AI 大模型开发工程师
之私有大模型开发常见问题

讲师：李希沅

01、项目升级

基于CVP架构的知识库

私有化的部署

开源、闭源

模型的大小

模型的参数

GPU和CPU

GPU和模型大小的关系

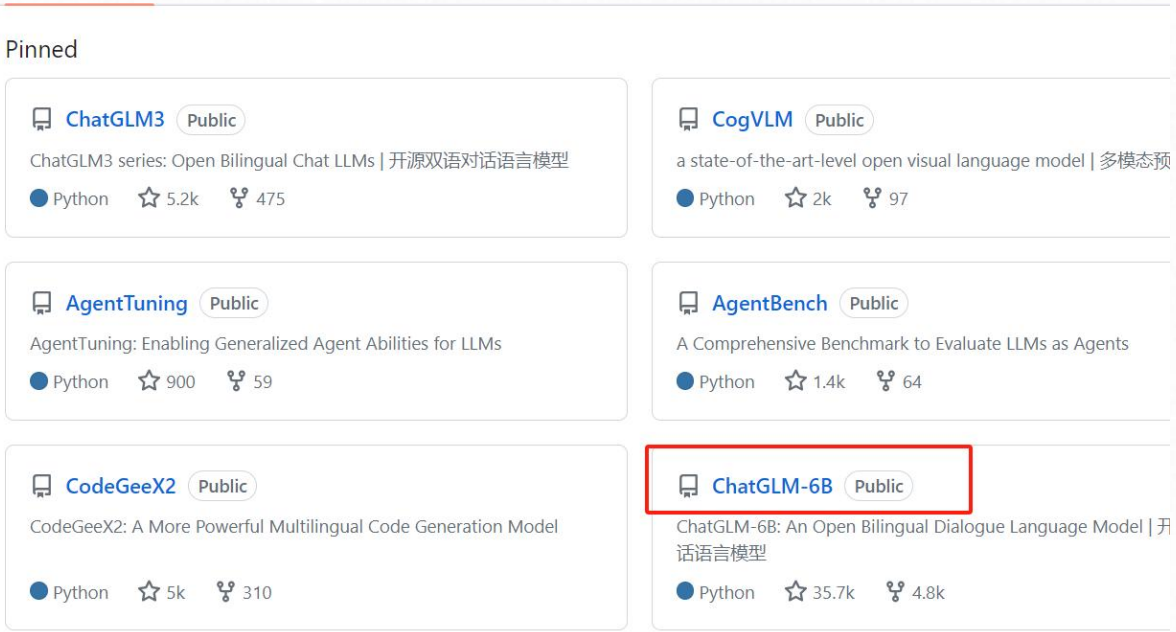
预训练和微调

02、开源和闭源

企业	大模型	开源	闭源
OpenAI	GPT-1、GPT-2、GPT-3	√	
	GPT-3.5、GPT-4		√
Meta	LLaMA	√（仅用于研究）	
	LLaMA2	√	
谷歌	PaLM 2		√
微软	Turning-NLG	√	
Anthropic	Claude		√
Cohere	Cohere		√
Stability AI	StableLM	√	
LMSYS	Vicuna	√	
Mosaic ML	MPT-30B	√	
阿联酋技术创新研究所	Falcon	√	
智谱	GLM-130B、ChatGLM-6B、ChatGLM2-6B	√	
	ChatGLM2-12B、ChatGLM2-32B、ChatGLM2-66B、ChatGLM2-130B		√
百度	文心		√
阿里	Qwen-7B、Qwen-7B-Chat	√	
华为	盘古		√
商汤	日日新		√
科大讯飞	星火		√
百川智能	Baichuan-7B、Baichuan 13B	√	
	Baichuan-53B		√

任何一家企业如果自己从零开发大模型，对算力、数据的要求极高，研发投入很大。根据Meta发布的数据，参数量最大的LLaMA-65B模型，使用2048块A100-80GB的GPU，训练数据量1.4万亿tokens，耗时为21天；如果采取租用云计算方式来训练算法，按照Microsoft Azure以1.36美元/小时提供A100租用价计算，训练成本约140万美元。

03、大模型的参数是什么意思?

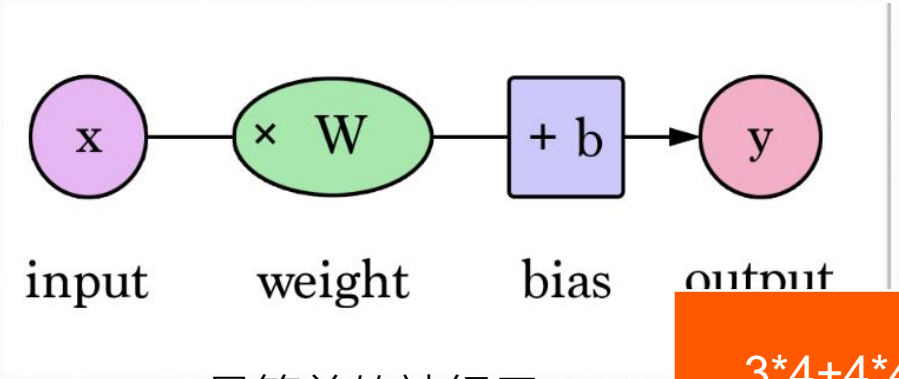


XXB 是一种用来表示模型的大小或者数据集的规模的缩写，其中 X 是一个数字，B 是 billion（十亿）的首字母。例如，6B 就是 60 亿，34B 就是 340 亿。

人工智能的大模型中，XXB 是用来看模型有多厉害的一个标准。一般来说，参数越多，模型就越厉害，能够做更多的任务，比如说话、写字、画画等。

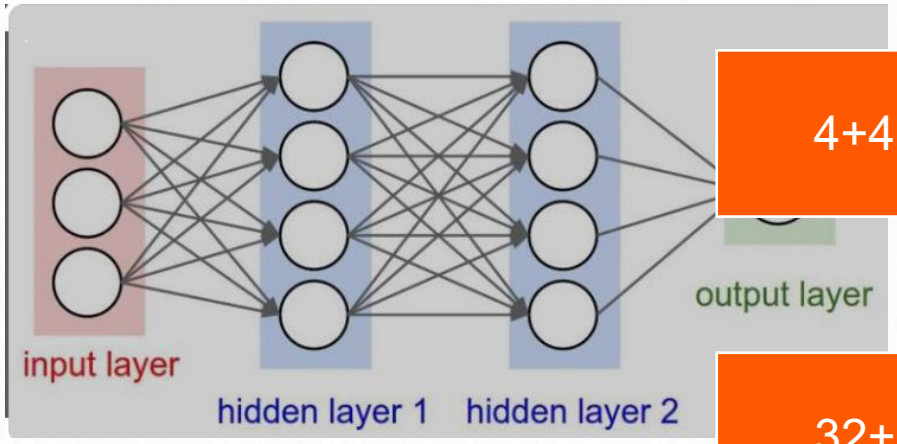
模型	发布时间	参数量	预训练数据量
GPT	2018 年 6 月	1.17 亿	约 5GB
GPT-2	2019 年 2 月	15 亿	40GB
GPT-3	2020 年 5 月	1,750 亿	45TB

04、大模型的参数有什么用?



最简单的神经元

3*4+4*4+4*1=32w



4+4+1=9神经元

32+9=41个参数

简单的全连接神经网络
3个输入
1个输出
隐藏层是用来计算的

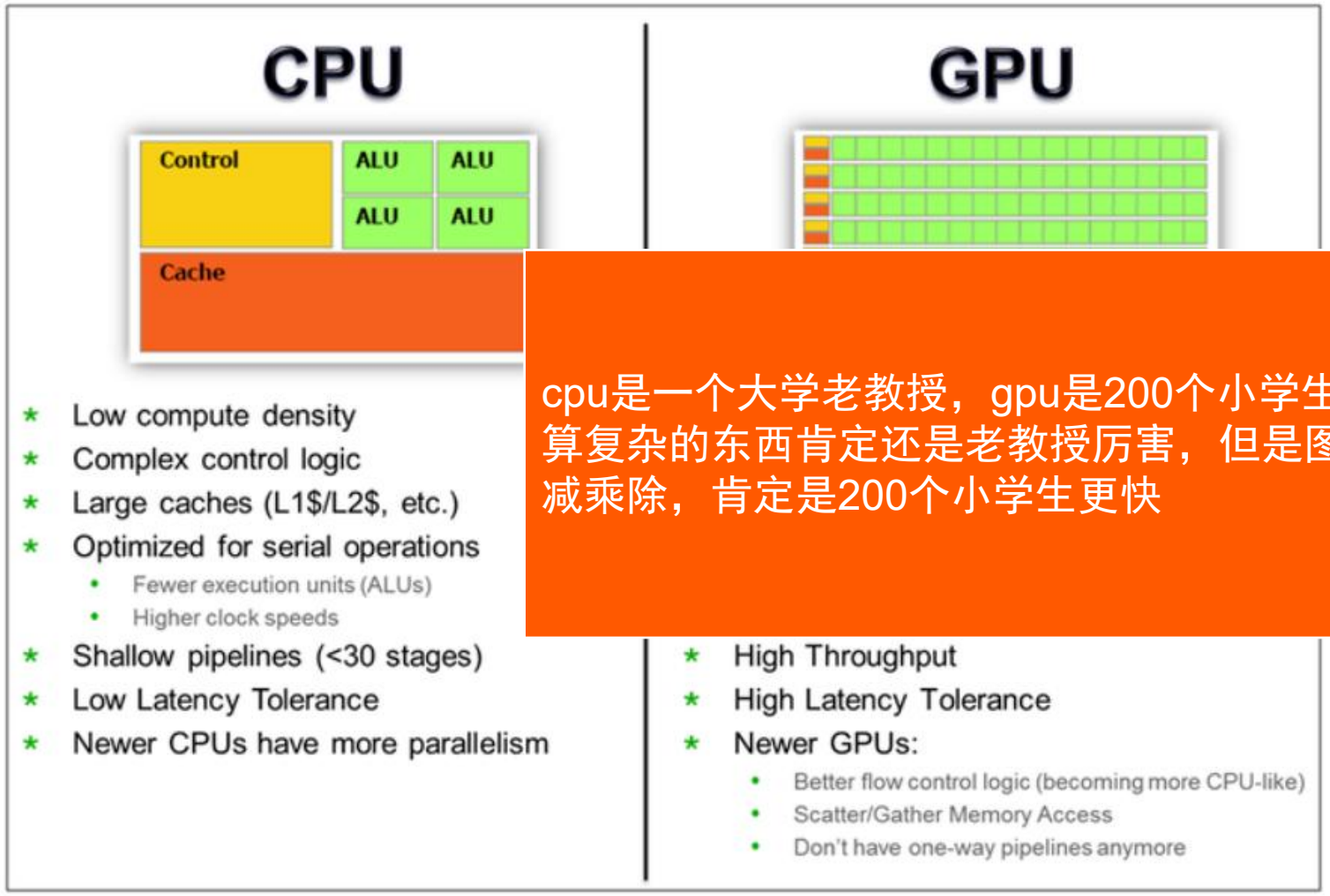
机器学习模型需要从大量的数据中学习，参数越多，它们在存储和处理信息方面的能力就越强大。大模型具有更多的参数，这意味着它们能够记住更多的信息和模式，并用于生成更准确、自然的输出。

举例来说，假设我们要训练一个模型来翻译不同语言之间的句子。小模型相当于一个只懂得基本单词和简单语法规则的翻译者，而大模型则像是一个非常精通多种语言、拥有大量词汇和语法知识的翻译专家。

W：影响的是X与Y的线性关系，多输入就会有多个W，W跟连接数有关。

B：绑定在神经元上跟连接数没关系，跟输出有关系。

05、GPU VS CPU



cpu是一个大学老教授，gpu是200个小学生。
算复杂的东西肯定还是老教授厉害，但是图形计算就是一堆加减乘除，肯定是200个小学生更快

CPU是对计算机的所有硬件资源进行控制调配、执行通用运算的核心硬件单元。

rocessing Unit）是一
理和并行计算的硬件
是加速计算和图形渲
PU（Central Processing Unit），GPU在并行计算和图形渲染方面都具有更高的性能和效率。

06、GPU VS 显卡

GPU不是显卡。

GPU是显卡上的一块芯片，也就是图像处理芯片，属于显卡的重要组成部分。

GPU使显卡减少了对CPU的依赖，并进行部分原本CPU的工作，尤其是在3D图形处理时GPU所采用的核心技术有硬件T&L（几何转换和光照处理）、立方环境材质贴图和顶点混合、纹理压缩和凹凸映射贴图、双重纹理四像素256位渲染引擎等，而硬件T&L技术可以说是GPU的标志

07、全球有哪些知名厂商

NVIDIA英伟达

上榜理由：始于1993年，1999年发明可编程

AMD

上榜理由：AMD始于1969年美国，全球知名

Intel英特尔

上榜理由：英特尔成立于1968年，是半导体行

高通Adreno

上榜理由：高通Adreno是高通推出的移动处理

Apple

上榜理由：创立于1976年美国，全球知名的高

景嘉微

上榜理由：景嘉微成立于2006年，2016年在深

N卡

A卡

08、英伟达GPU芯片销量最大的几个型号

1、Tesla系列

Tesla系列芯片是英伟达针对高性能计算和并行计算而设计的GPU芯片，其特点是高度可编程性和高性能。Tesla系列芯片的应用领域包括科学计算、石油勘探、气象预报、深度学习等领域。例如，Tesla V100是一款拥有640个张量核心的GPU芯片，能够实现高性能的深度学习计算。

2、Quadro系列

Quadro系列芯片是英伟达为计算机图形学和可视化而设计的GPU芯片，其特点是高度的图形性能和精度。Quadro系列芯片的应用领域包括建筑设计、影视制作、游戏开发等领域。例如，Quadro RTX 6000是一款拥有4864个CUDA核心的GPU芯片，能够实现高精度、高逼真的图形渲染。

3、GeForce系列

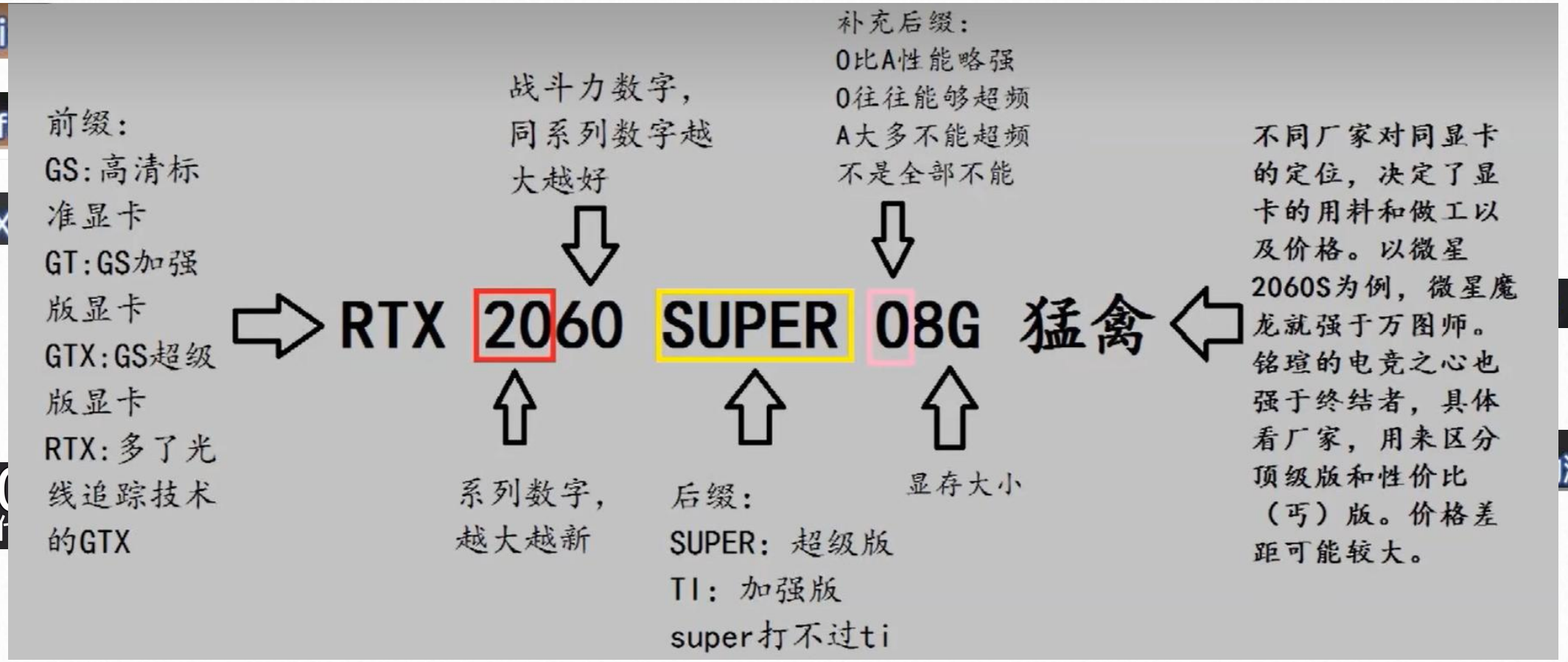
GeForce系列芯片是英伟达面向游戏玩家和计算机爱好者而设计的GPU芯片，其特点是出色的图形性能和较低的价格。GeForce系列芯片的应用领域包括游戏开发、虚拟现实、数字内容制作等领域。例如，GeForce RTX 2080 Ti是一款拥有4352个CUDA核心的GPU芯片，能够实现高速的游戏渲染和虚拟现实应用。

4、Titan系列

Titan系列芯片是英伟达面向专业用户和高端游戏玩家而设计的GPU芯片，其特点是超高的图形性能和精度。Titan系列芯片的应用领域包括游戏开发、计算机辅助设计、数字内容制作等领域。例如，Titan RTX是一款拥有4608个CUDA核心的GPU芯片，能够实现高精度、高逼真的图形渲染。

09、英伟达GPU型号命名规则

1
2
3
4
5



10、CUDA是干什么的?

CUDA（Compute Unified Device Architecture）是由NVIDIA开发的一种并行计算平台和编程模型。该平台利用GPU（图形处理器）的强大计算能力，使其更适用于高性能计算和数据并行计算任务。

CUDA的主要特点包括：

1. 并行计算：CUDA允许开发者利用GPU的并行处理能力，将计算任务划分为许多小的、可以独立执行的部分，并在多个处理器核心上同时执行。这种并行处理方式使得计算任务能够更快地完成。

2. 内存层次结构：CUDA对GPU的内存层次结构进行了优化，以提高并行计算的性能。具体来说，它提供了不同类型和大小的内存空间，如全局内存、共享内存和常量内存，以支持不同大小和性质的计算任务。

3. 编译器和编程语言：CUDA提供了自己的编译器和编程语言，以方便开发者编写高效的GPU代码。CUDA编程语言基于C/C++，但增加了一些用于GPU编程的特殊语法和函数。

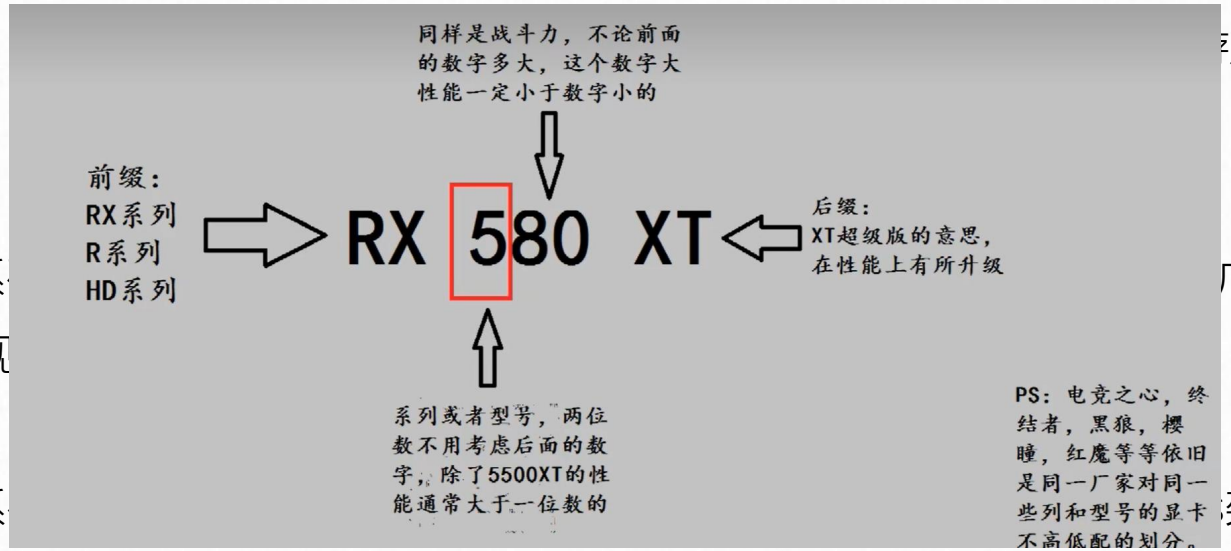
4. 通用计算：CUDA不仅适用于图形渲染，还可用于各种通用计算任务，如科学模拟、金融建模、深度学习等。这使得CUDA成为了一种强大的并行计算工具。

5. 可扩展性：CUDA不仅适用于NVIDIA的GPU，还支持其他厂商的GPU，如AMD和Intel。这使得CUDA成为了一种广泛使用的并行计算标准。

总的来说，CUDA是一种强大的并行计算工具，使得开发者能够更加方便地利用GPU的计算能力。它在许多领域都得到了广泛的应用，包括科学计算、人工智能、深度学习、图像处理等。

11、AMD显卡系列

1. Radeon Vega系列



戏玩家提供更加流畅的游戏体验，支持实时运行多项任务。

2. Radeon RX 500系

广大游戏爱好者的需求，同时还支持虚拟现实（VR）和高清视

3. Radeon RX 400系

架构，能够提供更高效的性能和更低的功耗。

A卡的命名规则不太规律，稍稍有些难以捉摸，所以给大家一个参考的不等差数列。

RX5700XT>RX5700>RX5600XT>RX590>
RX580>RX5500XT>RX480>RX570>RX470>
RX560XT=RX470D>RX560>RX560D

AMD系列显卡在游戏

无可挑剔的性能享受到更多的特色体验。

12、英伟达和AMD对比

英伟达

1. 强大的图形处理能力：英伟达gpu的核心是图形渲染和计算，因此它们比常规的cpu更适合用于处理高负荷的图形应用程序。
2. 高品质的视觉效果：英伟达gpu支持现代游戏和视频编辑软件中的高清晰度、高质量纹理、光线追踪和阴影等先进特性，为用户提供生动逼真的视觉体验。
3. 强大的机器学习能力：英伟达gpu很擅长处理大数据集。由于其高度可并行的架构，它们可以在几秒钟内处理数百万个数据点，并通过优秀的算法获得洞察。
4. 充足的技术支持：英伟达有强大的技术团队，他们通过各种方式（例如驱动更新、论坛帖子、在线聊天和客户支持电话）为用户提供良好的支持。

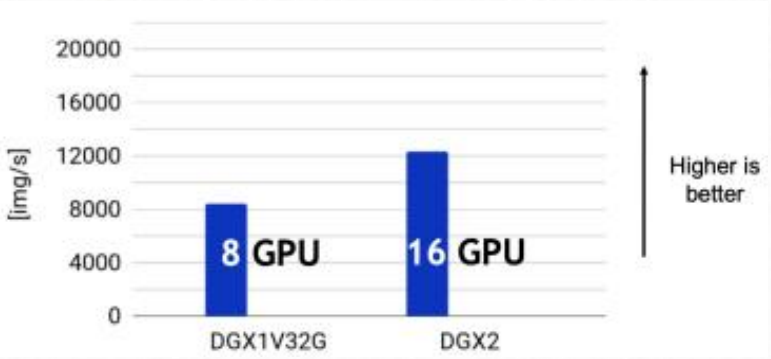
总之，英伟达独显具有强大的图形处理、高质量的视觉效果和出色的机器学习能力，使得它们成为游戏玩家、视频编辑者和数据科学家的首选。

AMD

1. 性能优异：AMD独显在性能方面表现出色，可提供顶级游戏图像质量和流畅运行体验。
2. 价格亲民：与其竞争对手相比，AMD独显的价格更为实惠，既能提供高性能，又不会使您的钱包肆虐。
3. 兼容性强：AMD独显可以与许多不同类型的计算机硬件兼容，从笔记本电脑到台式电脑，再到高端工作站。
4. 支持技术领先：AMD独显支持最新技术，如基于云的游戏流媒体、worksation gpu等，并且常常首次推出技术升级。
5. 能耗低：AMD独显通常比其他同类产品消耗更少的能源，并且通常都提供了强大的节能选项。

13、CPU和GPU如何配置

CPU非常重要！尽管CPU并不直接参与深度学习模型计算，但CPU需要提供大于模型训练吞吐的数据处理能力。比如，一台8卡NVIDIA V100的DGX服务器，训练ResNet-50 ImageNet图像分类的吞吐就达到8000张图像/秒，而扩展到16卡V100的DGX2服务器却没达到2倍的吞吐，说明这台DGX2服务器的CPU已经成为性能瓶颈了。



通常为每块GPU分配固定数量的CPU逻辑核心。理想情况下，模型计算吞吐随GPU数量线性增长，单GPU的合理CPU逻辑核心数分配可以直接线性扩展到多GPU上。AutoDL平台的算力实例提供了多种CPU分配规格。每块GPU应配备至少4~8核心的CPU，以满足多线程的异步数据读取。分配更多的核心通常不会再有很大的收益

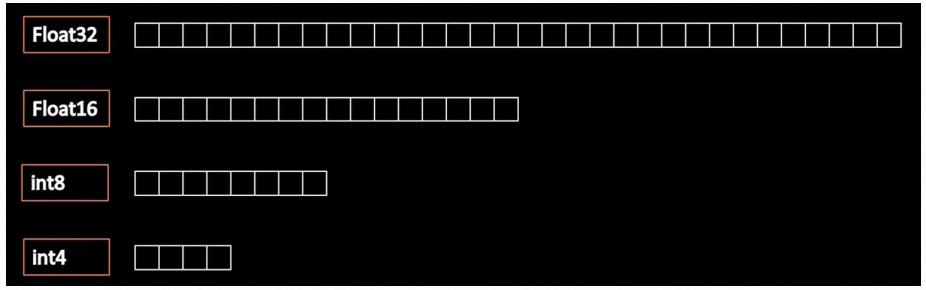
14、GPU与模型大小的关系

1

	Training Data	Params	Content Length
Llama 2	*A new mix of publicly available online data*	7B	4k
Llama 2	*A new mix of publicly available online data*	13B	4k
Llama 2	*A new mix of publicly available online data*	70B	4k

1B = 10亿

7B = 70亿



2

LLaMA-7b为例

1G = 1024 * 1024 *1024

Float32 7*10^9*4 / 1024^3 ~= 26.077 G

Float16 大约 13 G

int8量化 大约 6.5G

int4量化 大约 3.26G

模型规模	模型精度	所需显存
LLaMA-13b	Float32全精度	52G
LLaMA-13b	Float16半精度	26G
LLaMA-13b	Int8精度	13G
LLaMA-13b	Int4精度	6.5G
LLaMA-7b	Float32全精度	28G
LLaMA-7b	Float16半精度	14G
LLaMA-7b	Int8精度	7G
LLaMA-7b	Int4精度	3.5G

精度低模型准确损失越大

3

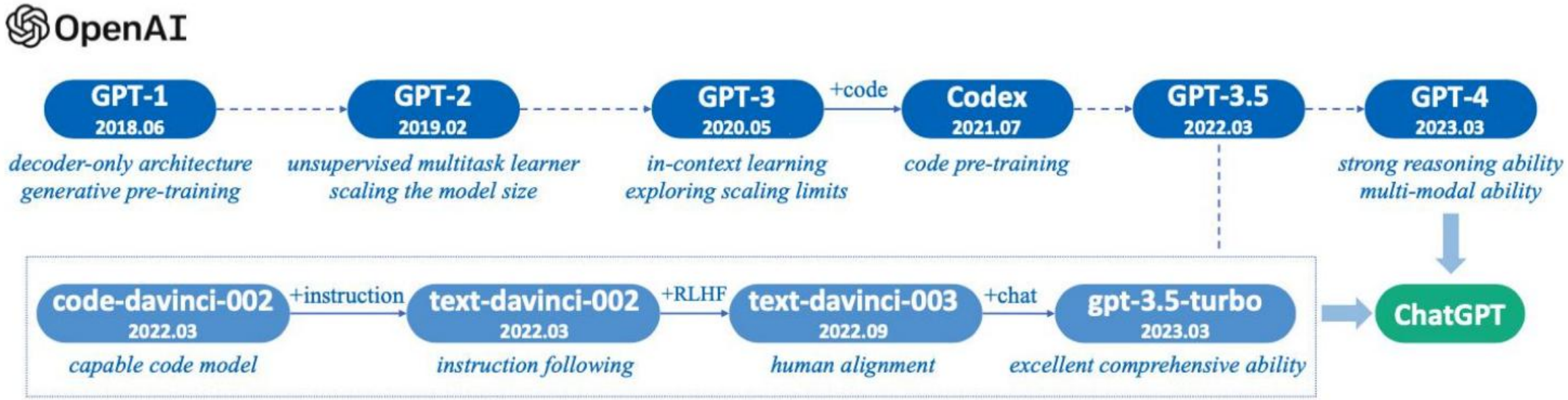
模型训练影响显存因素

模型参数
梯度
优化器参数
样本大小
BatchSize

训练需要的显存是推理的10几倍

15、Pretraining，Fine-Tuning，SFT，RLHF

预训练（Pre-training）是语言模型学习的初始阶段。 在预训练期间，模型会接触到大量未标记的文本数据，例如书籍、文章和网站。 在大量未标记文本数据上训练语言模型。 比如说在包含数百万本书、文章和网站的数据集上预训练像 GPT-3 这样的语



基于人类反馈的强化学习(Reinforcement Learning from Human Feedback)

人工先介入，通过对同一个Prompt生成答案的排序来训练一个Reward Model。再用Reward Model去反馈给SFT Model，通过评价生成结果的好坏，让模型更倾向于生成人们喜好的结果。RLHF是一种更复杂、更耗时的方法来微调LLM，但它比SFT更有效。（RLHF model）

谢谢观看