基座模型训练中的数据与模型架构

颜航*
上海人工智能实验室
yanhang@pjlab.org.cn

费朝烨*
复旦大学计算机学院
zyfei20@fudan.edu.cn

杨小珪
复旦大学计算机学院
yangxg21@m.fudan.edu.cn

高扬
上海人工智能实验室
gaoyang@pjlab.org.cn

邱锡鹏
复旦大学计算机学院
xpqiu@fudan.edu.cn

摘要

ChatGPT以对话形式的交互方式，降低了使用大模型的门槛，因此迅速在全球范围内流行起来。尽管OpenAI并未公开ChatGPT的技术路线，但一些后续的工作宣称已经在开源的基座模型上复现了ChatGPT的性能。然而，尽管这些模型在某些评测上表现出与ChatGPT相似的性能，但在实际的知识量和推理能力上，它们仍然不如ChatGPT。为了更接近ChatGPT甚至GPT4的性能，我们需要对基座模型的训练进行更深入的研究。本文针对基座模型训练的数据以及模型架构进行讨论，首先总结了当前预训练数据的来源以及基本处理流程，并针对目前关注较少的代码预训练数据和中文预训练数据进行了分析；然后对当前已有基座模型的网络架构进行了回顾，并针对这些架构调整背后的动机进行了阐述。

关键词： 基座模型数据 ； 基座模型架构

1 引言

从2022年11月底，美国OpenAI公司推出ChatGPT[1]后，大语言模型（Large Language Model，简称LLM）在学术界和工业界都引起了轰动。ChatGPT可以通过对话的形式完成各种任务，例如撰写代码、整理数据、润色论文等，并且当其没有输出预期结果时，还可以通过多轮对话逐步优化自身输出，这种通过对话交互的方式极大降低了模型的使用门槛，因此ChatGPT迅速在全球范围内出圈，热度扩散到了人工智能领域之外。OpenAI公司并未公开ChatGPT的技术路线，但他们在InstructGPT (Ouyang et al., 2022)论文中提到，可以通过一个基座语言模型，结合人类对齐（Human Alignment）训练来让模型跟随人类的指令完成特定任务。使用InstructGPT中类似的方法，后续的一些工作在开源的基座模型 (Nijkamp et al., 2023b; Touvron et al., 2023)上一定程度复现了ChatGPT的性能 (Sun et al., 2023; Taori et al., 2023)。不过最近来自美国伯克利大学的研究指出，尽管现在这些模型从一些评测上展现出与ChatGPT相似的性能，但从知识量及推理能力方面，它们均不及ChatGPT，只是由于回答形式上接近ChatGPT，才获得了不错的评测性能。为了能够更加接近ChatGPT乃至GPT4的性能，还需要在基座模型的训练上进行更深入的研究 (Gudibande et al., 2023)。

为了得到一个好的基座模型，我们首先需要大量的预训练数据，在图1中我们对比了近年来不同预训练模型的大小与使用的预训练数据大小。从GPT-2(Radford et al., 2019)到PaLM2(Anil et al., 2023)，模型的大小增长了200倍，但是预训练的数据量大小增长了450倍。因此，预训练模型对数据的需求是巨大的。表1中罗列了几个基座模型训练所需要的计算资源。除了对预训练数据进行总结之外，我们也将对比不同的基座模型的网络架构，并对各种基座模型的架构进行归纳。

2 预训练数据

数据是知识的载体，也是大规模预训练模型训练的基础，自深度学习技术发展以来，数据就是决定模型性能的重要因素。对于基座模型的训练而言，预训练数据在一定程度上决定了其能力的边界。Kaplan等人(2020)研究发现，训练基座模型的过程中，模型大小与数据规模是决

* 共同一作.

[1] https://chat.openai.com

第二十二届中国计算语言学大会论文集，第1页-第15页，哈尔滨，中国，2023年8月3日至5日。
卷2：前沿综述

模型	训练Token数	计算资源	训练时长
GPT3 (Brown et al., 2020)	300B	10000张V100	14天
GLM-130B (Zeng et al., 2023)	450B	992张A100	60天
LLaMA-65B (Touvron et al., 2023)	1.4T	2048张A100	21天
MPT-7B (Team, 2023)	1T	440张A100	10天

Table 1: 基座模型对训练资源需求

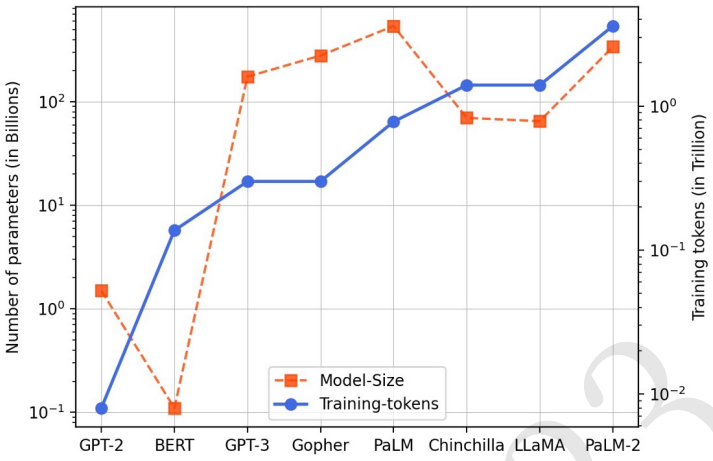

Figure 1: 近年的预训练模型大小以及它们使用的预训练数据大小

定预训练模型的关键因素，这也是GPT-3(Brown et al., 2020)乃至后面ChatGPT成功的理论基础。Hoffmann等人(2022)进行了类似的探究，发现在给定的计算量条件下，所需要的数据量相比较于Kaplan等人预计的要更多，并基于此训练了性能更好的Chichilla模型。本章将从数据入手，总结当前预训练数据的处理流程，并对其问题展开讨论。

2.1 预训练数据的来源

预训练数据的来源直接关系到语料的多样性。不同来源的语料往往是不同的主题，不同的格式以及不同的组织方式，补充不同来源的数据可以很好地提升模型的鲁棒性与泛化性(Longpre et al., 2023)。目前，大规模预训练语料主要来源于互联网中的文本信息。

互联网作为人类信息交换的重要方式，积累了巨量的信息。根据谷歌公司CEO Eric Schmidt预计，整个互联网数据数量高达5000 PB[2]。由于总量巨大且居于此的信息时刻在发生改变，因此如何获取这一份数据是一项巨大的挑战。受益于互联网爬虫计划Common Crawl [3]的开展，研究人员可以更加便利地收集网络中的数据，并将精力集中于数据处理阶段。Common Crawl是一项开放网络爬虫的数据存储库的开源项目，其爬取并保存了自2013年以来开放互联网中的各种数据，为研究人员提供了一个海量、非结构化、多语言的网页数据集。Common Crawl可自动爬取整个互联网上的数据，并采用时间作为刻度，每隔一段时间将会放出一部分数据集，不同时间片的URL以及内容尽量保证不重复。每一个时间片大概存有1.5 Billon个文档，包含该时间段内互联网中更新的绝大部分内容。

Common Crawl 数据集具有总量巨大和来源渠道多样化等特点，但正是这些特点导致从中提取高质量文本变得异常困难。因此，之前的几份工作均引入更多的预处理数据，或针对指定网站的数据进行定向收集。例如，在Brown等人(2020)提到，除了预处理Common Crawl的数据外，他们同样添加了高质量图书数据集以及维基百科等。Pile语料 (Gao et al., 2021)除了Common Crawl收集的数据外，还收集了将近21个站点的数据，其中包括学术论文网站PubMed、Arxiv，代码共享平台Github，编程交流平台Stack Exchange，预处理图书数据集BookCorpus 2、Books3以及其他相关高质量数据集等。悟道 (Yuan et al., 2021)同样的采用

[2] https://www.easytechjunkie.com/how-big-is-the-internet.htm
[3] https://commoncrawl.org/

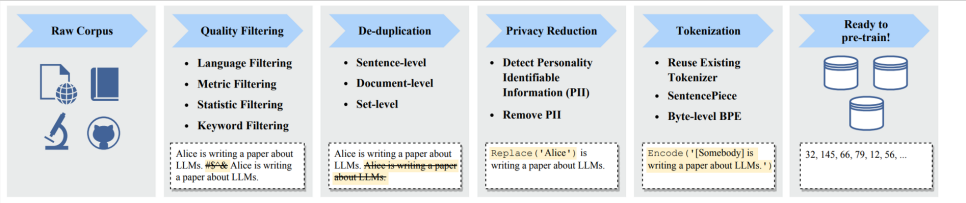

Figure 2: 预训练数据处理流程图示(Zhao et al., 2023)

了大量的网络语料，与前面不同的是，其主要筛选中文语料，并丢弃中文字符少于10个的网页。

此外，经过我们的调查发现，Pile(Gao et al., 2021)作为处理质量较高的数据集在开源后被广泛使用，后续的一系列工作倾向于在预训练语料中加入Pile或者Pile CC，以节省数据处理的时间，同时加入部分更新时间戳下的Common Crawl语料，以保证模型的质量与时效性。

2.2 预训练数据的处理

一般而言，从网页中获取的数据往往不能直接使用。一方面，这一类数据以HTML的格式组织在一起，一般的大语言模型无法通过训练从中提取纯文本信息；另一方面，预训练数据中往往充斥着各种虚假信息，模版信息，广告信息以及黄色暴力内容等，此外还有自动生成的信息，这一类信息我们统称为脏数据。脏数据的引入不利于语言模型对于语言建模任务的学习。因此，目前的主流模型通常会对收集到的网络数据进行处理，以使模型可以更好地学习其语言内部的分布。

如图2所示，预训练数据的处理流程主要分为四部分：质量筛选，数据去重，隐私信息删除与文本词元化。如果获取到的信息为网页信息，还需要对网页信息进行提取，尽量提取纯净的内容数据而避免模版数据。为此，除了采用一些网页信息抽取工具来对Common Crawl的信息进行抽取之外，Common Crawl本身同样提供了纯文本的.WET格式数据。在处理预训练语料的流程中，删除隐私信息可以避免大语言模型泄露隐私，文本词元化将文本转化为token，以作为预训练模型的输入。而数据质量筛选与数据去重将在内容层面直接关系预训练数据的属性与性质，从而影响基座模型的训练。本节将着重介绍数据质量筛选与数据去重，以及目前主流的处理方法。

2.2.1 质量筛选

预训练数据的质量直接关系到大语言模型的性能，而对于数据的质量筛选直接关系到预训练语言模型的性能。目前主流预训练模型采用的质量筛选方案包含两种，分别为规则过滤及训练分类器方法。

规则过滤方法，其最简单的方式是通过URL筛选数据。但由于数据量过于庞大，我们无法遍历所有的URL进行筛选，因此研究人员提出了多种自动化的方法以解决此类问题：

启发式规则过滤方法。人为设计一部分启发式规则，直接对文本进行筛选。C4即采用了这种方法，通过大量启发式规则对数据筛选，包括删除所有非停止符号结尾的段落，根据不良单词列表删除文档，删除过短的语句，删除带有“javascript”字样的段落等 (Raffel et al., 2020)。这一类筛选规则虽可以筛选出一些不流畅的语句以及不良语句，但无法避免数据重复或者广告等问题。为了筛选出重复性的数据，在MassiveText 语料构建过程中，研究人员提出通过计算不同gram的重复性 (Rae et al., 2021)，以及选择不同的筛选阈值来筛选重复性数据。

基于模型的数据过滤方法。之前的工作通常旨在减少直接的脏数据，例如数据中不符合人类规范的，重复多次的数据，带有不良词汇的数据，更多的脏数据可能是混入了HTML模版，机器自动生成的看起来是文字实际是乱码的数据。为了筛选掉这部分脏数据，主流的预训练方法提出引入基于模型的数据过滤方法。例如，GPT-3通过Wiki等高质量数据集训练了一个简单的分类器，并通过分类器对数据进行筛选，LLaMA等工作训练了一个KenLM的小型统计语言模型，用以筛选出脏数据。

2.2.2 数据去重

在网页数据中，数据可能会存在重复，Lee等人(2022)主张通过n-gram结合MinHash的方式

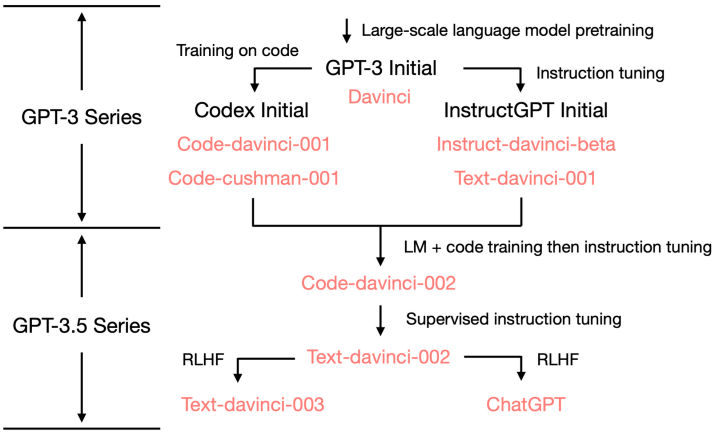

Figure 3: 符尧(2022)等人整理的OpenAI模型进化图

来删除数据中的重复数据，删掉重复数据可以让训练更加有效。Hernandez等人(2022)指出如果训练数据中存在10%的重复数据，将导致模型的有效大小减半，即400M参数量的模型其效果仅相当于一个200M参数量的模型。除此之外，PaLM模型(Chowdhery et al., 2022)也在论文中提到在新数据上训练将更有可能让模型获得更好的性能。尽管数据去重已经成为了基座训练中的标准一环(Zhang et al., 2022; Zeng et al., 2023; Scao et al., 2022; Touvron et al., 2023)，但也有研究表明适当的重复使用数据不会对模型的性能造成很严重的问题，在T5论文(Raffel et al., 2020)中，作者尝试了重复使用数据，模型的性能并没有受到特别严重的损害，不过由于作者使用的模型是编码器-解码器架构，其规律可能与语言模型的规律不一致。Muennighoff等人(2023)在9B参数量的语言模型上测试了重复数据对模型的影响，他们发现数据重复量只要在4次以内，就几乎不会对模型造成性能损害。综上所述，数据去重应避免留下重复过多次的数据，但如果重复量控制在4次以内，对训练产生的影响应该都不致命。

2.3 代码数据的引入

传统的大语言模型对于代码的关注度并不高，而对于训练纯文本模型十分感兴趣。OpenAI的CodeX (Chen et al., 2021)模型中首次将代码引入大模型预训练中。根据符尧等人(2022)的观察，引入代码，使得OpenAI的GPT系列模型有了重大性能突变。不过目前还没有非常直接的证据可证实这个猜想，但由于让语言模型能够生成代码本身也是一种非常重要的特性，且该方式也提供了一种未来大模型与现实世界发生交互的接口，因而将代码数据加入到预训练中也非常有必要。

在之前的模型训练中，代码数据的处理均较为粗粒度 (Touvron et al., 2023; Nijkamp et al., 2023b)，他们都只在代码文件粒度进行了去重，但实际上由于各种脚手架代码不断重复，代码数据中存在非常严重的重复问题。为了更好地使用代码数据，Li等人 (2023)发现，对于代码的精细化处理，可以使模型更好地学习代码中蕴含的能力与知识。例如，其筛选了一部分的Github中的数据，并对其分语言进行处理，处理了Jupyter Notebook，使其更符合人类语言的特点等，诸如此类的操作使得他们的模型在代码任务上的性能得到了显著提高。进一步地，Gunasekar等人(2023)认为大部分的代码是无用且低效的，他们提出，一个好的代码数据集应该是清晰的、独立的、有启发性的和平衡的。由此，他们对The Stack数据集 (Kocetkov et al., 2022)进行了过滤，同时采用ChatGPT生成了一部分数据，使其具有更强的启发性。

2.4 中文预训练数据

尽管之前已有部分工作开源了部分中文语料，例如Wudao (Yuan et al., 2021)、Yuan 1.0T (Wu et al., 2021)，但这部分数据一方面是时效性不足，另一方面是数量和质量都还不足以训练一个性能较好的基座模型，因此需要获取更多的中文语料。受到Pile语料的启发，我们也对中文语料进行了分类，并尝试寻找对应的数据来源，结果如表2所示。

我们对中文互联网上的数据进行了粗略的估计，结果如表3中所示，我们发现，目前可获取的中文token不足1T，对于目前动辄上T token级别的预训练，中文数据还是非常缺乏。除此

数据类别	英文来源	中文对照数据来源
Pile-CC	https://commoncrawl.org/the-data/	WuDao、CC中抽取中文数据
PubMed Central	https://pubmed.ncbi.nlm.nih.gov/	各类医学网站、开源数据库
Books3	https://bibliotik.me	豆瓣阅读、Kindle电子书、sobooks
OpenWebText2	https://www.reddit.com/	微博、百度贴吧、小红书
ArXiv	https://arxiv.org/	
Github	https://github.com/	
FreeLaw	https://www.freelaw.in/	中国法院网、北大法宝、威科先行法律信息库
Stack Exchange	https://stackexchange.com/	中文问答社区、百度知道、头条问答
USPTO Backgrounds	https://www.uspto.gov/	中国专利信息网、壹专利、开源专利数据库
PubMed Abstracts	https://pubmed.ncbi.nlm.nih.gov/	各类医学网站、开源数据库
EuroParl	https://www.statmt.org/europarl/	翻译数据
Gutenberg (PG-19)	https://www.gutenberg.org/	古典文学网、中国古典文学、古书房
OpenSubtitles	https://www.opensubtitles.org/en	字幕库、SubHD、诸神字幕组
Wikipedia(en)	https://www.wikipedia.org/	中文维基百科、百度百科、MBA智库百科
DM Mathematics	https://github.com/deepmind/mathematics_dataset	作业帮、各类数学题网站
Ubuntu IRC	https://webchat.freenode.net/	微博超话在线聊天、各类聊天机器人、多轮对话数据集
BookCorpus2	https://www.smashwords.com/	起点中文网、纵横中文网
HackerNews	https://news.ycombinator.com/	36氪、极客公园、虎嗅网
Youtube Subtitles	https://www.kaggle.com/datasets/wadzim/youtube-subtitles	Subscene网站
PhilPapers	https://philpapers.org/	哲学中国网、学术·哲学_爱思想、中国哲学书电子化计划
NIH Grant Abstracts	https://reporter.nih.gov/exporter/abstracts	财政部公开信息
Enron Emails	https://www.cs.cmu.edu/~enron/	互联网中免费共享的电子邮件数据库

Table 2: Pile中的各类语料对应的中文来源

之外，处理难度较大也是困扰中文大语言模型训练的一大难题。目前主流的数据清洗代码大都支持英文而不支持中文，需重新适配，例如上述提到的启发式规则，对于中文数据需要重新编写。中文数据中的广告处理相对于英文数据也存在难点，中文数据中，广告通常存在于语句内部，比较难将其筛选出来。

来源	Token数量
古文诗词	0.8B
百科	5B
各类小说	120B
社区问答	200B
新闻	100B
中文专利	9.5B
法律判决	90B
博客	64B
学术论文	9.5B
总计	598.8B

Table 3: 中文不同类型数据Token数量粗略统计

2.5 预训练语料质量的评估

由于大语言模型的训练成本巨大，对于处理好的预训练语料，研究人员希望对其进行质量评估。Gopher(Rae et al., 2021)论文中采用训练一个1.4B左右的语言模型的方式来评估不同处理流程对于下游任务的影响。他们发现，随着数据清洗与处理的不断深入，经由这些数据训练得到的模型的效果也越好。在经过质量筛选与数据去重之后，得到的数据训练的模型相比于原始数据与开源数据（OpenWebText与C4）在语言建模任务上性能存在明显的上升。采用同样的方法，Longpre等人(2023)在1.5B模型的基础上，研究了不同质量的数据对于大语言模型在处理不同下游任务上的性能差异。研究人员不仅希望评估语料质量对于模型训练的影响，更希望探究不同质量语料对于模型不同维度能力激发之间的差异。

这种采用小模型进行验证的方法成本较低，且可以在正式训练之前发现数据存在的诸多问题，以及对于基座模型的训练进行预测。但由于大语言模型涌现现象的存在，可能在某些能力上，模型需要达到一定量级才可以看出性能的差异。另外，生成式模型的能力评估也十分困难，由于Prompt的引入，评测本身也带有一定的不确定性，表4中展示了在评测过程中，如果使用不同的Prompt进行评测，得到的结论会大相径庭。因此，如何低成本地评估预训练语料的质量与不同数据对于模型能力的影响，仍然是自然语言处理社区活跃的研究问题。

训练Token数	验证集损失	Prompt #1	Prompt #2
20B	1.335	66.5	71.55
40B	1.334	65.6	73.88
60B	1.329	64.2	74.37
80B	1.328	64.8	76.17
100B	1.324	64.7	77.09

Table 4: 随着训练的进行，验证集损失在不断下降。如果使用Prompt #1的结果作为判断，模型的性能在变差，但如果使用Prompt #2的结果，则性能在变好。

2.6 开源预训练数据集

随着大语言模型在自然语言处理领域的广泛应用，高质量的开源模型与预训练数据集的需求迅速增长。正如Together公司[4]宣称的那样：“AI正在迎来Linux时代”，高质量大规模的开源

[4] https://together.ai/blog/redpajama

模型	位置编码	自注意力	归一化层位置	归一化层类型	损失函数
GPT3-175B	Absolute	Standard	PreNorm	LayerNorm	LM(+FIM) Loss
CodeGen-16B	RoPE	Standard	ParallelLayer	LayerNorm	LM Loss
PaLM-540B	RoPE	Multi-Query	ParallelLayer	LayerNorm	LM(+UL2) Loss
OPT-175B	Absolute	Standard	PreNorm	LayerNorm	LM Loss
GLM-130B	RoPE	Standard	PostNorm	LayerNorm	GLM+LM Loss
BLOOM-176B	ALiBi	Standard	PreNorm	LayerNorm	LM Loss
LLaMA	RoPE	Standard	PreNorm	RMSNorm	LM Loss
CodeGen2-16B	RoPE	Standard	ParallelLayer	LayerNorm	LM Loss
MPT-7B	ALiBi	Standard	PreNorm	LayerNorm	LM Loss
StarCoder	Absolute	Multi-Query	PreNorm	LayerNorm	LM+FIM Loss
Falcon	RoPE	Multi-Query	ParallelLayer	LayerNorm	LM Loss
ChatGLM2-6B	RoPE	Multi-Query	PreNorm	RMSNorm	-

Table 5: 各种基座模型

预训练数据集已经成为自然语言处理领域重要的基础设施与关键资源。2019年，Google开源了用于训练T5的C4数据集 (Raffel et al., 2020)，该数据集从Common Crawl中提取并对其进行了抽取与启发式的质量筛选。2020年，EleutherAI开源了Pile数据集(Gao et al., 2021)，该数据集不仅处理了Common Crawl的部分数据分片，还获取了部分高质量英文数据用以提升预训练数据的质量与多样性。此后，一些开源模型（如OPT (Zhang et al., 2022)和GLM-130B (Zeng et al., 2023)）均采用此数据集进行基座模型的训练。

受到LLaMA模型 (Touvron et al., 2023)的启发，Together公司处理并开源了一份大约3TB的预训练数据集RedPajama(Computer, 2023)。RedPajama采用了CCNet流水线处理方式 (Wenzek et al., 2020)，涵盖了2017年至2020年间的Common Crawl数据分片，并补充了如C4、Wikipedia等高质量开源数据集。借助RedPajama数据集，Together公司训练出了3B至7B参数规模的完全开源模型[5]。随后，在RePajama数据集的基础上，Together公司采用了更严格的数据处理方法得到了一个约600B Token的更高质量的SlimPajama数据集 (Soboleva et al., 2023)。此外，Falcon组织整理了2008年至2023年初的所有Common Crawl数据分片，并形成了一个约5T tokens的RefinedWeb数据集(Penedo et al., 2023)，其中600B数据子集可以公开获取到[6]。

此外，代码数据集也引起了NLP领域的广泛关注。开源社区组织的BigCode项目对预训练所需的代码数据进行了深入思考，他们收集并处理了网络中的开源代码库，对特殊代码，如Jupyter，Github issue等进行了特殊处理，并最终开源了大小达到6TB、包含358种编程语言的开源代码预训练数据集The Stack(Li et al., 2023)。

OpenLLaMA项目尝试了使用RedPajama、RefinedWeb和The Stack数据进行预训练，发现结合这三类数据可以取得很好的预训练效果 (Geng and Liu, 2023)。

3 基座模型架构

在这一节中，我们主要对当前的基座模型的模型架构进行讨论。在表5中对最近的基座模型的几个重要部件进行汇总，然后针对不同的部件进行简要介绍。

3.1 位置编码

位置编码从大类上来说可以分成绝对位置编码(Vaswani et al., 2017)和相对位置编码(Shaw et al., 2018; Su et al., 2021)两类。其中绝对位置编码一般有两类，正余弦函数的位置编码和可学习的位置编码，其中可学习的位置编码在之前的预训练模型中被广泛采用，例如BERT(Devlin et al., 2019)、RoBERTa(Liu et al., 2019)和GPT2(Radford et al., 2019)等，Wang (2020)将这几个预训练模型的位置编码两两位置计算了相似度，其结果如图4所示，可以看出不同位置

[5] https://huggingface.co/togethercomputer/RedPajama-INCITE-7B-Base
[6] https://huggingface.co/datasets/tiiuae/falcon-refinedweb

有一定的邻域关系，越靠近的两个位置，相似度也会越大，因此在Transformer中引入相对位置编码来体现这种归纳偏执应该是有益的。Shaw等人(2018)提出了相对位置编码的概念，苏剑林等人 (2021)从虚数的角度出发推导了乘性位置编码RoPE，这种位置编码提出之后便被广泛使用到了基座模型训练之中。尽管RoPE可以高效地表示相对位置，但是它的外推能力较差，即模型只能在训练数据长度以内的数据表现得不错，超出这个长度性能便大幅下降，如图5所示。ALiBi (Press et al., 2022)首先提出了相对位置编码应该具备良好外推性的概念，通过使用ALiBi可以实现在较短的训练语料上训练但在较长的语料上测试，使用ALiBi位置编码的MPT模型 (Team, 2023)甚至可以支持到65,000个词元的输入。

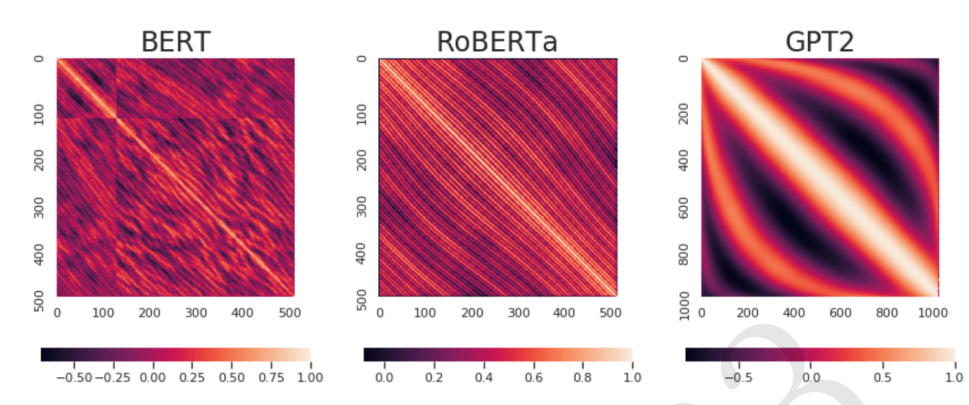

Figure 4: 不同预训练模型位置嵌入中两两位置的相似度

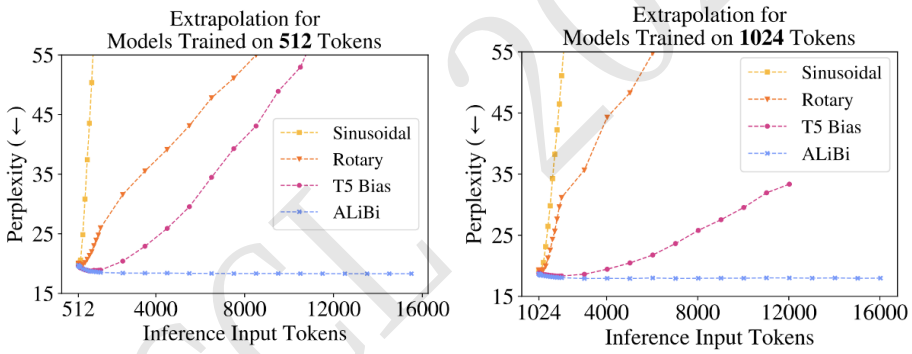

Figure 5: 不同位置编码在长度外推时的性能表现，ALiBi位置编码可以在训练时没有见过的长度上取得良好的性能

3.2 自注意力

自注意力是Transformer模型的核心模块，但其计算复杂度为$O(L^2)$，导致了其计算效率随着输入长度的增加二次增长，之前的工作提出了很多高效自注意力机制(Tay et al., 2023a)，例如Longformer(Beltagy et al., 2020)、Linformer(Wang et al., 2020)等。但在基座模型中这一类直接改良计算复杂度的工作没有被广泛采用，可能的原因有两个，第一个是这类高效Transformer一般都需要引入稀疏计算，这导致它们的计算可能不是GPU友好的，而预训练又是对计算效率非常敏感的任务；第二个原因是现在的预训练模型的隐藏层维度一般都是好几千以上，而目前的基座模型上下文多是两千左右，因此数据长度带来的二次增长，实际上并没有主导模型的计算量。

在基座模型中采用较多的用来降低计算量的方法是Multi-Query方法，Multi-Query方法通过在不同注意力头之间共享自注意力中的Key值和Value值减少显存占用，Multi-Query如图6所示。由于现在的基座模型大部分都是从左到右生成的语言模型，因此在生成过程中，需要使用KV缓存来缓存前面词元的Key值和Value值，在Multi-Query场景下，由于不同注意力头共享

了Key值和Value值，因此可以只保留一份缓存。这种方式可以极大地减少在推理过程中的显存占用，我们以65B模型为例，在图7中展现了在推理过程中Multi-Query的显存占用和常规自注意力的对比。随着输入长度的增加，常规自注意力机制的显存占用增长显著高于Multi-Query方案，因此未来如果想要将基座模型的输出长度扩充到更大长度，Multi-Query方法值得尝试。

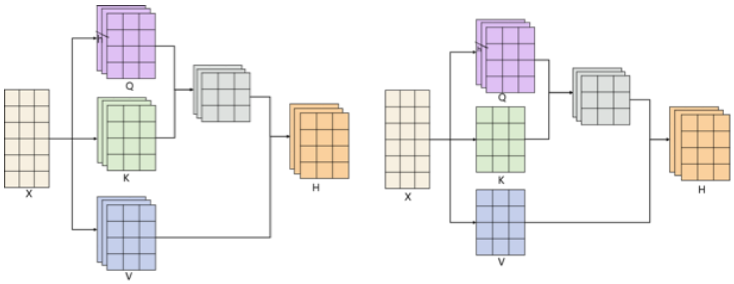

Figure 6: 常规自注意力和Multi-Query自注意力

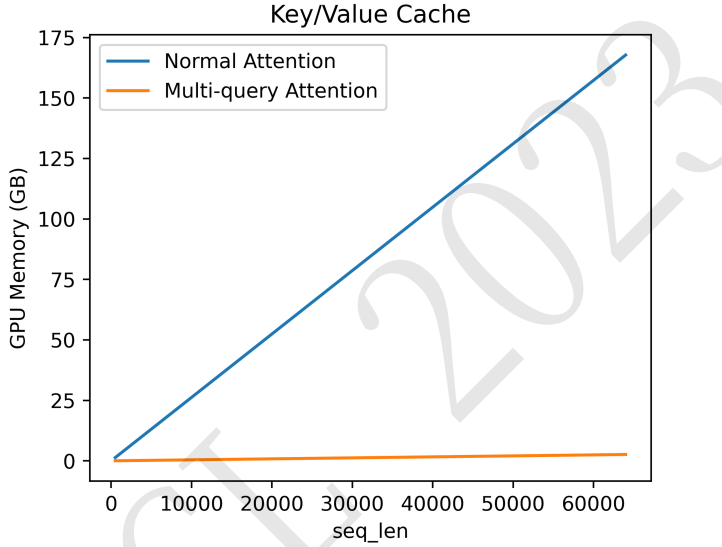

Figure 7: 常规自注意力和Multi-Query自注意力在推理时显存占用的对比

3.3 归一化层

不同的基座模型一方面在归一化层的位置上有所区别，另一方面也会使用不同的归一化层类型。目前基座模型使用的归一化层位置主要有三类，这三类的区别如图8所示。PreNorm在语言模型中被广泛采用(Radford et al., 2019; Brown et al., 2020)；PostNorm则是在BERT(Devlin et al., 2019)、RoBERTa(Liu et al., 2019)等编码器模型中采用较多；Parallel Layer由于将自注意力计算和前馈神经网络并行计算，因此在计算的时候，可以将两个模块的计算同步进行，这样可以让训练的速度更快。归一化层的类型在LLaMA模型发布之前大部分模型都采用的是LayerNorm层归一化，而在LLaMA之后，类似于ChatGLM2[7]、Baichuan[8]等都开始尝试RMSNorm (Zhang and Sennrich, 2019)的方案。

3.4 损失函数

目前大部分的基座模型都使用了语言模型损失作为优化的损失函数，但也有一些其他的尝试。

[7] https://github.com/THUDM/ChatGLM2-6B
[8] https://github.com/baichuan-inc/Baichuan-7B

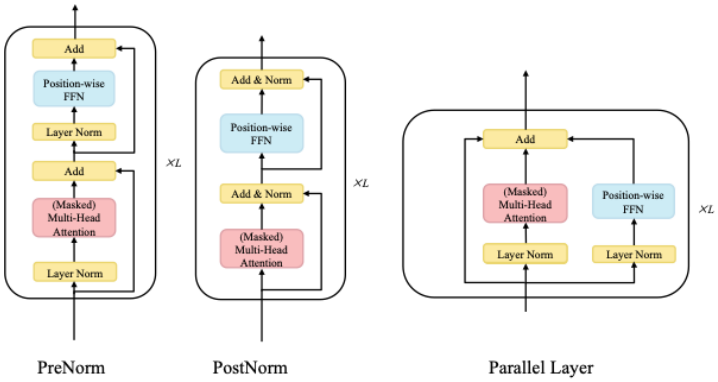

Figure 8: 三种不同的归一化层位置示例

在代码补全场景中存在给定上下文补齐中间代码的场景，因此有研究人员提出FIM（Fill in the Middle）损失函数(Bavarian et al., 2022)，该损失函数在训练的时候会将正常文本进行打乱，首先将文档随机分成三段，例如“document: (prefix, middle, suffix)”，之后在训练阶段将输入顺序调整为“(prefix, suffix, middle)”，这样即可让语言模型学会通过上下文预测中间缺失的地方。尽管这种训练损失函数和语言模型不一致，但加入这种损失并未影响语言模型的正常训练。不过近期也有论文指出，加入FIM损失函数会导致语言模型性能下降(Nijkamp et al., 2023a)。

Tay等人 (2023b)融合了编码器、解码器以及编码器-解码器预训练模型的损失函数，提出了一种融合的损失函数，这种损失函数包含了GPT (Radford et al., 2019)、BERT(Devlin et al., 2019)、T5(Raffel et al., 2020)、UniLM(Bao et al., 2020)等各种预训练模型的损失函数。通过这种损失函数的设计，作者发现可以训练得到性能最好的预训练模型。同时，在后续的研究中，作者发现如果将这种损失函数应用到一个已经训练过的语言模型上，可使得该语言模型仅需少量训练便大幅提升下游性能 (Tay et al., 2022)。

3.5 新的架构

除了基于Transformer的尝试，还有一些其他架构的尝试。例如基于状态空间模型（State Space Model，简称SSM）的模型(Gu et al., 2022; Dao et al., 2023)，这一类模型提出的目标是实现超长文本输入，仅需卷积神经网络量级的计算复杂度即可完成训练（避免了Transformer的二次计算复杂度），而在推理阶段只需固定的计算量，类似于循环神经网络（Transformer的计算量随着长度的增加而增加）。从数学上和实际效果上，状态空间模型在拟合长序列方面确实较Transformer有优势，但目前缺乏更大规模的模型验证，不确定能否成为新的基座模型。除了状态空间模型外，Peng等人(2023)提出了RWKV的方案，RWKV结合了RNN的思想，在Transformer状态更新的时候会融合上一个时刻的状态，在注意力计算过程中，模型将不再进行Query值和Key值的内积计算，而是设计为一个随着距离衰减的函数计算，从而在推理阶段，模型可将过去所有时刻的状态进行累加，无需进行类似于Transformer将Query值与过去的每个Key值做内积的计算，实现了推理时常数级计算复杂度。

总体而言，关于基座模型的架构选择方案仍未确定。一方面，即使是基于经典Transformer架构的基座模型，其在每个单元构件的设计上也并非一致，当前并未有实验表明哪种架构选型是性能更好的；另一方面，在Transformer之外的架构中，在当前的上下文长度下（10,000词元以内），还没有架构能够在运算效率以及性能上都达到与Transformer类模型匹配的效果。由于预训练的代价非常高，因此基座模型架构的运行效率非常关键；此外，由于自注意力机制的二次计算复杂度，未来针对超长序列，我们也许需要更高效的架构设计。

4 总结

本文首先对基座模型训练中的数据进行了回顾，介绍当前基座模型常用的数据来源，并讨论了这些数据的清洗方式，然后针对代码数据处理中的问题进行了讨论，此外，我们简要汇总了一些可能的中文预训练数据来源。如何评估这些收集到的数据质量是一个开放性的问题，不

同的评测Prompt可能使得结论发生反转，因此在选取数据质量评估方式时需格外小心。同时，我们还汇总了当前不同的预训练基座以及对应的模型架构，并针对这些架构做了初步说明，以期让读者可较快地把握当前基座模型架构的发展方向。

参考文献

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. Palm 2 technical report. *CoRR*, abs/2305.10403.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 642–652. PMLR.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022. Efficient training of language models to fill in the middle. *CoRR*, abs/2207.14255.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Tri Dao, Daniel Y. Fu, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. 2023. Hungry Hungry Hippos: Towards language modeling with state space models. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yao Fu, Hao Peng, and Tushar Khot. 2022. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Xinyang Geng and Hao Liu. 2023. Openllama: An open reproduction of llama, May.

Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *CoRR*, abs/2305.15717.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *CoRR*, abs/2306.11644.

Danny Hernandez, Tom B. Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Benjamin Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. 2022. Scaling laws and interpretability of learning from repeated data. *CoRR*, abs/2205.10487.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. 2022. The stack: 3 TB of permissively licensed source code. *CoRR*, abs/2211.15533.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8424–8445. Association for Computational Linguistics.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy V, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Moustafa-Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding,

Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023. Starcoder: may the source be with you! *CoRR*, abs/2305.06161.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2023. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *CoRR*, abs/2305.13169.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. *CoRR*, abs/2305.16264.

Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. 2023a. Codegen2: Lessons for training llms on programming and natural languages. *CoRR*, abs/2305.02309.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2023b. Codegen: An open large language model for code with multi-turn program synthesis. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. *CoRR*, abs/2306.01116.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran G. V., Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanislaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: reinventing rnns for the transformer era. *CoRR*, abs/2305.13048.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey

Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 464–468. Association for Computational Linguistics.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. 2023. Moss: Training conversational language models from synthetic data.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Yi Tay, Jason Wei, Hyung Won Chung, Vinh Q. Tran, David R. So, Siamak Shakeri, Xavier Garcia, Huaixiu Steven Zheng, Jinfeng Rao, Aakanksha Chowdhery, Denny Zhou, Donald Metzler, Slav Petrov, Neil Houlsby, Quoc V. Le, and Mostafa Dehghani. 2022. Transcending scaling laws with 0.1% extra compute. *CoRR*, abs/2210.11399.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2023a. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6):109:1–109:28.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023b. UL2: unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yu-An Wang and Yun-Nung Chen. 2020. What do position embeddings learn? an empirical study of pre-trained language model positional encoding. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6840–6849. Association for Computational Linguistics.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4003–4012. European Language Resources Association.

Shaohua Wu, Xudong Zhao, Tong Yu, Rongguo Zhang, Chong Shen, Hongli Liu, Feng Li, Hong Zhu, Jiangang Luo, Liang Xu, and Xuanwei Zhang. 2021. Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning. *CoRR*, abs/2110.04725.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12360–12371.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.