人工智能大模型应用案例实战

课程简介

大模型也称为大语言模型（LLM）是指使用大量文本数据训练的深度学习模型，可以生成自然语言文本或理解语言文本的含义。这些模型通常基于深度学习架构，如 Transformer 等，这有助于它们在各种 NLP 任务上取得令人惊叹的表现。目前的大语言模型（如 GPT 和 BERT）采用与小模型类似的 Transformer 架构和预训练目标（如 Language Modeling），与小模型的主要区别在于增加模型大小、多模态、训练数据和计算资源。

ChatGPT 是一款由 OpenAI 开发的大模型，它能够模拟人类的语言行为，与用户进行自然的交互。它的名称来源于它所使用的技术 GPT-4 架构，即生成式语言模型的第 4 代。

ChatGPT 的核心技术是 GPT-4 架构。它通过使用大量的训练数据来模拟人类的语音行为，并通过语法和语义分析，生成人类可以理解的文本。它可以根据上下文的语境，提供准确和恰当的回答，并模拟多种情绪和语气。这样，就可以让用户在与机器交互时,感受到更加真实和自然的对话体验。

ChatGPT 的应用场景也很广泛。它可以用于处理多种类型的对话，包过对话机器人、问答机器人和客服机器人等。它还可以用于各种自然语言处理任务，比如文本摘要、情感分析和信息提取等。例如，在一个问答系统中，ChatGPT 可以提供准确的答案，解决用户的疑惑；在一个客服机器人中,他可以帮主用户解决问题，提供更好的服务体验。

此课程是尹立庆老师多年人工智能工作经验的分享，重点介绍大模型以及揭开大模型的神秘面纱，大模型未来发展趋势和投资机会。

培训目标

1、 大模型软硬件架构、分层；

2、 介绍主流编程框架（和硬件结合）；

3、 各大主流大模型适用场景，优劣势；

4、 大模型调参；

5、 openai 接口介绍使用；

6、 针对企业应用和软件开发或工业场景的实战；

7、 大模型调小模型示例，全程演示；

8、 大模型幻觉介绍。

培训对象

1、 本课程适合于对大模型、ChatGPT 的原理感兴趣的人员；

2、 本课程适合于架构师、技术经理、高级工程师；

3、 适合于企业科技研发人员和人工智能科学家；

培训方式

以课堂讲解、演示、案例分析为主，内容偏实用，结合讲解与演示方式，循序渐进，辅以互动研讨、现场答疑、学以致用。

课程安排

课程时间：3 天

课程内容：

时间	内容	备注
第 1 天	第1个主题：**大模型软硬件架构、分层**（深入讲解大模型软硬件架构、分层）（90 分钟） 1、 大模型软硬件架构、分层 2、 大模型的工作原理 3、 大模型的软件架构 a） Transformer 架构 b） 深度学习架构 4、 大模型的硬件架构 5、 NVIDIA GPU 大语言模型架构 6、 NVIDIA 的 A100 或 H100 GPU 7、 NVIDIA 的 Megatron-LM 大模型框架	

8、 NVIDIA 大模型框架 TensorRT-LLM

9、 Google 大模型架构

10、 Google 的 TPU 大规模硬件架构

11、 主流大模型训练架构 GPU+PyTorch+Megatron-LM+DeepSpeed

12、 大模型的训练方法

a) 数据准备、模型训练、调优

第2个主题：**介绍主流编程框架(和硬件结合)**（深入讲解介绍主流编程框架(和硬件结合)）（90 分钟）

1、 介绍主流编程框架(和硬件结合)

2、 主流大模型训练架构 GPU+PyTorch+Megatron-LM+DeepSpeed

3、 大模型技术原理

4、 大模型分布式训练框架

5、 常用的分布式训练框架

a) Megatron-LM

b) DeepSpeed

6、 Megatron-LM

7、 DeepSpeed

a) 3D 并行化实现万亿参数模型训练

b) DeepSpeed 三种并行方法

i. 数据并行训练

ii. 模型并行训练

iii. 流水线并行训练

c) ZeRO 零冗余优化器

8、 如何选择一款分布式训练框架

9、 常见的分布式训练框架

a) TensorFlow

b) PyTorch

c) MindSpore

d) Oneflow

e) PaddlePaddle

f) Flax

g) Megatron-LM（张量并行）

h) DeepSpeed（Zero-DP）
i) Colossal-AI（高维模型并行）
j) Alpa（自动并行）

10、 训练超大规模语言模型主要技术路线
a) GPU + PyTorch + Megatron-LM + DeepSpeed
b) TPU + XLA + TensorFlow/JAX

11、 参数高效微调（PEFT）技术
12、 影响大模型性能的主要因素
13、 衡量大模型水平
14、 深度学习框架 Pytorch
15、 大模型编程选择 Pytorch 的理由
16、 Pytorch 的大模型应用案例
17、 深度学习算法设计通用流程
18、 PyTorch 与 Tensorflow 对比

第3个主题：**大模型分布式并行计算技术**（大模型分布式并行计算技术）（60 分钟）

1、 大模型分布式并行计算技术
2、 数据并行 DP（Data Parallel）
a) 分布式数据并行 DDP (Distribution Data Parallel)
b) 张量并行
c) 流水并行
i. G-pipe
ii. PipeDream
iii. virtual pipeline
d) 梯度累加
e) 激活检查点
f) ZeRO
3、 MPI、GLOO 和 NCCL 等通信策略
4、 大模型生态相关技术

第4个主题：**英伟达 GPU+CUDA 架构**（英伟达 GPU+CUDA 架构）（60 分钟）

1、 英伟达 GPU+CUDA 架构
2、 英伟达集合通信库 NCCL

	3、 通讯操作原语 a) 广播 Broadcast b) 数据散播 Scatter c) 规约运算 Reduce d) AllReduce e) 数据收集 Gather f) AllGather g) ReduceScatter 4、 Nvlink 5、 显存优化技术 a) 重计算(Recomputation) b) Activation checkpointing(Gradient checkpointing) c) 卸载（Offload）技术 d) ZeRO-Offload e) ZeRO-Infinity f) 混合精度（BF16/FP16） 第5个主题：**大模型分布式训练环境搭建**（大模型分布式训练环境搭建）（60 分钟） 1、 AI 大模型分布式集群 2、 AI 大模型分布式集群通信 3、 大模型分布式训练环境搭建 4、 GPU 服务器配置 5、 CPU 硬件配置清单 6、 GPU 硬件配置清单 7、 AI 处理器（加速卡） 8、 安装依赖包 9、 配置环境	
时间	内容	备注
第 2 天	第6个主题：**Pytorch 大模型实践案例**（深入剖析深度学习框架 Pytorch 大模型实践案例）（90 分钟） 1、 Pytorch 大模型实践案例	

2、 Tensor 以及相关的函数
3、 Autograd 机制以及相关函数
4、 Torch.nn 库
5、 Tensor 操作函数
6、 AutoGrad 自动求导
7、 神经网络相关函数
8、 导数，方向导数，偏导数，梯度等
9、 PyTorch 搭建深度神经网络
10、 使用 PyTorch 搭建手写数字识别
11、 数据处理
12、 模型搭建
13、 模型训练
14、 数据预测与识别

第7个主题：**各大主流大模型适用场景，优劣势**（深入讲解各大主流大模型适用场景，优劣势）（90 分钟）

1、 各大主流大模型适用场景，优劣势
2、 各大主流大模型适用场景，优劣势
   a) ChatGPT-4 大模型发展现状
   b) Sora 大模型发展现状
   c) 谷歌 PaLM 2 AI 大模型发展现状
   d) Claude 大模型发展现状
   e) LLaMA 大模型发展现状
   f) MidJournery 大模型发展现状
3、 各大主流大模型适用场景，优劣势
   g) 百度文心一言
   h) 百度文心一格
   i) 阿里巴巴通义千问
   j) 华为盘古
   k) 科大讯飞星火
4、 AIGC 大模型
   a) ChatGPT
   b) GPT4
   c) 文心一言

d) Google bard
e) DALL-E

5、 本地模型

f) 清华大学 ChatGLM
g) Facebook LLaMa
h) Stable Diffusion
i) 斯坦福 Alpaca
j) OpenJourney

6、 垂直领域产品

k) 方问智能中医辅助系统

第8个主题：**国内大模型应用建议**（深度解读国内大模型应用建议）（60 分钟）

1、 百度文心一言应用建议
2、 阿里巴巴通义千问应用建议
3、 华为盘古应用建议
4、 科大讯飞星火应用建议

第9个主题：**大模型调参**（深入讲解大模型调参）（90 分钟）

1、 大模型调参
2、 NVIDIA GPU 加速和优化大语言模型的性能
3、 大模型微调
4、 大模型微调的概念和意义
5、 预训练模型的优势和应用场景
6、 大模型微调基本原理
7、 大模型微调方法
8、 数据加载、模型训练、调参等常见操作的优化和加速方法
9、 使用可视化工具进行模型训练过程的分析和调试
10、 大模型微调的基本流程和关键步骤
11、 常用的深度学习框架和工具
12、 TensorFlow、PyTorch 等常见深度学习框架
13、 Parameter-Efficient Fine-Tuning (PEFT)
14、 预训练阶段
15、 目标任务准备

	16、 构建微调任务 17、 PEFT 微调 18、 常用的 PEFT 方法 a) Adapter Tuning b) Prefix Tuning c) Prompt Tuning d) P-Tuning e) LoRA 19、 案例剖析：应用大模型微调技术解决实际问题 第10个主题： **大模型微调技术与实践**（大模型微调技术与实践）（30 分钟） 1、 大模型微调技术与实践 2、 常见的大模型微调技术 a) 知识蒸馏 b) 迁移学习 c) 领域适应 3、 案例大模型微调的实践 a) 文本分类 b) 图像识别 c) 自然语言处理 4、 探讨大模型微调过程中可能遇到的问题和解决方案 5、 选择合适的预训练模型并进行微调 6、 如何评估微调效果和改进方案 7、 实际应用案例分享与讨论	
时间	内容	备注
第 3 天	第11个主题： **openai 接口介绍使用**（深入讲解 openai 接口介绍使用）（90 分钟） 1、 openai 接口介绍使用 2、 获取 OpenAI API 密钥 3、 选择 OpenAI API 4、 安装 OpenAI SDK 5、 调用 OpenAI API	

6、 处理 OpenAI API 响应
7、 优化 API 调用
8、 管理 API 使用
9、 反馈和改进
10、 实战案例：Python 调用 OpenAI API 实战案例

第12个主题：**针对企业应用和软件开发或工业场景的实战**（深入讲解针对企业应用和软件开发或工业场景的实战）（90 分钟）

1、 针对企业应用和软件开发或工业场景的实战
2、 大模型的企业应用场景
    a) 望闻问切
    b) 视觉、听觉、触觉、语言、思考
3、 文本生成
    c) 生成式对话、编写剧本、撰写论文
4、 文本理解
    d) 情感分析、主题分类、关系提取
    e) 语义理解、问答系统
5、 图像理解与图像生成
6、 语音识别与语音合成
    f) 企业智能客服
7、 视频理解与视频生成
8、 大模型应用的工业场景的实战
    a) 智能制造和质量控制
    b) 供应链优化
    c) 客户服务和支持
    d) 智能能源管理
    e) 产品推荐和个性化营销
    f) 安全监控和风险管理
    g) 生产优化和预测维护
    h) 企业提高生产效率、降低成本
    i) 改善产品质量和客户体验
    j) 增强竞争力并实现可持续发展
9、 大模型的应用中应该重点关注什么

第13个主题：**大模型调小模型示例，全程演示**（深入讲解大模型调小模型示例，全程演示）（90 分钟）

1、 大模型调小模型示例，全程演示
2、 模型压缩（蒸馏、剪枝）
    a) 知识蒸馏
    b) 剪枝大模型
    c) 大模型蒸馏
3、 提示语压缩
4、 联合推理
    a) 模型串联
    b) 数据采样
5、 迁移学习
6、 权值共享
7、 集成学习
8、 将小模型作为插件
9、 提示语压缩

第14个主题：**大模型幻觉介绍**（深入讲解大模型幻觉介绍）（90 分钟）

1、 大模型幻觉介绍
2、 什么是大模型幻觉
3、 大模型幻觉分类
    a) 事实性幻觉
    b) 忠实性幻觉
4、 大模型产生幻觉的来源
    a) 数据源、训练过程和推理
5、 预训练阶段导致大模型幻觉
    a) 架构缺陷
    b) 曝露偏差
    c) 能力错位
    d) 信念错位
6、 检测事实性幻觉的方法
    a) 检索外部事实
    b) 不确定性估计

	7、 基于内部状态的方法 8、 基于行为的方法 9、 检测忠实性幻觉的方法 a) 基于事实的度量 b) 分类器度量 c) 问答度量 d) 不确定度估计 e) 提示度量 第15个主题：**深度解读 glm2_6b 大模型**（深度解读 glm2_6b 大模型）（90 分钟） 1、 深度解读 glm2_6b 大模型 2、 glm2_6b 大模型的原理 3、 GPT（Generative Pre-trained Transformer）架构 4、 glm2_6b 大模型数据集 5、 glm2_6b 大模型的部署 a) 准备环境 b) 安装依赖库 c) 下载模型权重 d) 加载模型 e) 部署 API 或服务 f) 调优和监控 6、 glm2_6b 大模型的训练 7、 glm2_6b 大模型的应用 a) 自然语言处理 b) 文本生成 c) 机器翻译 d) 问答系统	