

基座模型

自然语言处理前沿——大语言模型的前世今生(tongji.edu.cn)

前身——基于Transformer架构的GPT/BERT等

分类: Base模型/Chat模型

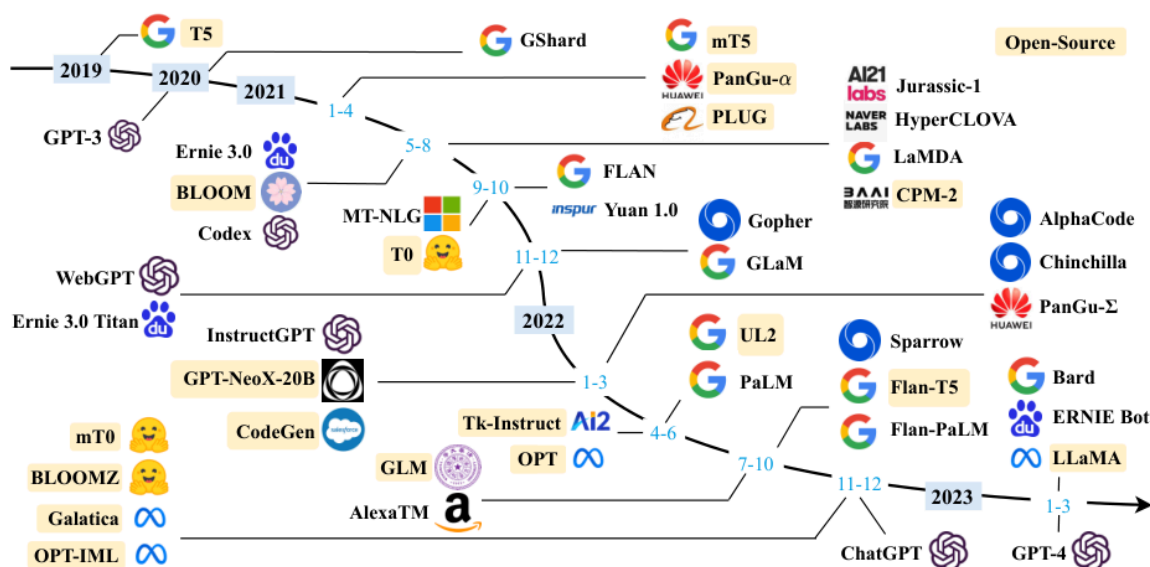


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

比较领先的大语言模型

GPT-4、Claude3、Gemini、Grok、ChatGLM4

可用的大语言模型（众多）

llama3 (8B/70B) ——填写问卷申请[meta-llama/Meta-Llama-3-8B](https://openai.com/index/meta-llama-3/)·Hugging Face、[meta-llama/Meta-Llama-3-70B](https://openai.com/index/meta-llama-3/)·Hugging Face

Gemma (7B) ——填写问卷申请[google/gemma-7b](https://openai.com/index/gemma/)·Hugging Face

GLM3 (6B) ——[THUDM\(github.com\)](https://github.com/THUDM)、[THUDM\(Knowledge Engineering Group\(KEG\) & Data Mining at Tsinghua University\)](https://github.com/THUDM)(huggingface.co)

Phi2 (2.7B) ——[microsoft/phi-2](https://openai.com/index/phi-2/)·Hugging Face

开源轻量级模型: [1-7B开源小型预训练语言模型整理汇总](https://github.com/1-7B) - 知乎([zhihu.com](https://www.zhihu.com))

模型汇总: [大模型综合评测对比](https://github.com/DataLearner) | [当前主流大模型在各评测数据集上的表现总榜单](https://github.com/DataLearner) | [数据学习\(DataLearner\)](https://github.com/DataLearner)

常用的通用模型

下面的很多模型都是基于通用模型在具体垂直领域进行微调。

模型	大小	机构	论文
LLaMA2	7B/7B-Chat 13B/13B-Chat 70B/70B-Chat	Meta	paper
ChatGLM3-6B	6B-Base/6B/6B-32K	清华大学	paper
Qwen	1.8B/1.8B-Chat 7B/7B-Chat 14B/14B-Chat 72B/72B-Chat	阿里云	paper
Baichuan2	7B/7B-Chat 13B/13B-Chat	百川智能	paper
InternLM	7B/7B-Chat 20B/20B-Chat	上海AI实验室	paper

垂直领域

已经很丰富的整理：

[luban-agi/Awesome-Domain-LLM](#): 收集和梳理垂直领域的开源模型、数据集及评测基准。
([github.com](#)) (截至[2023/11/26])

[层出不穷的垂域微调大模型非最全汇总：12大领域、57个领域微调模型概述及对垂直行业问答的一些讨论](#) (截至[2023/09/13])

[lonePatient/awesome-pretrained-chinese-nlp-models](#): Awesome Pretrained Chinese NLP Models, 高质量中文预训练模型&大模型&多模态模型&大语言模型集合 ([github.com](#))
(截至[2024/05/20])

[DSXiangLi/DecryptPrompt](#): 总结Prompt&LLM论文，开源数据&模型，AIGC应用
([github.com](#))

----整理----

医疗领域

中文医疗知识/对话/教育: [AlpaCare](#)、[BenTsao\(本草\)](#)、[BianQue\(扁鹊\)](#)、[CareGPT](#)、[ChatMed](#)、[ChiMed-GPT](#)、[Chinese-vicuna-med](#)、[DISC-MedLLM](#)、[DoctorGLM](#)、[HuatuogPT\(华佗\)](#)、[IvyGPT](#)、[MedicalGPT](#)、[Med-ChatGLM](#)、[MING](#)、[PULSE](#)、[QiZhenGPT](#)、[WiNGPT2](#)、[Sunsimiao\(孙思邈\)](#)、

英文医疗知识/对话: [ChatDoctor](#)、[medAlpaca](#)、[NHS-LLM](#)、[PMC-LLaMA](#)、

中医知识: [HuangDI\(皇帝\)](#)、[ShenNong-TCM-LLM\(神农\)](#)、[TCMLLM](#)、[ZhongJing\(仲景\)](#)、[Zhongjing-LLaMA\(仲景\)](#)、

心理健康: [ChatPsychiatrist](#)、[MentalLLaMA](#)、[MeChat](#)、[MindChat\(漫谈\)](#)

生物医学: [OpenBioMed\(多模态\)](#)、[SoulChat\(灵心\)](#)、[Taiyi\(太一\)](#)

胸部光片: [XrayGLM](#)

儿童陪伴: [QiaoBan\(巧板\)](#)

金融领域

知识问答/场景分析/计算检索: [BBT-FinCUGE-Applications](#)、[CFGPT](#)、[DeepMoney](#)、[DISC-FinLLM](#)、[PIXIU\(貔貅\)](#)、[Tongyi-Finance-14B](#)、[Cornucopia\(聚宝盆\)](#)、[XuanYuan\(轩辕\)](#)、[XuanYuan2.0](#)

英文: [FLANG](#)、[InvestLM](#)、[WeaverBird\(织工鸟\)](#)(双语对话)、

其他: [FinGLM](#)(解析上市公司年报)、[FinGPT](#)(多个金融大模型)、[InvestLM](#)(金融考试、投资问题等)、

法律领域

法律服务/知识: [ChatLaw](#)、[DISC-LawLLM](#)、[夫子·明察](#)、[JurisLMs](#)、[LaWGPT](#)、[LawGPT_zh\(獬豸\)](#)、[Lawyer LLaMA](#)、[LexiLaw](#)、[Lychee\(律知\)](#)、[HanFei\(韩非\)](#)、[wisdomInterrogatory\(智海-录问\)](#)、[XuanYuan](#)

编程领域

代码: [Aquila](#)、[ChatSQL](#)、[codegeex](#)、[codegeex2](#)、[codegemma-7b](#)、[codellama](#)、[CodeQwen1.5-7B-Chat](#)、[codeshell](#)、[DeepSeek-Coder](#)、[DeepSeekMoE](#)、[MFTCoder](#)、[stabelcode](#)、[SQLCoder](#)、[Starcoder](#)、[WaveCoder](#)

教育领域

教育服务: [EduChat](#)、[TuringMM-34B-Chat](#)

国际中文教育: [桃李 \(Taoli\)](#)

数学领域

讲题: [MathGPT](#)

解决问题: [MammoTH](#)、[MetaMath](#)、[Skywork-13B-Math](#)、[WizardMath](#)

其他领域

化学: [OpenDFM/ChemDFM-13B-v1.0](#)

地球科学: [K2](#)

植物科学: [PLLaMa](#)

天文学: [StarWhisper](#) (星语)

海洋学: [MarineGPT](#)(熟悉海洋动物知识, 能识图)、[OceanGPT](#)(海洋学领域专家)

农业: [AgriGPT](#)

自媒体: [MediaGPT](#)

电商: [EcomGPT](#)

网络安全: [AutoAudit](#)、[SecGPT](#)

科技: [Mozi](#) (墨子)(科技文献)、[TechGPT](#)(众多垂直领域)

交通: [TransGPT \(致远\)](#)(通用常识交通大模型)

故事生成: [ChatRWKV](#)

音乐生成: [facebook/musicgen-medium](#)

评估模型: [Auto-J](#)、[JudgeLM](#)

运维：[DevOps-Model](#)、[OWL](#)

舆情安全：[YaYi \(雅意\)](#)(覆盖媒体宣传、舆情分析、公共安全、金融风控、城市治理等五大领域)

更杂的一些工具：

工具描述	链接
GPT4v-ACT：基于JS DOM识别网页元素，服务于各类多模态webagent	https://github.com/ddupont808/GPT-4V-Act?tab=readme-ov-file
Deep-KE：基于LLM对数据进行智能解析实现知识抽取	https://github.com/zjunlp/DeepKE
IncarnaMind：多文档RAG方案，动态chunking的方案可以借鉴	https://github.com/junruxiong/IncarnaMind
Vectra：平台化的LLM Agent搭建方案，从索引构建，内容召回排序，到事实检查的LLM生成	https://vectara.com/tour-vectara/
Data-Copilot：时间序列等结构化数据分析领域的Agent解决方案	https://github.com/zwq2018/Data-Copilot
DB-GPT：以数据库为基础的GPT实验项目，使用本地化的GPT大模型与您的数据和环境进行交互	https://db-gpt.readthedocs.io/projects/db-gpt-docs-zh-cn/zh_CN/latest/index.html
guardrails：降低模型幻觉的python框架，prompt模板+validation+修正	https://github.com/shreyar/guardrails
guidance：微软新开源框架，同样是降低模型幻觉的框架，prompt+chain的升级版加入逐步生成和思维链路	https://github.com/guidance-ai/guidance
SolidGPT：上传个人数据，通过命令交互创建项目PRD等	https://github.com/AI-Citizen/SolidGPT
HR-Agent：类似HR和员工交互，支持多工具调用	https://github.com/stepanogil/autonomous-hr-chatbot
BambooAI：数据分析Agent	https://github.com/pgalko/BambooAI

工具描述	链接
AlphaCodium: 通过Flow Engineering完成代码任务	https://github.com/Codium-ai/AlphaCodium
REOR: AI驱动笔记软件	https://github.com/reorproject/reor
Vanna.AI: chat with sql database	https://vanna.ai/
ScrapeGraph: 融合了图逻辑和LLM	https://scrapegraph-doc.onrender.com/
OpenAct: Adapt-AI推出的和桌面GUI交互的Agent框架	https://github.com/OpenAdaptAI/OpenAdapt
LaVague: WebAgent框架, 偏低层指令交互性把指令转换成Selenium代码去和网页交互	https://github.com/lavague-ai/LaVague/tree/main
Tarsier: webagent的辅助工具把网站转换成可交互元素序号和描述	https://github.com/reworkd/tarsier?tab=readme-ov-file
RecAI: 微软推出的推荐领域LLM Agent	https://github.com/microsoft/RecAI

Diffusion Models

图像

DALL-E 2: [openai/DALL-E: PyTorch package for the discrete VAE used for DALL-E.](https://github.com/openai/DALL-E) (github.com)

Stable Diffusion: [Stability-AI/stablediffusion: High-Resolution Image Synthesis with Latent Diffusion Models](https://github.com/Stability-AI/stablediffusion) (github.com)

Disco Diffusion: [Alembics/Disco-diffusion](https://github.com/Alembics/Disco-diffusion) (github.com)

DDPM: [hojonathanho/diffusion: Denoising Diffusion Probabilistic Models](https://github.com/hojonathanho/diffusion) (github.com)

GLIDE: [openai/glide-text2im: GLIDE: a diffusion-based text-conditional image synthesis model](https://github.com/openai/glide-text2im) (github.com)

在Hugging Face上有很多开源扩散模型, 也有很多基于LoRA的微调。

音频

WaveGrad: [ivanvovk/WaveGrad: Implementation of WaveGrad high-fidelity vocoder from Google Brain in PyTorch. \(github.com\)](#)

DiffWave: [lmnt-com/diffwave: DiffWave is a fast, high-quality neural vocoder and waveform synthesizer. \(github.com\)](#)

音频处理

主要是基于wav2vec、HUBERT之类的模型。

[lmnt-com/diffwave: DiffWave is a fast, high-quality neural vocoder and waveform synthesizer. \(github.com\)](#)

[facebook/wav2vec2-large-xlsr-53 · Hugging Face](#)











































[facebook/hubert-base-ls960 · Hugging Face](#)

[TencentGameMate/chinese-wav2vec2-base · Hugging Face](#)

[TencentGameMate/chinese-hubert-base · Hugging Face](#)

多模态

模型	大小	时间	语言模型	非语言模型	语言	领域	下载	项目地址	机构/个人	文献
HunyuanDiT	1.5B	2024-05	multilingual T5 encoder	CLIP	中英	图文	🤗	HunyuanDiT	Tencent	Paper
CogVLM2		2024-05	Meta-Llama-3-8B-Instruct	/	中英	图文	🤗	CogVLM	Skip to content	
360VL	8/70B	2024-05	LLama3	CLIP-ViT	中英	图文	🤗	360VL	360CVGroup	
XVERSE-V	13B	2024-05	XVERSE-13B-Chat	clip-vit-large-patch14-224	中英	图文	🤗	XVERSE-V-13B	xverse-ai	
MiniCPM-V 2.0	2.8B	2024-04	MiniCPM-2.4B	SigLip-400M	中英	图文	🤗 🤗	MiniCPM-V	OpenBMB	Blog
Qwen-Audio	7B	2024-03	Qwen-7B	Whisper-large-v2	中英	语音	🤗 HF	Qwen-Audio 🌟 Star 1.2k	Qwen	Paper
DeepSeek-VL	1.3/7B	2024-03	DeepSeek	SigLip/SAM	中英	图文	🤗 HF	DeepSeek-VL 🌟 Star 1.8k	deepseek-ai	Paper
OmniLMM	3/12B	2024-02	MiniCPM	SigLip	中英	图文	🤗 HF	OmniLMM 🌟 Star 7.7k	[OpenBMB] (https://github.com/OpenBMB)	
MiniCPM-V	3B	2024-02	MiniCPM-2.4B	SigLip-400M	中英	图文	🤗 HF	OmniLMM 🌟 Star 7.7k	[OpenBMB] (https://github.com/OpenBMB)	
Yi-VL	6/34B	2024-01	Yi	CLIP-ViT	中英	图文	[🤗 HF]	Yi 🌟 Star 7.4k	01-ai	
Lyrics	14B	2023-12	/	/	中英	图文	[🤗 HF]	Fengshenbang-LM	IDEA研究院	

模型	大小	时间	语言模型	非语言模型	语言	领域	下载	项目地址	机构/个人	文献
Qwen-Audio	7B	2023-12	Qwen-7B	Whisper-large-v2	中英	语音	 HF	Qwen-Audio  Star 1.2k	Qwen	Paper
SPHINX	13B	2023-10	/	/	中英	图文	 HF	LLaMA2-Accessory  Star 2.6k	Alpha-VLLM	
Skywork-MM	13B	2023-10	/	/	中英	图文	 HF	Skywork	SkyworkAI	Paper
CogVLM	7/14B	2023-10	Qwen	ViT	中英	图文	 HF	/	CausalLM	
fuyu	8B	2023-10	/	/	中英	图文	 HF	/	Adept AI Labs	Blog
Ziya-Visual	14B	2023-10	LLaMA	InstructBLIP	中英	图文	 HF	Fengshenbang-LM  Star 4k	IDEA研究院	Paper
CogVLM	17B	2023-10	EVA2-CLIP-E	Vicuna-v1.5	中英	图文	TODO	CogVLM  Star 5.6k	THUDM	Paper
idefics	9/80B	2023-10	LLaMA	CLIP-ViT	中英	图文	 HF	/	HuggingFaceM4	log
InternLM-XComposer	7B	2023-10	InternLM	EVA-CLIP	中英	图文	 HF	InternLM-XComposer  Star 1.9k	InternLM	Report
WeMix-LLM	13B	2023-09	LLama2	/	中英	图文	 HF	WeMix-LLM  Star 17	Alpha-VLLM	
Vally	7/13B	2023-08	BelleGroup/BELLE-LLaMA-EXT	OFA-Sys/chinese-clip-vit-large-patch14	中英	图文	 HF  HF	Valley  Star 183	罗瑞璞	Paper
SALMONN	/	2023-08	/	/	中英	语音	TODO	SALMONN  Star 871	Bytedance	
IDEFICS	9/80B	2023-08	llama	CLIP-ViT	中英	图文-通用	 HF	m4-logs  Star 54	HuggingFaceM4	Paper
Qwen-VL	7B	2023-08	Qwen-7B	Openclip ViT-bigG	中英	通用	 HF	Qwen-VL  Star 4.2k	阿里云	
Qwen-VL-chat	7B	2023-08	Qwen-7B	Openclip ViT-bigG	中英	通用	 HF	Qwen-VL  Star 4.2k	阿里云	
LLaSM	7B	2023-07	Chinese-LLama2	whisper-large-v2	中英	语音	 HF	LLaSM  Star 493	北京灵珑	
Chinese-LLaVA	7B	2023-07	Chinese-LLama2	Clip-vit	中英	视觉	 HF	Chinese-LLaVA  Star 342	北京灵珑	
RemoteGLM	6B	2023-07	VisualGLM-6B	VisualGLM-6B	中文	遥感	TODO	RemoteGLM  Star 57	lzw-lzw	
VisualCLA	7B	2023-07	Chinese-Alpaca-Plus	CLIP-ViT-L/14	中文	视觉	 HF	Visual-Chinese-LLaMA-Alpaca  Star 385	Ziqing Yang	
yuren	7B	2023-07	baichuan-7B	CLIP	中英	视觉	 HF	yuren-baichuan-7b  Star 69	Pleisto	
VisCPM-Chat	10B	2023-06	CPM-Bee	Q-Former	中英	视觉	 HF	VisCPM  Star 1k	OpenBMB	
VisCPM-Paint	10B	2023-06	CPM-Bee	Stable Diffusion 2.1	中英	视觉	 HF	VisCPM  Star 1k	OpenBMB	
XrayPULSE	7B	2023-06	PULSE	MedCLIP	中文	医学	 HF	XrayPULSE  Star 165	OpenMEDLab	
SEEChat	6B	2023-06	ChatGLM	CLIP-ViT	中文	/	 HF	SEEChat  Star 95	360	

模型	大小	时间	语言模型	非语言模型	语言	领域	下载	项目地址	机构/个人	文献
Ziya-BLIP2-14B-Visual-v1	14B	2023-06	LLaMA-13B	BLIP2	中英	通用	 HF	Fengshenbang-LM  Star 4k	IDEA研究院	
Video-LLaMA-BiLLA	7B	2023-05	BiLLa-7B	MiniGPT-4	中英	通用	 HF	Video-LLaMA  Star 2.6k	达摩院多语言NLP	Paper
Video-LLaMA-Ziya	13B	2023-05	Ziya-13B	MiniGPT-4	中英	通用	 HF	Video-LLaMA  Star 2.6k	达摩院多语言NLP	Paper
XrayGLM	6B	2023-05	ChatGLM-6B	BLIP2-Qformer	中英	医学	 HF	XrayGLM  Star 840	澳门理工大学	
X-LLM		2023-05	ChatGLM	ViT-g	中文	/	TODO	X-LLM  Star 293	中科院自动化所	Paper
VisualGLM	6B	2023-05	ChatGLM-6B	BLIP2-Qformer	中英	视觉	 HF	VisualGLM-6B  Star 4k		

其他

非大模型，但是很全的网络模型：[Deep-Spark/DeepSparkHub: DeepSparkHub selects hundreds of application algorithms and models, covering various fields of AI and general-purpose computing, to support the mainstream intelligent computing scenarios.\(github.com\)](#)

大模型与GPU算力

重要参数

FLOPS：每秒执行的浮点运算次数，是衡量 GPU 计算能力的标准。

- **单精度 FLOPS (FP32)**：用于大部分深度学习模型训练。
- **双精度 FLOPS (FP64)**：用于科学计算和高精度任务。
- **半精度 FLOPS (FP16)**：用于加速训练过程，尤其是大规模模型训练。

显存(VRAM)：GPU 用于存储数据的内存，存储内容包括模型参数、激活值、中间计算结果等。

显存带宽：GPU 和显存之间的数据传输速度，以 GB/s 为单位。

几款GPU参数

GPU 型号	单精度 FLOPS (FP32)	显存容量	显存类型	显存带宽
RTX 3080	29.8 TFLOPS	10GB GDDR6X	GDDR6X	760.3 GB/s
RTX 3090	35.6 TFLOPS	24GB GDDR6X	GDDR6X	936.2 GB/s
RTX 4080	48.74 TFLOPS	16GB GDDR6X	GDDR6X	716.8 GB/s
RTX 4090	82.58 TFLOPS	24GB GDDR6X	GDDR6X	1,008 GB/s
T4	8.1 TFLOPS	16GB GDDR6	GDDR6	320 GB/s
A10	31.2 TFLOPS	24GB GDDR6	GDDR6	600 GB/s
A6000	38.7 TFLOPS	48GB GDDR6	GDDR6	768 GB/s
A100	19.5 TFLOPS	40GB / 80GB HBM2	HBM2	1.6 TB/s
V100	15.7 TFLOPS	16GB / 32GB HBM2	HBM2	900 GB/s
A800	20 TFLOPS	80GB HBM2e	HBM2e	2 TB/s
H100	60 TFLOPS	80GB HBM2e	HBM2e	2 TB/s

Space Hardware

Choose a hardware for your Space.
You'll be billed on a per minute basis.
View usage in your [billing settings](#).

Sleep time settings

Sleep after 48 hours of inactivity
Upgrade to a paid Hardware to set a custom sleep time.

Pause Space

Building something cool as a side project?
Apply for a [community GPU grant](#).

CPU basic
2 vCPU · 16 GB RAM
Current · Free

CPU upgrade
8 vCPU · 32 GB RAM
\$0.03/hour

Nvidia T4 small
4 vCPU · 15 GB RAM · 16 GB VRAM
\$0.40/hour

Nvidia T4 medium
8 vCPU · 30 GB RAM · 16 GB VRAM
\$0.60/hour

Nvidia 1xL4
8 vCPU · 30 GB RAM · 24 GB VRAM
\$0.80/hour

Nvidia 4xL4
48 vCPU · 190 GB RAM · 96 GB VRAM
\$3.80/hour

Nvidia A10G small
4 vCPU · 15 GB RAM · 24 GB VRAM
\$1.00/hour

Nvidia A10G large
12 vCPU · 46 GB RAM · 24 GB VRAM
\$1.50/hour

Nvidia 2xA10G large
24 vCPU · 92 GB RAM · 48 GB VRAM
\$3.00/hour

Nvidia 4xA10G large
48 vCPU · 184 GB RAM · 96 GB VRAM
\$5.00/hour

Nvidia A100 large
12 vCPU · 142 GB RAM · 40 GB VRAM
\$4.00/hour

Nvidia H100
24 vCPU · 250 GB RAM · 80 GB VRAM
\$10.00/hour

AI Accelerator

HPU · IPU · ...
Coming soon

Display price: per hour per month

- 显存与参数量关系计算：每个float32参数要占4字节，因此有（数量级）最小显存大小=模型参数量×4.

- 单个模型副本中每个参数量大约需要20倍于自身大小的空间占用，以175B模型训练为例，至少需要3.5TB的显存空间占用。模型推理中的显存压力相对小些， 只需1~2倍于模型参数的空间占用。

一些建议情况

模型	显卡要求	推荐显卡
Running Falcon-40B	运行 Falcon-40B 所需的显卡应该有 85GB 到 100GB 或更多的显存	See Falcon-40B table
Running MPT-30B	当运行 MPT-30B 时，显卡应该具有80GB 的显存	See MPT-30B table
Training LLaMA (65B)	对于训练 LLaMA (65B)，使用 8000 台 Nvidia A100 显卡。	Very large H100 cluster
Training Falcon (40B)	训练 Falcon (40B) 需要 384 台具有 40GB 显存的 A100 显卡。	Large H100 cluster
Fine tuning an LLM (large scale)	大规模微调 LLM 需要 64 台 40GB 显存的 A100 显卡	H100 cluster
Fine tuning an LLM (small scale)	小规模微调 LLM 则需要 4 台 80GB 显存的 A100 显卡。	Multi-H100 instance

具体开源模型建议

ChatGLM3-6B

微调：H100、A100

- SFT 全量微调: 4张显卡平均分配，每张显卡占用 48346MiB 显存。
- P-TuningV2 微调: 1张显卡，占用 18426MiB 显存。
- LORA 微调: 1张显卡，占用 14082MiB 显存。

推理：默认情况下，模型以 FP16 精度加载，运行需要大概 13GB 显存。

LLAMA2系列

通义千问输出

推理：

全精度（FP32）

- Llama2 7B最低显存要求为28GB。
- Llama2 13B最低显存要求为52GB。
- Llama2 70B最低显存要求高达280GB。

低精度：

- 对于16位精度（FP16 或其他半精度格式），Llama2 7B、13B、70B模型所需的最低显存分别约为14GB、26GB、140GB。
- 对于8位精度，这些数字进一步减小到7GB、13GB、70GB。

GPT-4o输出

模型	显存需求	推荐 GPU 型号
LLAMA2 7B	14GB	NVIDIA RTX 3080, NVIDIA RTX 3090, NVIDIA A6000, NVIDIA A100
LLAMA2 13B	26GB	NVIDIA RTX 3090, NVIDIA A6000, NVIDIA A100
LLAMA2 30B	60GB	NVIDIA A100, NVIDIA H100
LLAMA2 65B	128GB	NVIDIA H100（多卡并行），多块 NVIDIA A100（并行）

Inference Performance

This section provides the statistics of speed and memory of models in different precisions. The speed and memory profiling are conducted using [this script](#).

We measured the average inference speed (tokens/s) and GPU memory usage of generating 2048 with the models in BF16, Int8, and Int4.

Model Size	Quantization	Speed (Tokens/s)	GPU Memory Usage
1.8B	BF16	54.09	4.23GB
	Int8	55.56	3.48GB
	Int4	71.07	2.91GB
7B	BF16	40.93	16.99GB
	Int8	37.47	11.20GB
	Int4	50.09	8.21GB
14B	BF16	32.22	30.15GB
	Int8	29.28	18.81GB
	Int4	38.72	13.01GB
72B	BF16	8.48	144.69GB (2xA100)
	Int8	9.05	81.27GB (2xA100)
	Int4	11.32	48.86GB
72B + vLLM	BF16	17.60	2xA100

The profiling runs on a single A100-SXM4-80G GPU (except 2xA100 is mentioned) with PyTorch 2.0.1, CUDA 11.8, and

BenTsao(本草)

（基于LLAMA-7B）基于LLaMA模型的指令微调过程中，我们在一张A100-SXM-80GB显卡上进行了训练，训练总轮次10轮，耗时约2h17m。batch_size=128的情况下显存占用在40G左右。预计3090/4090显卡(24GB显存)以上显卡可以较好支持，根据显存大小来调整batch_size。

Chinese-vicuna-med

基座模型：LlaMa-7B/13B

- 训练：一张2080Ti即可。由于数据长度都在256（代码设置为cutoff_len，默认阶段长度）以内，大概占用9G显存。
 - 70w的数据，3个epoch，一张2080Ti大概200h
 - 13B需要18G左右显存（在3090上可以将数据长度开到2048）

- **推理**：一张2080Ti即可（7B），同时支持多卡推理（差不多均匀负载，某张卡会负载高一点）。

XrayGLM

基座模型：[VisualGLM-6B](#)

微调：4bit量化的情况下可以用7GB，否则需要十几个GB，全量微调的话需要50多个GB，使用4张A100可以跑起来。

LaWGPT

（参数量约7B，数据量较大）在通用中文基座模型（[ymcui/Chinese-LLaMA-Alpaca: 中文LLaMA&Alpaca大语言模型+本地CPU/GPU训练部署 \(Chinese LLaMA & Alpaca LLMs\)](#) ([github.com](#))) 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。

训练：8 张 Tesla V100-SXM2-32GB：二次训练阶段耗时约 24h / epoch，微调阶段耗时约 12h / epoch

codegeex2

基座模型：ChatGLM2-6B

推理：CodeGeeX2-6B 更好支持中英文输入，支持最大 8192 序列长度，推理速度较一代 CodeGeeX-13B 大幅提升，量化后仅需6GB显存即可运行，支持轻量级本地化部署。测试硬件为GeForce RTX-3090。

DeepSeekMoE

参数量：16B

推理：可以部署在具有 40GB 内存的单个 GPU 上，无需量化。

微调：在 8 个 A100 40GB GPU 上运行。也可以使用 4/8 位 qlora 微调模型，请随时尝试。对于此配置，可以在单个 A100 80G GPU 上运行。

EduChat

参数量：7B

推理：可在单张A100/A800或CPU运行，使用FP16精度时约占用15GB显存

AgriGPTs系列模型

- [AgriGPT-6B](#)，此版本为学术demo版，基于[ChatGLM2-6B](#)训练而来,所需显存约 $13225\text{MB}/1024=12.91\text{GB}$ 。
- [AgriGPT-13B](#)，此版本为学术demo版，基于[Baichuan2-13B](#)训练而来所需显存约 $30425\text{MB}/1024=29.7\text{GB}$ 。

YaYi (雅意)

推理：可在单张 A100/A800/3090 等GPU运行

微调：全参数微调建议使用 4*A100(80G) 以上硬件配置；LoRA微调使用单卡 A100(80G) 即可完成微调，学习率可调整为较大值。