

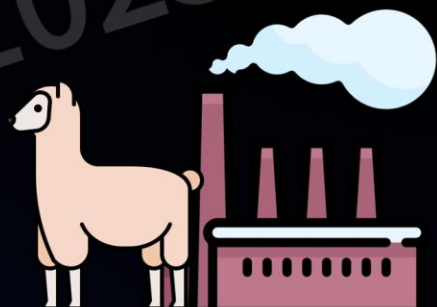


低代码大模型训练框架LLaMA Factory 赋能领域定制GPT

郑耀威

SeamlessAI CTO

LLaMA Factory 构建动机

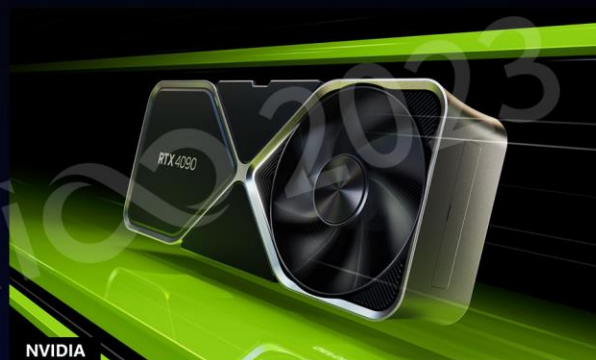


LLaMA Factory: 领域大模型高效训练框架

模型基座



有限算力



私域数据



LLaMA Factory 生态定位

推理

LLM



DeepSpeed MII



Text Generation Inference



TensorRT

训练

	百度千帆	DeepSpeed	LLaMA-Factory
直接偏好优化	×	×	✓
参数高效调优	✓	×	✓
本地私有部署	×	✓	✓

模型



LLaMA



百川



ChatGLM Alpha



千问

硬件



NVIDIA



HUAWEI

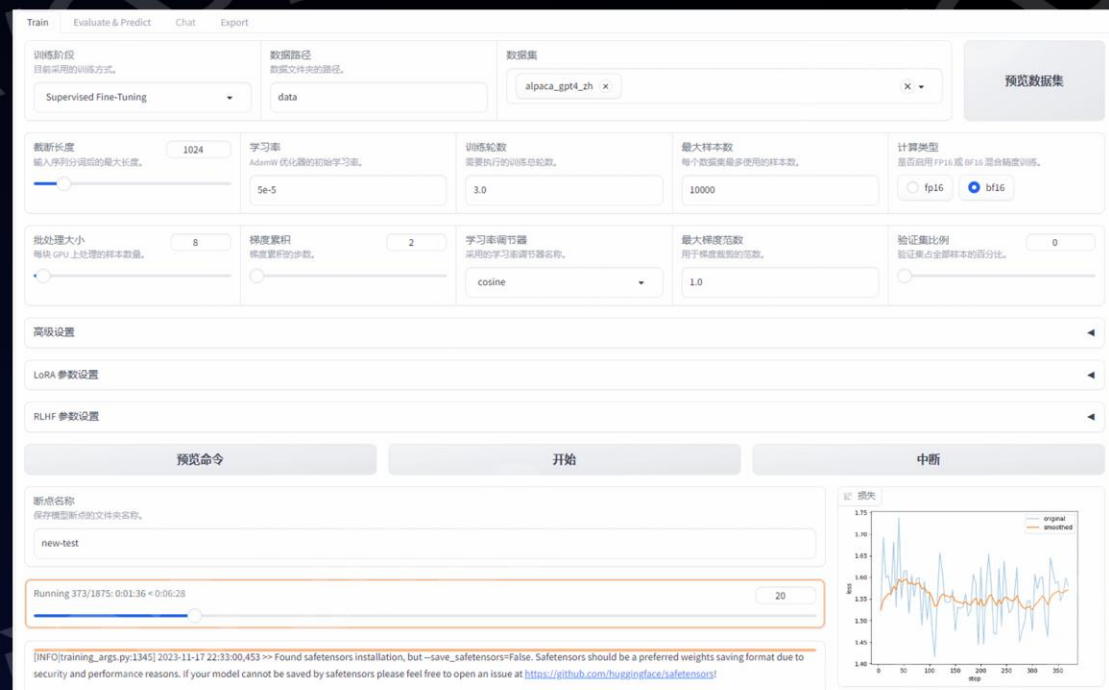
HYGON

LLaMA Factory 核心优势

✓ 易用

✓ 开源

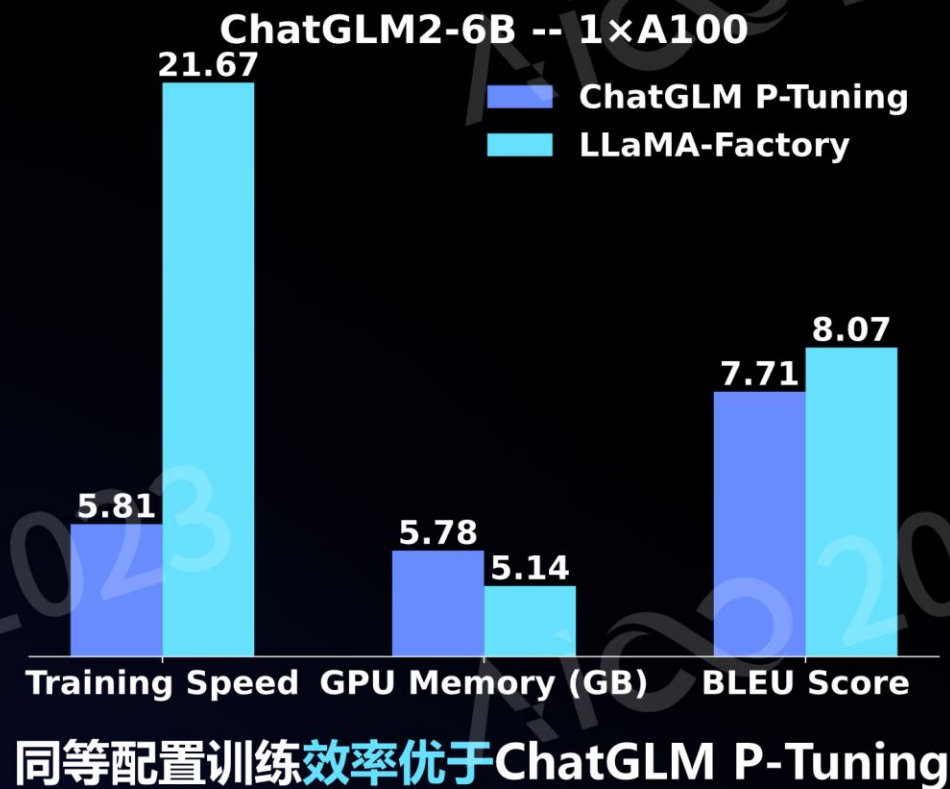
✓ 高效



The screenshot shows the LLaMA Factory web interface with the following configuration details:

- 训练阶段:** Supervised Fine-Tuning
- 数据路径:** data
- 数据集:** alpaca_gpt4_zh
- 截断长度:** 1024
- 学习率:** 5e-5
- 训练轮数:** 3.0
- 最大样本数:** 10000
- 计算类型:** bf16
- 批处理大小:** 8
- 梯度累积:** 2
- 学习率调度器:** cosine
- 最大梯度范数:** 1.0
- 验证集比例:** 0
- 高级设置:** LoRA 参数设置, RLHF 参数设置
- 预览命令:** 开始, 中断
- 断点名称:** new-test
- 运行状态:** Running 373/1875 0:01:36 + 0:06:28
- 损失图:** A line graph showing training loss over time, with 'original' and 'smoothed' lines.

提供可视化平台，一键配置模型训练



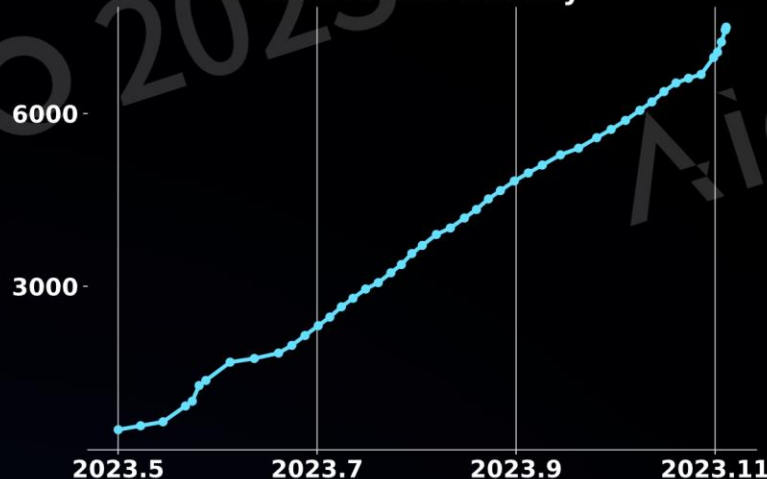
LLaMA Factory 社区影响力

月访问量**40000+**

月下载量**4000+**

关注人数**7000+**

GitHub Star History



使用了 LLaMA Factory 的开源项目:



复旦大学 DISC-LawLLM



华东理工大学 Sunsimiao



北京大学 DevOps Model



澳门理工大学 CareGPT

LLaMA Factory 整体架构

应用层



OpenAI API



LLaMA Board

算法层

训练模式 (20种)

评估模式 (3种)

高兼容

&

易扩展

基础层

模型类 (24个)

数据类 (34个)



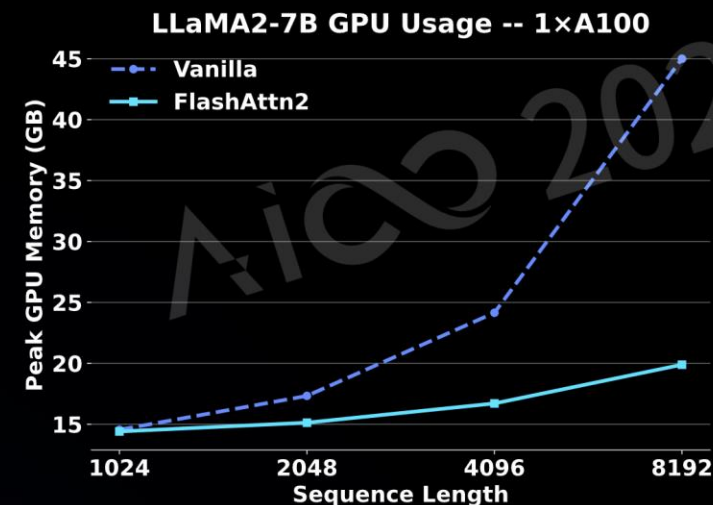
模型兼容性优化



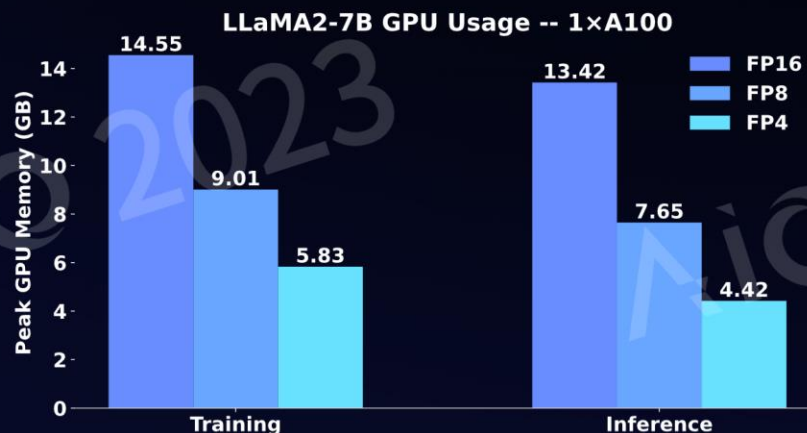
模型性能优化

FlashAttention-2: 缩放点积注意力的硬件高效实现

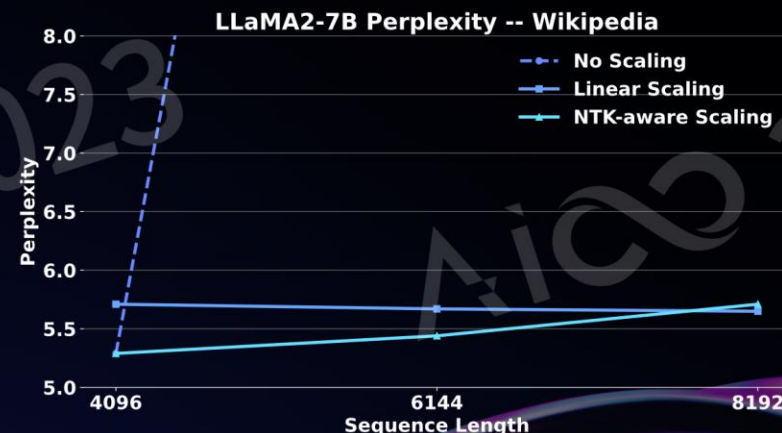
训练框架	TFLOPS	TGS	Latency	GPUs
NVIDIA Megatron-LM	163	-	-	3072 * A100
LLaMA-Factory	123	2095	3.91s	16 * A100
LLaMA-Factory + FA2	167	2957	2.77s	16 * A100



LLM.int8() 与 GPTQ 量化训练 (QLoRA)



位置编码插值与 LongLoRA



模型稳定性优化

■ 模型载入精度

- 混合精度训练优先选择 **BFLOAT16**
- 华为昇腾 NPU 选择 Float16

原始 \ 精调	fp16	bf16	fp32
fp16	✓	✓	✓
bf16	✗	✓	✓
fp32	✗	✓	✓

■ 层归一化参数精度

- 量化训练时将**层归一化**权重转为 Float32
- 避免舍入误差传播：层归一化输出保留 **32** 位精度

(weight * feature).half() ✗
weight * feature.half() ✓

■ Softmax 精度

- 保证 Softmax 计算精度是 32 位

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) = \text{softmax}\left(\left(\frac{QK^T}{\alpha\sqrt{d}} - \max\left(\frac{QK^T}{\alpha\sqrt{d}}\right)\right) \times \alpha\right) = \text{FP16}\left(\text{softmax}\left(\text{FP32}\left(\frac{QK^T}{\alpha\sqrt{d}}\right) \times \alpha\right)\right)$$

数据兼容性优化



数据利用率优化

多轮对话样本构造

A 问题 a 回答 # 结束符

方式1

输入	A	B	C	d	e	f	G	H	I	j	k	l	M	N	O	p	q	r	
标签																p	q	r	#

利用率17%

方式2

输入	A	B	C	d	e	f	#	G	H	I	j	k	l	#	M	N	O	p	q	r	
标签				d	e	f	#				j	k	l	#				p	q	r	#

利用率45%

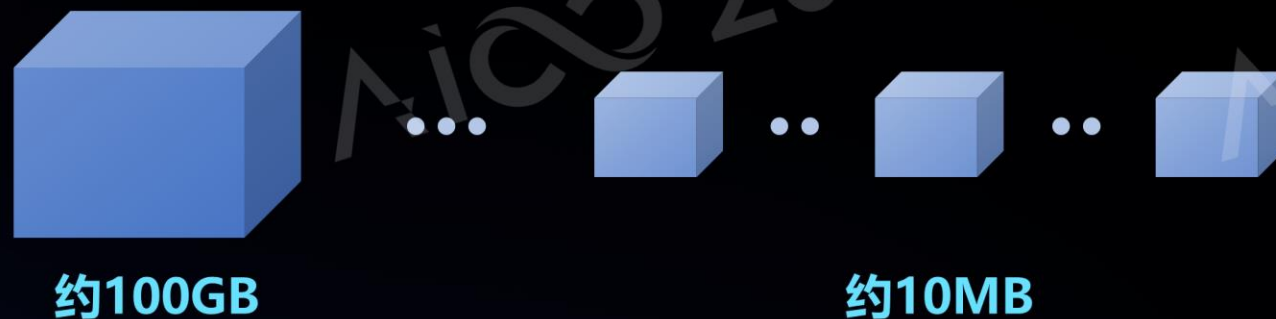
方式3

输入	A	B	C	d	e	f	G	H	I	j	k	l	M	N	O	p	q	r	
标签				d	e	f	#			j	k	l	#			p	q	r	#

利用率50%

数据性能优化

■ 内存高效的流式数据加载



■ 数据打包与分布式打包



- 分布式打包请参阅 [OpenChat/Multipack Sampler](#)

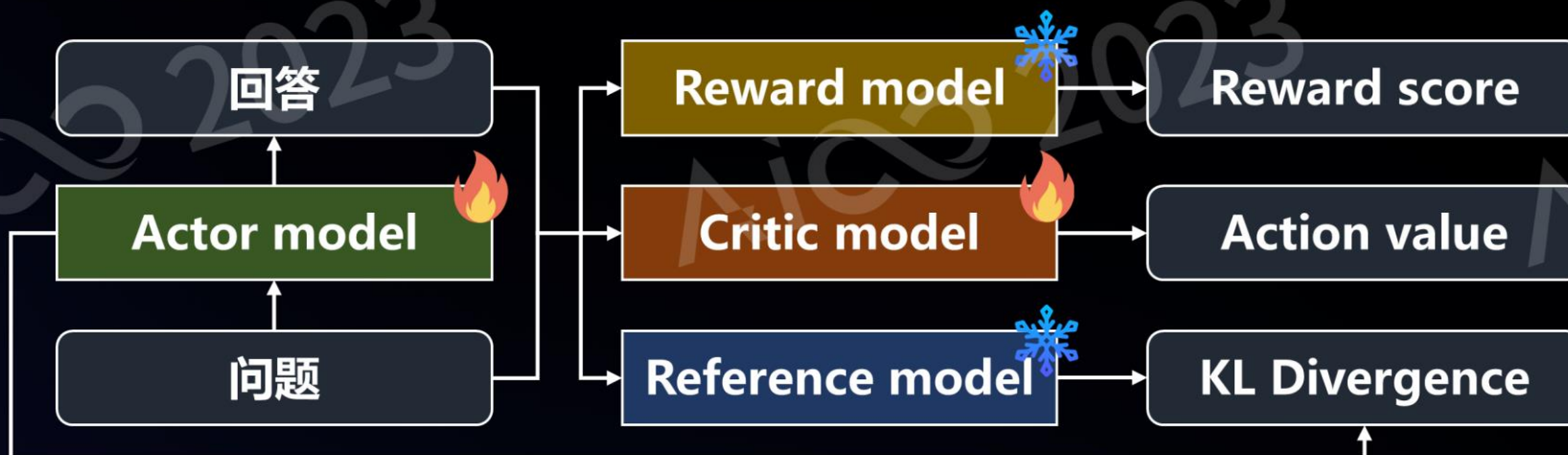
LLaMA-Factory 模型调优框架

(分布式) 命令行接口			LLaMA Board 接口	
自监督训练	有监督微调	奖励模型训练	近端策略优化	直接偏好优化
输入形态	损失函数		评估指标 (BLEU、Rouge)	
 Trainer				
 PyTorch			 deepspeed	

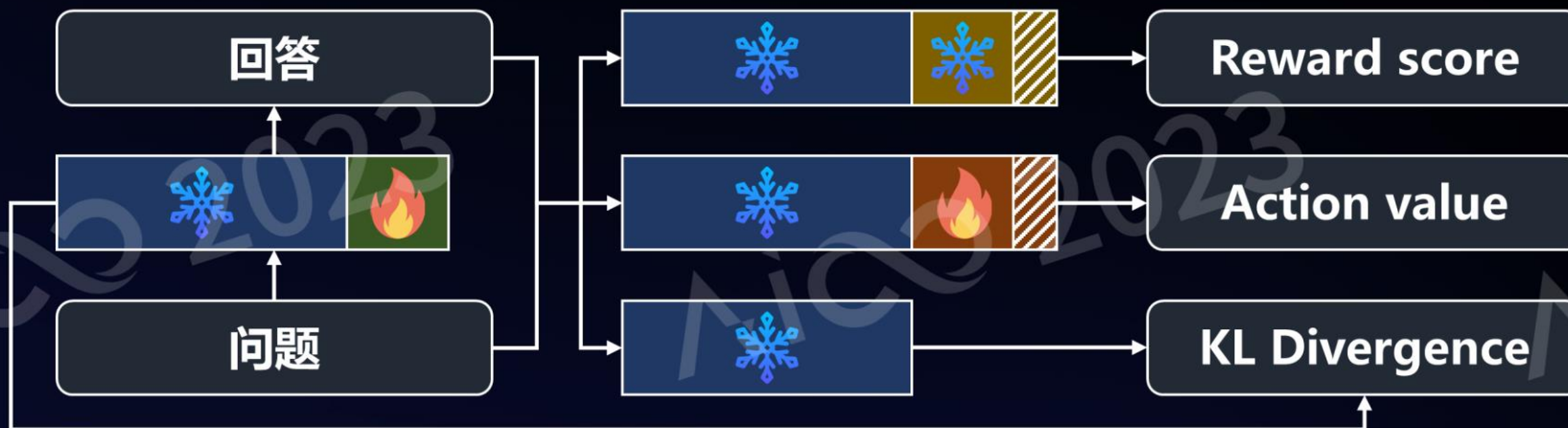
✓ 0-1对话模型训练

✓ 多机多卡支持

RLHF 性能优化



模型参数不共享
需要154GB显存



模型参数共享
需要14GB显存

关注 LLaMA Factory

开源仓库地址: <https://github.com/hiyouga/LLaMA-Factory>



加入微信交流群



在魔搭社区预览 LLaMA Board



AiC∞ 2023

AiC∞ 2023

AiC∞ 2023

谢 谢

AiC∞ 2023

AiC∞ 2023

AiC∞ 2023

