Datasheet

构建大语言模型 RAG 智能体

课程概述

大语言模型 (LLM) 的技术发展进程突飞猛进，其中检索系统在这场技术飞跃中尤为突出。大语言模型 (LLM) 驱动智能体的检索能力，在通过使用工具、查看文件和规划信息交互的过程中表现尤为卓越。这些模型不仅仅是自动化工具，更是提升生产力的得力助手，能够通过与各种工具和文档交互，形成有效对话。本课程将探索这些系统的潜力，重点讲解如何在实践中部署系统和高效实施，以满足用户和深度学习模型的大量需求。通过深入探索大语言模型的复杂性，您将了解包括内部推理、对话管理和有效工具策略在内的高级编排技术。

通过学习，您将掌握大语言模型 (LLM) 系统的构建，通过结合内部和外部推理组件，实现可预测的交互。在实战练习中，学习部署可以灵活扩展的代理系统，以满足用户和客户不断变化的需求。本课程重点强调设计强大的对话管理和文档推理系统维护状态，并将信息转换为易于理解的结构化格式。同时，您还将学习利用嵌入模型进行高效的相似性查询，增强内容检索和生成对话护栏，以及探讨检索增强生成 (RAG) 模型的实施和模块化，无需任何微调即可通过浏览研究论文提供答案。

学习目标

参加本次培训，您将学到：

- 构建 LLM 系统，利用内部和外部推理组件实现与用户的可预测交互
- 设计对话管理和文档推理系统，用于维护状态并将信息强制转换为结构化格式
- 利用嵌入模型对内容检索进行高效的相似性查询并生成对话护栏
- 开发、模块化和评估检索增强生成模型，无需任何微调即可回答有关研究论文数据集的问题

课程概要	
课程时长	8 小时 (课后 6 个月内可以继续访问课件和实验，资源用量有限额)
课程模式	讲师实时授课，每位学员使用完全配置的云端实验环境进行练习
预备知识	> 熟悉深度学习基础原理，特别需要掌握 PyTorch 和迁移学习；如**《深度学习新手入门》**或**《深度学习基础——理论与实践入门》**课程涵盖的内容，或拥有类似经验 > 如**《构建基于大语言模型 (LLM) 的应用》**课程涵盖的内容，或拥有类似经验 > 具备中级 Python 编程经验，包括面向对象编程和库的使用。参加过 **Python Tutorial** 或拥有类似经验
课程涵盖的技术	大语言模型 (LLM) 推理接口，训练数据集设计，LangChain，Gradio，LangServe，带有运行时状态的对话管理，文档处理，语义相似性和语意护栏的嵌入，RAG 智能体向量存储
课程测评问题类型	搭建 RAG 功能聊天机器人，可基于研究论文库回答问题
培训证书	成功完成本课程和测试后，学员将获得 NVIDIA DLI 培训证书，证明在相关领域的技能，为职业发展提供证明
课程语言	中文
学习此课程的硬件要求	您需要一台能够上网的笔记本电脑或台式机，且能够运行最新版 Chrome 或 Firefox 浏览器；我们为您提供在云端实验环境的专用访问权限
课程价格	微信添加 DLI 小助手 (微信号 NVIDIALearn)，沟通培训需求

课程大纲	
课程介绍 (15 分钟)	> 讲师介绍 > 登录 DLI 学习平台
LLM 推理接口和微服务 (60 分钟)	探索课程环境、微服务和 LLM 推理选项： > 熟悉课程环境，并了解用于软件模块化和资源交付的微服务 > 讨论推理用例的 LLM 服务选项，包括本地部署和可扩展部署策略及价值主张 > 熟悉 GPT4 和 NGC 托管的 NVIDIA AI Foundation 模型端的远程接入点
使用 LangChain, Gradio 和 LangServe 设计大语言模型训练数据集 (120 分钟)	使用开源框架编排 LLM 终端： > 学习如何使用 LangChain，通过功能性 LangChain Express Language (LCEL) 语法将多个 LLM 模块链接起来 > 将内部和外部推理规范化，并将它们模块化为可运行单元 > 使用 LangServe 通过接口发送 LLM 链，与 Gradio 前端交互
休息 (60 分钟)	
维护对话状态和集成知识提取 (75 分钟)	开发运行逻辑系统，存储信息并引导对话： > 了解运行状态逻辑，以便在链运行时保持该状态 > 利用槽位填充进行知识提取，维护智能知识库 > 集成一个对话管理聊天机器人，强制用户输入凭据、从数据库接口检索信息，并维护对话状态
长篇文档处理策略 (45 分钟)	学习如何处理超出上下文限制的长篇文档： > 了解文档分块、精简和重构策略 > 使用相同的 LLM 链接技能构建一个总结研究论文的系统，通过导出 while-loop 运行文件来实现
语义相似性和护栏的嵌入 (60 分钟)	探索嵌入模型在向量语义推理中的应用： > 理解编码器与解码器的优势和嵌入的工作方式 > 使用向量表示来推理段落的含义和相似性 > 设计一个通过定制化输入通道来回答问题或礼貌拒绝的护栏系统
休息 (15 分钟)	
实施高效文档检索的向量存储 (60 分钟)	将向量存储集成到代理系统中，辅助文档检索和推理： > 将向量存储转为有助于自动化向量推理的结构 > 将向量存储纳入检索增强生成工作流中，根据对话历史和预处理的文档池进行推理
评估测试和总结 (60 分钟)	使用 LLM 评估链对您的 RAG 系统进行评测： > 回顾所学要点 > 在前端部署检索组件并运行评估，完成评估并获得证书 > 填写培训调查表
下一步	学习更多 DLI 相关课程： > **构建基于大语言模型 (LLM) 的应用** > **高效定制大语言模型 (LLM)** > **构建基于扩散模型的生成式 AI 应用** > **模型并行 —— 构建和部署大型神经网络** > **构建基于 Transformer 的自然语言处理应用** > **深度学习基础 —— 理论与实践入门**

为何选择 NVIDIA 深度学习培训中心 (DLI) 的实战培训?

- 学习 NVIDIA 与技术专家和行业领导者合作开发的课程，获取全球同步、技术领先和现实可用的专业开发技能和经验。
- 学习使用行业通用、标准的软件、工具和框架进行端到端的应用开发，能够在广泛的行业中构建基于深度学习、加速计算、图形与仿真和数据科学的应用。
- 系统化地学习理论，并使用云端完全配置的实验环境同步边练，高效提升实战开发能力。
- 获得 NVIDIA 全球开发者培训证书，加持专业培训认证，助力职业发展。

准备好开始学习了吗?

查询更多 DLI 课程，请访问 **nvidia.cn/dli**。

如有疑问，请通过微信联系 DLI 小助手 (微信号 NVIDIALearn)。

© 2024 NVIDIA Corporation。保留所有权利。NVIDIA、NVIDIA 徽标、Clara、CUDA、DGX、DGX SuperPOD、Index 和 Triton 均为 NVIDIA Corporation 在美国和其他国家 / 地区的商标和 / 或注册商标。其他公司和产品名称可能是其各自关联公司的商标。其他所有商标均为其各自所有者的财产。3396006。2024 年 8 月