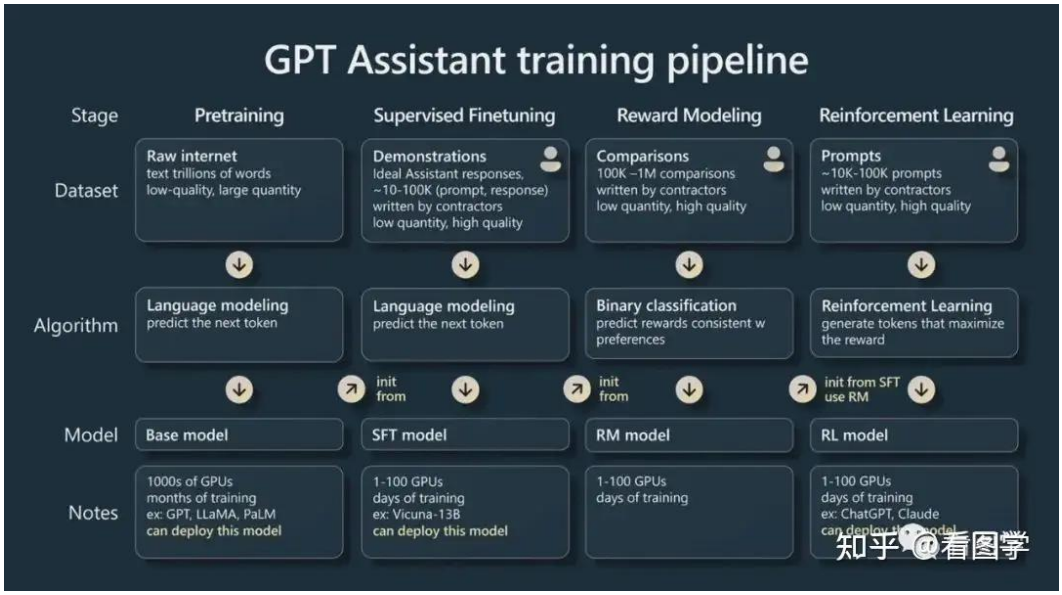
作者：看图学
链接：https://www.zhihu.com/question/608820310/answer/3091336166
来源：知乎

最近大模型发展卷的很，王慧文都被整抑郁了。想要研究学习大模型，应该从哪里开始呢？

目前大模型发展生态最好的当属 Meta 的 LLaMA 模型。如果 GPT 系列是 Windows 操作系统(巧了，OpenAI 的大东家目前就是微软)，那么 LLaMA 就是 Linux。如果 GPT 系列是苹果手机，那么 LLaMA 就是安卓。如果你想基于大模型做一些事情，无论是创业还是研究，最好选择一个生态好的模型，毕竟有人用才有市场。

ChatGPT 演化的路径如下图所示。



图片中分了 4 个阶段，但是第三个和第四个阶段一般都会放在一起，属于对齐阶段。所以一般会分成如下 3 个阶段：

- Stage 1: 预训练(Pretrain)
- Stage 2: 监督微调(SFT)
- Stage 3: 对齐(Reward Model + RLHF)

既然已经有了成功 ChatGPT 这一成功的案例，大家都想基于 LLaMA 把这条路再走一遍，以期望做出自己的 ChatGPT。

所以**基于 LLaMA 的模型虽然很多，但是基本都可以放到上面 3 个框架当中**。本文就沿着预训练、监督微调、对齐(RW+RLHF)这一路径来梳理一下 LLaMA 生态中的各个模型。

主要是点出这些模型处在大模型训练的那一个阶段，以及都做了哪些创新性的工作，方便你根据自己的兴趣和资源来选择使用哪一个，对中文支持比较好的也都有注明。

Stage1 预训练：LLaMA 复现

RedPajama

- 参考 LLaMA 论文中的训练数据，收集并且开源可商用。
- https://github.com/togethercomputer/RedPajama-Data

Baichuan-7B(支持中文)

- 采用 LLaMA 的相同架构，在中文上做预训练。可商用。
- 王小川这次做大模型的切入点其实挺不错的，绑定到 LLaMA 的生态上，然后在中文上有所突破。可能也在构思新三级火箭了吧。
- 目前 Baichuan 可以算是第一个 LLaMA 中文预训练模型，所以**后面的工作都可以在这上面都走一遍**，估计没多久 Baichuan-Alapca, Baichuan-Vicuna 就都出来了。
- https://github.com/baichuan-inc/baichuan-7B

OpenLLaMA

- 参考 LLaMA 的代码，在 Apache 2.0 license 下的重新实现和训练。使用了 RedPajama 训练集合。
- https://github.com/openlm-research/open_llama

Lit-LLaMA

- 参考 LLaMA，在 Apache 2.0 license 下的只有代码的重新实现。同时支持加载原始 LLaMA 和 OpenLLaMA 的权重。
- https://github.com/Lightning-AI/lit-llama

Stage 2：监督微调

因为预训练模型本质上还是个续写模型，所以并不能很好的满足人们的需求，所以监督微调的作用就是微调模型产生理想的回复。

在监督微调这里，大家目标都是一样的，但是做法有些不同，主要是有钱和没钱的区别。

有钱你可以全参数微调，没钱就只能使用一些低成本的方法，英文叫 PEFT(Parameter-Efficient Fine-Tuning)。

PEFT 确实是想我这种平民玩家的首选，但是有钱也可以用 PEFT，它可以让你微调更大的模型。比如我们就只能玩玩 10B 的，有点小钱用 PEFT 玩个几十 B 的问题不大。

2.1 LLaMA + Instruction Finetuning(全量参数)

Alpaca

- llama7b + self-instruct 数据指令微调。算是最早迈出 LLaMA+SFT 这一步的模型。最开始并没有提供权重，后来通过 diff 的方式给出，需要 LLaMA 原始模型才能恢复，github 上有教程。
- 当时他们采用 1 张 8 卡 A100(80G 显存)，52k 的数据，训练了 3 个小时。训练成本大概是 100 刀。
- https://github.com/tatsu-lab/stanford_alpaca

Alpaca 衍生模型

- BELLE(支持中文)：最早是基于 BLOOM 的，后来也支持 LLaMA https://github.com/LianjiaTech/BELLE
- openAlpaca：OpenLLaMA + databricks-dolly-15k dataset 进行指令微调 https://github.com/yxuansu/OpenAlpaca
- gpt4-x-alpaca：用 GPT4 的数据微调，数据集为 GPTeacher https://huggingface.co/chavinlo/gpt4-x-alpaca

Vicuna

- llama13b + ShareGPT 对话数据，微调
- 研发团队基于 Vicuna 发布了 FastChat 对话机器人。
- 和 Alpaca 一样，受协议限制，vicuna 模型公布的权重也是个 delta，每个参数要加上 llama 原来的权重才是模型权重。
- https://github.com/lm-sys/FastChat

Vicuna 衍生模型

- gpt4-x-vicuna-13b：用 GPT4 的数据微调，数据集为 GPTeacher https://huggingface.co/NousResearch/gpt4-x-vicuna-13b

WizardLM

- 采用了 Evol-Instruct 来构造指令，可以产生一些很难的指令.

1. 深度演化包括五种操作：添加约束、深化、具体化、增加推理步骤并使输入复杂化。
2. In-breadth Evolving 是突变，即根据给定的指令生成全新的指令
3. 进化是通过提示+LLM 来实现的。

- https://github.com/nlpxucan/WizardLM

TÜLU

- 使用 LLaMA + Human/GPT data mix 微调
- 验证了很多结论，论文值得一看。https://arxiv.org/abs/2306.04751
- https://github.com/allenai/open-instruct

GPT4ALL

- LLaMA 用 80w 的 GPT3.5 的数据(code, story, conversation)微调而来。
- https://github.com/nomic-ai/gpt4all

Koala

- LLaMA13B 基于 ChatGPT Distillation Data 和 Open Source Data 训练而来。
- 具体数据见下面：
    - https://bair.berkeley.edu/blog/2023/04/03/koala/

OpenBuddy(支持中文)

- 基于 LLaMA，Falcon, OpenLLaMA 微调的，只说用了对话数据，细节没透漏。
- https://github.com/OpenBuddy/OpenBuddy

Pygmalion 7B

- 给予 LLaMA 微调，使用了不同来源的 56MB 的对话数据，包含了人工和机器。
- https://huggingface.co/PygmalionAI/pygmalion-7b

2.2 LLaMA + PEFT

PEFT 目前最流行的是 LoRA，挺巧妙的架构，可以看看 https://arxiv.org/abs/2106.09685。

下面大多数的模型都是 LLaMA+lora 的架构，不只是文本，AIGC 的头部网站 http://civitai.com 上很多模型也都是基于 lora 的。

最近还出了 QLoRA，在 LoRA 的基础上加入了量化，进一步降低显存的使用。https://arxiv.org/abs/2305.14314。

Baize

- LLaMA + Lora
- https://github.com/project-baize/baize-chatbot

LLaMA-Adapter

- LLaMA + Adapter Layer
- https://github.com/OpenGVLab/LLaMA-Adapter

CalderaAI/30B-Lazarus

- 似乎是多个 LoRA 的 merge，但是没太公布太多细节。
- 在 huggingface 的 leaderboard 上排名还挺靠前。
- https://huggingface.co/CalderaAI/30B-Lazarus

Chinese-LLaMA-Alpaca(支持中文)

- https://arxiv.org/pdf/2304.08177.pdf
- LLaMA + 扩词表 + lora
- Chinese LLaMA 是属于局部参数预训练
    - Stage1: frozen encoder，只用来训练 Embedding 层。
    - Stage2: 只训练 Embedding, LM head, lora weights
- 在 Chinese LLaMA 的基础上，仿照 Alpaca 训练了 Chinese Alpaca
- https://github.com/ymcui/Chinese-LLaMA-Alpaca

Chinese-Vicuna(支持中文)

- 基于：https://github.com/tloen/alpaca-lora
- lora + 中文 instruction 数据
- chatv1 的数据使用了 50k 中文指令+对话混合数据。
- 并没有扩充词表，据说 Vicuna1.1 并没有扩充词表，但是中文效果不错。
- https://github.com/Facico/Chinese-Vicuna

Stage 3: 对齐(LLaMA + FT + RHLF)

- 这部分可以说是把 ChatGPT 的路径完整走了一遍。

StableVicuna

- Vicuna = LLaMA + FT
- StableVicuna = Vicuna + RLHF
- https://github.com/Stability-AI/StableLM

StackLLaMA

- SFT: LLaMA + Lora
- RM: LLaMA + Lora + 分类
- https://huggingface.co/blog/zh/stackllama

其他：LLaMA 推理优化

llama.cpp

- 用 C/C++实现的推理，不依赖显卡。
- https://github.com/ggerganov/llama.cpp

GPTQ-for-LLaMA

- 4 bits quantization of LLaMA using GPTQ.
- https://github.com/qwopqwop200/GPTQ-for-LLaMa

写在最后

上面的模型 github 中一般都有模型下载，但是国内的网络你懂得，有时候下载不下来。如果需要 LLaMA 模型的权重，可以看这一篇：**ChatGPT 平替模型：LLaMA（附下载地址，平民玩家和伸手党的福音！）**