Baby-Llama2-Chinese

Created by Limzero & Ambrose

介绍

本项目致力于构建一个小参数量的中文Llama2仓库。

包含：预训练、SFT指令微调、**奖励模型以及强化学习**（待做）完整流程。

除此之外，本项目还会梳理一套完整的LLM学习资料（正在进行中）。

希望该开源项目可以帮助LLM初学者以最快速度入门！

项目愿景

- 收集并汇总中文预训练语料，训练一个参数量500M-1B的Llama2-Chinese预训练模型，并在某个垂直领域可以表现不错
- 构建包含预训练、SFT指令微调、奖励模型以及强化学习整个完整流程的LLM代码仓库，包含DeepSpeed、Megatron等分布式训练技术
- 知识分享：梳理一套完整的LLM学习资料

Quick Start

```
# 1. 从“Baby-llama2-chinese Corpus”的百度网盘中下载分词处理后的预训练语料。（按需求下载-共634亿
# 2. 将下载好的数据放到./data/目录下
# 3. 根据下载的语料，修改data_process.py中的data_path_list部分
# 4. 运行data_process.py，在./data/目录下生成pretrain_data.bin文件
python data_process.py
# 5. 根据自身算力，修改 pretrain.py文件中的模型参数调整模型大小（max_seq_len、dim、n_layers、n
# 6. 预训练 pretrain.py—以下示例是基于4*3090
screen -S ambrose    #(创建新的名称为ambrose的screen)
screen -r ambrose    #(进入名称为ambrose的screen)
torchrun --standalone --nproc_per_node=4 pretrain.py
# 7. 运行结束后，预训练模型会保存在out/pretrain文件夹中
# 8. 针对alpaca-zh和bell两个SFT语料进行处理，如果新加SFT语料可以自行扩展。运行sft_data_process
python sft_data_process.py
# 9. 运行结束后，会在./sft_data目录下产生sft_data.csv文件
# 10. SFT微调
python sft.py
# 11. 运行结束后，SFT模型会保存在‘out/sft’文件夹中
```

```
# 12. 如果需要测试训练好的SFT模型，可以运行eval.py。（可以自定义问题）
python eval.py
```

📣 更新公告

- 2024年01月24日：新增了在84亿tokens预训练语料上的两个新模型Llama2-Chinese-92M-v1-smallvocab与Llama2-Chinese-218M-v1，与Llama2-Chinese-92M-v1进行对比分析模型大小和词表大小对预训练效果的影响！
- 2024年02月29日：新增了在634亿tokens预训练语料上的模型Llama2-Chinese-218M-v3，并以此为基座，使用医学垂直领域SFT数据进行finetune得到模型Llama2-Chinese-218M-v3-MedicalChat

预训练

一个好的预训练基座模型要具备**续写**的能力。

1. **分词器（Tokenizer）**：LLM分词器的构建方式有两种：一种是自己构造词表并训练一个分词器custom tokenizers，另一种是选择开源模型训练好的分词器，例如ChatGLM2-6B，Llama2等。

   由于llama官方所提供的词表中，中文的部分只有700个，这也是llama中文能力聊胜于无的原因。因此，为了方便使用，本项目选择ChatGLM2-6B的分词器，该词表大小为64793，值得注意的是：这是一个很妙的数字，因为它刚好在uint16的表示范围（0～65535的无符号整数），每一个token只需要两个字节即可表示，当我们的语料较大时候，相比常用的int32可以节省一半的存储空间。

2. **预训练语料（Corpus for pre-training ）**：从LLM技术革命以来，开源中文预训练语料越来越多。本项目本着拾人牙慧的精神，收集并处理了以下几个经典数据集：

中文预训练语料	描述
Wiki中文百科：wikipedia-cn-20230720-filtered	中文Wikipedia的数据
BaiduBaiKe：百度网盘 提取码: bwvb	中文BaiduBaiKe的数据
C4_zh：百度网盘 part1 提取码：zv4r；百度网盘 part2 提取码：sb83；百度网盘 part3 提取码：l89d	C4是可用的最大语言数据集之一，收集了来自互联网上超过3.65亿个域的超过1560亿个token。C4_zh是其中的一部分
WuDaoCorpora：智源研究院BAAI：WuDaoCorpora Text文本预训练数据集	中文悟道开源的200G数据
shibing624/medical：shibing624/medical	源自shibing624的一部分医学领域的预训练数据

同时，为了给大家节省数据预处理的时间，本项目开源了经过ChatGLM2-6B的分词器处理后的预训练语料，共计**634亿Tokens**的数据量，链接如下：Baby-llama2-chinese Corpus 提取码：6unr。将下载好的数据放到./data目录下即可。

【考虑到作者所持有机子的局限性（4张3090），目前634亿Tokens的预训练语料+300M参数量的模型已经是本人预训练的极限-注：没有使用DeepSpeed、Megatron等分布式训练架构】

预训练语料预处理

数据预处理采取GPT的通用做法，对语料进行提前分词，对一个样本做完分词后在末尾加上一个结束符号 <eos> ，与下一个样本区分开。然后将所有的训练语料拼接成一个数组（np.uint16）以.bin二进制格式存储到磁盘上。如果语料过大，避免内存溢出，可以选择mmap格式。

```
#脚本里面每一个函数对应一个语料库的预处理，搭建新加语料可以自行扩展。
python data_process.py
#运行结束后，会在./data目录下产生pretrain_data.bin文件
```

预训练

```
#考虑到预训练的运行时间非常久，需要采用程序后台运行的措施，本项目提供一种常用的程序后台运行的操作
screen -S ambrose    #(创建新的名称为ambrose的screen)
screen -r ambrose    #(进入名称为ambrose的screen)
#在该screen下执行预训练代码，如果你有四张卡，则nproc_per_node设置为4
torchrun --standalone --nproc_per_node=4 pretrain.py
#运行结束后，预训练模型会保存在‘out/pretrain’文件夹中
```

SFT指令微调

LLM微调的目的是将预训练模型中的知识引导出来的一种手段，通俗的讲就是教会模型说人话。

1. **微调方法**：自然语言处理目前存在一个重要的范式：一般领域数据的大规模预训练，对特定任务或领域的适应。因此，为了让预训练模型在特定任务或领域有不错的表现，需要对模型进行微调。目前主流的四种微调方法如下：

LLM微调方法

- **全面微调（Full Fine-tuning）**：使用任务特定数据调整LLM的所有参数。
- **参数高效精细调整（Parameter Efficient Fine-tuning）**：修改选定参数以实现更高效的适应。例如：LoRA、Adapter、Prefix-tuning、P-tuning以及P-tuning v2。
- **提示工程（Prompt Engineering）**：改进模型输入以指导模型输出理想结果。
- **检索增强生成（Retrieval Augmented Generation）**：将提示工程与数据库查询结合，以获得丰富的上下文答案。

其中Full Fine-tuning和Parameter Efficient Fine-tuning是需要基于特定任务或者垂直领域数据对模型（全部 or 部分）参数进行微调； Prompt Engineering和Retrieval Augmented Generation是通过设计模型输入的template，引导模型输出我们想要的内容，不需要对模型参数进行微调。其中RAG是通过外挂数据库的方式，为模型提供领域知识输入。

由于本项目模型参数（仅有218M左右，与bert-large-340M参数量差不多）并不大，因此选择Full Fine-tuning对特定任务或领域数据进行微调。后续有更大的预训练模型会补充其他微调方法。

2. **SFT微调数据**：LLM在垂直领域的适应已经是2023年的主格调，因此各个领域的SFT语料和微调模型层出不穷。目前已经有大佬整理并持续更新这方面的最新进展，大家有需要可以自己访问。

   本项目主要针对两类SFT语料进行模型微调，如下：

   日常问答SFT数据：

SFT语料	描述
alpaca-zh：alpaca-zh	源自shibing624的一部分SFT数据。该数据集是参考Alpaca方法基于GPT4得到的self-instruct数据，约5万条。
bell：bell	源自BelleGroup的一部分SFT数据。包含约100万条由BELLE项目生成的中文指令数据。

   医学垂直领域SFT数据：

SFT语料	描述
shibing624/medical：shibing624/medical	源自shibing624。该数据集不仅包含了预训练语料如上文所述，还包含一部分SFT数据。
HuatuoGPT-sft-data-v1：HuatuoGPT-sft-data-v1	源自HuatuoGPT的SFT数据
DISC-Med-SFT：HuatuoGPT-sft-data-v1	DISC-Med-SFT Dataset的子集
ChatMed_Consult-v0.3：michaelwzhu/ChatMed_Consult-v0.3	本数据集, ChatMed-Dataset, 中的query(或者是prompt)来自于互联网上的医疗问诊问题(549,326)，反映了真实世界的不同用户/患者的医疗问诊需求。目前response都是由OpenAI GPT-3.5引擎回答的。

SFT样本构建

因为SFT语料一般较小，我们没必要提前分词，而是在构建Dataloader的时候进行分词构建batch送给模型。所以自行参考dataset_sft.py即可！

基本逻辑如下：

- prompt和answer之间一定要有一个开始符 `<bos>` 隔开，然后answer后需要一个结束符 `<eos>` 。

- 计算loss的时候，对prompt部分的loss进行mask，只计算answer部分的loss即可。

```
#脚本里面针对alpaca-zh和bell两个SFT语料进行处理，搭建新加SFT语料可以自行扩展。
python sft_data_process.py
#运行结束后，会在./sft_data目录下产生sft_data.csv文件
```

全面微调（Full Fine-tuning）

```
#微调所需时间一般较短，如需要后台运行，本项目提供一种常用的程序后台运行的操作：
screen -S ambrose    #(创建新的名称为ambrose的screen)
screen -r ambrose    #(进入名称为ambrose的screen)
#在该screen下执行微调代码
python sft.py
#运行结束后，SFT模型会保存在‘out/sft’文件夹中
```

模型权重以及评测

1. **预训练模型**

模型名称	预训练语料	模型参数	下载地址
Llama2-Chinese-92M-v1	（82.78亿 Tokens） Wiki中文百科 +BaiduBaiKe +shibing624/medical	max_seq_len=512 dim=512 n_layers=8 n_heads=8	模型下载提取码：da7h
Llama2-Chinese-92M-v2	（140亿 Tokens） Wiki中文百科 +BaiduBaiKe +shibing624/medical +C4_zh	max_seq_len=512 dim=512 n_layers=8 n_heads=8	模型下载提取码：bjal
Llama2-Chinese-92M-v1-smallvocab **Notes:vocab size:21131**	（82.78亿 Tokens） Wiki中文百科 +BaiduBaiKe +shibing624/medical	max_seq_len=512 dim=512 n_layers=8 n_heads=8	模型下载提取码：ttst
Llama2-Chinese-218M-v1	（82.78亿 Tokens） Wiki中文百科 +BaiduBaiKe +shibing624/medical	max_seq_len=1024 dim=1024 n_layers=12 n_heads=8	模型下载提取码：c10m
Llama2-Chinese-218M-v2	（140亿 Tokens） Wiki中文百科 +BaiduBaiKe	max_seq_len=1024 dim=1024	模型下载提取码：dkne

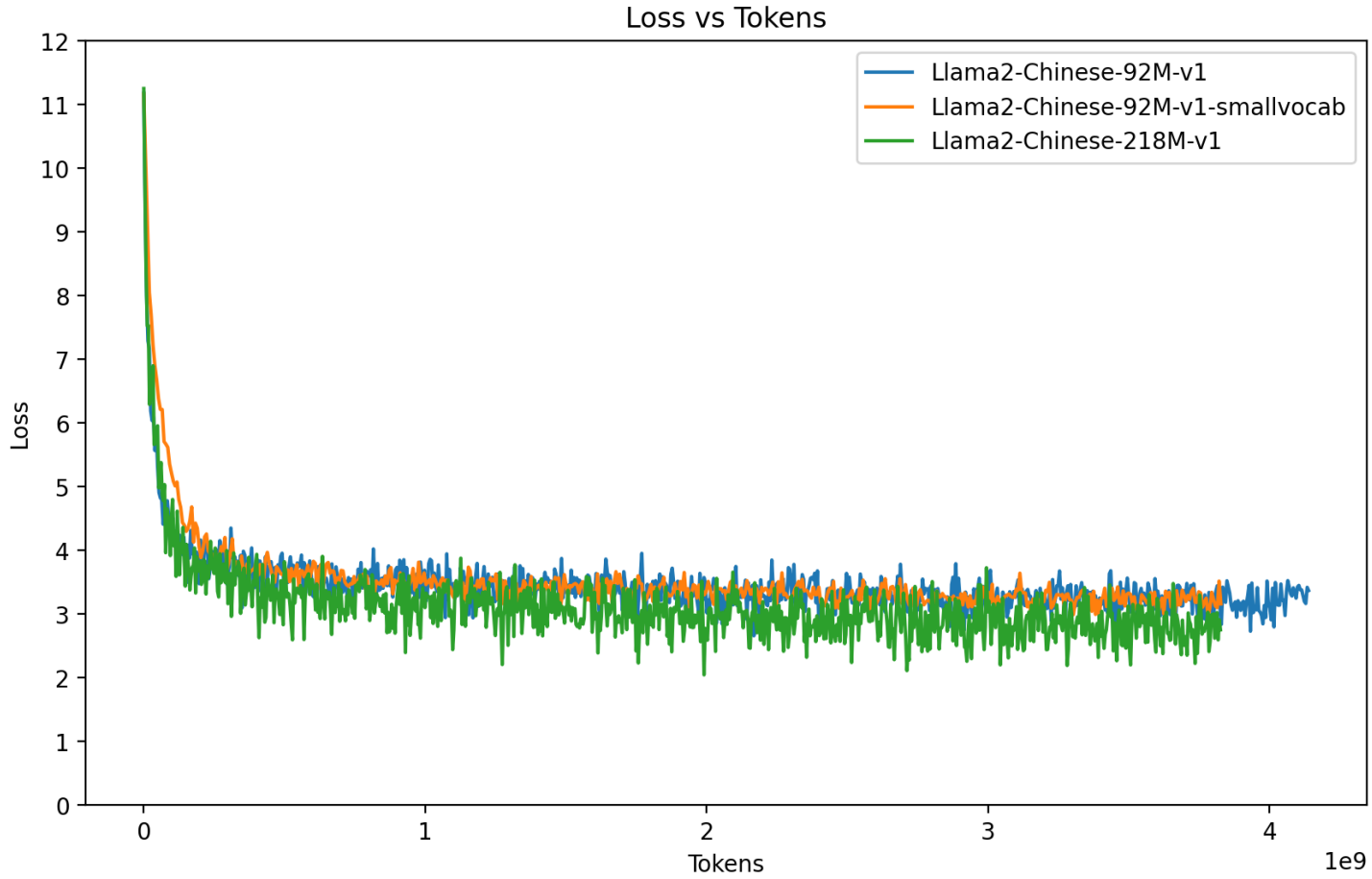
模型名称	预训练语料	模型参数	下载地址
	+shibing624/medical +C4_zh	n_layers=12 n_heads=8	
Llama2-Chinese-218M-v3	（634亿 Tokens） Wiki中文百科 +BaiduBaiKe +shibing624/medical +C4_zh +WuDaoCorpora	max_seq_len=1024 dim=1024 n_layers=12 n_heads=8	模型下载提取码：tpyy

各个预训练模型效果对比

预训练loss可视化展示：

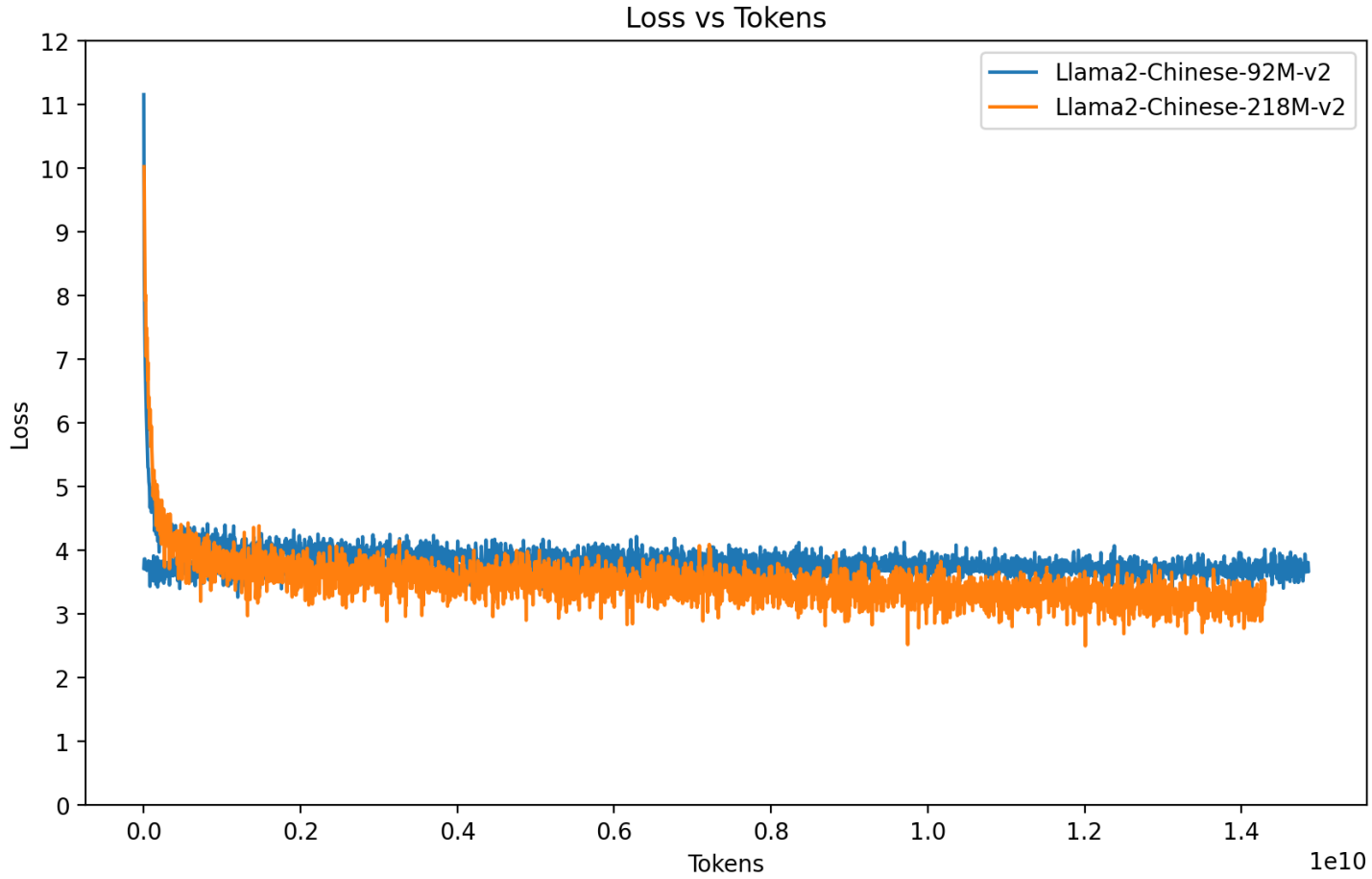
预训练语料v1：（82.78亿 Tokens） Wiki中文百科 + BaiduBaiKe + shibing624/medical

对比模型说明：**Llama2-Chinese-92M-v1** vs **Llama2-Chinese-92M-v1-smallvocab** vs **Llama2-Chinese-218M-v1**



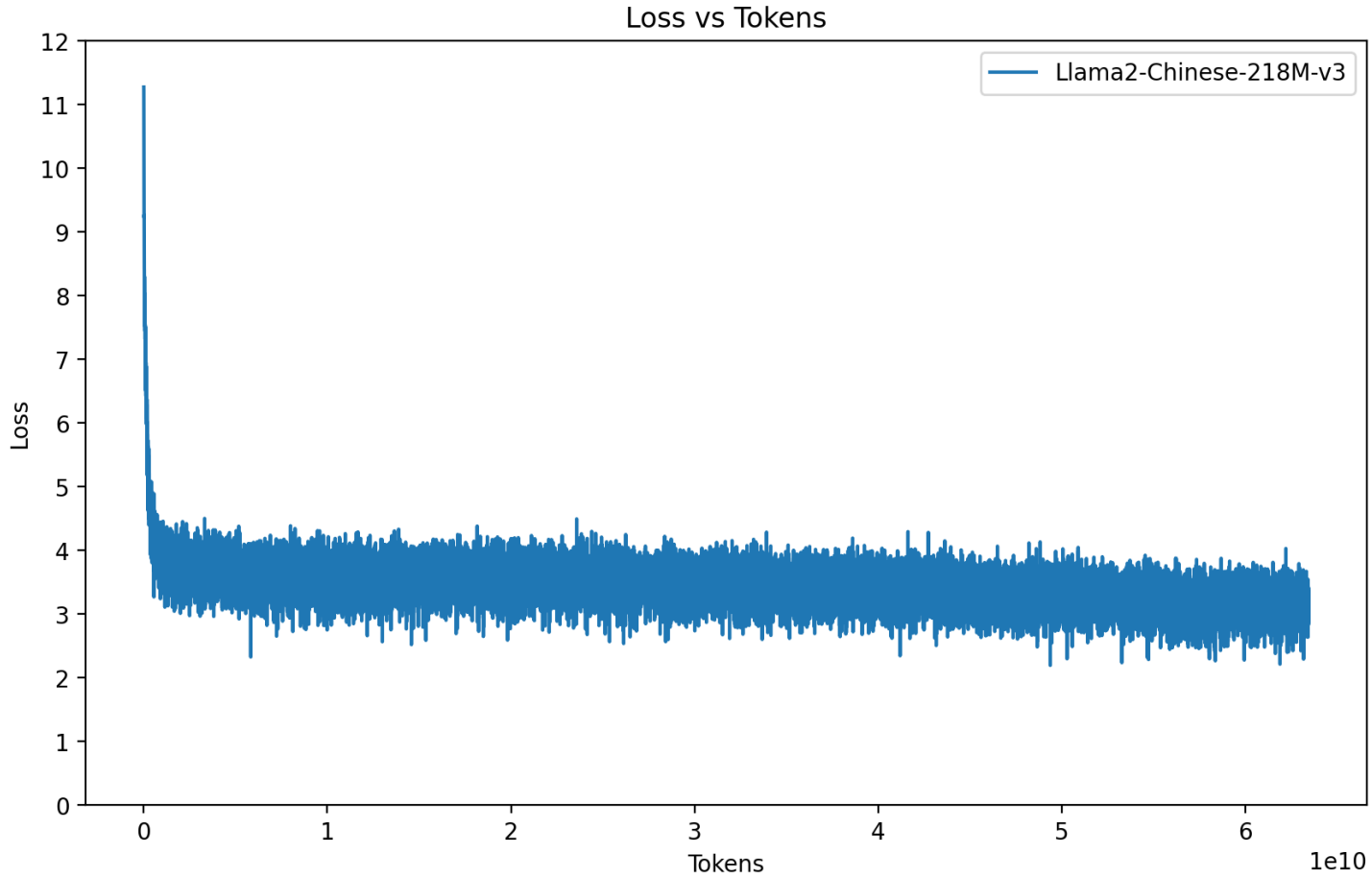
预训练语料v2：（140亿 Tokens） Wiki中文百科 + BaiduBaiKe + shibing624/medical + C4_zh

对比模型说明：**Llama2-Chinese-92M-v2** vs **Llama2-Chinese-218M-v2**



预训练语料v3：（634亿 Tokens） Wiki中文百科 + BaiduBaiKe + shibing624/medical + C4_zh + WuDaoCorpora

对比模型说明：**Llama2-Chinese-218M-v3**



主要从**续写能力**的角度对比各个预训练模型：

```
#预训练模型的推理：eval_pretrain.py。
python eval_pretrain.py
```

```
#示例一：Input：‘床前明月光，疑是地上霜。举头望明月，’
Llama2-Chinese-92M-v1 response：‘倚阑干照斜阳。● 早春何处无天，谁将归家归。谁愿去依花向，谁来教
Llama2-Chinese-92M-v2 response：‘却见天涯何处无。3、 想问你为什么对我这么好？1、 你是否在问自己
Llama2-Chinese-218M-v1 response：‘低头思故乡。2． 泪眼相对愁成丝，眉梢泪滴泪自流。3． 斜阳外，
Llama2-Chinese-218M-v2 response：‘低头思故乡。遥知兄弟登高处，遍插茱萸少一人。注释1、明月光：月
Llama2-Chinese-218M-v3 response：‘低头思故乡。人生如风，风雨来时裹住风，风雨后时披着雨。世事茫

#示例二：Input：‘《小王子》是一本畅销童话书，它讲述了：’
Llama2-Chinese-92M-v1 response：‘小王子和他的公主的故事，讲述的是一个善良、勇敢、智慧、善良的小
Llama2-Chinese-92M-v2 response：‘我有一座神奇的城堡。我有一个神奇的故事；我有一个美丽的童话；我
Llama2-Chinese-218M-v1 response：‘小王子和他的伙伴—小丑们要和人类结缘。在他们的帮助下，小王子
Llama2-Chinese-218M-v2 response：‘小王子和他的伙伴们在森林里玩耍。他们遇见了很多小动物，可是它
Llama2-Chinese-218M-v3 response：‘小王子去森林探望父亲。森林中发生了很多有趣的事情，可是小王子
```

2. **微调模型**

模型名称	SFT语料	模型参数	下载地址
Llama2-Chinese-92M-v1-NormalChat	alpaca-zh+bell	max_seq_len=512 dim=512 n_layers=8 n_heads=8	模型下载提取码：da7h
Llama2-Chinese-92M-v1-MedicalChat	shibing624/medical +HuatuoGPT-sft-data-v1 +DISC-Med-SFT +ChatMed_Consult-v0.3	max_seq_len=512 dim=512 n_layers=8 n_heads=8	模型下载提取码：da7h
Llama2-Chinese-92M-v2-NormalChat	alpaca-zh+bell	max_seq_len=512 dim=512 n_layers=8 n_heads=8	模型下载提取码：bjal
Llama2-Chinese-92M-v2-MedicalChat	shibing624/medical +HuatuoGPT-sft-data-v1 +DISC-Med-SFT +ChatMed_Consult-v0.3	max_seq_len=512 dim=512 n_layers=8 n_heads=8	正在加紧训练中！！！
Llama2-Chinese-218M-v1-NormalChat	alpaca-zh+bell	max_seq_len=1024 dim=1024 n_layers=12 n_heads=8	正在加紧训练中！！！
Llama2-Chinese-218M-v1-MedicalChat	shibing624/medical +HuatuoGPT-sft-data-v1 +DISC-Med-SFT +ChatMed_Consult-v0.3	max_seq_len=1024 dim=1024 n_layers=12 n_heads=8	正在加紧训练中！！！
Llama2-Chinese-218M-v2-NormalChat	alpaca-zh+bell	max_seq_len=1024 dim=1024 n_layers=12 n_heads=8	模型下载提取码：dkne
Llama2-Chinese-218M-v2-MedicalChat	shibing624/medical +HuatuoGPT-sft-data-v1 +DISC-Med-SFT +ChatMed_Consult-v0.3	max_seq_len=1024 dim=1024 n_layers=12 n_heads=8	正在加紧训练中！！！

模型名称	SFT语料	模型参数	下载地址
Llama2-Chinese-218M-v3-NormalChat	alpaca-zh+bell	max_seq_len=1024 dim=1024 n_layers=12 n_heads=8	正在加紧训练中！！！
Llama2-Chinese-218M-v3-MedicalChat	shibing624/medical +HuatuoGPT-sft-data-v1 +DISC-Med-SFT +ChatMed_Consult-v0.3	max_seq_len=1024 dim=1024 n_layers=12 n_heads=8	模型下载提取码：tpyy

目前没有在标准测试集上验证SFT模型效果。仅仅给出以下示例：

```
#SFT微调模型的推理：eval.py。
python eval.py
```

```
#示例一：Input：‘最近我在办公室坐久了会感到头晕，请问这是什么原因?有什么缓解办法吗？’—医学向问题
Llama2-Chinese-92M-v1-NormalChat response：‘是的，由于我的工作忙，我会很可能会坐在办公室里坐。
Llama2-Chinese-92M-v1-MedicalChat response：‘您的症状可能是由多种原因引起的，例如低血压、脑部
Llama2-Chinese-92M-v2-NormalChat response：‘我无法确定您所在的位置。您可以在公司办公室的电脑上
Llama2-Chinese-92M-v2-MedicalChat response：‘’
Llama2-Chinese-218M-v1-NormalChat response：‘’
Llama2-Chinese-218M-v1-MedicalChat response：‘’
Llama2-Chinese-218M-v2-NormalChat response：‘我无法确定您所描述的问题，因此无法回答。是一种可
Llama2-Chinese-218M-v2-MedicalChat response：‘’
Llama2-Chinese-218M-v3-NormalChat response：‘’
Llama2-Chinese-218M-v3-MedicalChat response：‘长时间保持同一姿势会导致血液循环不畅，影响血使用

#示例二：Input：‘前列腺囊肿的症状是什么？’—医学向问题
Llama2-Chinese-92M-v1-NormalChat response：‘人类大脑中的神经控制系统是研究人类大脑结构和功能的
Llama2-Chinese-92M-v1-MedicalChat response：‘尿频、尿急、尿痛；慢性前列腺炎；排尿不畅；遗精；
Llama2-Chinese-92M-v2-NormalChat response：‘影片光速是有可能划破天际的，可能带来无限可能的进展
Llama2-Chinese-92M-v2-MedicalChat response：‘’
Llama2-Chinese-218M-v1-NormalChat response：‘’
Llama2-Chinese-218M-v1-MedicalChat response：‘’
Llama2-Chinese-218M-v2-NormalChat response：‘也被称为“水花如水”，通常指在火上燃烧的面包或饭单
Llama2-Chinese-218M-v2-MedicalChat response：‘’
Llama2-Chinese-218M-v3-NormalChat response：‘’
Llama2-Chinese-218M-v3-MedicalChat response：‘尿频；脓性分泌物；血尿；尿急；前列腺痛；脓肿；尿
’

#示例三：Input：‘请问，世界上最大的动物是什么？’—科普向问题
Llama2-Chinese-92M-v1-NormalChat response：‘世界上最大的动物是蓝鲸。’
```

```
Llama2-Chinese-92M-v1-MedicalChat response：‘38%的动物都是天马科动物，而“天牛”的称法是 10%的
Llama2-Chinese-92M-v2-NormalChat response：‘世界上最大的动物是蓝鲸。它们是鲸类中最大的，可以长
Llama2-Chinese-92M-v2-MedicalChat response：‘’
Llama2-Chinese-218M-v1-NormalChat response：‘’
Llama2-Chinese-218M-v1-MedicalChat response：‘’
Llama2-Chinese-218M-v2-NormalChat response：‘世界上最大的动物是蓝鲸。它们的体重可以达到4000至
Llama2-Chinese-218M-v3-MedicalChat response：‘除了导致的，在一般情况下，保持适当的中毒处理方法
Llama2-Chinese-218M-v2-MedicalChat response：‘’
Llama2-Chinese-218M-v3-NormalChat response：‘’
```

可以明显看出，经过medical SFT数据微调后的模型在医学向问题的回答上比其他模型更加准确，但是对于日常科普向问题的回答遗忘性太大。

总而言之，模型越大，语料越多模型的性能越强。

号召

欢迎大家一起共建这个小项目，这对于希望入门LLM的同学来说，是一次不可多得的练手机会！感兴趣的小伙伴可以加QQ群: 716455397。

参考llama2.c