Colossal-AI：AI大模型的挑战与系统优化

尤洋

新加坡国立大学校长青年教授，潞晨科技创始人兼董事长

潞晨科技
HPC AI TECH

Colossal-AI

AI ! ” 战# $ 统优%

Git hub ” # $!

https://github.com/hpcaitech/ColossalAI

尤洋 创始人 & 董事长
- ✓ 伯克利博士
- ✓ 新加坡国立大学-校长青年教授
- ✓ IEEE杰出新人奖
- ✓ 2020全球HPC领域引用量最高的博士

Berkeley NUS IEEE

James Demmel 首席战略官
- ✓ 伯克利杰出教授
- ✓ 美国科学院院士、美国工程院院士
- ✓ 美国艺术与科学院院士
- ✓ IEEE/ACM Fellow

Berkeley NATIONAL ACADEMY OF SCIENCES IEEE

1

时. ！” 战# / 0

Colossal-AI

现状

👉https://www.youtube.com/watch?v=tgB671SFS4w

使用大型人工智能模型的挑战

PaLM@300 years by 1 NV A100 GPUsA$9.2M+

Training

- GPT-2 (2019): “What is the best rising stock this year?”
- GPT-J (2021): “What is the best rising stock this year?”

needs to re-train on new data repeatedly

Inference Fine-tuning

- A cluster of GPUs is required simply to load & make predictions
- GPT-3: 2400+ GB; NV A100 GPU: 80 GB

single GPU server is out-of-memory

Deployment

- A company needs 70 people building their internal tools for AI: $20M per year (impossible for startups)

Expensive Infrastructure and Systems

Colossal-AI：生成式AI和LLM的软件基础设施

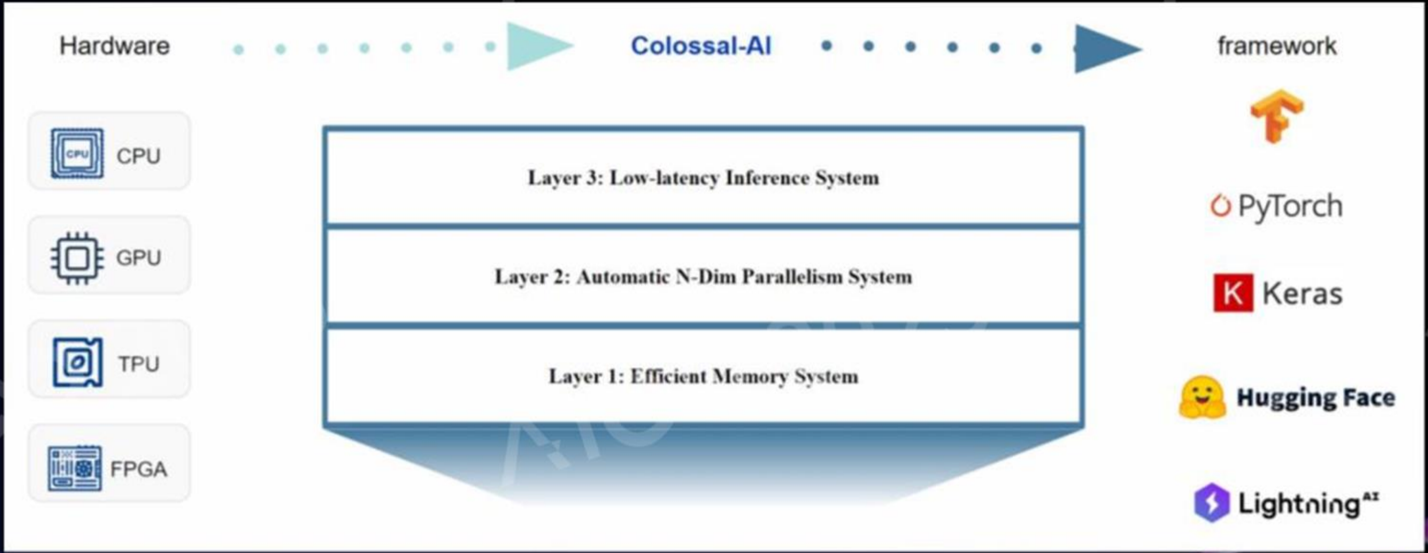
Platform	Colossal-AI Cloud Platform
Models	GPT, Diffusers, ResNet, Alpha-Fold, LLaMA
Deep learning libraries	PyTorch, TensorFlow, mosaic ML, Hugging Face, Lightning AI, NVIDIA APEX
Distributed training/inference	Colossal-AI
Cluster management	RAY, slurm workload manager, kubernetes
Compilation and runtime	NVIDIA CUDA
Hardware	NVIDIA, Ascend

Colossal-AI = 高效 + 低成本

A 面向大模型、桥接AI应用与底层部署的分布式AI系统

最大限度帮助企业级AI大模型应用和构建降本增效

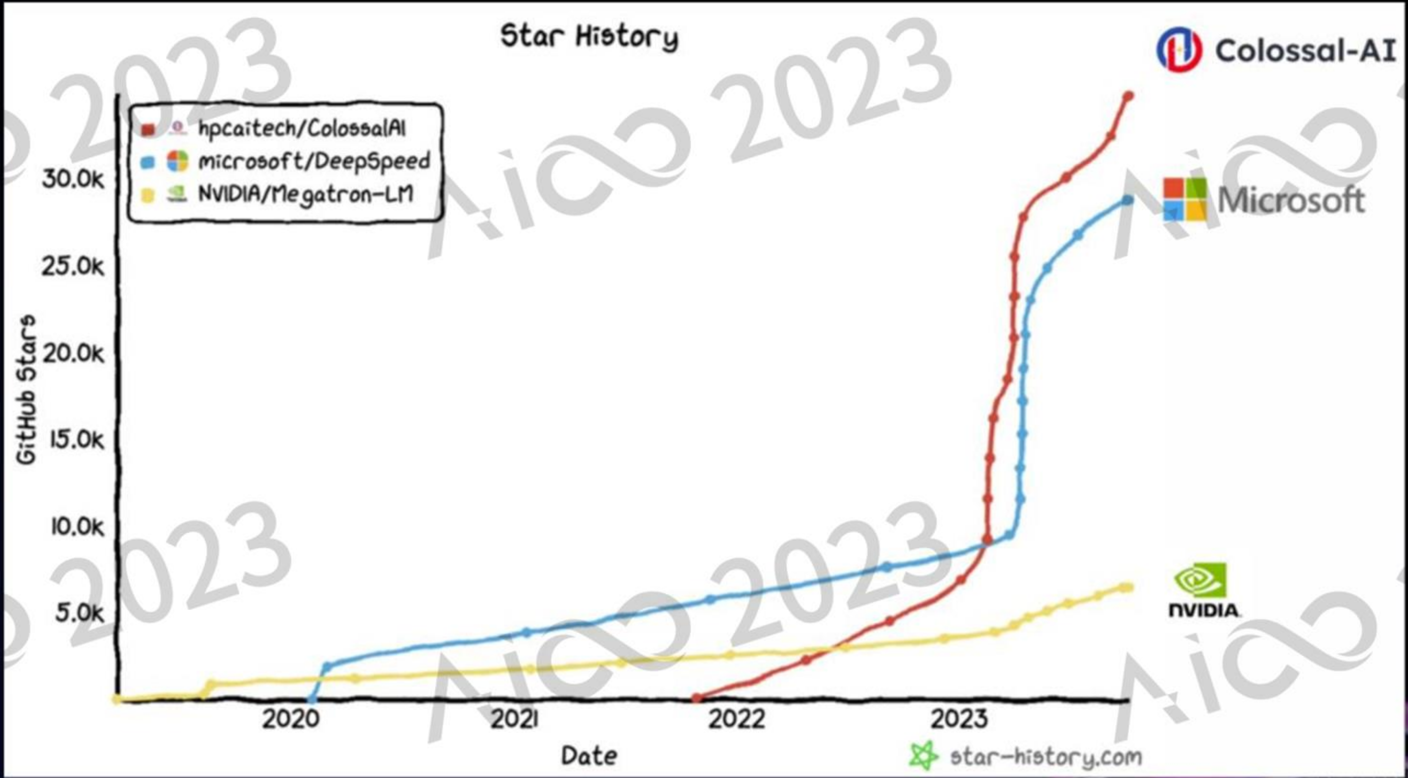
是全球第一个开源复现ChatGPT的RLHF训练过程的完整解决方案



Aico

Colossal-AI

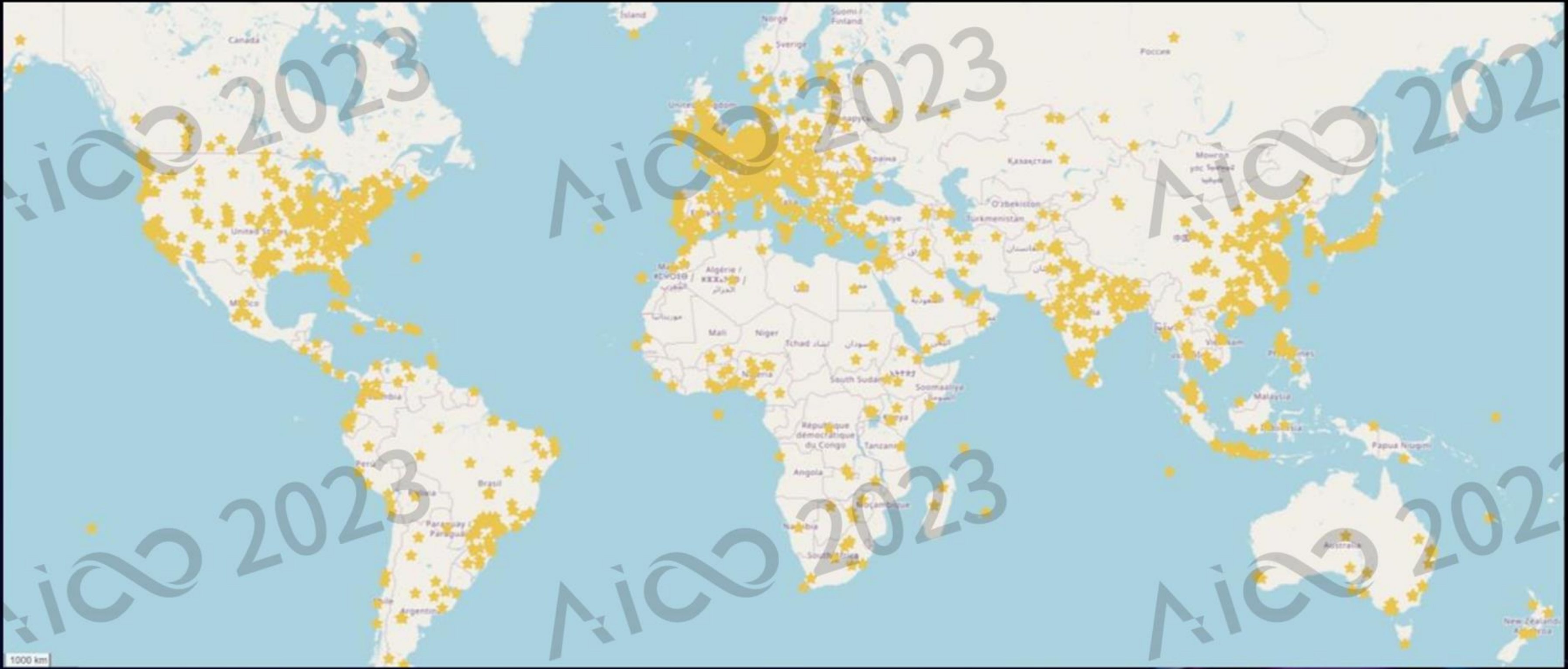
细分赛道世界第一



👉https://github.com/hpcaitech/ColossalAI

Colossal-AI

用户遍布全球



👉https://github.com/python-visualization/folium

Colossal-AI

集成大模型分布式所有SOTA方案

可满足不同实际场景的加速需求

功能	ColossalAI兼容性	DeepSpeed兼容性
Tensor Parallelism	支持1D/2D/2.5D/3D	只支持Megatron 1D
Pipeline Parallelism	支持1F1B/ Interleaved 1F1B	依赖Megatron支持
Zero data Parallelism	支持Zero 1/2/3	支持
TP+PP+Zero+DDP	完全兼容	依赖Megatron支持TP+PP+Zero/DP
Sequence Parallelism	支持	不支持
CPU Offloading	Adaptive offloading	全部offloading
Mixed Precision	支持01/02/03的FP16/BF16	只支持02的FP16/BF16
Activation Checkpointing	Selective recomputation	Block-wise
LAMB Optimizer	支持32/8bit	不支持8bit
Flash Attention 2	支持	不支持
RLHF	支持	支持有限模型
Paged Attention	支持	不支持
Inference Quantization	SmoothQuant/GPTQ/LUT（4/8bit）	DeepQuant（8bit）

AICO Colossal-AI

简单易用

提供从低到高不同粒度的封装接口

用户可以直接使用各种高度解耦的分布式策略

无论小白用户还是专业开发者都能快速上手

```python
from colossalai.booster import Booster, GeminiPlugin
from colossalai.lazy import LazyInitContext

# setup distributed strategy for booster
plugin = GeminiPlugin(placement_policy='auto', ...)
booster = Booster(plugin=plugin)

# initialize model in a lazy fashion
with LazyInitContext():
    model = LlamaModel()
# initialize optimizer
optimizer = AdamW(model.parameters(), ...)

# process model & optimizer according to
# the distributed strategy
model, optimizer = booster.boost(model, optimizer)

# train the model
...
```

AiCO

Colossal-AI

Colossal-AI 学术认可

NEURAL INFORMATION PROCESSING SYSTEMS

SC22 Dallas, TX hpc accelerates.

AAAI Association for the Advancement of Artificial Intelligence

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

ISC High Performance The HPC Event.

NVIDIA GTC GPU TECHNOLOGY CONFERENCE

B NeurIPS, SC, AAAI, PPoPP, CVPR, ISC, NVIDIA GTCC D 际顶级AI5 E F G 计H I 议录J K 为L M tutorialN K O 题P 讲

2

N维N Q $ 统

Colossal-AI

N维并行系统综述

Pipeline parallelism

Partition layers into stages

up to 47% faster

Tensor parallelism

1-D 2-D 3-D

1. Partition a model in 1/2/3 dimensions
2. Distribute a model to many processors
3. Minimize the communication

up to 130% faster

2.5-D

Balance memory & communication

up to 240% higher throughput

Sequence parallelism

Partition a data point

up to 50% longer sequence length, 50% faster

Data Parallelism powered by Large-Batch Algorithms

All in N-Dim Parallelism System from Colossal-AI

LARS/LAMB 的成就和行业影响

The world's leading deep learning training optimizer

LARS
World record for ImageNet training

LAMB
A leap breakthrough in performance

Tradition		LAMB
3 days	→	76 minutes
BERT		

Tradition		LAMB
16 chips	→	1024 chips
Google TPU		

State of the art	→	LAMB
		72 times increase
Source: Nvidia		

Tradition		LAMB
512	→	64K

Widely used in the industry

SimCLR

SEER

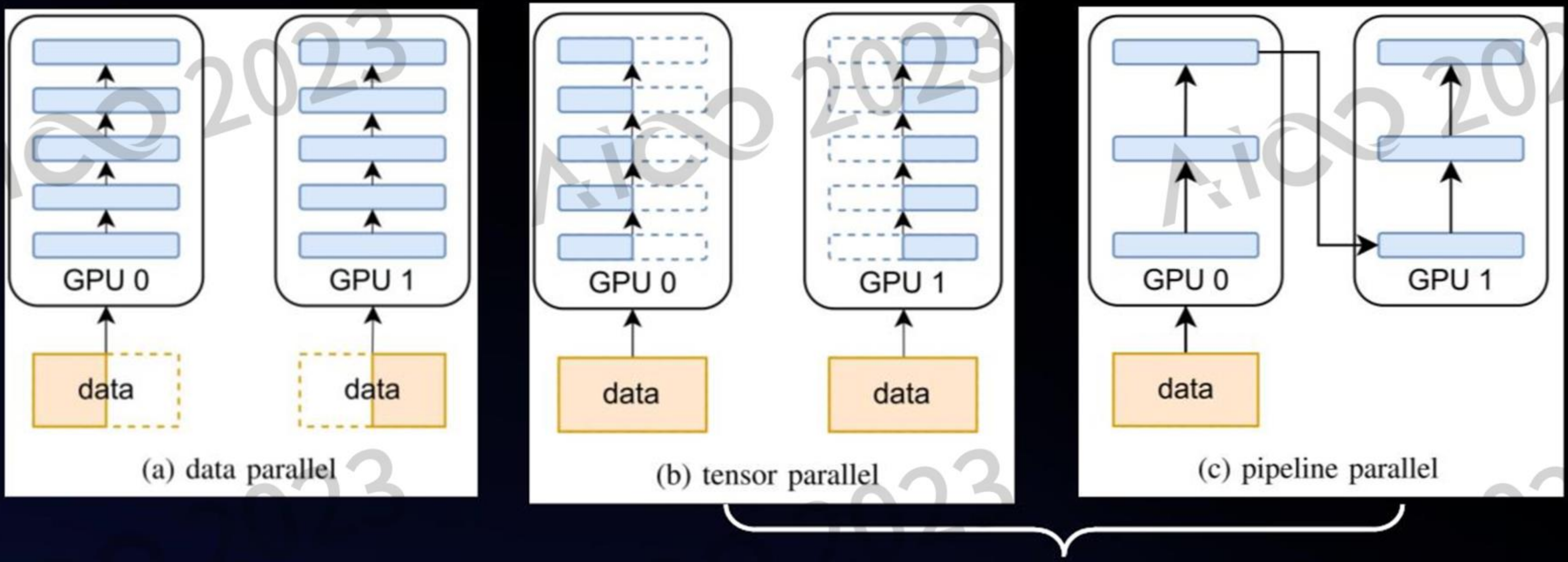
DeepSpeed

BYOL

Expand AI Scale

Industry AI Training Benchmark

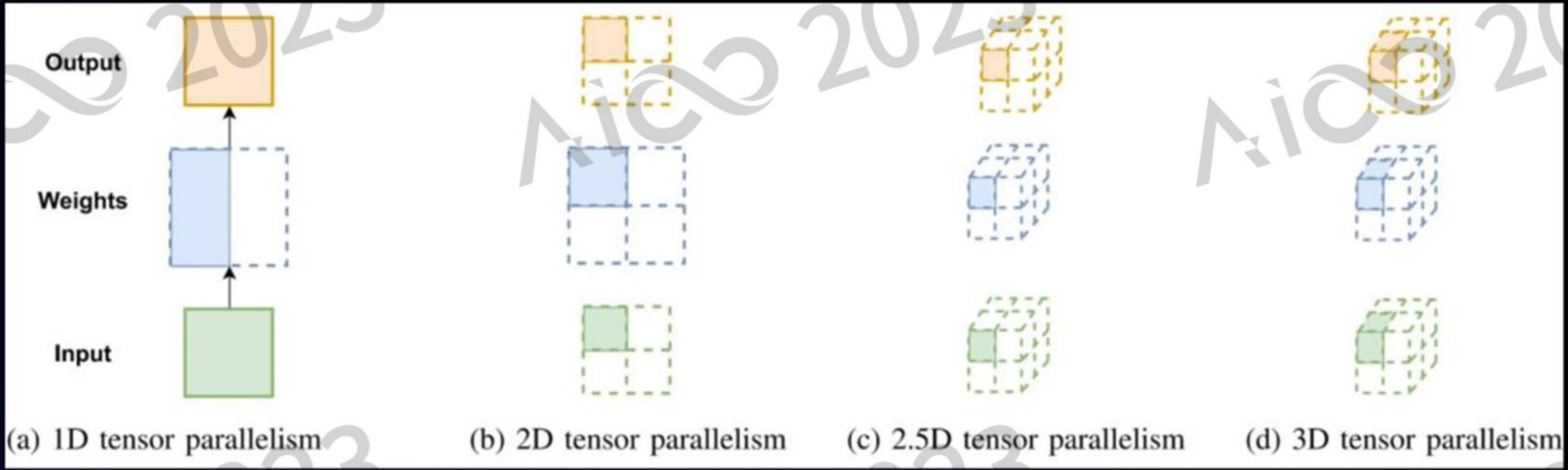
模型并行



Model Parallelism

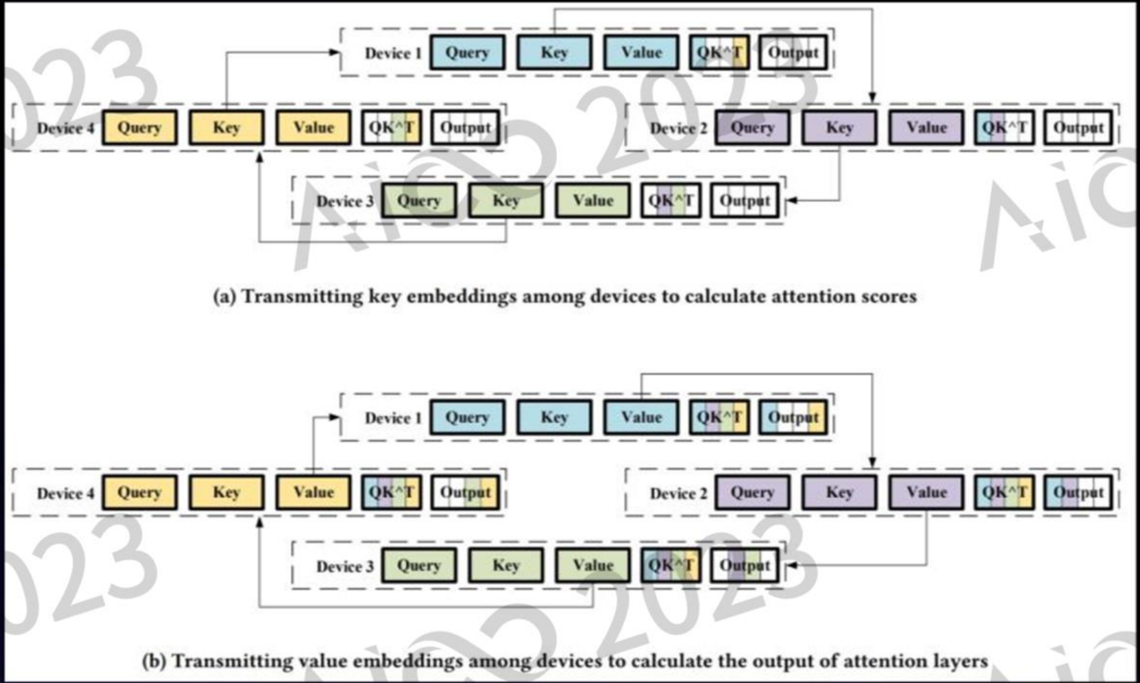
Existing parallelism for distributed training

张量并行



Tensor parallel illustration

Ring Self-Attention



- Inspired by Ring All-reduce
- Communicate query, key and value embeddings for self-attention calculation

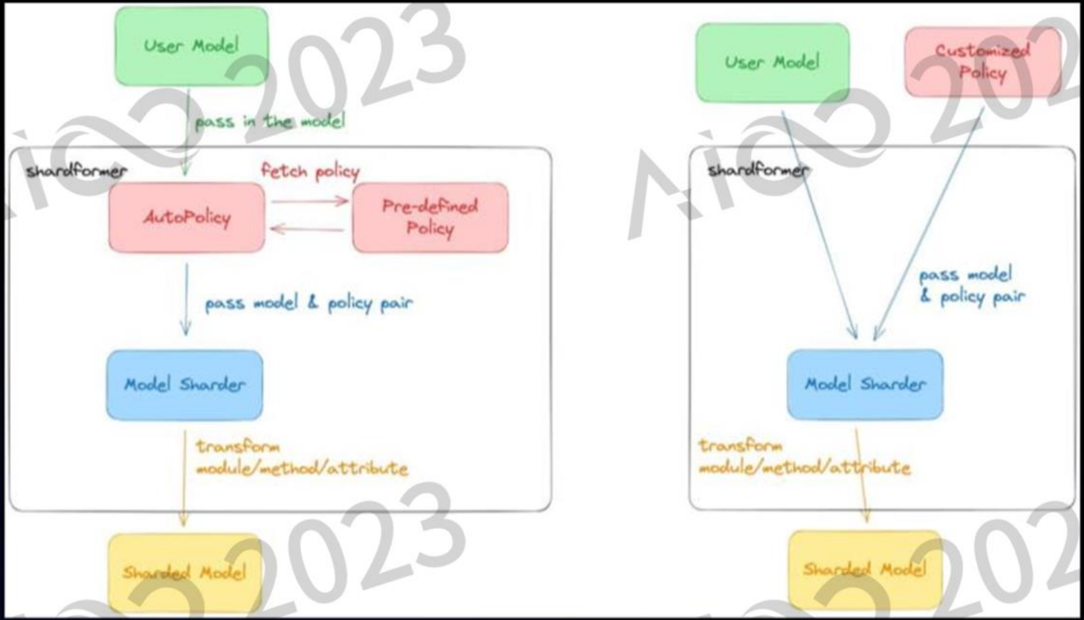
自动并行Shardformer

- 大模型自动扩展最佳并行策略
- 无缝衔接主流社区更新

PyTorch Hugging Face

- LLaMA1/2
- BLOOM
- OPT
- T5
- GPT-2
- BERT
- GLM
-

……

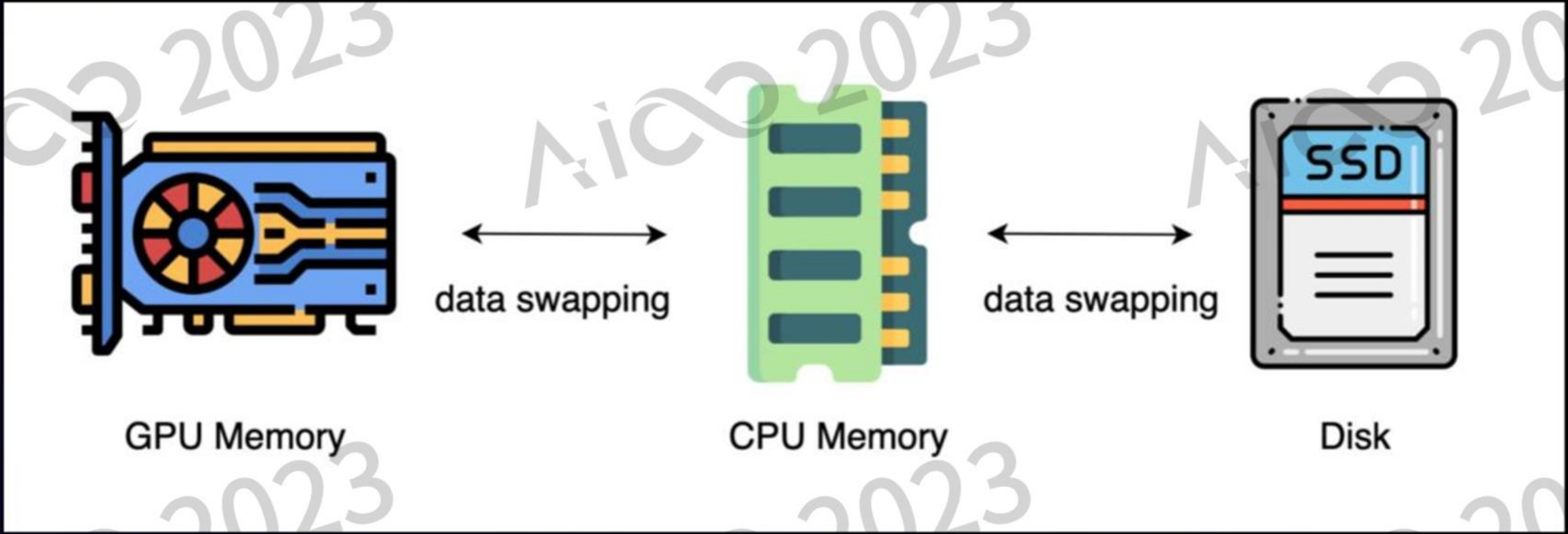


3

E R S T $ 统

Colossal-AI

异构内存系统上的数据移动

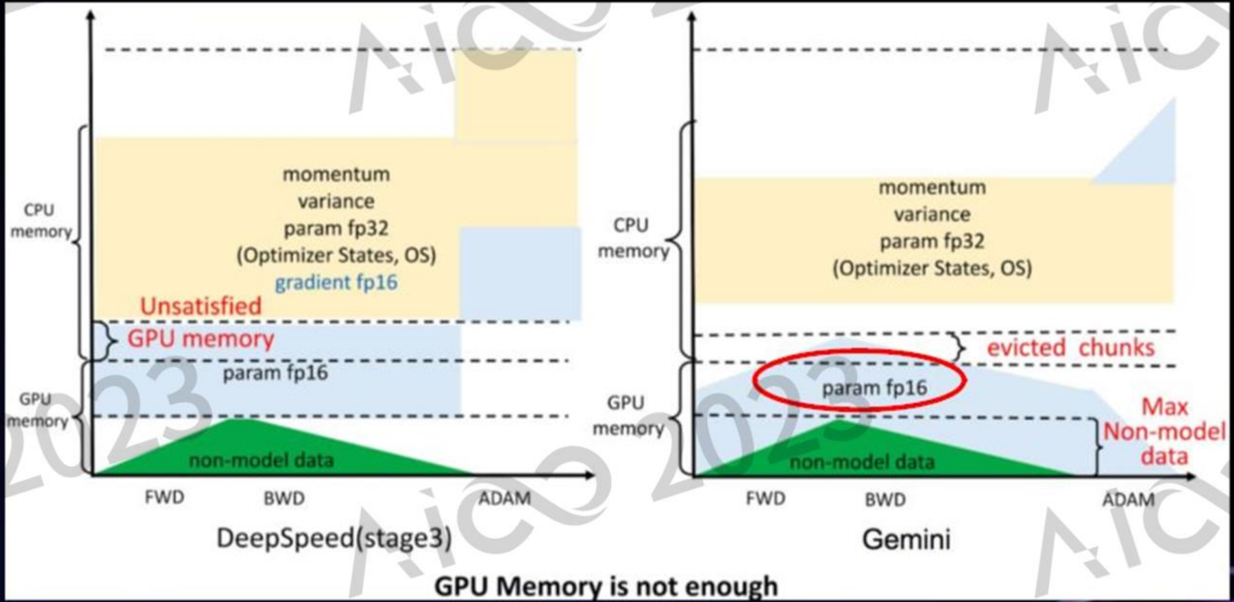


Heterogeneous system illustration

Colossal-AI：优化异构存储系统上的数据移动

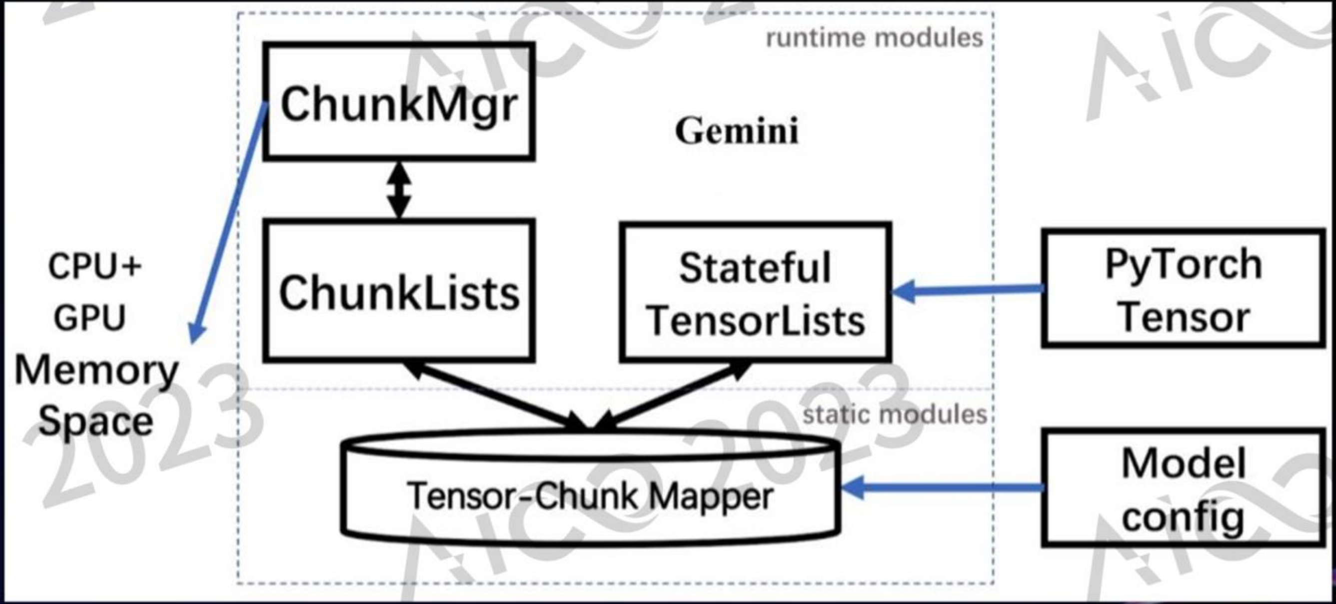
Gemini offload:

- 通过自适应tensor存储管理，来最小化数据移动



Colossal-AI：优化异构存储系统上的数据移动

- 通过将tensor聚合成chunk，来提高通信效率

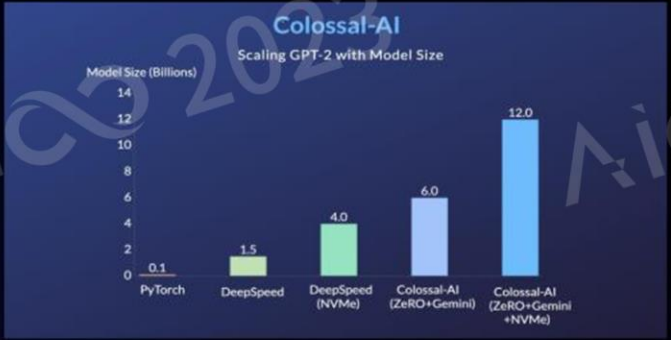


4

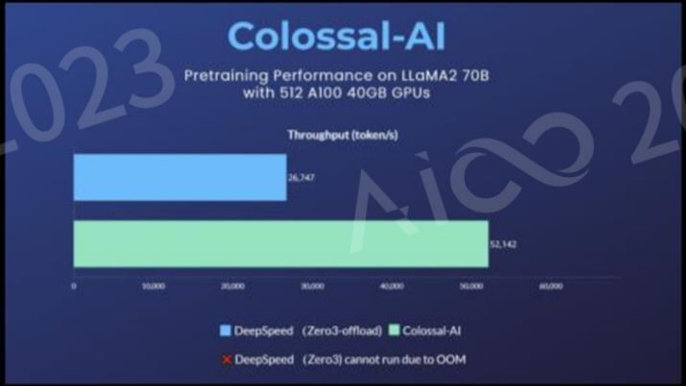
Benchmark & U V

Colossal-AI

Colossal-AI 应用案例

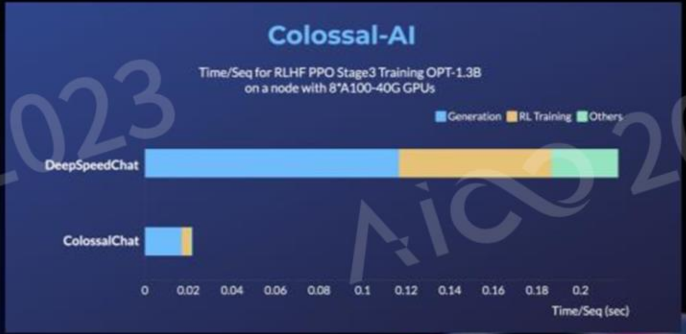


相同硬件的模型容量提升百倍



相同硬件更快速度

首个开源RLHF流程低成本ChatGPT复现方案



ChatGPT训练速度提升10倍

Colossal-LLaMA-2-7B-base：千元预算半天训练，效果媲美主流大模型，开源可商用中文LLaMA-2

Colossal-LLaMA-2

	Backbone	Tokens Consumed	MMLU	CMMLU	AGIEval	GAOKAO	CEval
Baichuan2-7B-Base	From scratch	2.6T	46.97	57.67	45.76	52.60	54.00
Baichuan2-13B-Base	From scratch	2.6T	54.84	62.62	52.08	58.25	58.10
ChatGLM2-6B	From scratch	1.4T	44.74	49.40	46.36	45.49	51.70
InternLM-7B	From scratch	1.6T	46.70	52.00	44.77	61.64	52.80
Qwen-7B	From scratch	2.2T	54.29	56.03	52.47	56.42	59.60
Llama-2-7B	From scratch	2.0T	44.47	32.97	32.60	25.46	-
Linly-AI/Chinese-LLaMA-2-7B-hf	Llama-2-7B	1.0T	37.43	29.92	32.00	27.57	-
TigerResearch/tigerbot-7b-base	Llama-2-7B	0.3T	43.73	42.04	37.64	30.61	-
FlagAlpha/Atom-7B	Lama-2-7B	0.1T	49.96	41.10	39.83	33.00	-
IDEA-CCNL/Ziva-LLaMA-13B-v1.1	Llama-13B	0.11T	50.25	40.99	40.04	30.54	-
Colossal-LLaMA-2-7b-base	Llama-2-7B	0.0085T	53.06	49.89	51.48	58.82	50.2

- 8.5B token数据、15小时、数千元RMB的训练成本
- 与开源社区同7B规模预训练SOTA模型媲美
- 克服英文的灾难性遗忘
  MMLU: 44.47 -> 53.06
- 中文能力极大提升
  CMMLU: 32.97->49.89

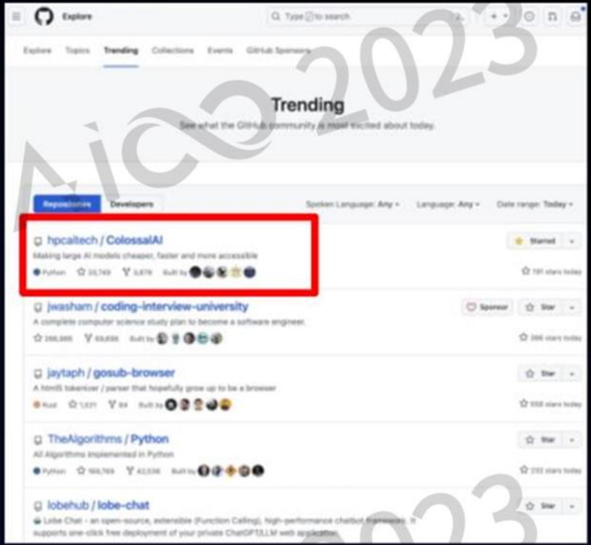
Colossal-LLaMA-2-7B-base：千元预算半天训练，效果媲美主流大模型，开源可商用中文LLaMA-2

- 通过SFT、LoRA等方式微调或Prompt工程，能有效注入基座模型的知识与能力十分有限
- 不能较好的满足高质量领域知识或垂类模型应用的构建的需求

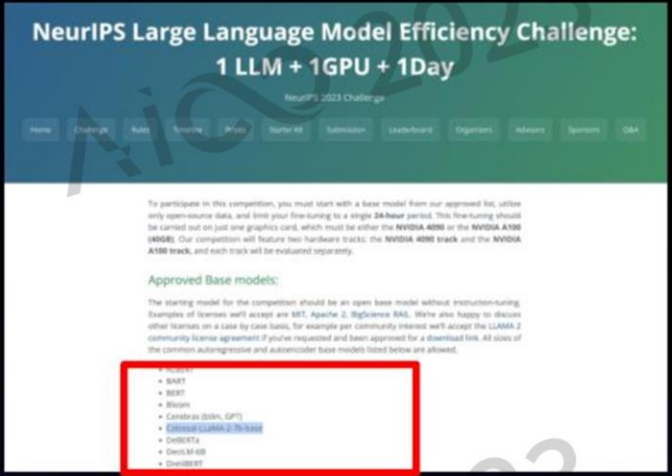
- 大规模预训练阶段
- 中文知识注入阶段
- 相关知识回放阶段

- 任意领域垂类大模型
- 极快开发速度与极低开发成本

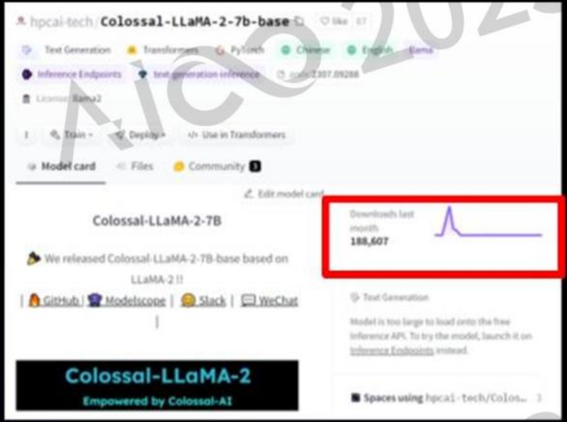
Colossal-LLaMA-2-7B-base 模型开源后成绩



GitHub Trending榜世界第一



2023 NeurIPS 大模型竞赛官方基座模型



Huggingface 三周 188K+下载量

HF 👉https://huggingface.co/hpcai-tech/Colossal-LLaMA-2-7b-base

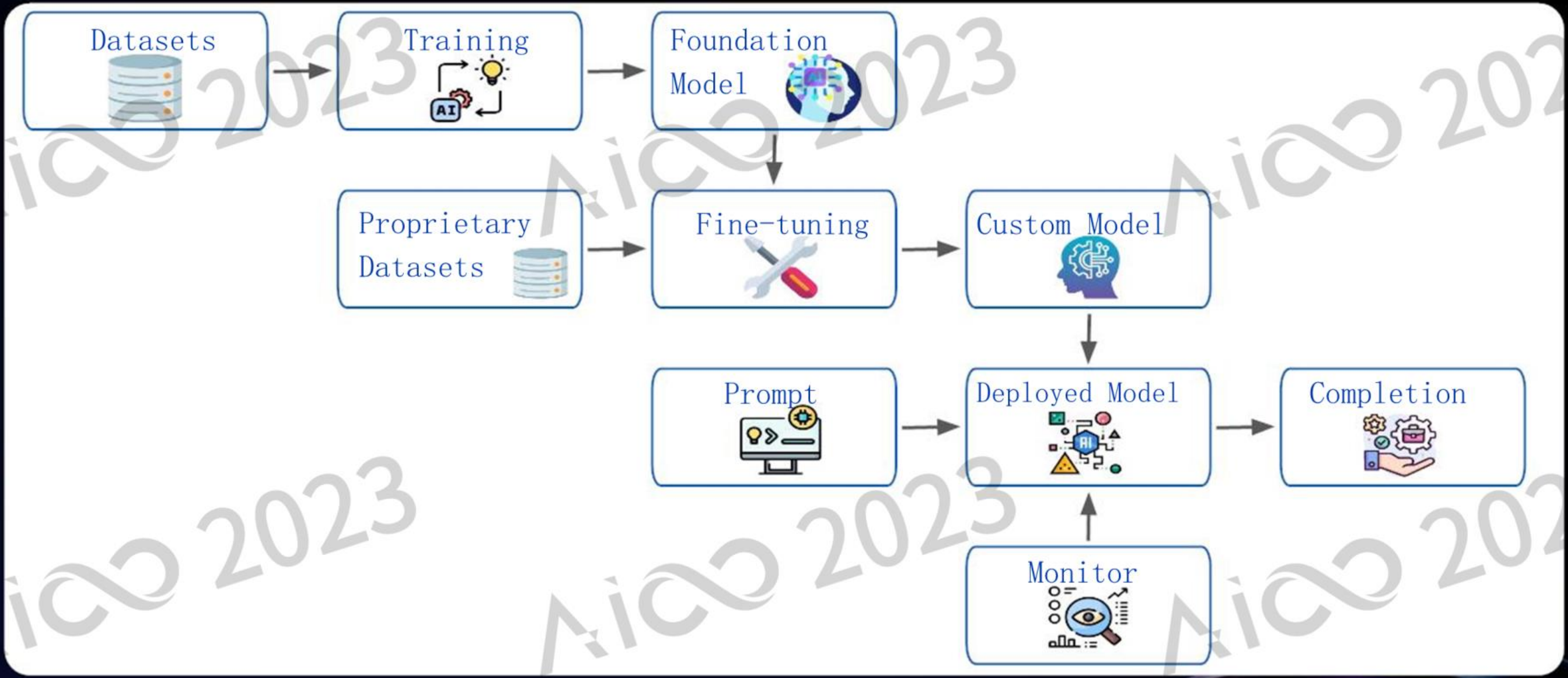
Modelscope 👉https://modelscope.cn/models/colossalai/Colossal-LLaMA-2-7b-base/summary

5

Colossal-AIW X Y &Z [/

Colossal-AI

LLMOps 全流程一站式低成本高效解决



Colossal Platform 解决AI大模型全流程痛点

训练

A \ ] \ 亿1 2 规 预训练A ˆ _ ` a

A \ b 预H c d e f 领g Q 业h i % j k

(调

A p q 6 7 ! r 3 s t u a

A v w x > & ' RLHFy z { | } ChatGPT

ˆ 现M ¤

l m

A LLMl m ˆ _ s t 10a

A Stable Diffusionl m ˆ _ s t 8a

n o

A Z 键n o § { | } ¨ 动弹F ° 缩

A l m API± • × |

https://platform.luchentech.com/

Colossal-AI 极速大模型工作站

Our Supports

ĭ E í ! 软7
A ě ī B Γ 优% ∧PyTorch M N Ξ O 领ŏ
$ 统
Colossal-AI A DeepSpeed A Megatron-
LMAFSDPC

1h

Z H ! 库5 Θ V
A ¨带GPTALLaMA,
FalconAVicunaASDC ! 训练A
n o A (调A l m C v y z
A & I K U @14 时 E | ! (调

10V

B Γ ! ˆ _ 5 R Δ
A 训练` ˆ 2-10a
A 304 时 E | LLaMA 7B ! 训练

100V

r 纳Π
A p q 6 7 ě ī 100a P ! A单
/ Σ G (调GPT

Our Survices

A 软7 à á # è 态$ 统@s é # à K ê ì í ò U ! ó ÷软7 à á § 库ù ú ü ā
A 扩ē F @ě ī ń í ¨ň /ō ň & ' A s t # 6 7 ū r F ā
A ě ī # ǎ 务@s é ĭ E 优ŏ 级ǔ 术ě ī Aû @2 ú ! ǔ ù # 标a § ! 训练/(调/l m g 训ǔ 术ˇ 导§H ¯ ! 调优ń í 5 ` ˆ A K C ā

Thanks for your time T

https://github.com/hpcaitech/ColossalAI ! ”# $ 商务合作

Welcome to contribute T