community and therefore it passes the bar for appearing in a top conference.
- However, there is room for improvement:
  - The paper describes different sections in detail which makes it hard to follow the big picture. Separating the details to appendices and maintaining a more focused main text would have been easier to read.
  - The training dataset uses OSCAR whereas prior work widely used The Pile. In a sub-paper, the results show using The Pile leads to a better performance. It is not clear why this choice has been made.

LLaMA: Open and Efficient Foundation Language Models

Meta AI

http://arxiv.org/abs/2302.13971

Presenter: Amirkeivan Mohtashami

All figures in this presentation are adapted from the original paper.

EPFL

Motivation

- Current large language models are trained with the goal of obtaining a certain performance in the shortest training time.
- Inference time is more important:
  - Deployment at scale
  - Usability and access for the research community

Goal

- Train LLMs to satisfy a limited inference computation budget.
- Use publicly available data to facilitate releasing the model.

What is the problem?

- How to develop LLMs that possibly take longer to train but are faster at inference?
- High performance LLMs for different inference budgets are needed.
- The current LLMs are created to ensure fastest training time.
- Used data must be public so the model can be released.

What are the insights?

- Longer training time of a smaller model might lead to the same performance but allows for faster inference.
- Performance of LLMs keep improving long after reaching the balancing threshold for efficient training.

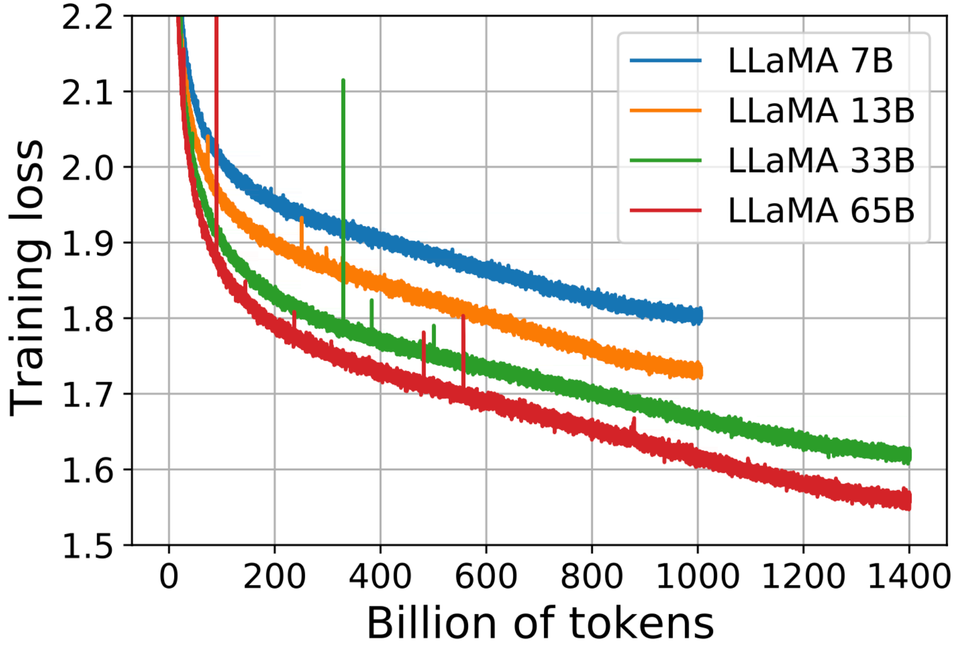

Figure 1: **Training loss over train tokens for the 7B, 13B, 33B, and 65 models.** LLaMA-33B and LLaMA-65B were trained on 1.4T tokens. The smaller models were trained on 1.0T tokens. All models are trained with a batch size of 4M tokens.

What are the insights? (Cont.)

- Various best practices based on observations and techniques that were successfully used to train a known LLM are deployed for dataset gathering and model training, such as:
  - Using SwiGLU activation function [PaLM]
  - Using Pre-normalization [GPT].
  - Using Rotary Embeddings [GPTneo].

What is the solution?

- Using a combination of various public datasets such as CommonCrawl, Github, and Wikipedia.
- Filtering is performed to ensure reliability of the content.
- The final dataset which has 1.4T tokens and allows for long training of the model.
- Trained LLaMa at different inference budgets for longer than typical:
  - 13B parameter model has performance comparable to GPT-3 with 175B parameters.

What is the takeaway message?

- It is important to consider inference budget as well instead of only optimizing for performance in the shortest training time.
- Outline of various steps needed to procure the dataset.
- An updated architecture based on prior work gathered in one place.
- LLaMA 13B can be used as a good accessible alternative to GPT-3.

Will this paper win the test of time award?

- The model is likely to be quickly outdated.
- The paper only provides high level overview of the process of training which limits reproducibility.
- Therefore, in my opinion, no!

One reason why this paper should have not appeared in top conferences

- The paper mainly describes the performance of the trained model.
- The training procedure is only explained from a high level view and details are missing.
- Most insights are taken from prior work.
- No ablation study is done between components adapted from different works.
- Therefore, seems more like a report on a trained model than a research paper.