尚硅谷

智能体速成班 V2.0

课程大纲

直播课程

01 大模型基础能力构建

大模型(LLM)认识与环境准备

- 大模型的起源与发展历程、大模型与AGI关系、AI应用场景
- 国际知名、国产主流的大模型的功能特点、优势与适用场景
- 大模型的发展历程、关键推动因素与趋势
- 训练三阶段:预训练->微调->推理

大模型架构原理

- Transformer 编码器/解码器结构、MoE模型
- 自注意力机制与多头注意力
- LLaMA、Qwen、GPT 等主流体系
- 多模态架构:文本+图像+音频

大模型调度平台

- Ollama定义与安装
- 如何调用私有大模型
- 云端部署(AWS、阿里云)、本地部署等模型部署流程与方法

提示词工程

- 提示词工程基础：核心原则、基础结构、基础编写方法。
- 高级技巧：深入上下文控制、任务分解与链式思维提示法、Few-shot 示例与模式引导
- 项目：电商、自媒体爆款文案生成、卖点提炼
- 评估提示效果与迭代优化、解决 Prompt 失效与偏差问题
- 多轮对话与记忆管理方法、提示词在Agent与工具调用中的应用

02 企业低代码平台开发与项目实战

Coze（扣子）平台

- Coze界面主要功能介绍、主要目标用户
- 核心功能模块（插件、知识库、工作流、智能体）的创建
- 提示词设计、模型选择与配置、发布渠道与多端部署能力
- Python调用Coze平台工作流

项目1：商户运营管家

- 模块1：一键生成行业调研PPT
- 模块2：复刻爆款视频
- 模块3：产品营销海报生成
- 模块4：商品营销卖点提炼
- 模块5：商品评论分析

Dify AI 平台

- 不同低代码平台对比、Dify核心功能与组件（工作流、Agent、知识库）
- 提示词编排、工具与Agent、模型集成、工作流基本结构
- Dify案例：客户投诉分类助手-钉钉
- Dify案例：一键生成行业调研报告

- Dify案例：客服对话记录分析
- Dify案例：商品评论分析
- Python调用Dify平台工作流

容器化技术

- 容器化技术Docker核心概念
- Docker安装、镜像、启动容器、常用命令

企业级大模型部署

- 部署核心方案
- 腾讯云/阿里云服务器部署、Docker的安装、Dify的下载与配置
- AutoDL服务器配置、Ollama下载与大语言模型的加载
- Xinference平台下载、嵌入模型&重排序模型的部署
- 低代码平台Coze Studio、Coze Loop的本地部署

AI代码编程工具-Trae AI 工具

- Trae的安装和使用
- Trae与多种第三方模型的 API Key
- Trae接入MCP

AI代码编程工具-Qoder

- Qoder的使用技巧
- Qoder的调试技巧
- 快速开发自己的项目

03 大模型核心开发框架

LangChain框架原理与应用

- LangChain框架概述、LangChain的安装与调用
- Model I/O：Message、Prompt Template、Output Parsers、FunctionCaling
- Chains：Chains的设计理念、SequentialChain、RouterChain
- Memory：Memory模块的设计理念、如何自定义Memory模块、内置的Memory模块
- Agents：Agent抽象、自定义基于ReAct范式的Agents、使用LangChain定义的ReAct策略
- Retrieval：Source 与 data loaders、Text Splitters、Text embedding models、vector store
- 电商平台商家对话助手案例：集成店铺运营数据库检索、平台政策实时查询和客户服务管理功能；开发多工具调用能力，记忆机制管理多会话上下文；本地知识库的搭建与调用

LangGraph框架原理与应用

- LangGraph入门：从链式到图状的思维转变
- 图的核心要素：State, Node, Edge - State (状态)
- Graphs：图的构建与编译
- Memory：图中的持久化状态与记忆
- Agents：使用LangGraph构建更鲁棒、更可控的智能体
- 高级应用与技巧 - 流式输出 (Streaming)

MCP从原理到实战

- Function Calling对比MCP的核心差异、功能定位、交互逻辑及适用场景
- MCP 应用场景、核心通信机制与传输逻辑
- MCP 关键组成要素，及各要素功能在体系中的作用
- MCP 从初始化到日志记录的完整工作流程，各环节核心动作
- 热门 MCP Server 推荐：主流工具，及各工具特点及适用场景
- 从底层逻辑剖析 MCP 在服务解耦、路由、容错等方面的核心原理
- 案例：多种具体环境MCP Server 部署与测试
- 案例：自定义MCP的开发步骤、功能验证要点与目标

跨Agent通信：A2A协议

- A2A协议定义与作用、与MCP协议关系、核心组件架构
- 工作流程机制、消息格式与数据结构、请求与响应流程
- 认证与授权机制、错误码与异常处理
- 性能优化与并发控制、典型业务场景示例

04 企业级RAG/Agent项目实战

项目2：掌柜智库

- 架构设计：基于 LangGraph 构建适配电商设备手册查询、商品售后咨询的可插拔 RAG 工作流
- 多模态解析：集成 MinerU 与 OCR，精准解析电商设备操作手册、商品售后指南类图文混排 PDF
- 检索机制：采用向量检索 + 稀疏检索 + Neo4j 电商知识图谱（设备故障 - 解决方案关联）多路召回
- 智能切片：支持滑动窗口、语义切分策略，适配电商设备故障排查步骤、商品参数说明的语义保留
- 深度优化：引入 HyDE 与 BGE-Rerank，提升 “打印机卡纸怎么办” “商品保修政策” 等电商疑问匹配精度
- 全链路评估：集成 RAGAS 框架，自动化评估电商售后问答准确性、设备操作指引合规性

项目3：电商小二

- 对话理解与意图解析：实现用户模糊咨询的多意图识别，支持上下文关联（如跨轮追问 “之前说的订单退款进度”），精准匹配用户真实需求
- 多源知识库联动：对接产品手册、售后工单库、常见问题库（FAQ），实现 “问题 - 答案” 智能映射，支持手册更新后的知识库自动同步与检索优化

- 实时交互体验优化：采用流式输出技术减少回复等待时长，配置常见问题快捷回复模板，针对高频咨询（如“物流查询”）实现 1 秒内响应，支持表情、链接等富文本回复
- 人机协同转人工机制：设置转人工触发条件（如复杂投诉、需求不明确），转人工时自动同步当前对话上下文至人工坐席，避免用户重复描述，提升协同效率
- 对话数据复盘与优化：自动采集对话日志，分析高频未解决问题、用户满意度低的回复场景，输出优化建议（如补充知识库内容、调整意图识别规则）
- 多渠道适配与监控：支持 APP、网页、小程序等多渠道接入，集成服务监控面板，实时查看响应耗时、意图识别准确率、转人工率等核心指标，保障服务稳定性

项目4：掌柜问数

- 自然语言转数据查询：用户用日常话术自动转化为精准数据查询指令
- 电商核心数据调取：实时关联订单量、库存水位、消费频次、类目 GMV 等核心维度数据，快速返回查询结果
- 数据结果解读：原始数据转化为易懂结论，避免用户解读数据成本
- 多轮上下文问数：记住用户历史查询，无需重复说明场景，支持递进式数据追问
- 数据权限适配：根据用户角色（如商家 / 运营）开放对应数据范围，仅查看权限内的店铺、类目或品牌数据
- 可视化结果输出：支持将查询结果自动生成折线图、柱状图或表格，可一键下载，适配汇报、分析等场景

项目5：市场罗盘

- 场景化任务拆解：结合问数、客服等智能体场景，布置真实业务需求（如电商订单数据查询、售后问题统计）
- 自主设计与开发：从 0 到 1 完成智能体架构设计、核心功能编码（如对话逻辑、数据接口集成）
- 阶段性目标划分：按“需求分析→原型设计→开发调试→功能验证”

拆分阶段，每个阶段设定明确产出（如需求文档、可运行 Demo）

- 态进度管控：通过项目看板同步进度，每周开展 1 次进度复盘，针对滞后情况调整任务难度或提供额外支持
- 维度过程检查：关键节点（如架构确定、核心功能完成）进行代码评审与功能测试，指出优化点（如交互流畅度、数据准确性）
- 成果落地与展示：提交可运行智能体及开发报告

05 大模型微调实践

大模型微调核心

- 大模型微调概述、核心要素、数据收集
- 大模型微调数据集处理、alpaca指令跟随格式、shareGPT多轮对话格式
- 大模型微调技术PEFT概述、prompt-tuning介绍、p-tuning介绍、zero-shot、few-shot
- 大模型量化算法、LoRA微调、QLoRA微调
- 大模型全参数微调技术详解、DeepSpeed分布式训练
- 大模型训练环境搭建、微调代码详解、微调参数详解
- 大模型合并、打包，vllm高性能部署
- 大模型评估方法与评估指标分析

企业级微调数据集构建方法论

- 公开数据集获取、私有数据采集
- 标注规范与质量管控
- 数据增强技术

基于 Llama-Factory 的高效微调落地

- 环境搭建、参数配置实战
- 本地GPU单卡/云端多卡训练步骤

- 适配部署的Safetensors/ONNX格式处理

调优案例

- 全程涉及多个调优案例

06 大厂开发规范

企业大模型研发流程

- 流程概述、技术前沿调研、行业实践调研、完整技术调研结构
- 自研方案输出、算法框架设计、RAG项目逻辑、对话系统分发与pipe-line
- 评估指标、业务方运营方与产品方角色、研发过程、行业趋势与能力培养
- 项目立项报告、产品需求、项目设计说明书等

大模型当下热点

- Agent/RAG项目研发主流技术，前沿大模型走向，热点跟踪

07 就业面试

简历指导

- 简历编写规范
- 线上简历投递技巧

模拟面试&职业规划

- 1v1 技术模拟面试、职业规划

企业真题(赠送)

- 赠送100+大模型企业面试真题

录播课程(赠送)

01 Python核心算法(录播视频)

数据结构与算法

- 常用数据结构:数组、链表、栈、队列、哈希表、树、图
- 常用算法:查找、排序、分治、动态规划、回溯、贪心
- 力扣经典面试150题

02 LangChain4j+Java大模型项目(录播视频)

LangChain4j

- LangChain4j概述、官网使用流程、架构分成说明、入门前置约定
- 工程入门搭建、POM依赖详解、环境脱敏处理、优化配置类说明
- Qwen/DeepSeek申请、多模型共存、SpringBoot整合对比、核心API说明
- Token用量计算、AIService深入理解、高阶API编码验证
- 模型参数详解、日志配置、监听配置、重试配置、请求超时配置
- 视觉理解、视觉编程、图像解析、万相模型实战
- 流式输出实战、ChatMemory理解、Eviction Policy、ChatMemory实战

- 持久化ChatMemoryStore、会话持久化编程实战
- FunctionCalling理解和整合、FunctionCalling实战（天气查询）、@Toolg进阶
- 扩展向量数据库、扩展RAG微调、MCP理论&实战

项目：小智医疗

- 对话理解与意图解析：识别医疗模糊咨询，支持上下文关联，精准匹配病症咨询、检查报告解读需求
- 多源知识库联动：对接电子病历库、用药指南、诊疗规范，实现“问题 - 答案”智能映射，自动同步新药信息、诊疗方案更新
- 实时交互体验优化：流式输出减等待，配置高频模板（如“降压药服用时间”“挂号流程”），1 秒响应，支持检查报告链接、用药提醒富文本
- 人机协同转人工机制：遇复杂病症、手术咨询触发转人工，同步患者基础病历片段，避免重复描述病情
- 对话数据复盘与优化：采集日志，分析高频问题（如“血糖高吃什么”“疫苗预约”），输出知识库补充、病症识别规则调整建议
- 多渠道适配与监控：支持医院 APP、公众号、小程序接入，监控响应耗时、病症识别准确率，同步保障患者隐私数据合规

预习内容

01 Python编程语言核心

python语言基础

- 变量以及数据类型、标识符和关键字、运算符、程序类型转换
- PyCharm 等开发环境的安装与配置
- 分支和循环、break、continue

- 字符串和列表、集合、字典和元组
- 函数的类型、函数参数、函数返回值、函数嵌套、递归函数、匿名函数、内置函数

python语言高级

- 面向对象：类和对象、实例属性、类属性、类方法、静态方法、封装、继承、多态、设计模式
- Object类、抽象类、异常、模块安装与使用、深拷贝、浅拷贝
- 生成器、迭代器、闭包、装饰器

02 学习链接

www.bilibili.com/video/BV1tDsgzxECr

说明：

结合实际情况，以上计划或有微调，以最终每月发的安排为准