



构建对话式 AI 应用程序

利用 NVIDIA Riva 框架，轻松快速地创建、部署和运行实时的对话式 AI 应用

随着世界不断向数字化发展，对话式 AI 越来越多地用于促进人机之间的通信。对话式 AI 为信息自动收发和语音应用程序提供技术支撑，被广泛应用于各行各业，以提升整体客户体验和客户服务效率。

如果从零开始，会话式 AI 应用的开发非常复杂和昂贵。在本课程中，您将学习如何使用 [NVIDIA Riva 框架](#) 构建对话式 AI 应用。Riva 提供了完整的 GPU 加速的软件堆栈，让开发者可以快速轻松地创建、部署和运行端到端、实时的对话式 AI 应用程序，这些应用程序能够理解各公司及其客户特有的术语。Riva 框架包括预先训练的对话式 AI 模型、工具和优化服务，可用于语音、视觉和自然语言理解 (NLU) 任务。通过 Riva，开发人员可以创建定制的基于语言的 AI 服务，比如虚拟智能助理、虚拟客户服务代理、实时转录、区分对话中不同的讲话者和聊天机器人等。

在此课程中，您将学习如何使用实时转录和自然语言处理 (NLP) 技术，快速构建和部署产品级的对话式 AI 应用程序。把 NVIDIA Riva 自动语音识别 (ASR) 和命名实体识别 (NER) 模型与基于 Web 的应用程序相集成，对输入的音频生成文本，并且能够高亮显示。然后，使用 [NVIDIA TAO 工具包 \(TAO\)](#) 定制 NER 模型，从而根据不同的针对性进行高亮显示。最后，您将探索采用 Helm 和 Kubernetes 集群的 Riva 服务的生产级部署性能，以及扩展注意事项。

课程时长	8 小时（课后 6 个月内可以继续访问和使用课件和实验资源）。
课程模式	讲师授课，及每位学员使用云端完全配置的 GPU 加速工作站实验练习。
课程价格	微信添加 DLI 小助手（微信号 NVIDIALearn），沟通培训需求。
学员评测方式	1) 编程技能评测，评估学员构建对话式 AI 应用程序的能力。 2) 多项选择题，评估学员工关于对话式 AI 概念的理解程度。



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。
预备知识	<ul style="list-style-type: none"> • 基础 Python 编程经验 • 基本了解 深度学习框架 (比如 TensorFlow、PyTorch 或 Keras) • 基本了解 神经网络
课程语言	英文
工具、库和框架	NVIDIA Riva 框架 , NVIDIA TAO 工具包 (TAO) , Kubernetes
学习此课程的硬件要求	您需要一台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完全配置的 GPU 加速工作站的专用访问权限。

学习目标

参加本次培训，您将学习到：

- 如何在 Riva 上部署和启用经过预训练的 ASR 和 NER 模型，以构建对话式 AI 应用程序。
- 如何使用 TAO 微调和部署特定领域的模型。
- 如何使用 Helm 部署产品级对话式 AI 应用程序，以在 Kubernetes 集群中扩展。

为何选择 NVIDIA 深度学习培训中心 (DLI) 的实战培训

- 随时随地访问云端完全配置的 GPU 加速工作站来动手实践。
- 获得实战经验指导，使用通用、行业标准的软件、工具和框架。
- 学习如何在广泛的行业中构建深度学习和加速计算应用程序，如自动驾驶汽车、数字内容创作、游戏开发、医疗医学及金融。
- 学习与行业领导者（例如洛杉矶儿童医院、梅奥医院和普华永道）合作设计的课程，获取现实应用的专业知识。
- 获得 NVIDIA 官方全球开发者培训证书，证明在相关领域的能力，助力职业发展。



课程大纲

议题	说明
课程介绍 (15 分钟)	<ul style="list-style-type: none"> > 讲师介绍 > 登录课件平台
对话式 AI 介绍 (120 分钟)	<p>探索对话式 AI 应用，深入了解 ASR 和 NLP 流水线的关键组件</p> <ul style="list-style-type: none"> > 完成一个 TAO 模型推理示例，语音识别 > 部署 Riva ASR 和 NER 模型 > 启动一个使用 ASR 和 NER 的联系应用程序
休息 (60 分钟)	
定制模型 (120 分钟)	<p>详细学习 Riva 架构，并讨论使用 TAO 部署微调模型时涉及到的工作流程</p> <ul style="list-style-type: none"> > 针对特定领域微调 NER > 在 Riva 中部署定制的 NER 模型 > 启动更新过模型的应用程序
休息 (15 分钟)	
推理和部署 (120 分钟)	<p>探索对话式 AI 应用程序在生产部署时的性能、优化和扩展方面的挑战</p> <ul style="list-style-type: none"> > 了解推理部署流程 > 分析非功能性需求及其影响 > 使用 Helm 和 Kubernetes 集群部署对话式 AI 应用程序
休息 (15 分钟)	



<p>总结 (15 分钟)</p>	<ul style="list-style-type: none">> 回顾所学要点并回答问题> 完成评估并获得证书> 填写培训调查表> 了解如何设置您自己的 AI 应用程序开发环境
-----------------------	--

相关培训课程

- 深度学习基础 —— 理论与实践入门
- 构建智能推荐系统
- 深度学习基础 —— 用多 GPU 训练神经网络
- 更多课程，请访问 www.nvidia.cn/dli

培训咨询和购买

- 微信扫码添加 DLI 小助手，微信号 **NVIDIALearn**。

