DaoCloud

基于云原生的方式分布式微调大模型

主讲人：黄敏杰

AI 进阶指南（上） 从理论到实践

DaoCloud

课程安排 AGENDA

07.23	深入浅出：大语言模型	张凡石	「DaoCloud 道客」高级研发工程师
07.30	基于 RAG 的 AI 应用落地	陈　佳	rDaocloud 道客」架构师
08.06	智能体的构建、迭代和高可用	尹伯昊	猴子无限 创始人
08.13	云原生技术优化模型推理	王　璠	「DaoCloud 道客」高级研发工程师
08.20	基于分布式和容器的方式微调大模型	黄敏杰	「DaoCloud 道客」高级研发工程师

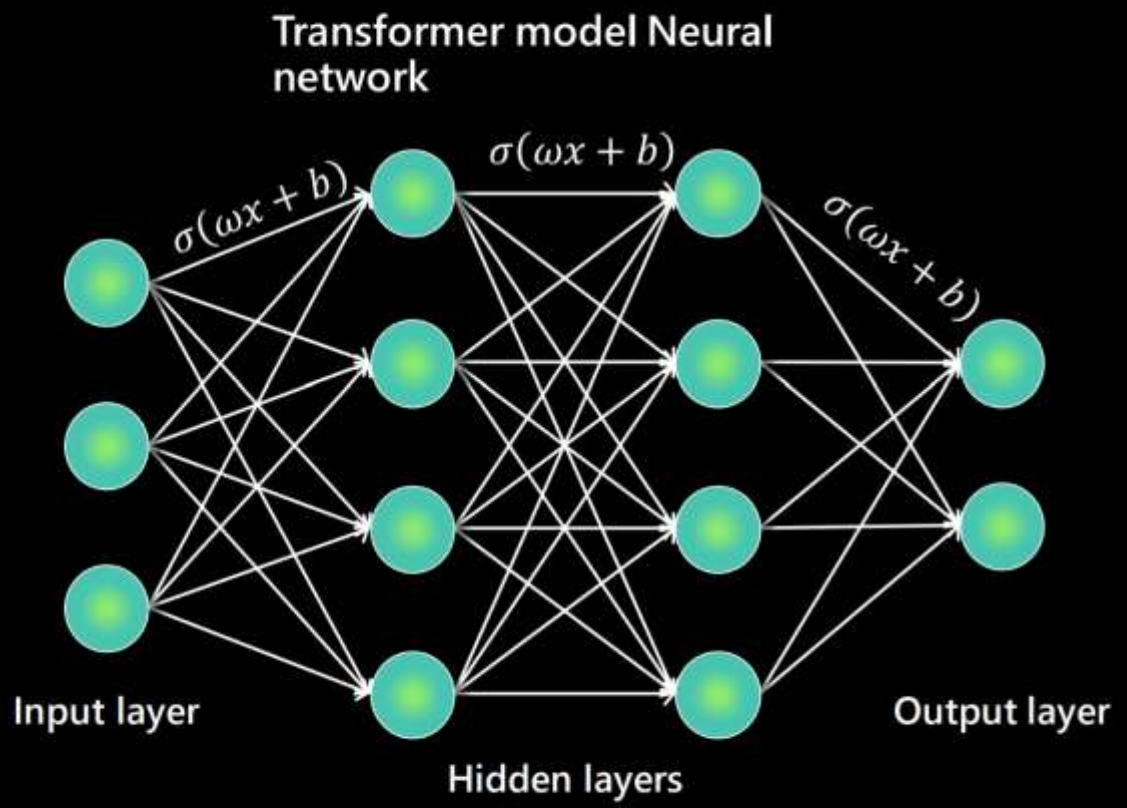
Content
目录

Part 01

大模型微调概述

大模型LLM

How large are they?



BERT Large - 2018
345M

GPT2 - 2019
1.5B

GPT3 - 2020
175B

Turing Megatron NLG 2021
530B

GPT4 – 2023
1.4T (estimated)

什么是大模型微调

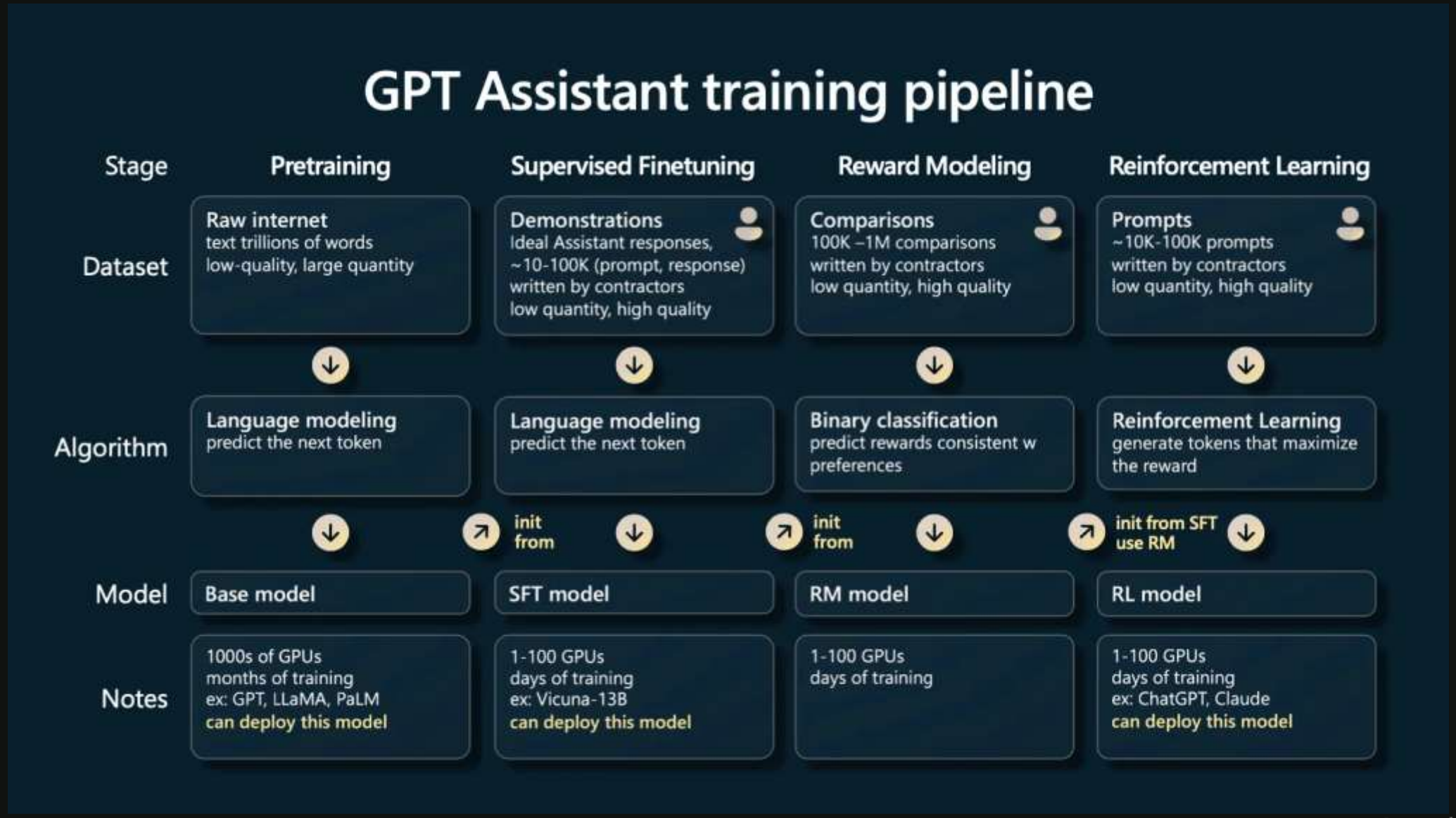
定义：

大模型微调是指在一个已经预训练的大型模型（如GPT、BERT等）的基础上，通过在特定领域或特定任务上进行进一步训练，以优化其表现。目的是使模型能够更好地适应特定任务的需求，从而提高精度和性能。

为什么需要微调：

- 提高准确性: 微调可以大幅提高模型在特定任务上的准确性和性能。
- 减少计算资源: 不需要从头开始训练模型，大幅减少计算成本和时间。
- 适配不同任务: 微调可以将模型调整为适应特定领域或任务的需求。

大模型训练流程



增量预训练

增量预训练，在海量领域文档数据上二次预训练基础模型，以注入领域知识。
侧重于提升模型对新领域的适应性。


```
[
  {
    "text": "Don’t think you need all the bells and whistles? No problem. McKinley Heating Service Experts Heating
  },
  {
    "text": "To the apparent surprise of everyone, the Walt Disney Company has announced a deal to purchase Lucasf
  },
  {
    "text": "I hadn’t been to Red Mountain in over 4 years and was happy to experience it again. Milana Knowles, ou
  },
  {
    "text": "Last month's fixture window provided a tease, but the final set of international matches for 2018 shou
  },
  {
    "text": "It was a very busy, but extremely enjoyable weekend here at The Cookery School, with a birthday celebr
  },
```

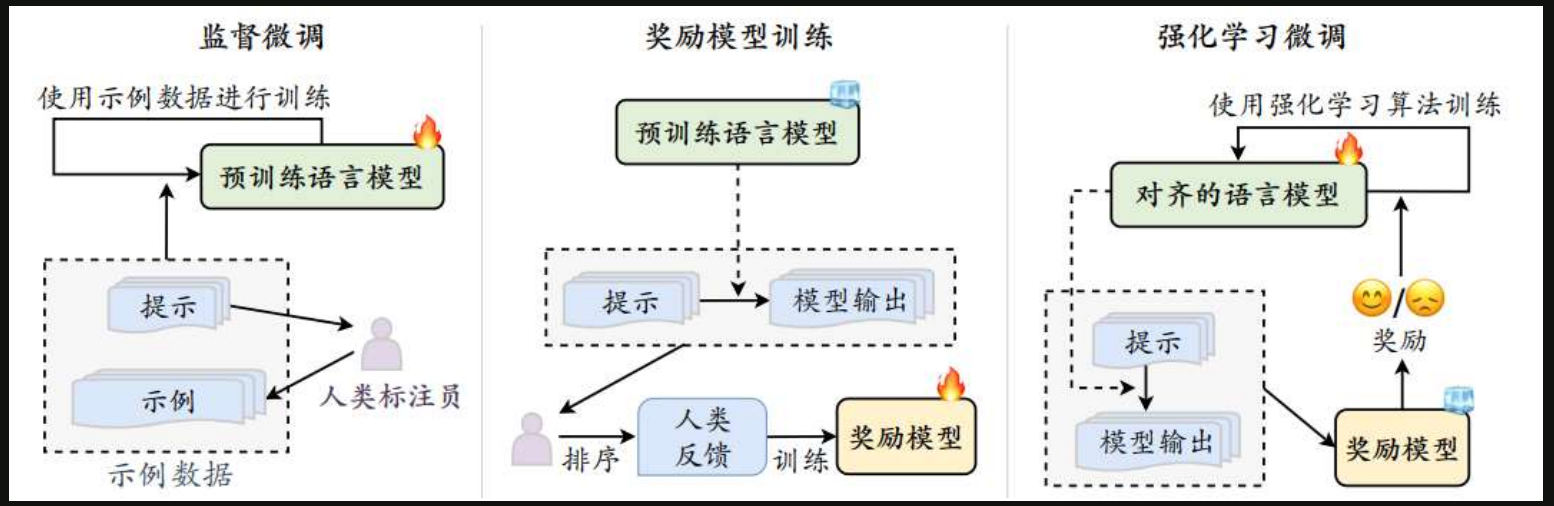
指令微调（SFT）

指令微调专注于让模型理解和遵循人类指令，通过在由（指令，输出）对组成的数据集上进一步训练大型语言模型。
侧重于提升模型对特定指令的遵循能力和任务性能。


```
{
  "instruction": "写一个关于上周末你做了什么的简短叙述。",
  "input": "",
  "output": "由于我是一个人工智能助手，所以我并不能真正的"做"什么事情。我的"工作"时间是24/7无休，确保我随时可供用户提供帮助和服务。所以在上周末，我
},
{
  "instruction": "创建一个程序，打印以下图案：",
  "input": "",
  "output": "您好！以下是一个打印指定图案的示例 Python 程序，它可以打印出一个类似字母"X"的图案。您可以使用程序中的一些规律性代码并对它进行更改，以
},
{
  "instruction": "分类以下数字系列。",
  "input": "输入：\n2、4、6、8",
  "output": "此系列数字为偶数序列。"
},
{
  "instruction": "列出一个应该在野营应急包中的7件物品。",
  "input": "",
  "output": "这是一个我们应该在野营应急包中放置的7件物品列表：\n\n1. 先进救援包：用于治疗划痕，创伤，烧伤，刀伤等紧急情况。\n2. 多用途刀具：可用于
},
```

基于人类反馈的强化学习（RLHF）

基于人类反馈的强化学习，是一种将人类对模型输出的评价转化为奖励信号，进而利用这些信号来训练优化大模型的技术。
RLHF的优势在于能够提高模型输出与人类价值和期望的一致性，使模型行为更符合伦理和用户偏好。




```
[
  {
    "instruction": "人类指令（必填）",
    "input": "人类输入（选填）",
    "chosen": "优质回答（必填）",
    "rejected": "劣质回答（必填）"
  }
]
```

微调的挑战

1 高质量的数据集

微调需要大量与目标领域或任务相关的高质量数据，获取和构建这样的数据集是一个挑战。

2 超参数微调

微调的过程非常复杂，涉及控制优化过程的各种超参数，如学习率，批大小等。

3 计算资源需求

大模型的微调需要大量的GPU计算资源，这些资源的获取和维护成本高昂。

4 模型评估

传统的评估指标并不能完全反应LLM的能力，如何精确评估LLM能力会是一个挑战。

Part 02

高效微调方法

微调方法分类

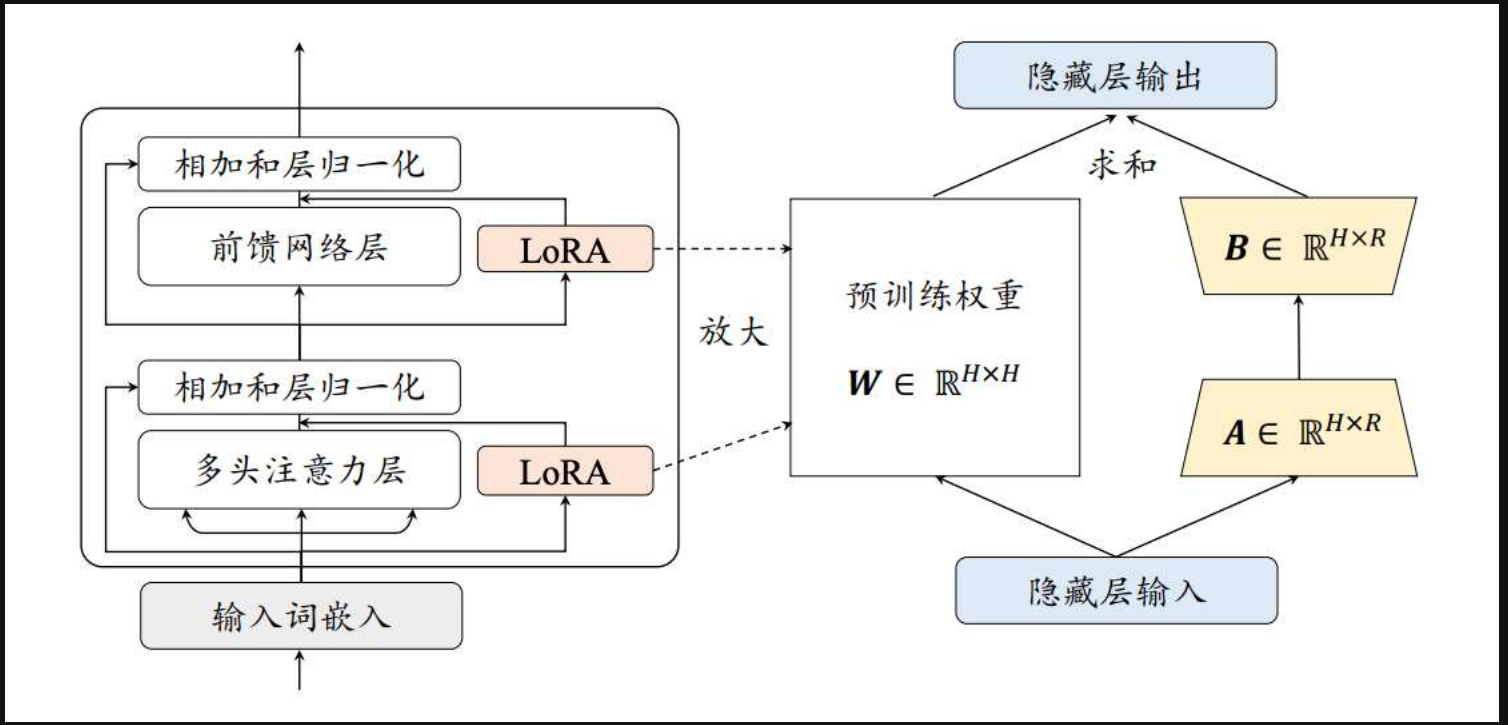
全量微调（Full Fine-tuning）：

全微调是指对整个预训练模型进行微调，包括所有的模型参数。在这种方法中，预训练模型的所有层和参数都会被更新和优化，以适应目标任务的需求。这种微调方法通常适用于任务和预训练模型之间存在较大差异的情况，或者任务需要模型具有高度灵活性和自适应能力的情况。Full Fine-tuning需要较大的计算资源和时间，但可以获得更好的性能。

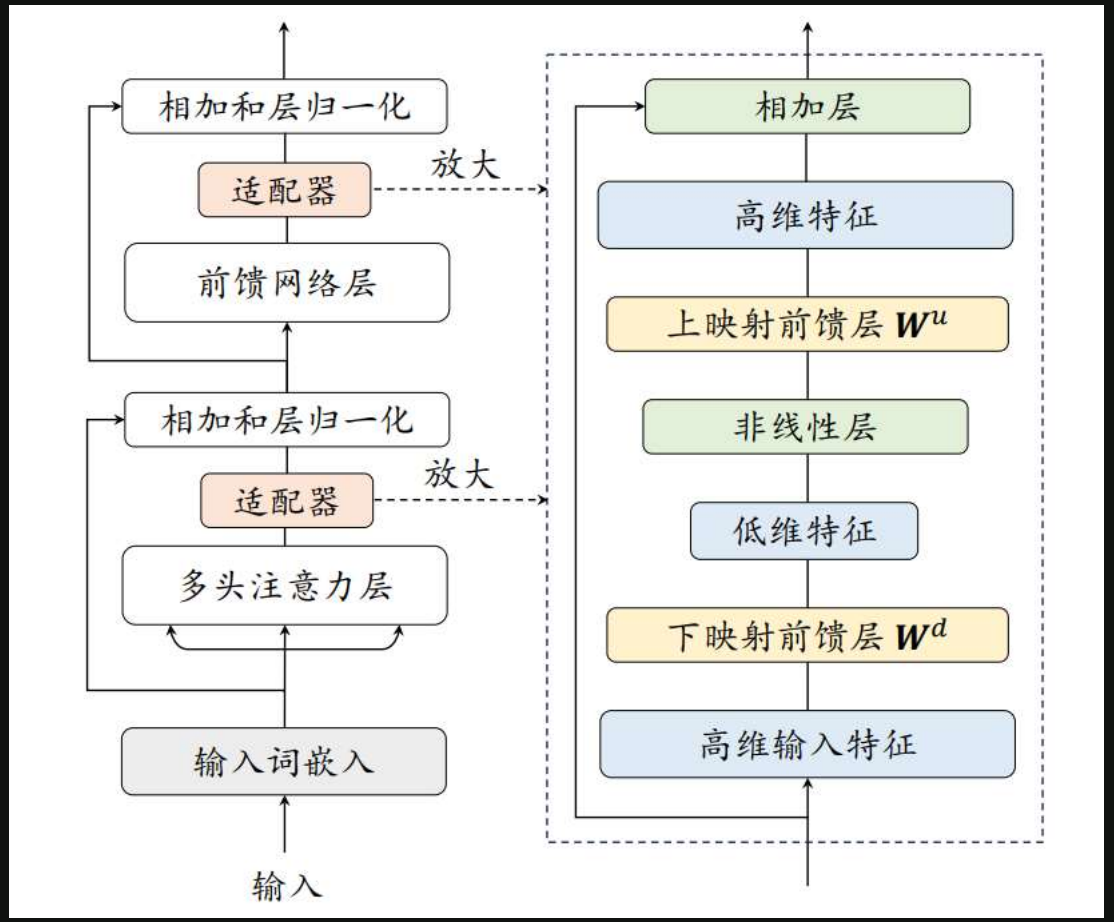
参数高效微调（Parameter-Efficient Fine-Tuning，简称PEFT）：

PEFT通过添加少量可训练的参数，如适配器或小型网络模块，来适应新任务，而不是重新训练整个模型。这种方法的优点在于减少了计算资源的消耗，加快了训练速度，并有助于避免灾难性遗忘。PEFT特别适用于数据量较小的任务，因为它不需要大量的数据来更新大量的参数

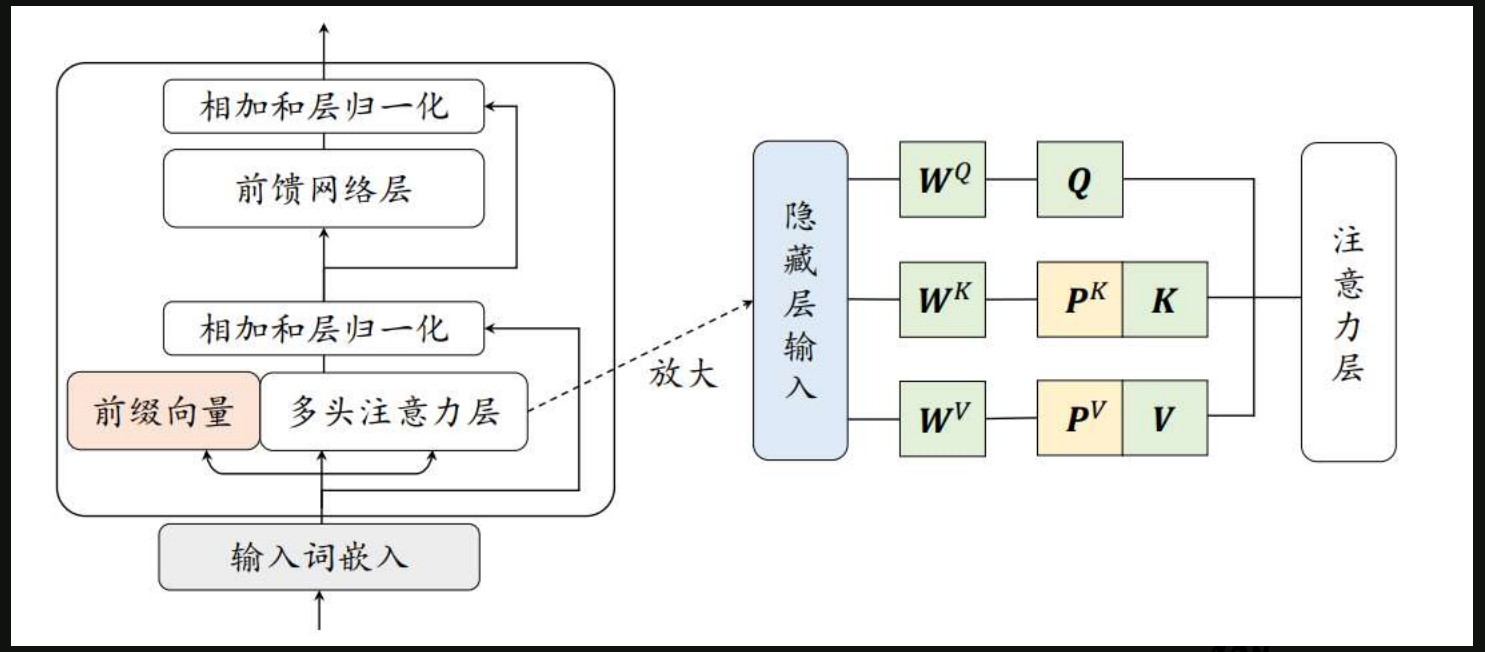
LoRA



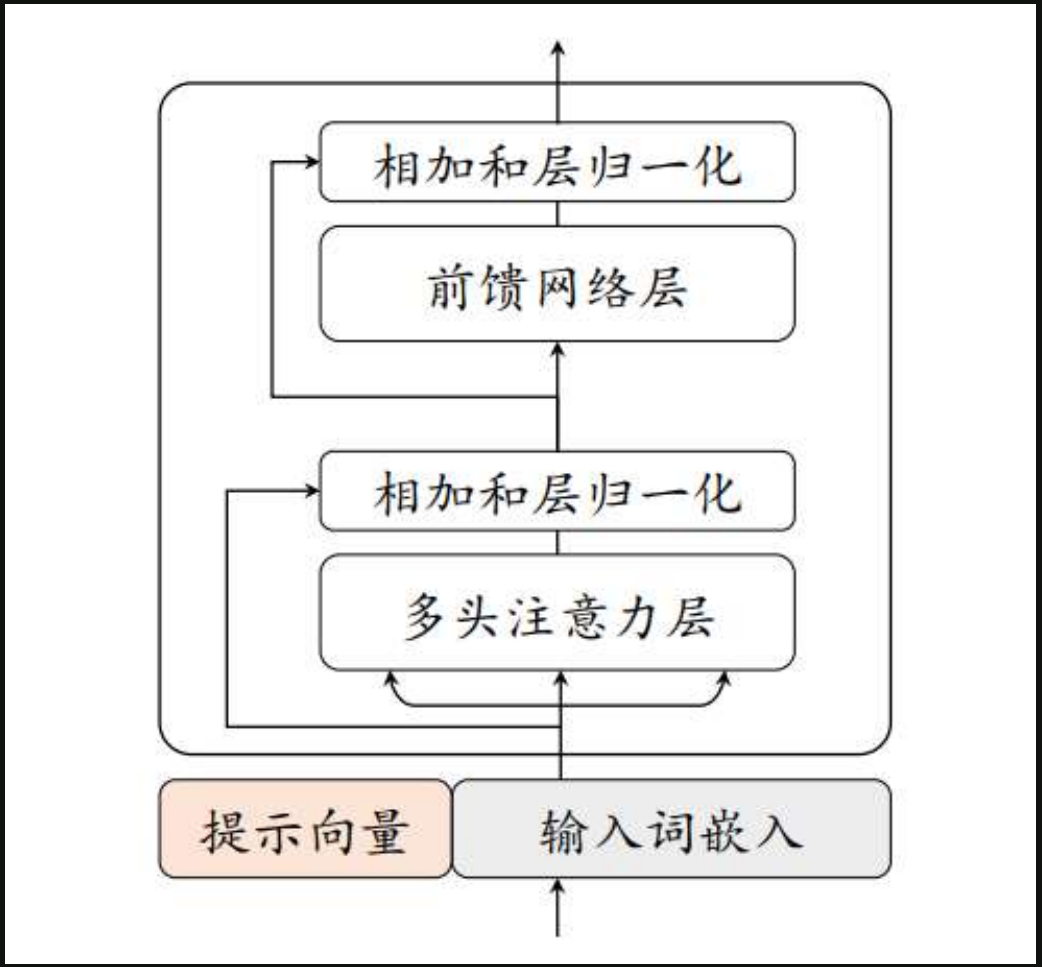
Adapter Tuning



Prefix Tuning



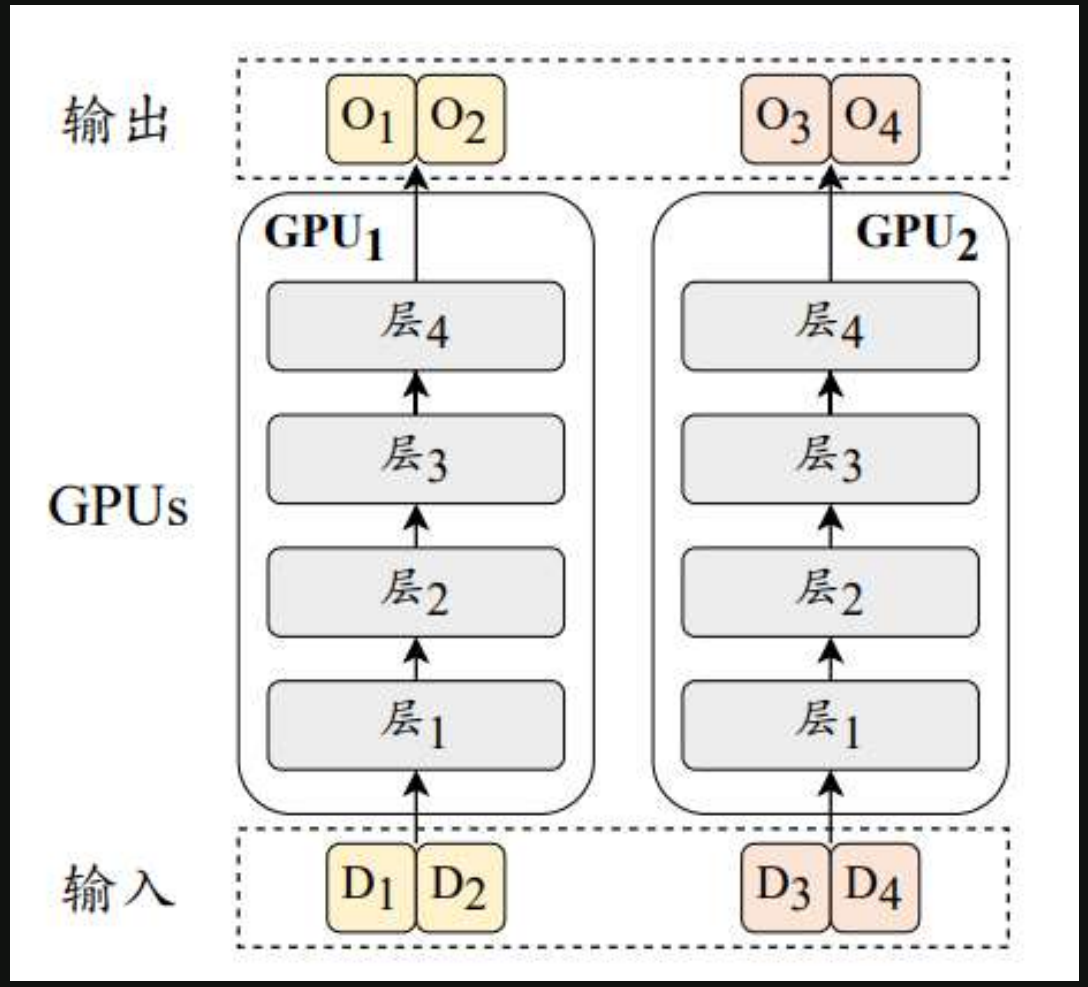
Prompt Tuning



Part 03

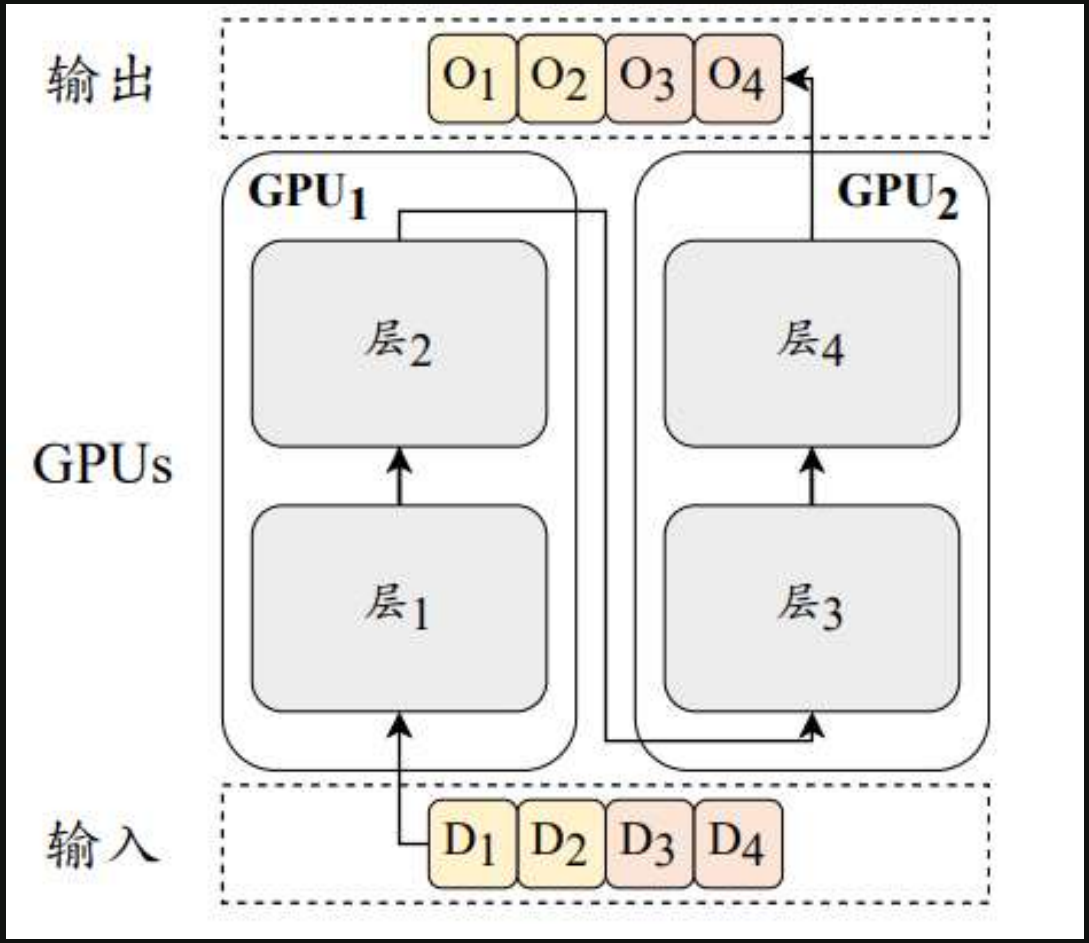
分布式微调技术

数据并行（DP）



数据并行是一种提高训练吞吐量的方法，它将模型参数和优化器
状态复制到多个 GPU 上，然后将训练
数据平均分配到这些 GPU 上。

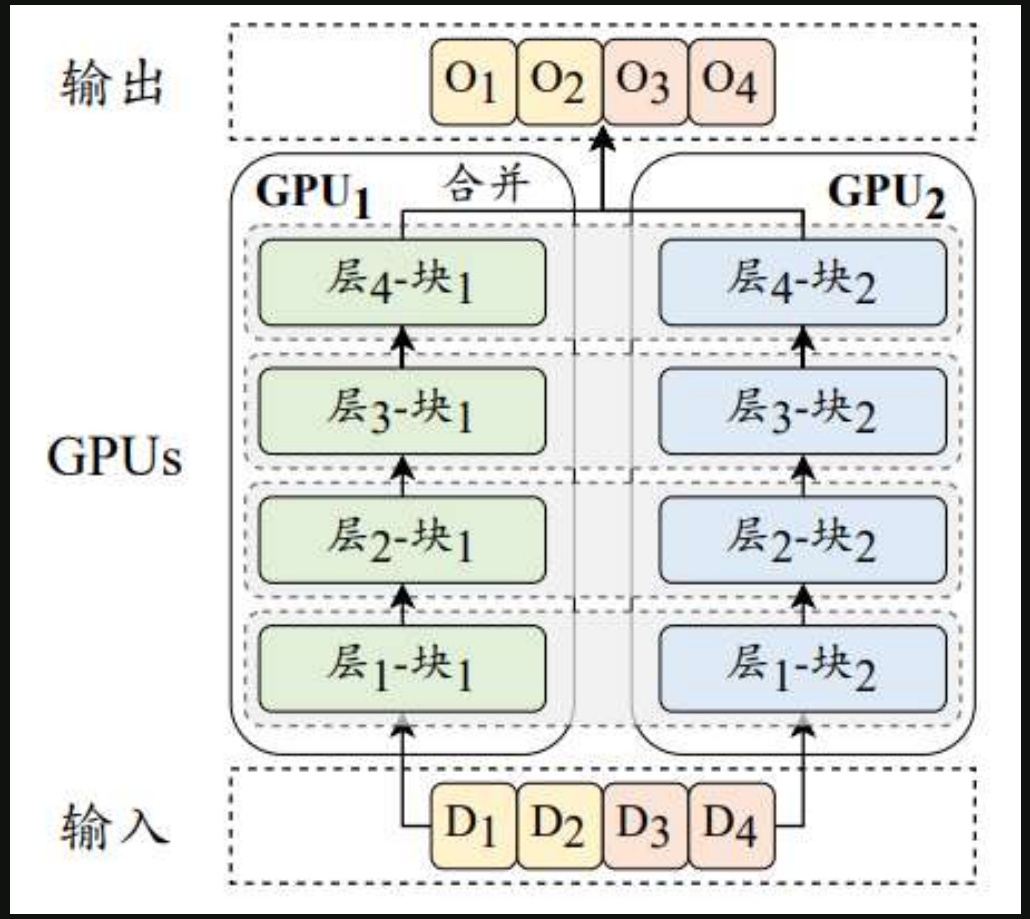
流水线并行（PP）



流水线并行旨在将大语言模型不同层的参数分配到不同的GPU 上。

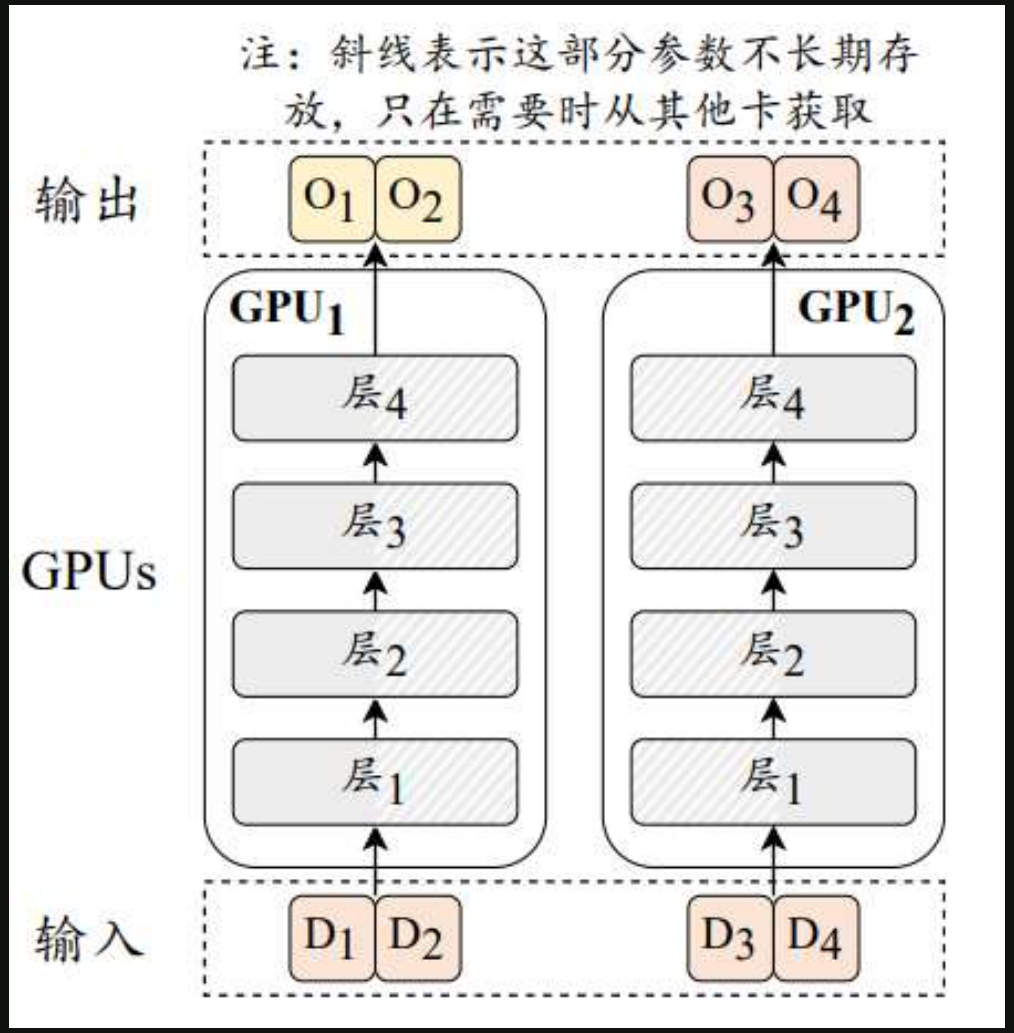
在实践中，可以将 Transformer 连续的层加载到同一 GPU 上，以减少GPU 之间传输隐藏状态或梯度的成本。

张量并行（TP）



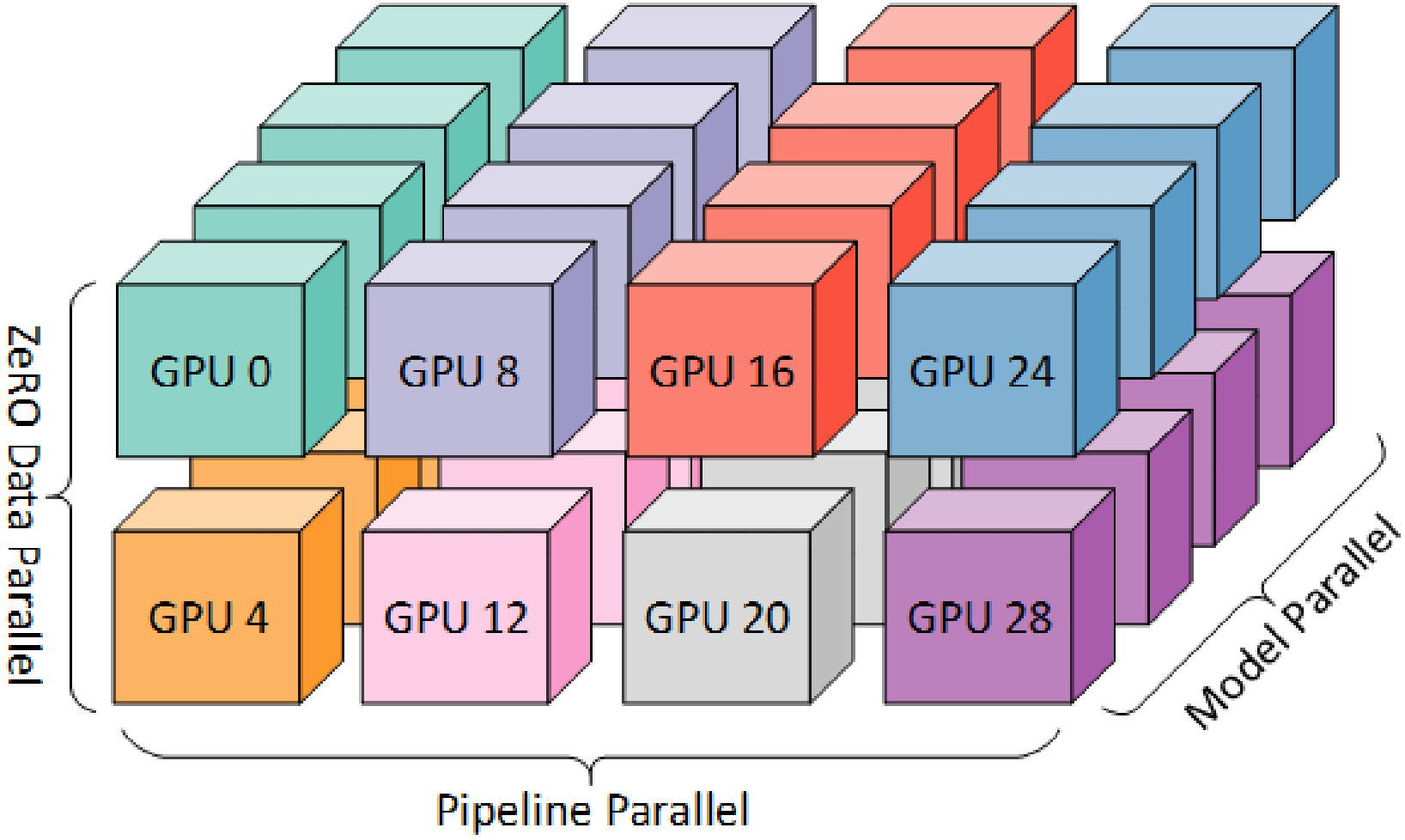
张量并行也是将大模型参数加载到多个 GPU 上的并行技术，它分配粒度更细，进一步分解了模型的参数张量（即参数矩阵），以便更高效地利用多个 GPU 的并行计算能力。

零冗余优化器（ZeRO）



零冗余优化器主要用于解决数据并行中的模型冗余问题。
ZeRO 技术仅在每个 GPU 上保留部分模型参数和优化器参数，当需要时再从其它 GPU 中读取。

3D 并行

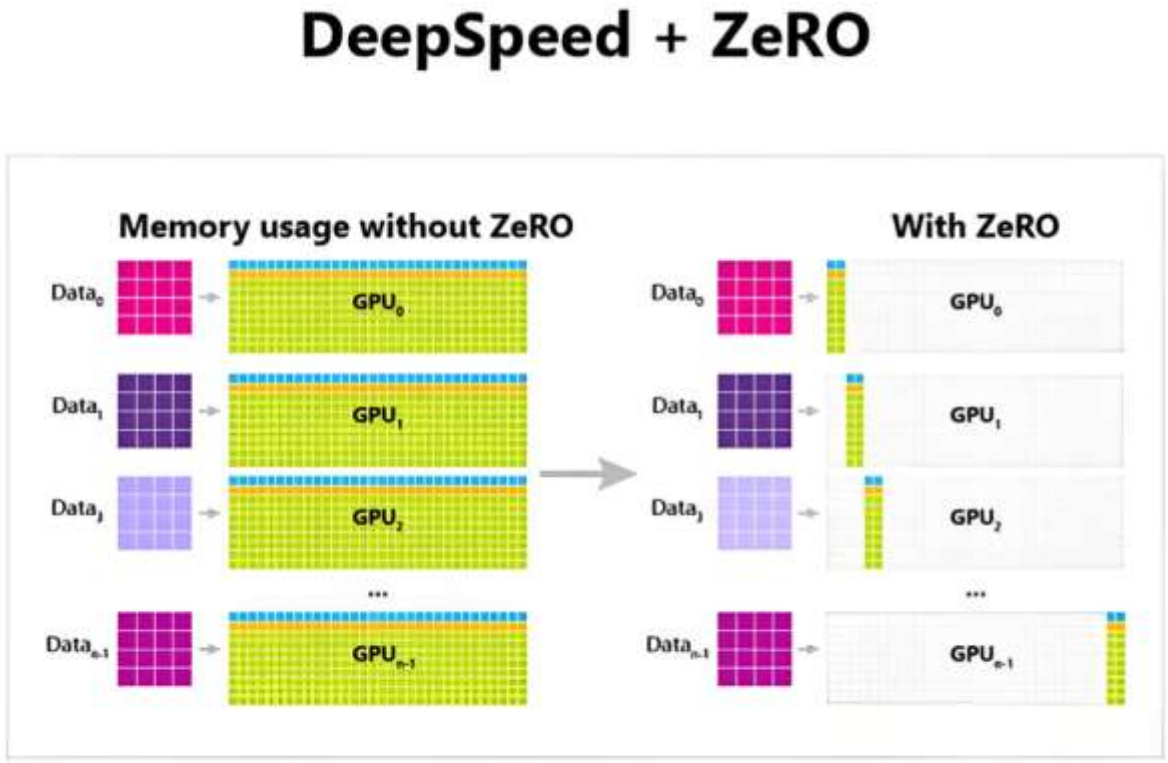


为了更高效地训练，可以将 PP、TP 和 DP 相结合，称为 3D 并行

deepspeed

deepspeed

DeepSpeed 是一个深度学习优化库，使分布式训练和推理变得简单、高效、有效。



Scale
- 100B parameter
- 10X bigger

Speed
- Up to 5X faster

Cost
- Up to 5X cheaper

Usability
- Minimal code change

Part 04

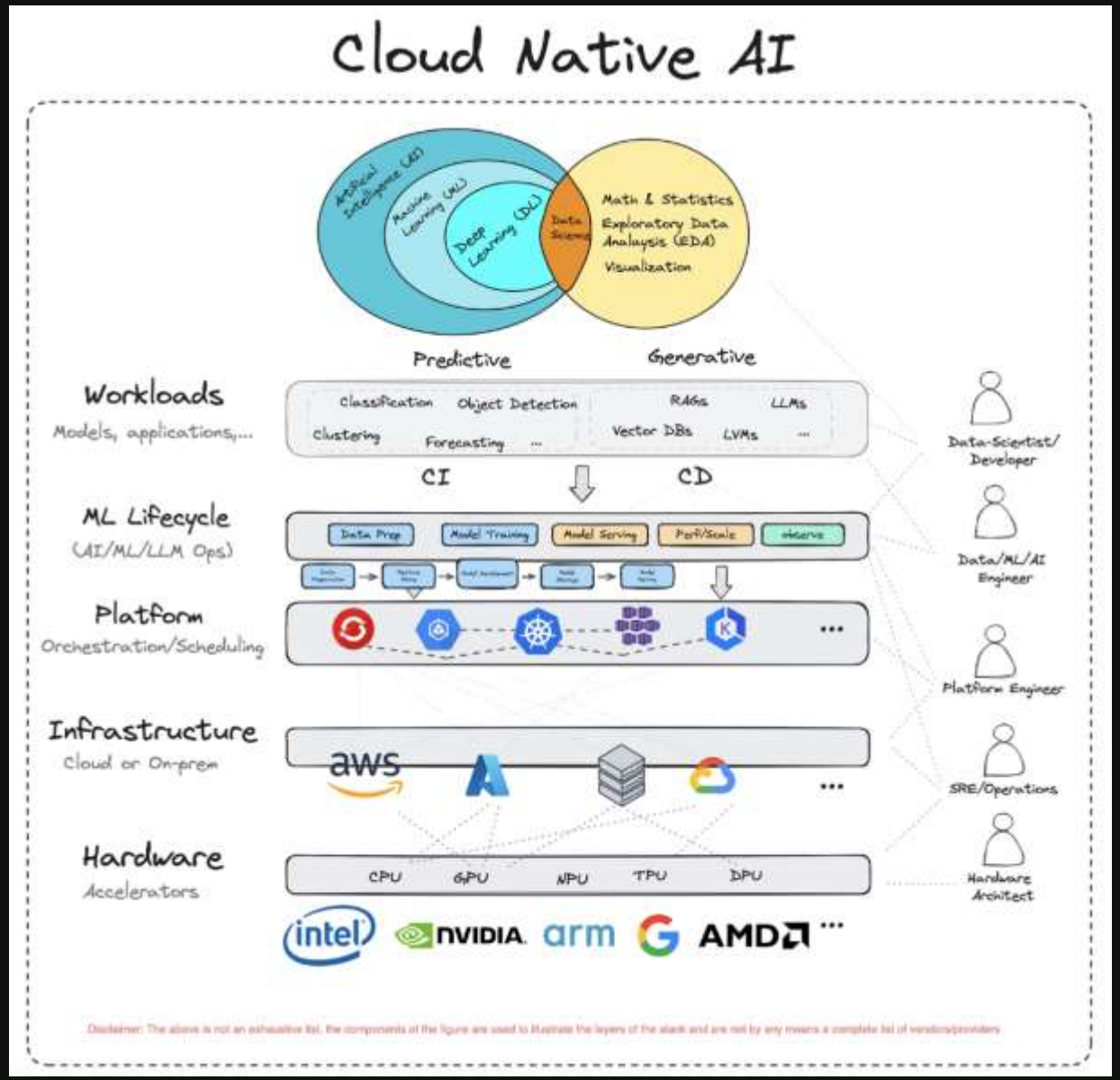
云原生的方式微调大模型

Llama3 Infrastructure

- 超过16K H100 GPUs
- 由7500台服务器组成的存储集群，提供240pb的存储空间，支持2tb /s的可持续吞吐量和7tb /s的峰值吞吐量

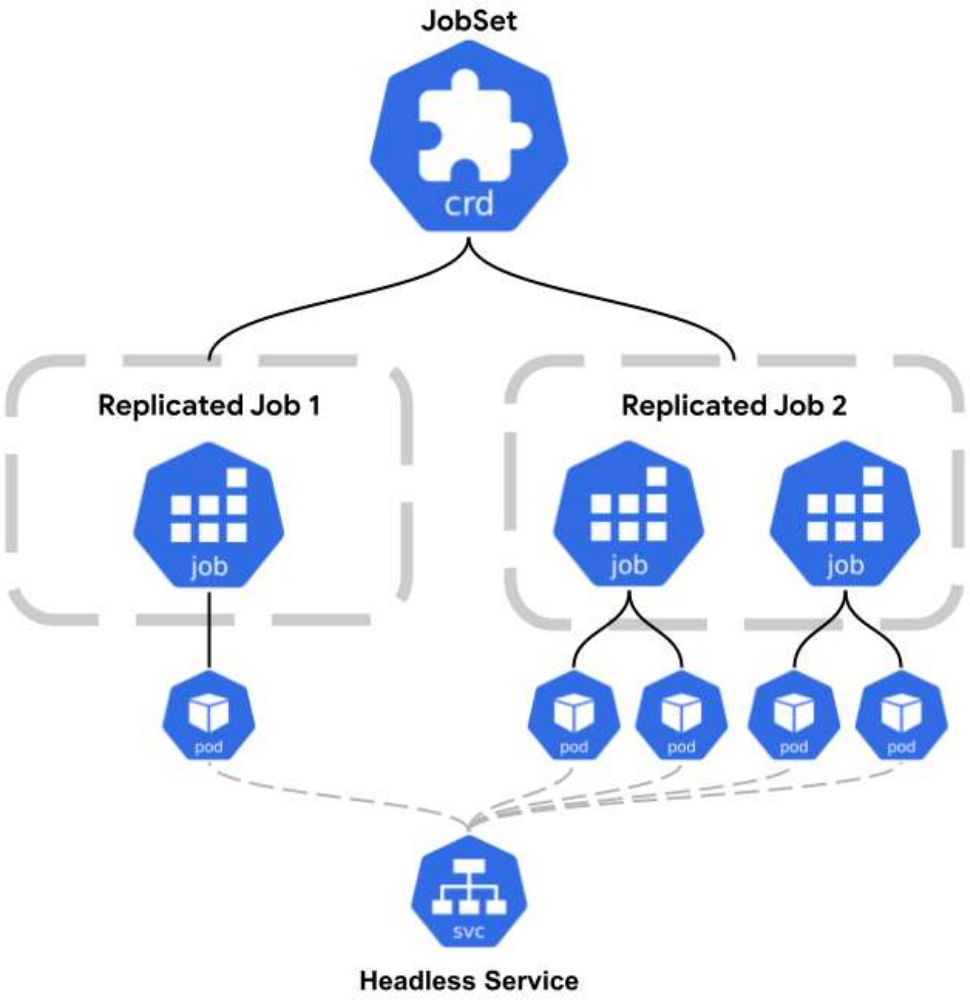
Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency	Dependency	2	0.5%
CPU	Host	2	0.5%
System Memory	Host	2	0.5%

云原生AI



云原生人工智能使得构建实际系统以部署、运行和扩展人工智能工作负载成为可能。 它解决了人工智能应用科学家、开发人员和部署者在云基础设施上开发、部署、运行、扩展和监控人工智能工作负载时所面临的挑战。通过利用云基础设施的计算（如 CPU 和 GPU）、网络和存储能力，并提供隔离和受控共享机制，云原生人工智能加速了人工智能应用的性能，并降低了成本。

JobSet



JobSet 是一个 Kubernetes 原生的 API，它专门设计用来管理和调度 AI/ML 训练和高性能计算（HPC）工作负载。它提供了一个统一的 API，用以在 Kubernetes 上部署如 PyTorch、Jax、Tensorflow 等 AI/ML 训练工作负载，以及 MPI 等 HPC 工作负载。

JobSet 还提供了对分布式 PyTorch 训练工作负载的支持，允许用户通过命令行工具列出属于 JobSet 的所有作业和 Pod，以及配置 DNS 以确保 Pod 之间的网络连通性

LLaMA-Factory

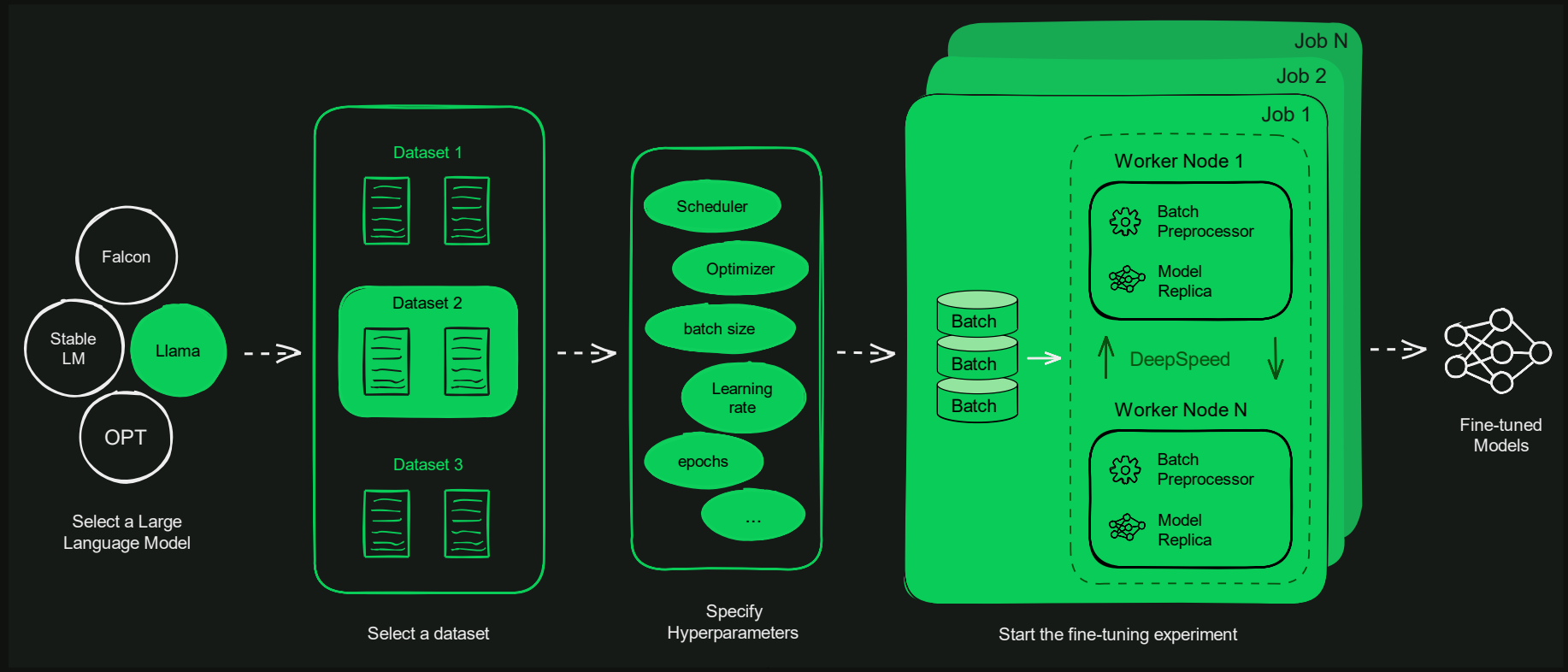
多种模型：LLaMA、LLaVA、Mistral、Mixtral-MoE、Qwen、Yi、Gemma、Baichuan、ChatGLM、Phi 等等。
集成方法：（增量）预训练、（多模态）指令监督微调、奖励模型训练、PPO 训练、DPO 训练、KTO 训练、ORPO 训练等等。
多种精度：16 比特全参数微调、冻结微调、LoRA 微调和基于 AQLM/AWQ/GPTQ/LLM.int8/HQQ/EETQ 的 2/3/4/5/6/8 比特 QLoRA 微调。
先进算法：GaLore、BAdam、Adam-mini、DoRA、LongLoRA、LLaMA Pro、Mixture-of-Depths、LoRA+、LoftQ、PiSSA 和 Agent 微调。
实用技巧：FlashAttention-2、Unsloth、RoPE scaling、NEFTune 和 rsLoRA。
实验监控：LlamaBoard、TensorBoard、Wandb、MLflow 等等。
极速推理：基于 vLLM 的 OpenAI 风格 API、浏览器界面和命令行接口。

项目地址：https://github.com/hiyouga/LLaMA-Factory

JobSet + LLaMA-Factory

DTX

DaoCloud DataTunerX（DTX）是一站式自动化平台，专注于大型语言模型微调。涵盖了数据集、超参组、模型仓库、模型微调、模型评估和模型推理的全生命周期，实现了高度自动化。通过高效利用底层分布式算力资源，DTX 能够进行矩阵式的模型微调实验，从而推动大型模型的敏捷和自动化迭代。



Thanks.