

CSDN 博客 下载 学习 社区 插件 GitCode InsCode 搜CSND 登录 会员中心 消息 历史 创作中心 +发布

后端 前端 移动开发 编程语言 Java Python 人工智能 AIGC

登录后您可以：
免费复制代码
下载海量资源
关注/点赞/评论/收藏
写文章/发动态/加入社区

云原生 云平台

“HAI域探秘”澎湃算力即开即用
高性能应用服务 HAI 新品先锋 立即登录 2.25 立即报名 广告

头条

OpenAI ChatGPT

ChatGPT 不愿多写一行代码、偷懒变笨，网友：承诺给它“小费”试试
LLM 爱好者偶然发现，不知是 Bug，还是 OpenAI 有此意图？

GPT-4 未通过图灵测试
Docker 1号员工亲述：我们曾犯...
“没有10-15年，一门新编程语言...
DATA+AI，生产效率至少+30%?
生成式 AI 工具迎来大升级，应...

大学生用 AI 帮 11 名失踪儿童回家
Django 5.0 发布 | 极客头条
美图发布AI视觉大模型4.0
MiracleVision(奇想智能)4.0版本、主...
ThreadX 操作系统正式开源！
百度曾出价8500万挖“AI 教父”被拒
Meta推出AI音频模型 | 极客头条
巴西都发明两门流行的编程语言了！
Lua以其简单性、小尺寸和可移植性、...

Apache Flink FLINK FORWARD ASIA 2023 点击报名 12月8-9日 北京

推荐 资讯 热榜 自荐 动态 有红包 排行榜 向你推荐 所有活动 >

语言大模型的分布式训练与高效微调指南
最近语言大模型 (LLM) 异常火爆，一个非常特别的开源社区正在探索在消费级硬件上微调、提供服务和进行推理的最佳方式。为满足上述需求，出现了许多出色的开源代码库，以HuggingFace生态系统为中心，这些代码库还包括FastChat、Axolotl和...
1 赞 1 踩 作者: OneFlow_Official

NCCL源码解析⑦：机器间Channel连接
上节课中完成了单机内部的channel搜索，仍然以ringGraph为例的话，相当于在单台机器内部搜索出来了一系列的环，接下来需要将机器之间的环连接起来。为了方便理解，假设两机十六卡的情况下第一台机器的一个ring为：graph->intra: GPU/0 GPU/7...
1 赞 1 踩 作者: OneFlow_Official

基于Amazon Bedrock的企业级生成式AI平台
Amazon Bedrock 是一项新的 AWS 服务，可让企业通过 API 轻松利用和自定义生成式 AI 模型。公司现在可以构建和扩展人工智能应用程序，而无需管理运行这些模型本身所需的复杂基础设施和维护。Amazon Bedrock 充当“平台”。一个关键的好处是...
2 赞 1 踩 作者: chszs

NCCL源码解析⑥：Channel搜索
上节课讲到已经计算出GPU和NIC节点到其他任意节点的最优路径了，本节课看下NCCL中channel的搜索过程。NCCL中channel的概念表示一个通信路径，为了更好地利用带宽和网卡，以及同一块数据可以通过多个channel并发通信，另外后续可以看到一个...
2 赞 1 踩 作者: OneFlow_Official

可复现的语言大模型推理性能指标
LLMPerf是一个开源项目，旨在帮助用户对语言模型进行基准测试，并使其性能具有可复现性。它能够帮助用户评估不同LLM的性能，并根据具体任务做出明智的决策。该项目选择了多个指标来衡量LLM的性能，包括吞吐量、时延、内存使用和成本等。本文...
1 赞 1 踩 作者: OneFlow_Official

国产720亿参数开源免费模型来了！对标Llama2 70B，一手实测在此
鱼羊 发自 凹非寺量子位 | 公众号 QbitAI最强开源大模型，再次易主！就在刚刚，阿里云通义千问又双収开源了，并且直接开大：甩出了720亿参数版本——在中国的开源大模型中，少见地直接对标最大号羊驼Llama2-70B。此番登场，这个代号为Qwen-72B的模型在10个权威基准评测中刷新开源模型最优成绩。在部分...
0 赞 1 踩 作者: QbitAI

漫画 | 20年了，走投无路的CPU终于躺平了！
很多年前，电脑的世界有个叫CPU的小伙子。他一出生就野心勃勃，梦想行走江湖，征服全世界。CPU对内存和硬盘的嘲讽毫不理会。他继续前行。有一天，他遇到了一个神秘人，送了他一本武林秘籍。CPU非常高兴，千恩万谢之后，操练起来。他孜孜不倦地练习，果然，如同秘籍中所说那样，每隔18个月，CPU...
18 赞 1 踩 作者: 码农翻身

开源语言大模型演进史：向LLaMA 2看齐
本文是开源 LLM 发展史系列文章的第三部分。此前，第一部分《开源语言大模型演进史：早期革新》回顾了创建开源 LLM 的最初尝试。第二部分《开源语言大模型演进史：高质量基础模型竞赛》研究了目前可用的最受欢迎的开源基础模型（即已进行预训...
1 赞 1 踩 作者: OneFlow_Official

华东理工李洪林课题组开发 Macformer，加速大环类药物发现
华东理工大学的李洪林课题组基于 Transformer 开发了 Macformer。Macformer 成功将无环药物菲卓替尼大环化，得到了药效更强的新化合物，为药物开发提供了新方法。

直播 更多 >

GitHub Universe 2023 Watch Party in Shanghai 12/10 13:00

揭秘GPTs，领略个性化AI面试追问技术 12/14 19:00

Prompt Engineering Conf (上海) 12/09 13:30

一起学习生成式人工智能 (一) | 生成式人工智能 12/06 19:30

1 赞 0 踩 作者: HyperAI



NCCL源码解析④：建图过程

上次分析了NCCL对机器PCI系统进行拓扑分析的过程，产出的结果为xml格式，接下来，NCCL会根据这个xml进图的建立过程以便之后进行路径搜索。ncclTopoGetSystem的最后会执行ncclTopoGetSystemFromXml将xml格式转成图格式。ncclResult t...

1 赞 0 踩 作者: OneFlow_Official



Transformer作者：指令型智能体的构建之法

2017年，Google发布的《Attention Is All You Need》论文提出了Transformer架构，这成为过去十年神经网络领域最具影响力的技术创新之一，并被广泛应用于NLP、计算机视觉、蛋白折叠等诸多领域。更重要的是，它成为后来包括ChatGPT在内的诸多大...

1 赞 0 踩 作者: OneFlow_Official



从意义中恢复，而不是从数据包中恢复

IM其实是一种反自然的交流手段，是一种“计算机式”的手段，存储转发式的手段，发出信息后，对方没回复，发送者不得不再发一遍“在吗”，后加入的“已读”功能将问题进一步复杂化而不是解决了问题，“已读”可能意味着“已读已忘”，日常交流不会这...

1 赞 0 踩 作者: dog250



微调语言大模型选LoRA还是全参数？基于LLaMA 2深度分析

本文对比了全参数微调和LoRA，并分析了这两种技术各自的优势和劣势。作者使用了三个真实用例来训练LLaMA 2模型，这提供了比较特定任务的性能、硬件要求和训练成本的基准。本文证明了使用LoRA需要在serving效率和模型质量之间做出权衡，而这...

1 赞 0 踩 作者: OneFlow_Official



golang channel执行原理与代码分析

从源码的角度分析channel的数据结构、发送数据、接收数据和关闭这些基本操作，业务中对性能要求比较高建议不要使用chan。

12 赞 0 踩 作者: 一名路过的小码农



时延抖动和通信的本质

网络中或大或小的buffer，网络边缘的无线wifi，再加上使用了tcp，都是抖动的根源，端到端时延不可能稳定，解决问题的方法很简单，在应用层加buffer，这才是问题的实质，buffer带来的问题通过再加一个buffer就能解决，重读上面的3个段落，可...

3 赞 0 踩 作者: dog250



龙芯重磅发布新一代处理器，全力打造IT产业新生态

该芯片采用异构大小核结构，集成DDR3内存、GMAC、OTG等多种功能模块，具有打印数据接收、解析和处理，打印引擎控制、扫描时序控制、数据扫描、图像处理、马达控制等功能，单芯片即可满足打印、扫描、复印等多种典型应用需求。胡伟武在介绍...

25 赞 0 踩 作者: CSDN资讯



开源语言大模型演进史：高质量基础模型竞赛

本文是开源LLM发展史系列文章的第二部分。第一部分《开源语言大模型演进史：早期革新》回顾了创建开源LLM的最初尝试。本文将研究目前可用的最受欢迎的开源基础模型（即已进行预训练但尚未微调或对齐的语言模型）。（本文作者为Rebuy公...

1 赞 0 踩 作者: OneFlow_Official



为什么多数情况下GPT-3.5比LLaMA 2更便宜？

本文旨在为用户选择合适的开源或闭源语言模型提供指导，以便在不同任务需求下获得更高的性价比。通过测试比较LLaMA-2和GPT-3.5的成本和时延，本文作者分别计算了二者的1000词元成本，证明在大多数情况下，选择GPT-3.5的成本更低、速度...

2 赞 0 踩 作者: OneFlow_Official



开源语言大模型的正确姿势

如今，很多公司都被迫加快步伐，参与到开源语言大模型（LLM）的竞争之中。发布某种形式的开源语言大模型已成为机器学习公司实力的象征。最近，Mistral AI完成资金筹集，发布了一款拥有70亿参数的强大语言模型。尽管更多人参与到开源机器学习...

1 赞 0 踩 作者: OneFlow_Official

活动日历

8 北京 Flink Forward 峰会 | Flink Forwar
12月 d Asia 2023

9 北京 NPCon (新程序员大会) 云原生实
12月 践峰会

10 线下 Google DevFest 2023 上海站
12月

14 线上 揭秘GPTs、领略个性化AI面试追
12月 问技术

16 无锡 2023开放原子开发者大会·无锡
12月

19 线上 2023 英特尔On技术创新大会中国
12月 站

20 北京 大会直击 | 2023数据资产管理大会
12月 议程公布

20 北京 2023百度云智大会-智算大会
12月

竞赛平台

亚马逊云科技
『云上探索实验室』
Amazon Codewhisperer 挑战周赛
解锁Codewhisperer，马上实战等你来
① 2023年10月7日-10月13日
[立即报名](#)
【云上探索实验室】Amazon CodeWhisperer
挑战周赛
主办方：亚马逊云科技 协办方：CSDN

会员精选



Java8编程实战



面试之排序算法



实用数据分析：数据分析
师从小白到精通



Java之路



韦东山嵌入式Linux第一
期视频



编程可以这样学

推荐专题



浪潮信息云峦服务器操作
系统征文挑战赛



免费领取阿里云云服务器
ECS，部署Java W...



以网强算，中国移动提出
“算力网络”全新技术方案

