# 国内外开源大语言模型一览表

国内外开源大语言模型一览表

2023-07-19 收藏 分享

20秒
国
大
推荐
· 最新
· 链接
向量
· Chin
· 开源
Chat
· 国内
· 国外
，而
· 一个

## 1 Chinese Open Source Language Models

### 本草

https://zhuanlan.zhihu.com/p/626536996

https://github.com/scir-hi/huatuo-llama-med-chinese

基于中文医学知识的LLaMa指令微调模型

在生物医学领域，LLM模型（如LLaMa，ChatGLM）因为缺乏一定的医学专业知识语料而表现不佳。该项目通过医学知识图谱和GPT3.5API构建了中文医学指令数据集，并对LLaMa模型进行了指令微调得到了一个针对医学领域的智能问诊模型HuaTuo，相比于未经过医学数据指令微调的原LLaMa而言，HuaTuo模型在智能问诊层面表现出色，可生成一些更为可靠的医学知识回答；与此同时，基于相同医学数据，该项目还训练了医疗版本的ChatGLM模型: ChatGLM-6B-Med，

该团队还即将发布扁鹊模型PienChueh(同为基于医学数据训练的大模型)，欢迎大家届时使用体验。

### 百川 Baichuan-7B

https://github.com/baichuan-inc/baichuan-7B

https://huggingface.co/baichuan-inc/baichuan-7B

baichuan-7B 是由百川智能开发的一个开源可商用的大规模预训练语言模型。基于 Transformer 结构，在大约1.2万亿 tokens 上训练的70亿参数模型，支持中英双语，上下文窗口长度为4096。在标准的中文和英文权威 benchmark（C-EVAL/MMLU）上均取得同尺寸最好的效果。

原始数据包括开源的中英文数据和自行抓取的中文互联网数据，以及部分高质量知识性数据。

参考相关数据工作，频率和质量是数据处理环节重点考虑的两个维度。 我们基于启发式规则和质量模型打分，对原始数据集进行篇章和句子粒度的过滤。在全量数据上，利用局部敏感哈希方法，对篇章和句子粒度做滤重。

**百川 Baichuan-13B（可商用）**

https://githuB.com/Baichuan-inc/Baichuan-13B

更大尺寸、更多数据:Baichuan-13B 在 Baichuan-7B 的基础上进一步扩大参数量到130亿，并且在高质量的语料上训练了1.4万亿 tokens，超过 LLaMA-13B40%，是当前开源13B 尺寸下训练数据量最多的模型。支持中英双语，使用 ALiBi 位置编码，上下文窗口长度为4096。
同时开源预训练和对齐模型:预训练模型是适用开发者的『 基座 』，而广大普通用户对有对话功能的对齐模型具有更强的需求。因此本次开源我们同时发布了对齐模型（Baichuan-13B-Chat），具有很强的对话能力，开箱即用，几行代码即可简单的部署。
更高效的推理:为了支持更广大用户的使用，我们本次同时开源了 int8和 int4的量化版本，相对非量化版本在几乎没有效果损失的情况下大大降低了部署的机器资源门槛，可以部署在如 Nvidia3090这样的消费级显卡上。
开源免费可商用:Baichuan-13B 不仅对学术研究完全开放，开发者也仅需邮件申请并获得官方商用许可后，即可以免费商用。

### 华佗

https://mp.weixin.qq.com/s/lwJb8N420xfMTvXJPM2gtg

https://arxiv.org/pdf/2305.15075.pdf

https://github.com/FreedomIntelligence/HuatuoGPT

https://www.huatuogpt.cn/

该论文提出的语言模型训练方法可以结合医生和 ChatGPT 的数据，充分发挥它们的互补作用，既保留真实医疗数据的专业性和准确性，又借助 ChatGPT 的多样性和内容丰富性的特点。

### 扁鹊

https://github.com/scutcyr/BianQue

基于主动健康的主动性、预防性、精确性、个性化、共建共享、自律性六大特征，华南理工大学未来技术学院-广东省数字孪生人重点实验室开源了中文领域生活空间主动健康大模型基座ProactiveHealthGPT，包括：

经过千万规模中文健康对话数据指令微调的生活空间健康大模型扁鹊（BianQue）

经过百万规模心理咨询领域中文长文本指令与多轮共情对话数据联合指令微调的心理健康大模型灵心（SoulChat）

我们期望，生活空间主动健康大模型基座ProactiveHealthGPT 可以帮助学术界加速大模型在慢性病、心理咨询等主动健康领域的研究与应用。本项目为 生活空间健康大模型扁鹊（BianQue） 。

### 灵心（SoulChat）

https://github.com/scutcyr/SoulChat

我们调研了当前常见的心理咨询平台，发现，用户寻求在线心理帮助时，通常需要进行较长篇幅地进行自我描述，然后提供帮助的心理咨询师同样地提供长篇幅的回复，缺失了一个渐进式的倾诉过程。但是，在实际的心理咨询过程当中，用户和心理咨询师之间会存在多轮次的沟通过程，在该过程当中，心理咨询师会引导用户进行倾诉，并且提供共情，例如："非常棒"、"我理解你的感受"、"当然可以"等等。

考虑到当前十分欠缺多轮共情对话数据集，我们一方面，构建了超过15万规模的 单轮长文本心理咨询指令与答案（SoulChatCorpus-single_turn），回答数量超过50万（指令数是当前的常见的心理咨询数据集 PsyQA 的6.7倍），并利用ChatGPT与GPT4，生成总共约100万轮次的多轮回答数据（SoulChatCorpus-multi_turn） 。特别地，我们在预实验中发现，纯单轮长本文驱动的心理咨询模型会产生让用户感到厌烦的文本长度，而且不具备引导用户倾诉的能力，纯多轮心理咨询对话数据驱动的心理咨询模型则弱化了模型的建议能力，因此，我们混合SoulChatCorpus-single_turn和SoulChatCorpus-multi_turn构造成超过120万个样本的 单轮与多轮混合的共情对话数据集SoulChatCorpus 。所有数据采用"用户：xxx\n心理咨询师：xxx\n用户：xxx\n心理咨询师："的形式统一为一种指令格式。

我们选择了 ChatGLM-6B 作为初始化模型，进行了全量参数的指令微调，旨在提升模型的共情能力、引导用户倾诉能力以及提供合理建议的能力。更多训练细节请留意我们后续发布的论文。

### 启真医学大模型

https://github.com/CMKRG/QiZhenGPT

本项目利用启真医学知识库构建的中文医学指令数据集，并基于此在Chinese-LLaMA-Plus-7B、CaMA-13B、ChatGLM-6B模型上进行指令精调，大幅提高了模型在中文医疗场景下效果，首先针对药品知识问答发布了评测数据集，后续计划优化疾病、手术、检验等方面的问答效果，并针对医患问答、病历自动生成等应用展开拓展。

### 【貔貅】FinMA & PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance

https://github.com/chancefocus/PIXIU

https://arxiv.org/abs/2306.05443

https://huggingface.co/spaces/ChanceFocus/FLARE

The advancement of Natural Language Processing (NLP) and machine learning (ML) techniques in financial technology (FinTech) has enabled a diverse set of capabilities from predicting stock price movements to advanced financial analytics. However, to effectively understand the complex financial language and concepts, domain-specific LLMs are necessary.

Despite prior efforts, there is a lack of open-source financial LLMs and benchmarks to evaluate them. Additionally, these models are not fine-tuned to follow natural language instructions, limiting their performance in downstream financial tasks.

To address these gaps, we introduce PIXIU, providing:

Open-source LLMs tailored for finance called FinMA, by fine-tuning LLaMA with the dataset constructed in PIXIU.

Large-scale, high-quality multi-task and multi-modal financial instruction tuning data FIT.

Holistic financial evaluation benchmarks FLARE for assessing financial LLMs.

Key Features

Open resources: PIXIU openly provides the financial LLM, instruction tuning data, and datasets included in the evaluation benchmark to encourage open research and transparency.

Multi-task: The instruction tuning data in PIXIU cover a diverse set of financial tasks, including four financial NLP tasks and one financial prediction task.

Multi-modality: PIXIU's instruction tuning data consist of multi-modality financial data, including time series data from the stock movement prediction task. It covers various types of financial texts, including reports, news articles, tweets, and regulatory filings.

Diversity: Unlike previous benchmarks focusing mainly on financial NLP tasks, PIXIU's evaluation benchmark includes critical financial prediction tasks aligned with real-world scenarios, making it more challenging.

### 中文Alpaca模型Luotuo

https://sota.jiqizhixin.com/project/luotuo

https://github.com/LC1332/Luotuo-Chinese-LLM

Alpaca 是斯坦福团队基于 LLaMA 7B 在 52k 指令上微调得到的模型，能出色适应多种自然语言应用场景。近日来自商汤科技和华中科技大学开源中文语言模型 Luotuo，基于 ChatGPT API 翻译 Alpaca 微调指令数据，并使用 lora 进行微调得到。目前该项目已公开训练的语料和模型权重文件（两个型号），供开发者可使用自己各种大小的语料，训练自己的语言模型，并适用到对应的垂直领域。

### 中文LLaMA&Alpaca大模型

https://github.com/ymcui/Chinese-LLaMA-Alpaca

以ChatGPT、GPT-4等为代表的大语言模型（Large Language Model, LLM）掀起了新一轮自然语言处理领域的研究浪潮，展现出了类通用人工智能（AGI）的能力，受到业界广泛关注。然而，由于大语言模型的训练和部署都极为昂贵，为构建透明且开放的学术研究造成了一定的阻碍。

为了促进大模型在中文NLP社区的开放研究，本项目开源了中文LLaMA模型和经过指令精调的Alpaca大模型。这些模型在原版LLaMA的基础上扩充了中文词表并使用了中文数据进行二次预训练，进一步提升了中文基础语义理解能力。同时，在中文LLaMA的基础上，本项目使用了中文指令数据进行指令精调，显著提升了模型对指令的理解和执行能力。

### 中文对话式大语言模型Firefly

https://mp.weixin.qq.com/s/tyH9Ifcvw4DKqoIoYjT6Kg

https://github.com/yangjianxin1/Firefly

Firefly（流萤） 是一个开源的中文对话式大语言模型，使用指令微调（Instruction Tuning）在中文数据集上进行调优。同时使用了词表裁剪、ZeRO、张量并行等技术，有效降低显存消耗和提高训练效率。 在训练中，我们使用了更小的模型参数量，以及更少的计算资源。

我们构造了许多与中华文化相关的数据，以提升模型这方面的表现，如对联、作诗、文言文翻译、散文、金庸小说等。

### 凤凰

https://mp.weixin.qq.com/s/beAAh_MdqssV8bEKsccElg

https://github.com/FreedomIntelligence/LLMZoo

LLM Zoo is a project that provides data, models, and evaluation benchmark for large language models.

### 【复旦】MOSS

https://github.com/OpenLMLab/MOSS

https://mp.weixin.qq.com/s/LjToZVWjQ-ot5KJFCFtA3g

MOSS是一个支持中英双语和多种插件的开源对话语言模型，moss-moon系列模型具有160亿参数，在FP16精度下可在单张A100/A800或两张3090显卡运行，在INT4/8精度下可在单张3090显卡运行。MOSS基座语言模型在约七千亿中英文以及代码单词上预训练得到，后续经过对话指令微调、插件增强学习和人类偏好训练具备多轮对话能力及使用多种插件的能力。

### 【复旦】MOSS-RLHF

https://mp.weixin.qq.com/s/BjXtnEEVCQiPOy-_qCNM4g

https://openlmlab.github.io/MOSS-RLHF/paper/SecretsOfRLHFPart1.pdf

https://openlmlab.github.io/MOSS-RLHF/

FudanNLP 团队通过大量、详实工作，设计实验充分探索了大模型 RLHF 的完整工作流程，仔细剖析了 RLHF 中的强化学习 PPO 算法的内部工作原理以及它在整个 RLHF 中的作用，并研究各种优化方法如何影响训练过程。通过这些努力，确定了使得 PPO 算法在大模型人类对齐方面行之有效的关键因素。

综合上述发现，该团队进一步总结出在大模型上训练更稳定的 PPO 算法版本：PPO-max。并使用 Helpful 和 Harmless 数据集全面评估，结果显示经过 PPO-max 算法训练的模型展现出了出色的人类对齐性能！

综合上述发现，该团队进一步总结出在大模型上训练更稳定的 PPO 算法版本：PPO-max。并使用 Helpful 和 Harmless 数据集全面评估，结果显示经过 PPO-max 算法训练的模型展现出了出色的人类对齐性能！

### 【度小满】轩辕-首个千亿级中文金融对话模型

https://arxiv.org/pdf/2305.12002.pdf

https://huggingface.co/xyz-nlp/XuanYuan2.0

https://github.com/Duxiaoman-DI/XuanYuan

https://huggingface.co/xyz-nlp/XuanYuan2.0

https://zhuanlan.zhihu.com/p/632780608

轩辕是国内首个开源的千亿级中文对话大模型，同时也是首个针对中文金融领域优化的千亿级开源对话大模型。轩辕在BLOOM-176B的基础上针对中文通用领域和金融领域进行了针对性的预训练与微调，它不仅可以应对通用领域的问题，也可以解答与金融相关的各类问题，为用户提供准确、全面的金融信息和建议。

### 悟道·天鹰（Aquila）

https://github.com/FlagAI-Open/FlagAI/tree/master/examples/Aquila

这是首个具备中英双语知识、支持商用许可协议、支持国内数据合规要求的开源语言大模型。悟道·天鹰（Aquila）系列模型包括 Aquila基础模型（7B、33B），AquilaChat对话模型（7B、33B）以及 AquilaCode“文本-代码”生成模型。

### 桃李：国际中文教育大模型

https://github.com/blcuicall/taoli

随着ChatGPT引起全社会的关注，及各类大语言模型（Large Language Model）争相亮相，通用领域自然语言处理任务已获得巨大成功，引起了国际中文教育领域的普遍关注。

国际中文教育人士纷纷展开了对大模型的探讨： 大模型是否可以根据学习者的水平，提供合适的语言表达，或根据学习者的问题给出详细的解答，从而在一定程度上辅助甚至充当学习伙伴、语言教师？ 然而，目前通用领域的大模型在垂直领域的效果仍有限。

为解决上述问题，我们全面推出适用于国际中文教育领域的大模型“桃李”（Taoli）1.0，一个在国际中文教育领域数据上进行了额外训练的模型。

我们基于目前国际中文教育领域流通的500余册国际中文教育教材与教辅书、汉语水平考试试题以及汉语学习者词典等，构建了国际中文教育资源库。 我们设置了多种形式的指令来充分利用知识，构造了共计 88000 条的高质量国际中文教育问答数据集，并利用收集到的数据对模型进行指令微调，让模型习得将法律知识应用到具体场景中的能力。

### 情感大模型PICA

https://mp.weixin.qq.com/s/E37EFe10185THHa3pSqBig

https://github.com/NEU-DataMining/PICA

https://huggingface.co/NEUDM/PICA-V1

PICA 以清华大学开源的ChatGLM2-6B为基础，采用Prompt tuning技术在4 卡 A6000 训练大约15个小时得到。我们和SoulChat 进行了对比（最后部分），我们的模型在体验和安全上更有优势。我们只使用了2K的数据进行了p-tuning 微调，这充分说明了我们构造的数据质量比较高。模型权重可以在 HuggingFace 访问，欢迎各位使用并提出宝贵的意见。

### Anima：基于QLoRA的33B中文大语言模型

https://github.com/lyogavin/Anima

AI Community从来都是非常开放的，AI发展到今天，离不开很多以前的重要开源工作，开放共享的Paper，或者的开源数据和代码。我们相信AI的未来也一定是开放的。希望能为开源社区做一些贡献。

为什么33B模型很重要？QLoRA是个Game Changer?

之前大部分开源可finetune的模型大都是比较小的模型7B或者13B，虽然可以在一些简单的chatbot评测集上，通过finetune训练有不错的表现。但是由于这些模型规模还是有限，LLM核心的reasoning的能力还是相对比较弱。这就是为什么很多这种小规模的模型在实际应用的场景表现像是个玩具。如这个工作中的论述：chatbot评测集比较简单，真正比较考验模型能力的复杂逻辑推理及数学问题上小模型和大模型差距还是很明显的。

因此我们认为QLoRA 的工作很重要，重要到可能是个Game Changer。通过QLoRA的优化方法，第一次让33B规模的模型可以比较民主化的，比较低成本的finetune训练，并且普及使用。我们认为33B模型既可以发挥大规模模型的比较强的reasoning能力，又可以针对私有业务领域数据进行灵活的finetune训练提升对于LLM的控制力。

### BayLing: Bridging Cross-lingual Alignment and Instruction Following through Interactive Translation for Large Language Models

https://github.com/ictnlp/BayLing

https://arxiv.org/abs/2306.10968

BayLing (百聆, bǎi líng) is an instruction-following large language model equipped with advanced language alignment, showing superior capability in English/Chinese generation, instruction following and multi-turn interaction. BayLing can be effortlessly deployed on a consumer-grade GPU with 16GB of memory, and assists users with tasks such as translation, writing, creation, suggestion...

### BBT-FinCUGE-Applications

https://github.com/ssymmetry/BBT-FinCUGE-Applications

https://arxiv.org/abs/2302.09432

https://bbt.ssymmetry.com/index.html

1.目前最大规模的中文金融领域开源语料库BBT-FinCorpus。预训练语料库的规模与多样性对PLM的性能和泛化能力具有重要作用，所以为了更好的训练PLM，首先需要搜集大规模多样性的语料库。然而，目前中文金融领域缺乏大规模多样性开源语料库，已有的中文金融领域模型多数基于小规模的私有语料库，严重限制了中文金融PLM的能力提升。为此，我们构建了BBT-FinCorpus，一个包含有从四种异质性来源获取的约300GB文本的大规模多样性语料库。针对如何确定语料库的覆盖范围和语料来源集合的问题，我们首先搜集了中文互联网上可获取的所有中文金融NLP任务数据集，并根据其文本来源分布来确定所需要爬取的文本来源集合。在确认好需要爬取的文本来源集合之后，我们使用基于代理的分布式爬虫技术实现大规模爬取网页上的文本。

2.目前最大规模的中文金融领域知识增强型预训练语言模型BBT-FinT5。PLM的架构与参数量对其性能有重要影响。现有的中文金融领域PLM都基于较为原始的BERT模型架构，参数量也相对较小，不能满足日益丰富的领域NLP需求。因此，我们基于T5模型架构构建了一个拥有十亿参数量的目前最大规模的中文金融领域预训练语言模型BBT-FinT5。为了在有限的硬件算力条件下，尽可能高效地利用好硬件算力，我们使用DeepSpeed加速框架对预训练过程进行效率优化。此外，我们还针对T5模型设计了独特的知识增强预训练方法，通过实验证明了该方法的有效性。

3.首个中文金融领域自然语言处理评测基准CFLEB。现有的自然语言处理评估基准多是通用领域的，没有公开可用的中文金融领域评测基准。这导致中文金融领域现有的预训练语言模型在不同的任务集合上进行评测，难以相互比较，阻碍了中文金融领域PLM性能的快速提升。为此，我们首先构建了首个中文金融领域自然语言处理评测基准CFLEB，包含六种不同的任务，涵盖对PLM理解与生成能力的评估。针对评测基准任务的选择及其选择标准问题，我们认为领域评测基准应当着重强调任务的实用性，以更好的反映学术界改进PLM对现实世界的帮助。为此，我们首先邀请金融领域专家对所有可获取的中文金融任务进行了实用性评价，筛选出具有较高实用性评分的任务。之后，我们综合任务数据集的开源情况确定了六个任务数据集作为最终的评测基准。该评测基准的早期版本命名为FinCUGE，包含八个任务，该版本目前已舍弃。

### BELLE: Bloom-Enhanced Large Language model Engine

https://huggingface.co/BelleGroup

https://github.com/LianjiaTech/BELLE

https://zhuanlan.zhihu.com/p/616079388

本项目目标是促进中文对话大模型开源社区的发展，愿景做能帮到每一个人的LLM Engine。现阶段本项目基于一些开源预训练大语言模型（如BLOOM），针对中文做了优化，模型调优仅使用由ChatGPT生产的数据（不包含任何其他数据）。

本项目基于 Stanford Alpaca ，Stanford Alpaca 的目标是构建和开源一个基于LLaMA的模型。 Stanford Alpaca 的种子任务都是英语，收集的数据也都是英文，因此训练出来的模型未对中文优化。

本项目目标是促进中文对话大模型开源社区的发展。本项目针对中文做了优化，模型调优仅使用由ChatGPT生产的数据（不包含任何其他数据）。

### Bloom

https://huggingface.co/blog/bloom

https://huggingface.co/bigscience/bloom

BLOOM is an autoregressive Large Language Model (LLM), trained to continue text from a prompt on vast amounts of text data using industrial-scale computational resources. As such, it is able to output coherent text in 46 languages and 13 programming languages that is hardl y distinguishable from text written by humans. BLOOM can also be instructed to perform text tasks it hasn't been explicitly trained for, by casting them as text generation tasks.

### BiLLa: A Bilingual LLaMA with Enhanced Reasoning Ability

https://zhuanlan.zhihu.com/p/628688680

https://github.com/Neutralzz/BiLLa

BiLLa是开源的推理能力增强的中英双语LLaMA模型。模型的主要特性有：

较大提升LLaMA的中文理解能力，并尽可能减少对原始LLaMA英文能力的损伤；

训练过程增加较多的任务型数据，利用ChatGPT生成解析，强化模型理解任务求解逻辑；

全量参数更新，追求更好的生成效果。

### BLOOMChat176B

https://mp.weixin.qq.com/s/cY6ORD8CUyXRL0l20EjwqQ

https://sambanova.ai/blog/introducing-bloomchat-176b-the-multilingual-chat-based-llm/

https://huggingface.co/spaces/sambanovasystems/BLOOMChat

https://github.com/sambanova/bloomchat

开源对话模型一直跟闭源模型在多语言能力上存在差距。SambaNova 和斯坦福 Together Computer 开源可商用的多语言聊天模型 BLOOMChat 176B，支持中文。BLOOMChat 在SambaNova 自研芯片 RDU 上完成训练，借助 SambaNova 的独特可重构数据流架构，利用 BLOOM 开源模型的核心能力，通过在 OpenChatKit、Dolly 2.0 和 OASST1 的 OIG 上进行微调。在基于六种语言的早期双盲测试中，BLOOMChat 在66%的测评数据上产生的对话表现优于近期的开源对话模型。同时在与 GPT4 的基于六种语言的人工测评对比中，BLOOMChat 得到 45%对55%的胜率，大大缩小开源和闭源模型的多语言对话能力差距。当前 BLOOMChat 开源模型文件，支持在 huggingface 在线推理试用。

### ChatLaw 法律大模型

https://www.chatlaw.cloud/

https://github.com/PKU-YuanGroup/ChatLaw

https://arxiv.org/pdf/2306.16092.pdf

但愿世间不纷争，何惜法典卷生尘

ChatGPT浪潮下，人工智能的不断扩展和发展为LLM的扩散提供了肥沃的土壤，目前医疗、教育、金融领域已逐渐有了各自的模型，但法律领域迟迟没有明显进展。

为了促进LLM在法律甚至其他垂直应用落地的开放研究，本项目开源了中文法律大模型，并针对LLM和知识库的结合问题给出了法律场景下合理的解决方案。

ChatLaw法律大模型目前开源的仅供学术参考的版本底座为姜子牙-13B、Anima-33B，我们使用大量法律新闻、法律论坛、法条、司法解释、法律咨询、法考题、判决文书等原始文本来构造对话数据。

基于姜子牙-13B的模型是第一版模型，得益于姜子牙的优秀中文能力和我们对数据清洗、数据增强过程的严格要求，我们在逻辑简单的法律任务上表现优异，但涉及到复杂逻辑的法律推理任务时往往表现不佳。

随后基于Anima-33B，我们增加了训练数据，做成了ChatLaw-33B，发现逻辑推理能力大幅提升，由此可见，大参数的中文LLM是至关重要的。

我们的技术报告在这里: arXiv: ChatLaw

基于可商用的模型训练而成的版本会作为我们后续产品内部接入的版本，对外不开源，可以在这里进行开源版本模型的试用

### Chinese-Vicuna-medical

https://github.com/Facico/Chinese-Vicuna/blob/master/docs/performance-medical.md

在cMedQA2上使用我们的checkpoint-11600 continue finetune

目前从2个epoch的Vicuna开始continue finetune，效果比3个epoch的在医疗问答数据更具有专业性，同时由于数据集构建的问题，会更加规范，比如经常性的加上“到正规医院检查”等等

同时验证了指令微调的有效性

使用单指令continue-finetune能保留原来更多的性能

### Cornucopia-LLaMA-Fin-Chinese

https://github.com/jerry1993-tech/Cornucopia-LLaMA-Fin-Chinese

聚宝盆(Cornucopia): 基于中文金融知识的LLaMA微调模型 本项目开源了经过中文金融知识指令精调/指令微调(Instruct-tuning) 的LLaMA-7B模型。通过中文金融公开数据+爬取的金融数据构建指令数据集，并在此基础上对LLaMA进行了指令微调，提高了 LLaMA 在金融领域的问答效果。

基于相同的数据，后期还会利用GPT3.5 API构建高质量的数据集，另在中文知识图谱-金融上进一步扩充高质量的指令数据集

陆续会发布研发的新模型（next-pretrain、multi-task SFT、RLHF Optimize），欢迎大家届时使用体验。

### chatglm-maths

https://github.com/yongzhuo/chatglm-maths

chatglm-6b微调/LORA/PPO/推理, 样本为自动生成的整数/小数加减乘除运算, 可gpu/cpu。

### ChatRWKV

https://github.com/BlinkDL/ChatRWKV

ChatRWKV is like ChatGPT but powered by my RWKV (100% RNN) language model, which is the only RNN (as of now) that can match transformers in quality and scaling, while being faster and saves VRAM. Training sponsored by Stability EleutherAI :)

### ChatYuan

https://github.com/clue-ai/ChatYuan

https://modelscope.cn/models/ClueAI/ChatYuan-large

元语功能型对话大模型, 这个模型可以用于问答、结合上下文做对话、做各种生成任务，包括创意性写作，也能回答一些像法律、新冠等领域问题。它基于PromptCLUE-large结合数亿条功能对话多轮对话数据进一步训练得到。

PromptCLUE-large在1000亿token中文语料上预训练，累计学习1.5万亿中文token，并且在数百种任务上进行Prompt任务式训练。针对理解类任务，如分类、情感分析、抽取等，可以自定义标签体系；针对多种生成任务，可以进行采样自由生成。

### ChatGLM-6B

https://github.com/THUDM/ChatGLM-6B

https://github.com/THUDM/ChatGLM-6B/tree/main/ptuning

ChatGLM-6B 是一个开源的、支持中英双语的对话语言模型，基于 General Language Model (GLM) 架构，具有 62 亿参数。结合模型量化技术，用户可以在消费级的显卡上进行本地部署（INT4 量化级别下最低只需 6GB 显存）。 ChatGLM-6B 使用了和 ChatGPT 相似的技术，针对中文问答和对话进行了优化。经过约 1T 标识符的中英双语训练，辅以监督微调、反馈自助、人类反馈强化学习等技术的加持，62 亿参数的 ChatGLM-6B 已经能生成相当符合人类偏好的回答。更多信息请参考我们的博客。

### ChatGLM2-6B

https://github.com/THUDM/ChatGLM2-6B

ChatGLM2-6B 是开源中英双语对话模型 ChatGLM-6B 的第二代版本，在保留了初代模型对话流畅、部署门槛较低等众多优秀特性的基础之上，ChatGLM2-6B 引入了如下新特性：

更强大的性能：基于 ChatGLM 初代模型的开发经验，我们全面升级了 ChatGLM2-6B 的基座模型。ChatGLM2-6B 使用了 GLM 的混合目标函数，经过了 1.4T 中英标识符的预训练与人类偏好对齐训练，评测结果显示，相比于初代模型，ChatGLM2-6B 在 MMLU（+23%）、CEval（+33%）、GSM8K（+571%） 、BBH（+60%）等数据集上的性能取得了大幅度的提升，在同尺寸开源模型中具有较强的竞争力。

更长的上下文：基于 FlashAttention 技术，我们将基座模型的上下文长度（Context Length）由 ChatGLM-6B 的 2K 扩展到了 32K，并在对话阶段使用 8K 的上下文长度训练，允许更多轮次的对话。但当前版本的 ChatGLM2-6B 对单轮超长文档的理解能力有限，我们会在后续迭代升级中着重进行优化。

更高效的推理：基于 Multi-Query Attention 技术，ChatGLM2-6B 有更高效的推理速度和更低的显存占用：在官方的模型实现下，推理速度相比初代提升了 42%，INT4 量化下，6G 显存支持的对话长度由 1K 提升到了 8K。

> 更开放的协议：ChatGLM2-6B 权重对学术研究完全开放，在获得官方的书面许可后，亦允许商业使用。如果您发现我们的开源模型对您的业务有用，我们欢迎您对下一代模型 ChatGLM3 研发的捐赠。

### Chinese-Transformer-XL

> https://github.com/THUDM/Chinese-Transformer-XL

本项目提供了智源研究院"文汇" 预训练模型Chinese-Transformer-XL的预训练和文本生成代码。

### ChatMed-TCM & ChatMed-Consult

> https://github.com/michael-wzhu/ChatMed

🚀 ChatMed-Consult : 基于中文医疗在线问诊数据集ChatMed_Consult_Dataset的50w+在线问诊+ChatGPT回复作为训练集。模型主干为LlaMA-7b,融合了Chinese-LlaMA-Alpaca的LoRA权重与中文扩展词表，然后再进行基于LoRA的参数高效微调。我们将全部代码都进行了公开。我们也将部署一个在线Gradio demo, 敬请关注。

⏳ ChatMed-TCM : 大模型赋能中医药传承。这一模型的训练数据为中医药指令数据集ChatMed_TCM_Dataset。以我们开源的中医药知识图谱为基础，采用以实体为中心的自指令方法(entity-centric self-instruct)，调用ChatGPT得到2.6w+的围绕中医药的指令数据。ChatMed-TCM模型也是以LlaMA为底座，采用LoRA微调得到。

### ChatGLM-Med

> https://github.com/SCIR-HI/Med-ChatGLM

基于中文医学知识的ChatGLM模型微调，本项目开源了经过中文医学指令精调/指令微调(Instruct-tuning) 的ChatGLM-6B模型。我们通过医学知识图谱和GPT3.5 API构建了中文医学指令数据集，并在此基础上对ChatGLM-6B进行了指令微调，提高了ChatGLM在医疗领域的问答效果。

### CPM-Bee

> https://mp.weixin.qq.com/s/UCW1BT60Lr9x24Rj0cLuxw
>
> https://huggingface.co/openbmb/cpm-bee-10b
>
> https://github.com/OpenBMB/CPM-Bee

CPM-Bee 是一个 完全开源、允许商用 的百亿参数中英文基座模型。它采用 Transformer 自回归架构（auto-regressive），使用万亿级高质量语料进行预训练，拥有强大的基础能力。CPM-Bee 的特点可以总结如下：

开源可商用：OpenBMB 始终秉承“让大模型飞入千家万户”的开源精神，CPM-Bee 基座模型将完全开源并且可商用，以推动大模型领域的发展。如需将模型用于商业用途，只需企业实名邮件申请并获得官方授权证书，即可商用使用。

中英双语性能优异：CPM-Bee 基座模型在预训练语料上进行了严格的筛选和配比，同时在中英双语上具有亮眼表现，具体可参见评测任务和结果。

超大规模高质量语料：CPM-Bee基座模型在万亿级语料上进行训练，是开源社区内经过语料最多的模型之一。同时，我们对预训练语料进行了严格的筛选、清洗和后处理以确保质量。

OpenBMB大模型系统生态支持：OpenBMB 大模型系统在高性能预训练、适配、压缩、部署、工具开发了一系列工具，CPM-Bee 基座模型将配套所有的工具脚本，高效支持开发者进行进阶使用。

强大的对话和工具使用能力：结合OpenBMB 在指令微调和工具学习的探索，我们在 CPM-Bee 基座模型的基础上进行微调，训练出了具有强大对话和工具使用能力的实例模型，现已开放定向邀请内测，未来会逐步向公众开放。

### * 【Data-Copilot】

> https://github.com/zwq2018/Data-Copilot
>
> https://arxiv.org/abs/2306.07209
>
> https://huggingface.co/spaces/zwq2018/Data-Copilot

Data-Copilot 是一个基于 LLM 的系统，用于处理与数据相关的任务，连接了数十亿条数据和多样化的用户需求。它独立设计接口工具，以高效地管理、调用、处理和可视化数据。在接收到复杂请求时，Data-Copilot 会自主调用这些自设计的接口，构建一个工作流程来满足用户的意图。在没有人类协助的情况下，它能够熟练地将来自不同来源、不同格式的原始数据转化为人性化的输出，如图形、表格和文本。

### DoctorGLM

https://github.com/xionghonglin/DoctorGLM

DoctorGLM，基于 ChatGLM-6B的中文问诊模型。

## EduChat

https://github.com/icalk-nlp/EduChat

教育是影响人的身心发展的社会实践活动，旨在把人所固有的或潜在的素质自内而外激发出来。因此，必须贯彻“以人为本”的教育理念，重点关注人的个性化、引导式、身心全面发展。为了更好地助力”以人为本“的教育，华东师范大学计算机科学与技术学院的EduNLP团队探索了针对教育垂直领域的对话大模型EduChat相关项目研发。该项目主要研究以预训练大模型为基底的教育对话大模型相关技术，融合多样化的教育垂直领域数据，辅以指令微调、价值观对齐等方法，提供教育场景下自动出题、作业批改、情感支持、课程辅导、高考咨询等丰富功能，服务于广大老师、学生和家长群体，助力实现因材施教、公平公正、富有温度的智能教育。

## EVA: 大规模中文开放域对话系统

https://github.com/thu-coai/EVA

EVA 是目前最大的开源中文预训练对话模型，拥有28亿参数，主要擅长开放域闲聊，目前有 1.0 和 2.0 两个版本。其中，1.0版本在 WudaoCorpus-Dialog 上训练而成，2.0 版本在从 WudaoCorpus-Dialog 中清洗出的更高质量的对话数据上训练而成，模型性能也明显好于 EVA1.0。

## GPT2 for Multiple Language

https://github.com/imcaspar/gpt2-ml

简化整理 GPT2 训练代码（based on Grover, supporting TPUs）

移植 bert tokenizer，添加多语言支持

15亿参数 GPT2 中文预训练模型( 15G 语料，训练 10w 步 )

开箱即用的模型生成效果 demo #

15亿参数 GPT2 中文预训练模型( 30G 语料，训练 22w 步 )

## InternLM 书生・浦语

https://github.com/InternLM
https://mp.weixin.qq.com/s/oTXnvWZJVdoOpFLHngbTYQ
https://intern-ai.org.cn/home

InternLM has open-sourced a 7 billion parameter base model and a chat model tailored for practical scenarios. The model has the following characteristics:

It leverages trillions of high-quality tokens for training to establish a powerful knowledge base. It supports an 8k context window length, enabling longer input sequences and stronger reasoning capabilities.

It provides a versatile toolset for users to flexibly build their own workflows. Additionally, a lightweight training framework is offered to support model pre-training without the need for extensive dependencies. With a single codebase, it supports pre-training on large-scale clusters with thousands of GPUs, and fine-tuning on a single GPU while achieving remarkable performance optimizations. InternLM achieves nearly 90% acceleration efficiency during training on 1024 GPUs.

## LaWGPT

https://github.com/pengxiao-song/LaWGPT

LaWGPT 是一系列基于中文法律知识的开源大语言模型。

该系列模型在通用中文基座模型（如 Chinese-LLaMA、ChatGLM 等）的基础上扩充法律领域专有词表、大规模中文法律语料预训练，增强了大模型在法律领域的基础语义理解能力。在此基础上，构造法律领域对话问答数据集、中国司法考试数据集进行指令精调，提升了模型对法律内容的理解和执行能力。

## Lawyer LLaMA

https://github.com/AndrewZhe/lawyer-llama

Lawyer LLaMA 首先在大规模法律语料上进行了continual pretraining，让它系统的学习中国的法律知识体系。 在此基础上，我们借助ChatGPT收集了一批对中国国家统一法律职业资格考试客观题（以下简称法考）的分析和对法律咨询的回答，利用收集到的数据对模型进行指令微调，让模型习得将法律知识应用到具体场景中的能力。

我们的模型能够：

> 掌握中国法律知识： 能够正确的理解民法、刑法、行政法、诉讼法等常见领域的法律概念。例如，掌握了刑法中的犯罪构成理论，能够从刑事案件的事实描述中识别犯罪主体、犯罪客体、犯罪行为、主观心理状态等犯罪构成要件。模型利用学到的法律概念与理论，能够较好回答法考中的大部分题目。
>
> 应用于中国法律实务：能够以通俗易懂的语言解释法律概念，并且进行基础的法律咨询，涵盖婚姻、借贷、海商、刑事等法律领域。
>
> 为了给中文法律大模型的开放研究添砖加瓦，本项目将开源一系列法律领域的指令微调数据和基于LLaMA训练的中文法律大模型的参数。

### LexiLaw

> https://github.com/CSHaitao/LexiLaw

LexiLaw 是一个经过微调的中文法律大模型，它基于 ChatGLM-6B 架构，通过在法律领域的数据集上进行微调，使其在提供法律咨询和支持方面具备更高的性能和专业性。

该模型旨在为法律从业者、学生和普通用户提供准确、可靠的法律咨询服务。无论您是需要针对具体法律问题的咨询，还是对法律条款、案例解析、法规解读等方面的查询，LexiLaw 都能够为您提供有益的建议和指导。

同时，我们将分享在大模型基础上微调的经验和最佳实践，以帮助社区开发更多优秀的中文法律大模型，推动中文法律智能化的发展。

### LawGPT_zh 中文法律大模型（獬豸）

> https://mp.weixin.qq.com/s/Pk4NdFQq5G6iZ3QmcyyFUg
>
> https://github.com/LiuHC0428/LAW-GPT

我们的愿景是为让所有人在遇到法律问题时能第一时间获得专业可靠的回答。因为专业的律师服务只有真正触手可及，才会让人们习惯运用，一如二十年前的搜索引擎，十年前的快递业务。我们希望让法律走进日常生活，为构建法治社会贡献我们的力量。项目海报由Midjourney生成。

本项目开源的中文法律通用模型由ChatGLM-6B LoRA 16-bit指令微调得到。数据集包括现有的法律问答数据集和基于法条和真实案例指导的self-Instruct构建的高质量法律文本问答，提高了通用语言大模型在法律领域的表现，提高了模型回答的可靠性和专业程度。

### Linly伶荔说

> https://github.com/CVI-SZU/Linly
>
> https://mp.weixin.qq.com/s/zSxsArP1pxYNubNDZua7iA

“伶荔说”模型具有以下优势：1. 在32*A100 GPU上训练了不同量级和功能的中文模型，对模型充分训练并提供强大的baseline。据我们所知33B的Linly-Chinese-LLAMA是目前最大的中文LLaMA模型。2. 公开所有训练数据、代码、参数细节以及实验结果，确保项目的可复现性，用户可以选择合适的资源直接用于自己的流程中。3. 项目具有高兼容性和易用性，提供可用于CUDA和CPU的量化推理框架，并支持Huggingface格式。

目前公开可用的模型有：

Linly-Chinese-LLaMA：中文基础模型，基于LLaMA在高质量中文语料上增量训练强化中文语言能力，现已开放 7B、13B 和 33B 量级，65B正在训练中。

Linly-ChatFlow：中文对话模型，在400万指令数据集合上对中文基础模型指令精调，现已开放7B、13B对话模型。

Linly-ChatFlow-int4 ： ChatFlow 4-bit量化版本，用于在CPU上部署模型推理。

进行中的项目： Linly-Chinese-BLOOM：基于BLOOM中文增量训练的中文基础模型，包含7B和175B模型量级，可用于商业场景。

### Linly伶荔说-Chinese-Falcon

> https://mp.weixin.qq.com/s/AuAG3tw4JI8lHyLkSdM18g
>
> https://github.com/CVI-SZU/Linly

近期，阿联酋阿布扎比的技术创新研究所（TII）开源了 Falcon 系列模型，使用经过筛选的 1 万亿 tokens 进行预训练，并以 Apache 2.0 协议开源，可能是目前效果最好且许可协议最宽松（允许商用）的开源模型。

然而，Falcon 模型在使用上面临和 LLaMA 模型类似的问题：由于模型主要在英文数据集上训练，因此它理解和生成中文的能力偏弱。此外，Falcon 在构建词表时没有加入中文字/词，中文字会被拆分成多个 token 的组合，这导致中文文本会被拆分成更长的 tokens 序列，降低了编码和生成效率。

针对以上问题，“伶荔（Linly）”项目团队以 Falcon 模型为底座扩充中文词表，利用中文和中英平行增量预训练将模型的语言能力迁移学习到中文，实现 Chinese-Falcon。本文从模型结构上分析 Falcon、LLaMA 与传统 GPT 的异同，代码实现细节。并介绍我们的中文 Falcon 训练方案，包括中文字词扩充、数据集构建和训练参数等。

### MeChat (Mental Health Support Chatbot)

> https://github.com/qiuhuachuan/smile
>
> https://huggingface.co/qiuhuachuan/MeChat
>
> https://mechat.fly.dev/

我们的愿景是为让所有人在遇到心理健康问题时能够获得及时、有效的倾听和支持。我们相信，心理健康是每个人的权利，而不是奢侈品。我们的使命是为人们提供平等、全面、易于访问的心理健康服务，无论他们身在何处、面临何种挑战。我们的愿景还包括推动社会对心理健康问题的认识和理解，打破心理健康问题带来的污名和歧视，为创建一个更加健康、包容和平等的社会做出贡献。项目海报取自 flaticon 。

### MedicalGPT

> https://github.com/shibing624/MedicalGPT

MedicalGPT 训练医疗大模型，实现包括二次预训练、有监督微调、奖励建模、强化学习训练。

基于ChatGPT Training Pipeline，本项目实现了领域模型--医疗模型的四阶段训练：

第一阶段：PT(Continue PreTraining)增量预训练，在海量领域文档数据上二次预训练GPT模型，以注入领域知识

第二阶段：SFT(Supervised Fine-tuning)有监督微调，构造指令微调数据集，在预训练模型基础上做指令精调，以对齐指令意图

第三阶段：RM(Reward Model)奖励模型建模，构造人类偏好排序数据集，训练奖励模型，用来对齐人类偏好，主要是"HHH"原则，具体是" helpful, honest, harmless"

第四阶段：RL(Reinforcement Learning)基于人类反馈的强化学习(RLHF)，用奖励模型来训练SFT模型，生成模型使用奖励或惩罚来更新其策略，以便生成更高质量、更符合人类偏好的文本

### MedicalGPT-zh

> github.com/MediaBrain-SJTU/MedicalGPT-zh

该开源了基于ChatGLM-6B LoRA 16-bit指令微调的中文医疗通用模型。基于共计28科室的中文医疗共识与临床指南文本，我们生成医疗知识覆盖面更全，回答内容更加精准的高质量指令数据集。

### OpenKG-KnowLLM

> https://github.com/zjunlp/KnowLLM

Knowledgable Large Language Model Series.

With the rapid development of deep learning technology, large language models such as ChatGPT have achieved significant success in the field of natural language processing. However, these large models still face some challenges and issues in learning and understanding knowledge, including the difficulty of knowledge updating, and issues with potential errors and biases within the model, known as knowledge fallacies. The Deep Model series aims to release a series of open-source large models to mitigate these knowledge fallacy issues. The first phase of this project released a knowledge extraction large model based on LLaMA, named Zhishi. To provide Chinese capabilities without disrupting the original model's distribution, we firstly (1) use Chinese corpora for the full-scale pre-training of LLaMA (13B), in order to improve the model's understanding of Chinese and knowledge reserve as much as possible while retaining its original English and code capabilities; Then (2) we fine-tune the model from the first step using an instruction dataset, to enhance the language model's understanding of human extraction instructions.

### OpenMEDLab 浦医

> https://github.com/OpenMEDLab
>
> https://github.com/openmedlab/PULSE
>
> https://stcsm.sh.gov.cn/xwzx/kjzl/20230630/c783c30d8e62494e83073535f841675f.html

OpenMEDLab is an open-source platform to share medical foundation models in multi-modalities, e.g., medical imaging, medical NLP, bioinformatics, protein, etc. It targets promoting novel approaches to long-tail problems in medicine, and meanwhile, it seeks solutions to achieve lower cost, higher efficiency, and better generalizability in training medical AI models. The new learning paradigm of adapting foundation models to downstream applications makes it possible to develop innovative solutions for cross-domain and cross-modality diagnostic tasks efficiently. OpenMEDLab is distinguished by several features:

- World's first open-source platform for medical foundation models.
- 10+ medical data modalities targeting a variety of clinical and research problems.
- Pioneering works of the new learning paradigm using foundation models, including pre-trained models, code, and data.
- Releasing multiple sets of medical data for pre-training and downstream applications.
- Collaboration with top medical institutes and facilities.

### PromptCLUE

- https://github.com/clue-ai/PromptCLUE

PromptCLUE：大规模多任务Prompt预训练中文开源模型。

中文上的三大统一：统一模型框架，统一任务形式，统一应用方式。

支持几十个不同类型的任务，具有较好的零样本学习能力和少样本学习能力。针对理解类任务，如分类、情感分析、抽取等，可以自定义标签体系；针对生成任务，可以进行采样自由生成。

千亿中文token上大规模预训练，累计学习1.5万亿中文token，亿级中文任务数据上完成训练，训练任务超过150+。比base版平均任务提升7个点+；具有更好的理解、生成和抽取能力，并且支持文本改写、纠错、知识图谱问答。

### SkyText-Chinese-GPT3

- https://github.com/SkyWorkAIGC/SkyText-Chinese-GPT3

SkyText是由奇点智源发布的中文GPT3预训练大模型，可以进行聊天、问答、中英互译等不同的任务。 应用这个模型，除了可以实现基本的聊天、对话、你问我答外，还能支持中英文互译、内容续写、对对联、写古诗、生成菜谱、第三人称转述、创建采访问题等多种功能。

### ShenNong-TCM-LLM

- https://github.com/michael-wzhu/ShenNong-TCM-LLM

为推动LLM在中医药领域的发展和落地，提升LLM的在中医药方面的知识与回答医学咨询的能力，同时推动大模型赋能中医药传承，我们现推出ShenNong中医药大规模语言模型:

🚀 ShenNong-TCM :

- 这一模型的训练数据为中医药指令数据集ShenNong_TCM_Dataset。
- ChatMed_TCM_Dataset以我们开源的中医药知识图谱为基础；
- 采用以实体为中心的自指令方法entity-centric self-instruct，调用ChatGPT得到11w+的围绕中医药的指令数据；
- ShenNong-TCM模型也是以LlaMA为底座，采用LoRA (rank=16)微调得到。微调代码与ChatMed代码库相同

### TableGPT

- https://github.com/ZJU-M3/TableGPT-techreport

TableGPT is a specifically designed for table analysis. By unifying tables, natural language, and commands into one model, TableGPT comprehends tabular data, understands user intent through natural language, dissects the desired actions, and executes external commands on the table. It subsequently returns the processed results in both tabular and textual explanations to the user. This novel approach simplifies the way users engage with table data, bringing an intuitive feel to data analysis.

### TechGPT

- https://mp.weixin.qq.com/s/nF1He7jhAHfh7PzhjqHoZg
- https://huggingface.co/neukg/TechGPT-7B
- https://github.com/neukg/TechGPT

2023年6月26日，“东北大学知识图谱研究组”正式发布大语言模型TechGPT。

TechGPT的名字主要来源于小组在2018年推出的TechKG大规模中文学术多领域的知识库。

与当前其他各类大模型相比，TechGPT主要强化了以“知识图谱构建”为核心的关系三元组抽取等各类信息抽取任务、以“逻辑推理”为核心的机器阅读理解等各类智能问答任务、以“文本理解”为核心的关键词生成等各类序列生成任务。

在这三大自然语言处理核心能力之内，TechGPT还具备了对计算机科学、材料、机械、冶金、金融和航空航天等十余种垂直专业领域自然语言文本的处理能力。

### TigerBot

> https://github.com/TigerResearch/TigerBot

TigerBot 是一个多语言多任务的大规模语言模型(LLM)。根据 OpenAI InstructGPT 论文在公开 NLP 数据集上的自动评测，TigerBot-7B 达到 OpenAI 同样大小模型的综合表现的 96%，并且这只是我们的 MVP，在此我们将如下探索成果开源：

> 模型：TigerBot-7B, TigerBot-7B-base，TigerBot-180B (research version)，
>
> 代码：基本训练和推理代码，包括双卡推理 180B 模型的量化和推理代码，
>
> 数据：预训练 100G，从 2TB 过滤后的数据中经过去噪去重清洗而得；监督微调 1G 或 100 万条数据，按比例涵盖用户指令常见的 10 大类 120 小类任务，
>
> API: chat, plugin, finetune, 让用户能在半小时内无代码的训练和使用专属于自己的大模型和数据，
>
> 领域数据：涵盖金融，法律，百科，广邀大模型应用开发者，一起打造中国的世界级的应用。

我们在 BLOOM 基础上，在模型架构和算法上做了如下优化：

> 指令完成监督微调的创新算法以获得更好的可学习型(learnability)，
>
> 运用 ensemble 和 probabilistic modeling 的方法实现更可控的事实性(factuality)和创造性(generativeness)，
>
> 在并行训练上，我们突破了 deep-speed 等主流框架中若干内存和通信问题，使得在千卡环境下数月无间断，
>
> 对中文语言的更不规则的分布，从 tokenizer 到训练算法上做了更适合的算法优化。

### YuLan-Chat

> https://github.com/RUC-GSAI/YuLan-Chat
>
> https://mp.weixin.qq.com/s/nPS4N3stAAG_51fnZANbMA

中国人民大学高瓴人工智能学院相关研究团队（由多位学院老师联合指导）展开了一系列关于指令微调技术的研究，并发布了学院初版大语言对话模型——YuLan-Chat，旨在探索和提升大语言模型的中英文双语对话能力。

我们分别开源了13B和65B的YuLan-Chat模型文件及相关代码，并采用量化技术使其分别可以在单张RTX3090-24G和A800-80G显卡上部署。YuLan-Chat模型基于LLaMA底座模型，采用精心优化的高质量中英文混合指令进行微调，其中YuLan-Chat-65B模型目前能够在中英文相关评测数据集上显著超越已有开源模型效果。后续我们会继续优化指令微调方法与底座模型，持续更新YuLan-Chat模型。

### Ziya-LLaMA

> https://huggingface.co/IDEA-CCNL/Ziya-LLaMA-13B-v1
>
> https://github.com/IDEA-CCNL/Fengshenbang-LM
>
> https://mp.weixin.qq.com/s/IeXgq8blGoeVbpIlAUCAjA

姜子牙通用大模型V1是基于LLaMa的130亿参数的大规模预训练模型，具备翻译，编程，文本分类，信息抽取，摘要，文案生成，常识问答和数学计算等能力。目前姜子牙通用大模型已完成大规模预训练、多任务有监督微调和人类反馈学习三阶段的训练过程。

The Ziya-LLaMA-13B-v1 is a large-scale pre-trained model based on LLaMA with 13 billion parameters. It has the ability to perform tasks such as translation, programming, text classification, information extraction, summarization, copywriting, common sense Q&A, and mathematical calculation. The Ziya-LLaMA-13B-v1 has undergone three stages of training: large-scale continual pre-training (PT), multi-task supervised fine-tuning (SFT), and human feedback learning (RM, PPO).

## 2 训练/推理

### 高效对齐算法RAFT「木筏」

> https://github.com/OptimalScale/LMFlow
>
> https://arxiv.org/abs/2304.06767
>
> https://optimalscale.github.io/LMFlow/examples/raft.html

An extensible, convenient, and efficient toolbox for finetuning large machine learning models, designed to be user-friendly, speedy and reliable, and accessible to the entire community.

### Alpaca-LoRA

https://github.com/tloen/alpaca-lora

Low-Rank LLaMA Instruct-Tuning

This repository contains code for reproducing the Stanford Alpaca results using low-rank adaptation (LoRA). We provide an Instruct model of similar quality to text-davinci-003 that can run on a Raspberry Pi (for research), and the code can be easily extended to the 13b, 30b, and 65b models.

In addition to the training code, which runs within five hours on a single RTX 4090, we publish a script for downloading and inference on the foundation model and LoRA, as well as the resulting LoRA weights themselves. To fine-tune cheaply and efficiently, we use Hugging Face's PEFT as well as Tim Dettmers' bitsandbytes.

Without hyperparameter tuning or validation-based checkpointing, the LoRA model produces outputs comparable to the Stanford Alpaca model. (Please see the outputs included below.) Further tuning might be able to achieve better performance; I invite interested users to give it a try and report their results.

### AlpacaFarm

https://mp.weixin.qq.com/s/CIF2F5Vx_RSN1-LwU_ppOQ

https://tatsu-lab.github.io/alpaca_farm_paper.pdf

https://github.com/tatsu-lab/alpaca_farm

主流的大型语言模型训练都离不开RLHF(人工反馈强化学习)，其主要思想是使用人类专家提供的反馈示例来指导模型的学习过程，它可以加速强化学习过程，提高大模型的性能，但「目前RLHF这个过程既复杂又昂贵」。

针对RLHF这个问题，学术界目前主要有两种解决方法：「1）避开RLHF」，比如Meta最近研究的“Meta最新模型：LIMA-65B，没有RLHF，模型效果远胜Alpaca！！”，验证了精心制作的少量标注数据同样能达到不错的效果。2）「简化RLHF」，就是今天给大家分享的这篇文章：斯坦福发布了一个名为AlpacaFarm（羊驼农场）的模拟器，旨在降低训练语言模型的成本，且比人工成本低45倍，并表现出与人类反馈的高度一致性，同时也为RLHF的研究开辟了新的道路。

### ColossalAI

https://github.com/hpcaitech/ColossalAI

Colossal-AI: Making large AI models cheaper, faster and more accessible

Colossal-AI provides a collection of parallel components for you. We aim to support you to write your distributed deep learning models just like how you write your model on your laptop. We provide user-friendly tools to kickstart distributed training and inference in a few lines.

### ChatLLaMA

https://github.com/nebuly-ai/nebullvm/tree/main/apps/accelerate/chatllama

ChatLLaMA 🦙 has been designed to help developers with various use cases, all related to RLHF training and optimized inference.

ChatLLaMA is a library that allows you to create hyper-personalized ChatGPT-like assistants using your own data and the least amount of compute possible. Instead of depending on one large assistant that “rules us all”, we envision a future where each of us can create our own personalized version of ChatGPT-like assistants. Imagine a future where many ChatLLaMAs at the "edge" will support a variety of human's needs. But creating a personalized assistant at the "edge" requires huge optimization efforts on many fronts: dataset creation, efficient training with RLHF, and inference optimization.

### Chinese-Guanaco

https://github.com/jianzhnie/Chinese-Guanaco

This is the repo for the Chinese-Guanaco project, which aims to build and share instruction-following Chinese LLaMA/Pythia/GLM model tuning methods which can be trained on a single Nvidia RTX-2080TI, multi-round chatbot which can be trained on a single Nvidia RTX-3090 with the context len 2048.

Chinese-Guanaco uses bitsandbytes for quantization and is integrated with Huggingface's PEFT and transformers libraries.

## DPO (Direct Preference Optimization)

https://arxiv.org/abs/2305.18290

https://zhuanlan.zhihu.com/p/641045324

https://huggingface.co/lyogavin/Anima33B-DPO-Belle-1k-merged

https://github.com/lyogavin/Anima/tree/main/rlhf

DPO的核心原理是：PPO训练难度核心是因为需要通过reward model来表达偏好，进行强化学习。

为了不再依赖于reward model进行强化学习，他进行了一系列的数学变换，直接推导出了基于Policy Language Model的标注偏好的概率表达形式，从而可以直接求解一个Language Model的最大似然估计。不再需要复杂繁琐的reward model和强化学习。

While large-scale unsupervised language models (LMs) learn broad world knowledge and some reasoning skills, achieving precise control of their behavior is difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability collect human labels of the relative quality of model generations and fine-tune the unsupervised LM to align with these preferences, often with reinforcement learning from human feedback (RLHF). However, RLHF is a complex and often unstable procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the large unsupervised LM using reinforcement learning to maximize this estimated reward without drifting too far from the original model. In this paper, we leverage a mapping between reward functions and optimal policies to show that this constrained reward maximization problem can be optimized exactly with a single stage of policy training, essentially solving a classification problem on the human preference data. The resulting algorithm, which we call Direct Preference Optimization (DPO), is stable, performant and computationally lightweight, eliminating the need for fitting a reward model, sampling from the LM during fine-tuning, or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds RLHF's ability to control sentiment of generations and improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train.

## DialogADV：Evaluate What You Can't Evaluate: Unassessable Generated Responses Quality

https://github.com/misonsky/DialogADV

https://mp.weixin.qq.com/s/Ga0a6a1L6CmCXgk6WDz0Xg

https://arxiv.org/abs/2305.14658

我们构建了两个具有挑战的元验证对话数据集，通过实验分析表明

大型语言模型作为评估器评估对话文本生成质量仍然存在很多问题：1）LLMs无法识别与事实不一致的、虚构的回复，对不合理的回复仍然给出较高的评价；2） LLMs自身的知识有限，对于依赖知识的样例大语言模型无法依靠自身的知识给出合理的判断；3）LLMs利用外部知识的能力有待提高。在给定外部知识的情况下，LLMs仍然会对不合理的回复给出较高的评价。

## DeepSpeed-Chat

https://mp.weixin.qq.com/s/t3HA4Hu61LLDC3h2Njmo_Q

https://github.com/microsoft/DeepSpeed

微软宣布开源 DeepSpeed-Chat，帮助用户轻松训练类 ChatGPT 等大语言模型。

据悉，Deep Speed Chat 是基于微软 Deep Speed 深度学习优化库开发而成，具备训练、强化推理等功能，还使用了 RLHF（基于人类反馈的强化学习）技术，可将训练速度提升 15 倍以上，而成本却大大降低。

## FlexGen

https://github.com/FMInference/FlexGen

FlexGen is a high-throughput generation engine for running large language models with limited GPU memory. FlexGen allows high-throughput generation by IO-efficient offloading, compression, and large effective batch sizes.

Limitation. As an offloading-based system running on weak GPUs, FlexGen also has its limitations. FlexGen can be significantly slower than the case when you have enough powerful GPUs to hold the whole model, especially for small-batch cases. FlexGen is mostly optimized for throughput-oriented batch processing settings (e.g., classifying or extracting information from many documents in batches), on single GPUs.

## FlagAI and FlagData

https://github.com/FlagAI-Open/FlagAI

FlagAI (Fast LArge-scale General AI models) is a fast, easy-to-use and extensible toolkit for large-scale model. Our goal is to support training, fine-tuning, and deployment of large-scale models on various downstream tasks with multi-modality.

https://github.com/FlagOpen/FlagData

FlagData, a data processing toolkit that is easy to use and expand. FlagData integrates the tools and algorithms of multi-step data processing, including cleaning, condensation, annotation and analysis, providing powerful data processing support for model training and deployment in multiple fields, including natural language processing and computer vision.

### Guanaco & QloRA

https://mp.weixin.qq.com/s/SGJQHsEJTNB6hiVqdc87sg

https://arxiv.org/abs/2305.14314

https://github.com/artidoro/qlora

https://huggingface.co/blog/hf-bitsandbytes-integration

Integration: https://colab.research.google.com/drive/1ge2F1QSK8Q7h0hn3YKuBCOAS0bK8E0wf?usp=sharing

Training: https://colab.research.google.com/drive/1VoYNfYDKcKRQRor98Zbf2-9VQTtGJ24k?usp=sharing

We present QLoRA, an efficient finetuning approach that reduces memory usage enough to finetune a 65B parameter model on a single 48GB GPU while preserving full 16-bit finetuning task performance. QLoRA backpropagates gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters (LoRA). Our best model family, which we name Guanaco, outperforms all previous openly released models on the Vicuna benchmark, reaching 99.3% of the performance level of ChatGPT while only requiring 24 hours of finetuning on a single GPU. QLoRA introduces a number of innovations to save memory without sacrificing performance: (a) 4-bit NormalFloat (NF4), a new data type that is information theoretically optimal for normally distributed weights (b) Double Quantization to reduce the average memory footprint by quantizing the quantization constants, and (c) Paged Optimizers to manage memory spikes. We use QLoRA to finetune more than 1,000 models, providing a detailed analysis of instruction following and chatbot performance across 8 instruction datasets, multiple model types (LLaMA, T5), and model scales that would be infeasible to run with regular finetuning (e.g. 33B and 65B parameter models). Our results show that QLoRA finetuning on a small high-quality dataset leads to state-of-the-art results, even when using smaller models than the previous SoTA. We provide a detailed analysis of chatbot performance based on both human and GPT-4 evaluations showing that GPT-4 evaluations are a cheap and reasonable alternative to human evaluation. Furthermore, we find that current chatbot benchmarks are not trustworthy to accurately evaluate the performance levels of chatbots. We release all of our models and code, including CUDA kernels for 4-bit training.

### GPT4All

https://github.com/nomic-ai/gpt4all

Demo, data and code to train an assistant-style large language model with ~800k GPT-3.5-Turbo Generations based on LLaMa

### HugNLP

https://mp.weixin.qq.com/s/IpgOQJ8vrIvnjdrmGCT2FA

https://github.com/HugAILab/HugNLP

https://arxiv.org/abs/2302.14286

华师大HugAILab团队研发了HugNLP框架，这是一个面向研究者和开发者的全面统一的NLP训练框架，可支持包括文本分类、文本匹配、问答、信息抽取、文本生成、小样本学习等多种NLP任务模型搭建和训练。

HugNLP还集成了大量最新的Prompt技术，例如Prompt-Tuning、In-Context Learning、Instruction-tuning，未来还将引入Chain-of-thought

HugAILab团队还研发了一系列的应用，例如CLUE&GLUE刷榜工具，可支持ChatGPT类模型训练和部署产品HugChat，以及统一信息抽取产品HugIE等。

HugNLP是一个分层式框架，遵循“高内聚低耦合”的开发模式，其核心包括模型层（Models）、处理器层（Processors）、评估器层（Evaluators）和应用层（Applications）四部分。

### INSTRUCTEVAL

https://mp.weixin.qq.com/s/E6hq0AUy_hItA5HGo2tCAQ

https://github.com/declare-lab/instruct-eval

https://arxiv.org/abs/2306.04757

本文引入了一个名为INSTRUCTEVAL的新型评估套件。该套件专用于对指令调优大型语言模型的全面评估，相比之前对LLMs的评估方法，该评估策略不仅详细评估了模型解决问题的能力、文字写作能力，而且还严格评估了模型与人类价值的对齐能力。

### LOw-Memory Optimization (LOMO)

https://arxiv.org/abs/2306.09782

https://github.com/OpenLMLab/LOMO

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP) but demand massive GPU resources for training. Lowering the threshold for LLMs training would encourage greater participation from researchers, benefiting both academia and society. While existing approaches have focused on parameter-efficient fine-tuning, which tunes or adds a small number of parameters, few have addressed the challenge of tuning the full parameters of LLMs with limited resources. In this work, we propose a new optimizer, LOw-Memory Optimization (LOMO), which fuses the gradient computation and the parameter update in one step to reduce memory usage. By integrating LOMO with existing memory saving techniques, we reduce memory usage to 10.8% compared to the standard approach (DeepSpeed solution). Consequently, our approach enables the full parameter fine-tuning of a 65B model on a single machine with 8 RTX 3090, each with 24GB memory.

### llama.cpp

https://github.com/ggerganov/llama.cpp

Inference of LLaMA model in pure C/C++

The main goal is to run the model using 4-bit quantization on a MacBook

Plain C/C++ implementation without dependencies

Apple silicon first-class citizen - optimized via ARM NEON

AVX2 support for x86 architectures

Mixed F16 / F32 precision

4-bit quantization support

Runs on the CPU

### MeZO: Fine-Tuning Language Models with Just Forward Passes

https://github.com/princeton-nlp/MeZO

https://arxiv.org/abs/2305.17333

https://mp.weixin.qq.com/s/3RLCVQg2QJGSiDUtx9DgPg

This is the implementation for the paper Fine-Tuning Language Models with Just Forward Passes. In this paper we propose a memory-efficient zeroth-order optimizer (MeZO), adapting the classical zeroth-order SGD method to operate in-place, thereby fine-tuning language models (LMs) with the same memory footprint as inference.

With a single A100 80GB GPU, MeZO can train a 30-billion parameter OPT model, whereas fine-tuning with Adam can train only a 2.7B LM. MeZO demonstrates comparable performance to fine-tuning with backpropagation across multiple tasks, with up to 12× memory reduction. MeZO is also compatible with both full-parameter and parameter-efficient tuning techniques such as LoRA and prefix tuning. We also show that MeZO can effectively optimize non-differentiable objectives (e.g., maximizing accuracy or F1).

### MLC LLM

https://github.com/mlc-ai/mlc-llm

MLC LLM is a universal solution that allows any language models to be deployed natively on a diverse set of hardware backends and native applications, plus a productive framework for everyone to further optimize model performance for their own use cases.

Our mission is to enable everyone to develop, optimize and deploy AI models natively on everyone's devices.

Everything runs locally with no server support and accelerated with local GPUs on your phone and laptops. Supported platforms include:

iPhone, iPad

Metal GPUs and Intel/ARM MacBooks;

AMD, Intel and NVIDIA GPUs via Vulkan on Windows and Linux;

NVIDIA GPUs via CUDA on Windows and Linux;

WebGPU on browsers (through companion project WebLLM).

### PKU-Beaver 河狸 (Safe RLHF)

https://github.com/PKU-Alignment/safe-rlhf

https://mp.weixin.qq.com/s/ZpkgszXbisl5xf63EfTNjQ

北京大学团队开源了名为 PKU-Beaver（河狸）项目，其开源地址为：https://github.com/PKU-Alignment/safe-rlhf。该项目首次公开了 RLHF 所需的数据集、训练和验证代码，是目前首个开源的可复现的 RLHF 基准。同时，为解决人类标注产生的偏见和歧视等不安全因素，北京大学团队首次提出了带有约束的价值对齐技术 CVA（Constrained Value Alignment）。该技术通过对标注信息进行细粒度划分，并结合带约束的安全强化学习方法，显著降低了模型的偏见和歧视，提高了模型的安全性。Beaver使用GPT4进行Evaluation，结果表明，在原有性能保持不变的情况下，Beaver回复的安全性大幅度提升。

### PaLM + RLHF (Pytorch)

https://github.com/lucidrains/PaLM-rlhf-pytorch

Implementation of RLHF (Reinforcement Learning with Human Feedback) on top of the PaLM architecture. Maybe I'll add retrieval functionality too, à la RETRO

### RL4LMs

https://github.com/allenai/RL4LMs

https://rl4lms.apps.allenai.org/

A modular RL library to fine-tune language models to human preferences

We provide easily customizable building blocks for training language models including implementations of on-policy algorithms, reward functions, metrics, datasets and LM based actor-critic policies

### Reinforcement Learning with Language Model

https://github.com/HarderThenHarder/transformers_tasks/tree/main/RLHF

在这个项目中，我们将通过开源项目 trl 搭建一个通过强化学习算法（PPO）来更新语言模型（GPT-2）的几个示例，包括：

基于中文情感识别模型的正向评论生成机器人（No Human Reward）

基于人工打分的正向评论生成机器人（With Human Reward）

基于排序序列（Rank List）训练一个奖励模型（Reward Model）

排序序列（Rank List）标注平台

### SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression

https://github.com/Vahe1994/SpQR

https://arxiv.org/pdf/2306.03078.pdf

https://mp.weixin.qq.com/s/819L-dY54BaVM1vub9OSpQ

SpQR 通过识别和隔离异常权重来工作，这些异常权重会导致特别大的量化误差，研究者将它们以更高的精度存储，同时将所有其他权重压缩到 3-4 位，在 LLaMA 和 Falcon LLMs 中实现了不到 1% 的困惑度相对准确率损失。从而可以在单个 24GB 的消费级 GPU 上运行 33B 参数的 LLM，而不会有任何性能下降，同时还能提高 15% 的速度。

### Scikit-LLM: Sklearn Meets Large Language Models

https://github.com/iryna-kondr/scikit-llm

Seamlessly integrate powerful language models like ChatGPT into scikit-learn for enhanced text analysis tasks.

### Transformer Reinforcement Learning

https://github.com/lvwerra/trl

With trl you can train transformer language models with Proximal Policy Optimization (PPO). The library is built on top of the transformers library by 🤗 Hugging Face. Therefore, pre-trained language models can be directly loaded via transformers. At this point most of decoder architectures and encoder-decoder architectures are supported.

### Train_Transformers_with_INT4

https://mp.weixin.qq.com/s/pyEJJ5AvQqfyncO7CA8eNA

https://arxiv.org/abs/2306.11987

https://github.com/xijiu9/Train_Transformers_with_INT4

Quantizing the activation, weight, and gradient to 4-bit is promising to accelerate neural network training. However, existing 4-bit training methods require custom numerical formats which are not supported by contemporary hardware. In this work, we propose a training method for transformers with all matrix multiplications implemented with the INT4 arithmetic. Training with an ultra-low INT4 precision is challenging. To achieve this, we carefully analyze the specific structures of activation and gradients in transformers to propose dedicated quantizers for them. For forward propagation, we identify the challenge of outliers and propose a Hadamard quantizer to suppress the outliers. For backpropagation, we leverage the structural sparsity of gradients by proposing bit splitting and leverage score sampling techniques to quantize gradients accurately. Our algorithm achieves competitive accuracy on a wide range of tasks including natural language understanding, machine translation, and image classification. Unlike previous 4-bit training methods, our algorithm can be implemented on the current generation of GPUs. Our prototypical linear operator implementation is up to 2.2 times faster than the FP16 counterparts and speeds up the training by up to 35.1%.

### Transformer Reinforcement Learning X

https://github.com/CarperAI/trlx

trlX is a distributed training framework designed from the ground up to focus on fine-tuning large language models with reinforcement learning using either a provided reward function or a reward-labeled dataset.

Training support for 🤗 Hugging Face models is provided by Accelerate-backed trainers, allowing users to fine-tune causal and T5-based language models of up to 20B parameters, such as facebook/opt-6.7b, EleutherAI/gpt-neox-20b, and google/flan-t5-xxl. For models beyond 20B parameters, trlX provides NVIDIA NeMo-backed trainers that leverage efficient parallelism techniques to scale effectively.

### vLLM

https://github.com/vllm-project/vllm

vLLM is a fast and easy-to-use library for LLM inference and serving.

vLLM is fast with:

State-of-the-art serving throughput

Efficient management of attention key and value memory with PagedAttention

Dynamic batching of incoming requests

Optimized CUDA kernels

vLLM is flexible and easy to use with:

Seamless integration with popular HuggingFace models

High-throughput serving with various decoding algorithms, including parallel sampling, beam search, and more

Tensor parallelism support for distributed inference

Streaming outputs

OpenAI-compatible API server

## 3 可参考的其它开源模型

### Cerebras（可商用）

https://www.cerebras.net/blog/cerebras-gpt-a-family-of-open-compute-efficient-large-language-models/

https://huggingface.co/cerebras

开源7个可商用GPT模型，含数据集和可直接下载的预训练模型权重: Cerebras 开源 7 个 GPT 模型，均可商用，参数量分别达到 1.11 亿、2.56 亿、5.9 亿、13 亿、27 亿、67 亿和 130 亿。其中最大的模型参数量达到 130 亿，与 Meta 最近开源的 LLaMA-13B 相当。该项目开源数据集和预训练模型权重，其中预训练模型权重文件大小近50G可直接下载，并且可用于商业和研究用途。与此前的 GPT-3 模型相比，Cerebras 开源的模型具有更高的可用性和透明度，研究人员和开发者可以使用少量数据对其进行微调，构建出高质量的自然语言处理应用。

### ChatDoctor

https://github.com/Kent0n-Li/ChatDoctor

https://arxiv.org/abs/2303.14070

Recent large language models (LLMs) in the general domain, such as ChatGPT, have shown remarkable success in following instructions and producing human-like responses. However, such language models have yet to be adapted for the medical domain, resulting in poor accuracy of responses and an inability to provide sound advice on medical diagnoses, medications, etc. To address this problem, we fine-tuned our ChatDoctor model based on 100k real-world patient-physician conversations from an online medical consultation site. Besides, we add autonomous knowledge retrieval capabilities to our ChatDoctor, for example, Wikipedia or a disease database as a knowledge brain. By fine-tuning the LLMs using these 100k patient-physician conversations, our model showed significant improvements in understanding patients' needs and providing informed advice. The autonomous ChatDoctor model based on Wikipedia and Database Brain can access real-time and authoritative information and answer patient questions based on this information, significantly improving the accuracy of the model's responses, which shows extraordinary potential for the medical field with a low tolerance for error.

### Dolly 1&2（可商用）

https://github.com/databrickslabs/dolly

https://huggingface.co/databricks/dolly-v2-12b

https://www.databricks.com/blog/2023/03/24/hello-dolly-democratizing-magic-chatgpt-open-models.html

We show that anyone can take a dated off-the-shelf open source large language model (LLM) and give it magical ChatGPT-like instruction following ability by training it in 30 minutes on one machine, using high-quality training data. Surprisingly, instruction-following does not seem to require the latest or largest models: our model is only 6 billion parameters, compared to 175 billion for GPT-3. We open source the code for our model (Dolly) and show how it can be re-created on Databricks. We believe models like Dolly will help democratize LLMs, transforming them from something very few companies can afford into a commodity every company can own and customize to improve their products.

### FinGPT

https://github.com/ai4finance-foundation/fingpt

https://arxiv.org/pdf/2306.06031v1.pdf

https://mp.weixin.qq.com/s/A9euFin675nxGGciiX6rJQ

Large language models (LLMs) have shown the potential of revolutionizing natural language processing tasks in diverse domains, sparking great interest in finance. Accessing high-quality financial data is the first challenge for financial LLMs (FinLLMs). While proprietary models like BloombergGPT have taken advantage of their unique data accumulation, such privileged access calls for an open-source alternative to democratize Internet-scale financial data.

In this paper, we present an open-source large language model, FinGPT, for the finance sector. Unlike proprietary models, FinGPT takes a data-centric approach, providing researchers and practitioners with accessible and transparent resources to develop their FinLLMs. We highlight the importance of an automatic data curation pipeline and the lightweight low-rank adaptation technique in building FinGPT. Furthermore, we showcase several potential applications as stepping stones for users, such as robo-advising, algorithmic trading, and low-code development. Through collaborative efforts within the open-source AI4Finance community, FinGPT aims to stimulate innovation, democratize FinLLMs, and unlock new opportunities in open finance.

### Falcon（可商用）

https://mp.weixin.qq.com/s/mKx0ZiTB28khj4U7EVJiVw

https://falconllm.tii.ae/

https://huggingface.co/tiiuae/falcon-40b

Falcon LLM is a foundational large language model (LLM) with 40 billion parameters trained on one trillion tokens. TII has now released Falcon LLM – a 40B model.

The model uses only 75 percent of GPT-3’s training compute, 40 percent of Chinchilla’s, and 80 percent of PaLM-62B’s.

### Facebook/Meta LLaMA/LLaMA2

https://github.com/facebookresearch/llama

https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/

#### LLaMA1

LLaMA: Open and Efficient Foundation Language Models

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community.

**LLaMA2**

We are unlocking the power of large language models. Our latest version of Llama is now accessible to individuals, creators, researchers and businesses of all sizes so that they can experiment, innovate and scale their ideas responsibly.

This release includes model weights and starting code for pretrained and fine-tuned Llama language models — ranging from 7B to 70B parameters.

This repository is intended as a minimal example to load Llama 2 models and run inference. For more detailed examples leveraging HuggingFace, see llama-recipes.

## GALACTICA

https://github.com/paperswithcode/galai

https://arxiv.org/pdf/2211.09085.pdf

https://galactica.org/

GALACTICA is a general-purpose scientific language model. It is trained on a large corpus of scientific text and data. It can perform scientific NLP tasks at a high level, as well as tasks such as citation prediction, mathematical reasoning, molecular property prediction and protein annotation. More information is available at galactica.org.

## Goar-7B for Arithmetic Tasks

https://mp.weixin.qq.com/s/_haINkHNV4bMszm9F41yXA

https://arxiv.org/pdf/2305.14201.pdf

https://github.com/liutiedong/goat

在本文介绍了一种微调的语言模型：Goat。不同于以往对算术计算的研究，该模型在 LLaMA上采用端到端监督指令微调范式，利用包含约100万个样本的综合生成数据集进行训练得到。它非常擅长算术任务。Goat 在初等算术（包括整数的加法、减法、乘法和除法）中实现了最先进的性能。实验结果表明，仅通过监督微调而不应用任何特殊技术，「Goat模型能够在Zero-shot设置中以近乎完美的精度为大数加法和减法生成答案」。这种出色的算术能力归因于 LLaMA 对数字的一致标记化，并表明这对于以前的 LLM 来说几乎是不可能实现的，例如 Bloom 、OPT、GPT-NeoX 、Pythia等。

然而，该模型在面对乘除运算时遇到了很大的挑战。为了克服这一挑战，本文提出了一种方法，即「将各种算术任务分为可学习和不可学习任务」，随后利用基本算术原理将不可学习任务（例如多位数乘法和除法）分解为一系列可学习任务。本文方法确保促进模型学习的中间监督也很容易被人类理解，即通过模型微调在生成最终答案之前生成合适的CoT。「本文方法大大优于 GPT-4 的长乘法和长除法」。最终使用 BIG-bench (Srivastava et al., 2022) 算术子任务评估模型的性能，并对本文方法的有效性进行综合评估。实验结果表明，该模型可以学习计算模式并将其泛化到看不见的数据，而不仅仅是纯粹权重记忆计算。此外，Goat-7B 可以在24GB VRAM GPU上使用LoRA低秩适应技术进行训练，可以「很容易复现论文成果」。

## HuggingChat

https://huggingface.co/chat/

Making the community's best AI chat models available to everyone.

## Koala: A Dialogue Model for Academic Research

https://bair.berkeley.edu/blog/2023/04/03/koala/

In this post, we introduce Koala, a chatbot trained by fine-tuning Meta’s LLaMA on dialogue data gathered from the web. We describe the dataset curation and training process of our model, and also present the results of a user study that compares our model to ChatGPT and Stanford’s Alpaca. Our results show that Koala can effectively respond to a variety of user queries, generating responses that are often preferred over Alpaca, and at least tied with ChatGPT in over half of the cases.

## LongLLaMA

https://mp.weixin.qq.com/s/XzaET7WfrNpOf-zdiSxrig

https://arxiv.org/pdf/2307.03170.pdf

https://github.com/CStanKonrad/long_llama

https://huggingface.co/syzymon/long_llama_3b

This repository contains the research preview of LongLLaMA, a large language model capable of handling long contexts of 256k tokens or even more.

LongLLaMA is built upon the foundation of OpenLLaMA and fine-tuned using the Focused Transformer (FoT) method. We release a smaller 3B variant of the LongLLaMA model on a permissive license (Apache 2.0) and inference code supporting longer contexts on Hugging Face. Our model weights can serve as the drop-in replacement of LLaMA in existing implementations (for short context up to 2048 tokens). Additionally, we provide evaluation results and comparisons against the original OpenLLaMA models. Stay tuned for further updates.

### LLaMA复刻版OpenLLaMA

https://github.com/openlm-research/open_llama

In this repo, we release a permissively licensed open source reproduction of Meta AI's LLaMA large language model. In this release, we're releasing a public preview of the 7B OpenLLaMA model that has been trained with 200 billion tokens. We provide PyTorch and Jax weights of pre-trained OpenLLaMA models, as well as evaluation results and comparison against the original LLaMA models. Stay tuned for our updates.

### Llama-X: Open Academic Research on Improving LLaMA to SOTA LLM

https://github.com/AetherCortex/Llama-X

This is the repo for the Llama-X, which aims to:

Progressively improve the performance of LLaMA to SOTA LLM with open-source community.

Conduct Llama-X as an open academic research which is long-term, systematic and rigorous.

Save the repetitive work of community and we work together to create more and faster increment.

### Lit-LLaMA

https://github.com/Lightning-AI/lit-llama

Lit-LLaMA is:

Simple: Single-file implementation without boilerplate.

Correct: Numerically equivalent to the original model.

Optimized: Runs on consumer hardware or at scale.

Open-source: No strings attached.

### MPT-7B（可商用）

https://www.mosaicml.com/blog/mpt-7b

https://huggingface.co/mosaicml/mpt-7b

MPT-7B is a decoder-style transformer pretrained from scratch on 1T tokens of English text and code. This model was trained by MosaicML.

MPT-7B is part of the family of MosaicPretrainedTransformer (MPT) models, which use a modified transformer architecture optimized for efficient training and inference.

Introducing MPT-7B, the latest entry in our MosaicML Foundation Series. MPT-7B is a transformer trained from scratch on 1T tokens of text and code. It is open source, available for commercial use, and matches the quality of LLaMA-7B. MPT-7B was trained on the MosaicML platform in 9.5 days with zero human intervention at a cost of ~$200k. Starting today, you can train, finetune, and deploy your own private MPT models, either starting from one of our checkpoints or training from scratch. For inspiration, we are also releasing three finetuned models in addition to the base MPT-7B: MPT-7B-Instruct, MPT-7B-Chat, and MPT-7B-StoryWriter-65k+, the last of which uses a context length of 65k tokens!

### OpenGPT

https://github.com/CogStack/OpenGPT

A framework for creating grounded instruction based datasets and training conversational domain expert Large Language Models (LLMs).

NHS-LLM：A conversational model for healthcare trained using OpenGPT. All the medical datasets used to train this model were created using OpenGPT and are available below.

### Orca

https://aka.ms/orca-lm

https://arxiv.org/pdf/2306.02707.pdf

https://mp.weixin.qq.com/s/RRdrSeI2ux5QE6MqJ8opSg

Recent research has focused on enhancing the capability of smaller models through imitation learning, drawing on the outputs generated by large foundation models (LFMs). A number of issues impact the quality of these models, ranging from limited imitation signals from shallow LFM outputs; small scale homogeneous training data; and most notably a lack of rigorous evaluation resulting in overestimating the small model's capability as they tend to learn to imitate the style, but not the reasoning process of LFMs. To address these challenges, we develop Orca (We are working with our legal team to publicly release a diff of the model weights in accordance with LLaMA's release policy to be published at this https URL), a 13-billion parameter model that learns to imitate the reasoning process of LFMs. Orca learns from rich signals from GPT-4 including explanation traces; step-by-step thought processes; and other complex instructions, guided by teacher assistance from ChatGPT. To promote this progressive learning, we tap into large-scale and diverse imitation data with judicious sampling and selection. Orca surpasses conventional state-of-the-art instruction-tuned models such as Vicuna-13B by more than 100% in complex zero-shot reasoning benchmarks like Big-Bench Hard (BBH) and 42% on AGIEval. Moreover, Orca reaches parity with ChatGPT on the BBH benchmark and shows competitive performance (4 pts gap with optimized system message) in professional and academic examinations like the SAT, LSAT, GRE, and GMAT, both in zero-shot settings without CoT; while trailing behind GPT-4. Our research indicates that learning from step-by-step explanations, whether these are generated by humans or more advanced AI models, is a promising direction to improve model capabilities and skills.

### OpenChatKit

https://www.together.xyz/blog/openchatkit

https://huggingface.co/spaces/togethercomputer/OpenChatKit

https://github.com/togethercomputer/OpenChatKit

OpenChatKit uses a 20 billion parameter chat model trained on 43 million instructions and supports reasoning, multi-turn conversation, knowledge and generative answers.

OpenChatKit provides a powerful, open-source base to create both specialized and general purpose chatbots for various applications. The kit includes an instruction-tuned 20 billion parameter language model, a 6 billion parameter moderation model, and an extensible retrieval system for including up-to-date responses from custom repositories. It was trained on the OIG-43M training dataset, which was a collaboration between Together, LAION, and Ontocord.ai. Much more than a model release, this is the beginning of an open source project. We are releasing a set of tools and processes for ongoing improvement with community contributions.

### Open-Assistant

https://github.com/LAION-AI/Open-Assistant

https://open-assistant.io/zh

Open Assistant is a project meant to give everyone access to a great chat based large language model.

We believe that by doing this we will create a revolution in innovation in language. In the same way that stable-diffusion helped the world make art and images in new ways we hope Open Assistant can help improve the world by improving language itself.

### MedLLaMA-13B & PMC-LLaMA: Continue Training LLaMA on Medical Papers

https://github.com/chaoyi-wu/PMC-LLaMA

https://huggingface.co/chaoyi-wu/PMC_LLAMA_7B

https://arxiv.org/abs/2304.14454

We have release a new model MedLLaMA-13B finetuned with LLaMA-13B on some medical corpus, termed as MedLLaMA-13B. It have been proved to be more powerful than both LLaMA-13B and PMC-LLaMA, refering to our benchmark for detail comparison.

### RedPajama（可商用）

https://www.together.xyz/blog/redpajama

https://github.com/togethercomputer/RedPajama-Data

RedPajama, a project to create leading open-source models, starts by reproducing LLaMA training dataset of over 1.2 trillion tokens.

### StableLM

https://zhuanlan.zhihu.com/p/623542189

https://github.com/Stability-AI/StableLM

StableLM: Stability AI Language Models

This repository contains Stability AI's ongoing development of the StableLM series of language models and will be continuously updated with new checkpoints. The following provides an overview of all currently available models. More coming soon.

### StableVicuna

https://github.com/Stability-AI/StableLM

StableVicuna基于小羊驼Vicuna-13B的进一步指令微调和RLHF训练的版本。Vicuna-13B是LLaMA-13B的一个指令微调模型。

### Stanford Alpaca

https://crfm.stanford.edu/2023/03/13/alpaca.html

https://alpaca-ai.ngrok.io/

https://github.com/tatsu-lab/stanford_alpaca

Alpaca: A Strong, Replicable Instruction-Following ModelAl

We introduce Alpaca 7B, a model fine-tuned from the LLaMA 7B model on 52K instruction-following demonstrations. On our preliminary evaluation of single-turn instruction following, Alpaca behaves qualitatively similarly to OpenAI’s text-davinci-003, while being surprisingly small and easy/cheap to reproduce (<600$).

### UltraLM-13B

https://github.com/thunlp/UltraChat

UltraLM is a series of chat language models trained on UltraChat. Currently, we have released the 13B version, which ranks #1 among open-source models and ranks #4 among all models on AlpacaEval Leaderboard. UltraLM-13B is based upon LLaMA-13B.

This project aims to construct open-source, large-scale, and multi-round dialogue data powered by Turbo APIs to facilitate the construction of powerful language models with general conversational capability. In consideration of factors such as safeguarding privacy, we do not directly use any data available on the Internet as prompts. To ensure generation quality, two separate ChatGPT Turbo APIs are adopted in generation, where one plays the role of the user to generate queries and the other generates the response. We instruct the user model with carefully designed prompts to mimic human user behavior and call the two APIs iteratively. The generated dialogues undergo further post-processing and filtering.

### Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality

https://chat.lmsys.org/

https://vicuna.lmsys.org/

https://github.com/lm-sys/FastChat

An open platform for training, serving, and evaluating large language model based chatbots.

### Wombat

https://mp.weixin.qq.com/s/xoPKmOzjlNZ2qGdcKeGARw

https://mp.weixin.qq.com/s/UI-ij5o43ct1efYoNVdQDg

https://arxiv.org/abs/2304.05302v1

https://github.com/GanjinZero/RRHF

This is the repository for RRHF (Rank Response to align Human Feedback) and open-sourced language models Wombat. RRHF helps align large language models with human perference easier.

Reinforcement Learning from Human Feedback (RLHF) enables the alignment of large language models with human preference, improving the quality of interactions between humans and language models. Recent practice of RLHF uses PPO to enable the large language model optimization of such alignment. However, implementing PPO is non-trivial (where the training procedure requires interactive between policy, behavior policy, reward, value model) and it is also tedious to tuning many hyper-parameters. Our motivation is to simplify the alignment between language models with human preference, and our proposed paradigm RRHF (Rank Response from Human Feedback) can achieve such alignment as easily as conventional fine-tuning. It is simpler than PPO from the aspects of coding, model counts, and hyperparameters.

### XGen-7B

https://blog.salesforceairesearch.com/xgen/

https://github.com/salesforce/xgen

We trained a series of 7B LLMs named XGen-7B with standard dense attention on up to 8K sequence length for up to 1.5T tokens. We also fine tune the models on public-domain instructional data. The main take-aways are:

On standard NLP benchmarks, XGen achieves comparable or better results when compared with state-of-the-art open-source LLMs (e.g. MPT, Falcon, LLaMA, Redpajama, OpenLLaMA) of similar model size.

Our targeted evaluation on long sequence modeling benchmarks show benefits of our 8K-seq models over 2K- and 4K-seq models.

XGen-7B archives equally strong results both in text (e.g., MMLU, QA) and code (HumanEval) tasks.

Training cost of $150K on 1T tokens under Google Cloud pricing for TPU-v4.

## 4 评价

### 天秤（FlagEval）

https://flageval.baai.ac.cn/#/home

大语言评测体系及开放平台：构建“能力-任务-指标”三维评测框架，细粒度刻画模型的认知能力边界。

### 獬豸（Xiezhi）Benchmark

https://arxiv.org/abs/2306.05783

https://github.com/MikeGu721/XiezhiBenchmark

Xiezhi是一个综合的、多学科的、能够自动更新的领域知识评估Benchmark。Xiezhi包含了哲学、经济学、法学、教育学、文学、历史学、自然科学、工学、农学、医学、军事学、管理学、艺术学这13个学科门类，24万道学科题目，516个具体学科，249587道题目。这 516 个学科以及分类方式源自中国教育部颁布的学科分类法。作者从中国研究生入学考试中手动选择并注释了 20,000 道多选题，涵盖了这 516 个标签，以形成Xiezhi-Meta数据集。Xiezhi-Meta被用来训练一个能够计算题目和学科标签之间相关性的标注模型。作者们随后收集了来自不同考试的 150,000 个多项选择题，以及来自学术Survey的 70,000 个多项选择题，并使用标注模型对所有这些问题进行了注释。

为了方便进行实验，并能够有效地评估LLM对于跨学科知识的处理能力，作者们提出了Xiezhi-Specialty和Xiezhi-Interdiscipline，这两个数据集都提供了中英文的版本，并由 15,000 个更平衡、更不敏感、更不以中国为中心的多选题组成。 Xiezhi-Specialty 包含可以使用单一领域的知识解决的问题，而 Xiezhi-Interdiscipline 包含需要来自多个领域的知识才能解决的问题。

### C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models

https://arxiv.org/abs/2305.08322

https://cevalbenchmark.com/

https://github.com/SJTU-LIT/ceval

C-Eval is a comprehensive Chinese evaluation suite for foundation models. It consists of 13948 multi-choice questions spanning 52 diverse disciplines and four difficulty levels.

### HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models

https://mp.weixin.qq.com/s/cuoO2V4X-GQOuWyA-e9BeQ

https://arxiv.org/abs/2305.11747

https://github.com/RUCAIBox/HaluEval

为了进一步研究大模型幻象的内容类型和大模型生成幻象的原因，本文提出了用于大语言模型幻象评估的基准——HaluEval。我们基于现有的数据集，通过自动生成和手动标注的方式构建了大量的幻象数据组成HaluEval的数据集，其中包含特定于问答、对话、文本摘要任务的3000

0条样本以及普通用户查询的5000条样本。在本文中，我们详细介绍了HaluEval数据集的构建过程，对构建的数据集进行了内容分析，并初步探索了大模型识别和减少幻象的策略。

### KoLA: Carefully Benchmarking World Knowledge of Large Language Models

https://mp.weixin.qq.com/s/xVj1blhRtpO-Y1HgQ8WI-A

https://arxiv.org/pdf/2306.09296.pdf

https://kola.xlore.cn

KoLA基于19个关注实体、概念和事件的任务。参考了Bloom认知体系，KoLA从知识的记忆、理解、应用和创造4个层级，从深度而非广度去衡量大语言模型处理世界知识的能力。实验结果表明，GPT-4虽然很强，但依然未能霸榜，在知识创造层次的测试中仅排第三名。

### Multiscale Positive-Unlabeled Detection of AI-Generated Texts

https://mp.weixin.qq.com/s/KBN8TMwXD1bcE2X_dImXVg

https://arxiv.org/abs/2305.18149

https://github.com/mindspore-lab/mindone/tree/master/examples/detect_chatgpt

https://github.com/YuchuanTian/AIGC_text_detector

Recent releases of Large Language Models (LLMs), e.g. ChatGPT, are astonishing at generating human-like texts, but they may get misused for fake scholarly texts, fake news, fake tweets, et cetera. Previous works have proposed methods to detect these multiscale AI-generated texts, including simple ML classifiers, pretrained-model-based training-agnostic methods, and finetuned language classification models. However, mainstream detectors are formulated without considering the factor of corpus length: shorter corpuses are harder to detect compared with longer ones for shortage of informative features. In this paper, a Multiscale Positive-Unlabeled (MPU) training framework is proposed to address the challenge of multiscale text detection. Firstly, we acknowledge the human-resemblance property of short machine texts, and rephrase text classification as a Positive-Unlabeled (PU) problem by marking these short machine texts as "unlabeled" during training. In this PU context, we propose the length-sensitive Multiscale PU Loss, where we use a recurrent model in abstraction to estimate positive priors of scale-variant corpuses. Additionally, we introduce a Text Multiscaling module to enrich training corpuses. Experiments show that our MPU method augments detection performance on long AI-generated text, and significantly improves short-corpus detection of language model detectors. Language Models trained with MPU could outcompete existing detectors by large margins on multiscale AI-generated texts.

### PandaLM

https://github.com/WeOpenML/PandaLM

https://zhuanlan.zhihu.com/p/630173415

https://mp.weixin.qq.com/s/HE6jez3G9aEO5qLkvwtKXg

This is the official repository for PandaLM: ReProducible and Automated Language Model Assessment.

PandaLM aims to provide reproducible and automated comparisons between different large language models (LLMs). By giving PandaLM the same context, it can compare the responses of different LLMs and provide a reason for the decision, along with a reference answer. The target audience for PandaLM may be organizations that have confidential data and research labs with limited funds that seek reproducibility. These organizations may not want to disclose their data to third parties or may not be able to afford the high costs of secret data leakage using third-party APIs or hiring human annotators. With PandaLM, they can perform evaluations without compromising data security or incurring high costs, and obtain reproducible results. To demonstrate the reliability and consistency of our tool, we have created a diverse human-annotated test dataset of approximately 1,000 samples, where the contexts and the labels are all created by humans. On our test dataset, PandaLM-7B has achieved 94% ChatGPT's evaluation ability in terms of accuracy. The papers and more features are coming soon.

## 5 其它

### Alpaca-CoT

https://github.com/PhoebusSi/Alpaca-CoT

https://mp.weixin.qq.com/s/Q5Q3RpQ80XmpbfhSxq2R1Q

An Instruction Fine-Tuning Platform with Instruction Data Collection and Unified Large Language Models Interface

Alpaca-CoT项目旨在探究如何更好地通过instruction-tuning的方式来诱导LLM具备类似ChatGPT的交互和instruction-following能力。为此，我们广泛收集了不同类型的instruction（尤其是Chain-of-Thought数据集），并基于LLaMA给出了深入细致的实证研究，以供未来工作参考。据我们所知，我们是首个将CoT拓展进Alpaca的工作，因此简称为"Alpaca-CoT"。

### Auto-GPT

> https://github.com/torantulino/auto-gpt

Auto-GPT is an experimental open-source application showcasing the capabilities of the GPT-4 language model. This program, driven by GPT-4, chains together LLM "thoughts", to autonomously achieve whatever goal you set. As one of the first examples of GPT-4 running fully autonomously, Auto-GPT pushes the boundaries of what is possible with AI.

### ChatPiXiu

> https://github.com/catqaq/ChatPiXiu

我们是羡鱼智能【xianyu.ai】，主要成员是一群来自老和山下、西湖边上的咸鱼们，塘主叫作羡鱼，想在LLMs时代做点有意义的事！我们的口号是：做OpenNLP和OpenX！希望在CloseAI卷死我们之前退出江湖！

也许有一天，等到GPT-X发布的时候，有人会说NLP不存在了，但是我们想证明有人曾经来过、热爱过！在以ChatGPT/GPT4为代表的LLMs时代，在被CloseAI卷死之前，我们发起了OpenNLP计划，宗旨是OpenNLP for everyone!

ChatPiXiu项目为OpenNLP计划的第2个正式的开源项目，旨在Open ChatGPT for everyone！在以ChatGPT/GPT4为代表的LLMs时代，在被OpenAI卷死之前，做一点有意义的事情！未来有一天，等到GPT-X发布的时候，或许有人会说NLP不存在了，但是我们想证明有人曾来过！

### Gorilla

> https://mp.weixin.qq.com/s/p9tx3q3Lpr4fNqdyxWhzyA
>
> gorilla.cs.berkeley.edu
>
> arxiv.org/abs/2305.15334
>
> https://github.com/ShishirPatil/gorilla/

大型语言模型性能强大，但为了更好地用于解决实际问题，各式各样的 API 是必不可少的。

加利福尼亚大学伯克利分校和微软研究院造出了一只「大猩猩」Gorilla，该模型能根据用户输入的自然语言为用户选择合适的 API 来执行对应任务。理论上讲，这个模型可以根据用户需求调用其它各种 AI 模型，因此 Gorilla 有望成为一个统御其它 AI 的 AI 模型。该项目的代码、模型、数据和演示都已发布。

### HuggingGPT

> https://mp.weixin.qq.com/s/o51CmLt2JViJ4nsKfBJfwg
>
> https://arxiv.org/pdf/2303.17580.pdf

HuggingGPT利用ChatGPT作为控制器，连接HuggingFace社区中的各种AI模型，来完成多模态复杂任务。

这意味着，你将拥有一种超魔法，通过HuggingGPT，便可拥有多模态能力，文生图、文生视频、语音全能拿捏了。

### LLMPruner：大语言模型裁剪工具

> https://mp.weixin.qq.com/s/u0UcCxzJOkF4fO_JI6ToQA
>
> https://github.com/yangjianxin1/LLMPruner

在许多下游任务中，我们往往只需要使用到一两种语言，例如在中文场景中，一般只会用到中英文。 所以我们可以对大语言模型的词表进行裁剪，只留下所需的部分词表，这样不仅能够充分保留模型的预训练知识，并且减少模型参数量，降低显存占用，提升训练速度，使用更少的显卡进行下游任务的finetune训练。

基于上述原因，笔者开发了LLMPruner项目，目前主要包含裁剪后的各种参数规模的Bloom模型。对Bloom进行词表裁剪，保留常用的中英文token，词表由250880将至46145，缩减为原来的18.39%。

### LLM-Pruner: On the Structural Pruning of Large Language Models

> https://github.com/horseee/LLM-Pruner
>
> https://arxiv.org/abs/2305.11627
>
> https://mp.weixin.qq.com/s/feqFfy4n31eztoZfodMieQ

在本文中，我们提出了 LLM-Pruner，一种用于大型语言模型的结构化剪枝方法。LLM-Pruner 旨在以任务无关的方式压缩庞大的语言模型，同时尽量减少对原始训练语料库的依赖，并保留 LLM 的语言能力。LLM-Pruner 通过迭代地检查模型中的每个神经元作为识别依赖组的触发器，从而构建 LLM 的依赖图。随后，LLM-Pruner 使用参数级和权重级估计来评估这些组的重要性。

最后，我们利用 LoRA 对被剪枝模型进行快速恢复和调整。我们使用多个 zero-shot 数据集评估了 LLM-Pruner 在三个不同模型（LLaMA，Vicuna 和 ChatGLM）上的有效性。我们的实验结果表明，LLM-Pruner 成功地剪枝了模型，在保留 zero-shot 能力的同时减轻了计算负担。

### LLM for Recommendation Systems

https://github.com/WLiK/LLM4Rec

https://arxiv.org/abs/2305.19860

https://mp.weixin.qq.com/s/WCUjCahiak4STbb0QjJlnQ

Large Language Models (LLMs) have emerged as powerful tools in the field of Natural Language Processing (NLP) and have recently gained significant attention in the domain of Recommendation Systems (RS). These models, trained on massive amounts of data using self-supervised learning, have demonstrated remarkable success in learning universal representations and have the potential to enhance various aspects of recommendation systems by some effective transfer techniques such as fine-tuning and prompt tuning, and so on. The crucial aspect of harnessing the power of language models in enhancing recommendation quality is the utilization of their high-quality representations of textual features and their extensive coverage of external knowledge to establish correlations between items and users. To provide a comprehensive understanding of the existing LLM-based recommendation systems, this survey presents a taxonomy that categorizes these models into two major paradigms, respectively Discriminative LLM for Recommendation (DLLM4Rec) and Generative LLM for Recommendation (GLLM4Rec), with the latter being systematically sorted out for the first time. Furthermore, we systematically review and analyze existing LLM-based recommendation systems within each paradigm, providing insights into their methodologies, techniques, and performance. Additionally, we identify key challenges and several valuable findings to provide researchers and practitioners with inspiration.

### Self-Instruct

https://github.com/yizhongw/self-instruct

https://arxiv.org/abs/2212.10560

Self-Instruct is a framework that helps language models improve their ability to follow natural language instructions. It does this by using the model's own generations to create a large collection of instructional data. With Self-Instruct, it is possible to improve the instruction-following capabilities of language models without relying on extensive manual annotation.

### ToolBench

https://github.com/OpenBMB/ToolBench

https://arxiv.org/pdf/2304.08354.pdf

https://mp.weixin.qq.com/s/DuoQJj1OBl5iFPvjidDiCg

This project aims to construct open-source, large-scale, high-quality instruction tuning SFT data to facilitate the construction of powerful LLMs with general tool-use capability. We provide the dataset, the corresponding training and evaluation scripts, and a capable model ToolLLaMA fine-tuned on ToolBench.

### Wanda (Pruning by Weights and activations)

https://github.com/locuslab/wanda

https://mp.weixin.qq.com/s/UoQLCQiFnKZUQPedDM_MCQ

https://arxiv.org/pdf/2306.11695.pdf

A Simple and Effective Pruning Approach for Large Language Models

As their size increases, Large Languages Models (LLMs) are natural candidates for network pruning methods: approaches that drop a subset of network weights while striving to preserve performance. Existing methods, however, require either retraining, which is rarely affordable for billion-scale LLMs, or solving a weight reconstruction problem reliant on second-order information, which may also be computationally expensive. In this paper, we introduce a novel, straightforward yet effective pruning method, termed Wanda (Pruning by Weights and activations), designed to induce sparsity in pretrained LLMs. Motivated by the recent observation of emergent large magnitude features in LLMs, our approach prune weights with the smallest magnitudes multiplied by the corresponding input activations, on a per-output basis. Notably, Wanda requires no retraining or weight update, and the pruned LLM can be used as is. We conduct a thorough evaluation of our method on LLaMA across various language benchmarks. Wanda significantly outperforms the established baseline of magnitude pruning and competes favorably against recent methods involving intensive weight update.

**这个创作者的更多内容**

检索增强生成 (RAG):What, Why and How?

为什么说数智化可以帮助中小企业降本增效？

ChatGPT提示工程5篇合集 - 吴恩达和OpenAI出品

查看更多

评论

0 评论