LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron,* Thibaut Lavril,* Gautier Izacard,* Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave,* Guillaume Lample*

Meta AI

Paper available on Meta Research, not published yet.

LLaMA (Large Language Model Meta AI)

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

- 380 tokens/sec/GPU
- 2048 x A100-80GB
- 1.4T tokens —— 21 days

成本分析

$$\text{cost} \propto \text{computation(FLOPs)}$$

$$\underset{\text{FLOPs}}{\text{computation}} \propto \underset{\text{FLOPs/sec/GPU}}{\text{GPU FLOPS}} \times (\#\text{GPU}) \times \text{time}$$

$$\underset{\text{FLOPs}}{\text{computation}} \propto \underset{\text{FLOPs/token}}{(\#\text{params})} \times (\#\text{tokens})$$

- 训练成本：
  - 参数量 × 语料库大小
- 部署成本（推理成本）：
  - 参数量 × 用户访问量

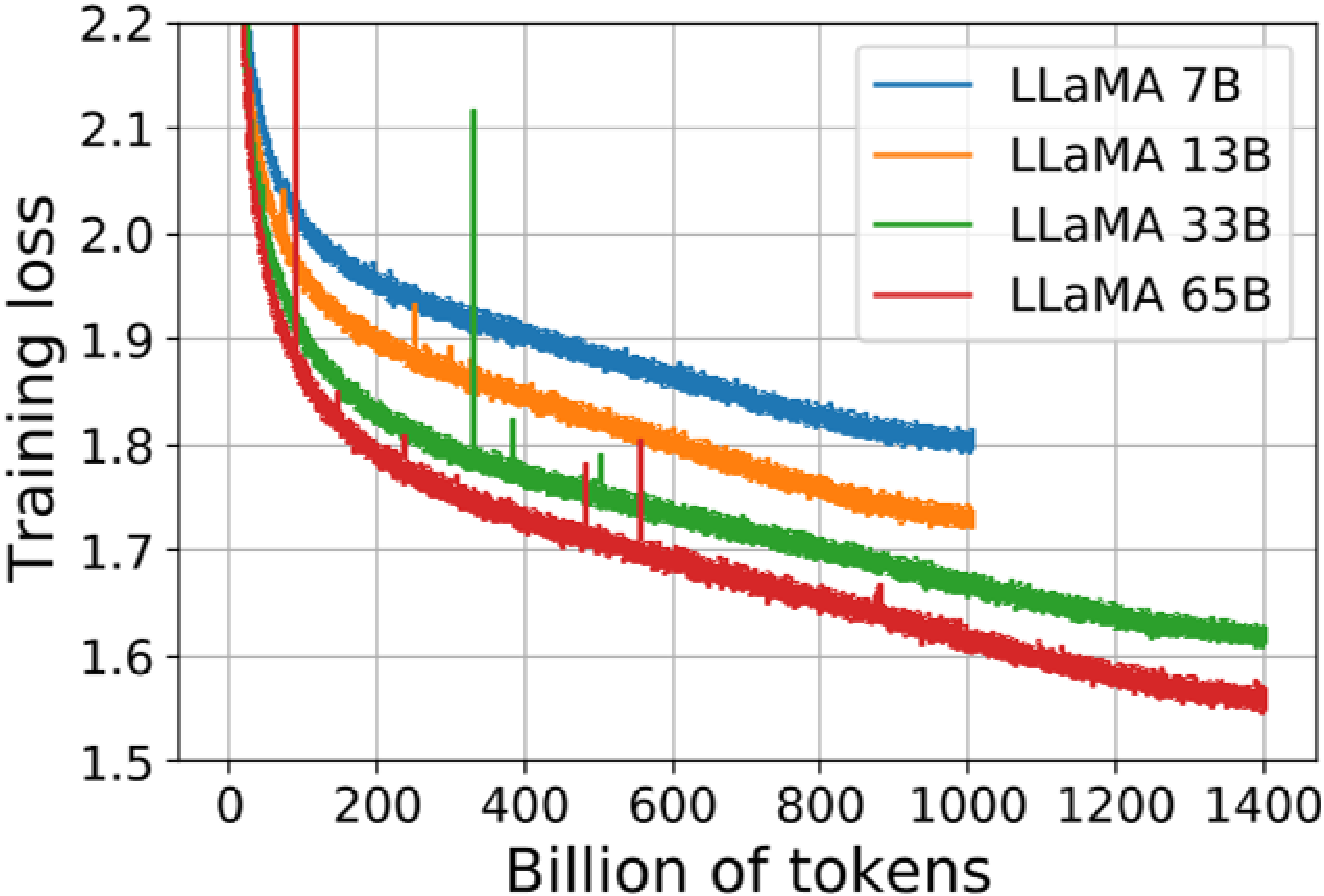
Meta：降低部署成本 $\Rightarrow$ 减少参数量，扩大训练语料库

- "the performance of a 7B model continues to improve even after 1T tokens."

一些大模型参数

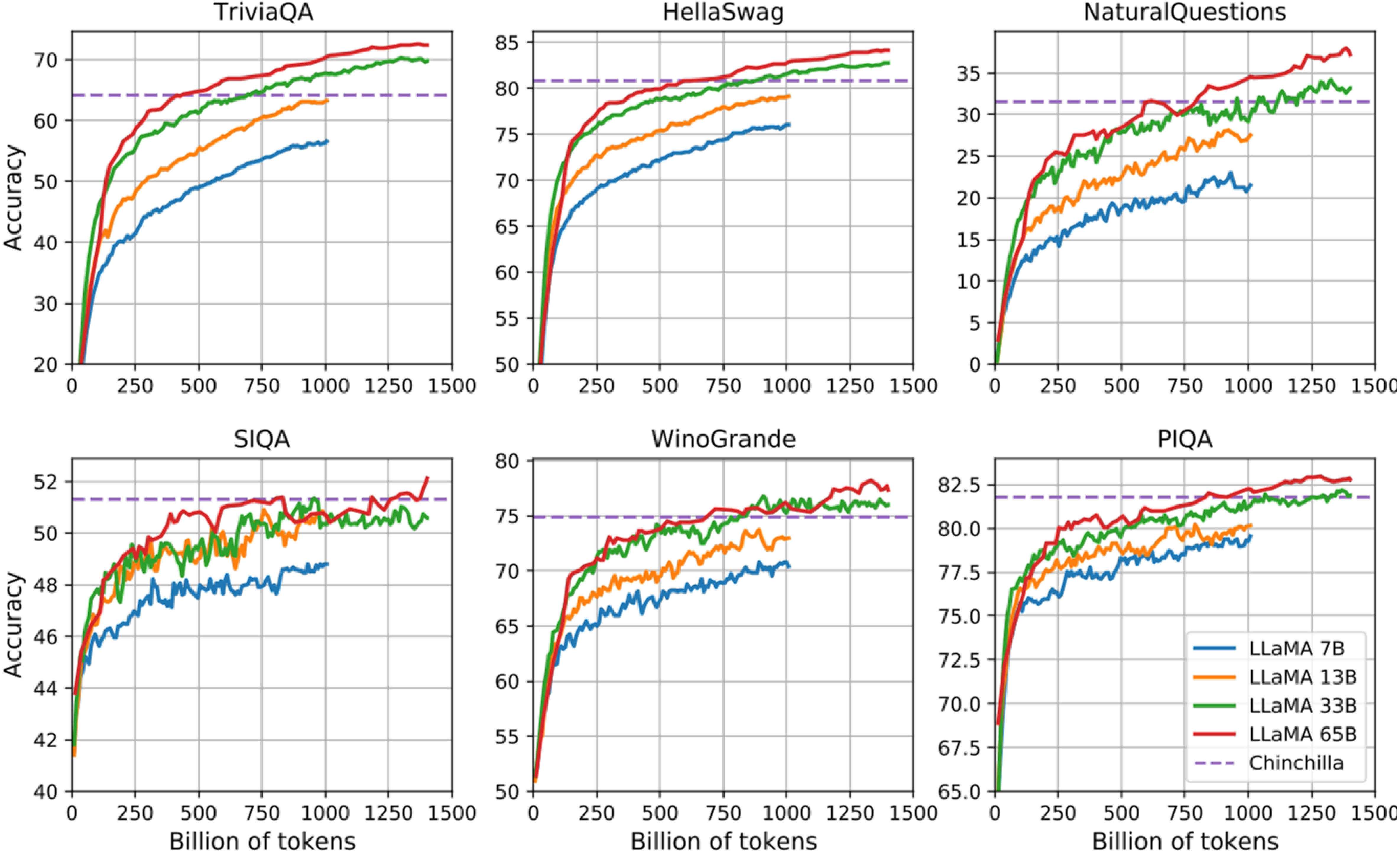
Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
Chinchilla	70 Billion	1.4 Trillion

数据来自：Chinchilla (Training Compute-Optimal Large Language Models, 2022)


LLaMA 7B
LLaMA 13B
LLaMA 33B
LLaMA 65B
Training loss
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
0
200
400
600
800
1000
1200
1400
Billion of tokens

Evaluation

- Common Sense Reasoning:
  - BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, ARC easy and challenge, OpenBookQA
- Closed-book Question Answering:
  - Natural Questions, TriviaQA
- Reading Comprehension:
  - RACE
- Mathematical reasoning:
  - MATH, GSM8k
- Code generation:
  - HumanEval, MBPP
- Massive Multitask Language Understanding(MMLU)


TriviaQA
HellaSwag
NaturalQuestions
SIQA
WinoGrande
PIQA
Accuracy
Accuracy
Billion of tokens
Billion of tokens
Billion of tokens
LLaMA 7B
LLaMA 13B
LLaMA 33B
LLaMA 65B
Chinchilla

Performance on TriviaQA(L) and NaturalQuestions(R)

		0-shot	1-shot	5-shot	64-shot
Gopher	280B	43.5	-	57.0	57.2
Chinchilla	70B	55.4	-	64.1	64.6
LLaMA	7B	50.0	53.4	56.3	57.6
	13B	56.6	60.5	63.1	64.0
	33B	65.1	67.9	69.9	70.4
	65B	**68.2**	**71.6**	**72.6**	**73.0**

		0-shot	1-shot	5-shot	64-shot
GPT-3	175B	14.6	23.0	-	29.9
Gopher	280B	10.1	-	24.5	28.2
Chinchilla	70B	16.6	-	31.5	35.5
PaLM	8B	8.4	10.6	-	14.6
	62B	18.1	26.5	-	27.6
	540B	21.2	29.3	-	39.6
LLaMA	7B	16.8	18.7	22.0	26.1
	13B	20.1	23.4	28.1	31.9
	33B	**24.9**	28.3	32.9	36.0
	65B	23.8	**31.0**	**35.0**	**39.9**

对比Chinchilla

数据集

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

模型架构

transformer+:

- Pre-normalization (GPT3)
- SwiGLU activation function (PaLM)
- Rotary Embeddings (GPTNeo)

代码优化（高效训练）

- 高效的“因果(causal)多头注意力”算子实现，降低内存用量和计算量
  - 不存储掩码覆盖的注意力权重，不计算掩码覆盖的 query/key 值
  - 开源： https://github.com/facebookresearch/xformers
- 手动实现 Transformer 层的反向传播，提升计算性能
  - model and sequence parallelism：通过增加 checkpoint，减少反向传播时的 activation 的重新计算
    - Megatron-LM
  - 尽量并行处理网络通信和 activation 的计算

总结

- 训练了 4 个模型（训练目标为语言模型）
- 达到相同的性能，模型大小缩小 10 倍
- 开源模型和参数，数据集来自开源社区