

李俊颖

云南财经大学 · 智能科学与技术 · 本科 (2026 届) | ✉ 2710200510@qq.com | 📞 18288718934

求职意向: 大模型应用开发 / AI Agent 研发 / 智能算法工程师

教育背景

云南财经大学 智能科学与技术 本科 2022.09 – 2026.06

主修: 机器学习、深度学习、自然语言处理、计算机视觉、数据挖掘、智能系统设计、Python/C++ 程序设计

项目经历

大模型本地部署与 API 服务 个人项目 2025.06 – 2025.07

技术栈: Python · FastAPI · vLLM · Ollama · deepseek-r1-7b · Docker · Redis

- 部署实践: 在 AutoDL 云 GPU (RTX 4090) 及本地 RTX 4060 上, 基于 vLLM 部署 deepseek-r1-7b 实现推理服务; Ollama 部署 14B 模型用于本地调试; llama.cpp 量化 (Q4_K_M) 适配低显存场景
- API 网关: FastAPI 封装 OpenAI 兼容 /v1/chat/completions 接口, 实现到网页前端渲染的实时对话系统。
- 容器化: Docker Compose 编排模型服务+Nginx+Redis, 编写 Dockerfile, 确保环境可复现

AI Agent 智能助手开发 个人全栈项目 2024.09 – 至今

技术栈: LangChain · LangGraph · FastAPI · React · Chroma 向量库 · Function Calling

- Agent 架构: 基于 LangGraph 状态图实现主控 Agent 调度多个子 Agent (规划/执行/反思), ReAct 策略驱动自主推理-行动循环
- 工具系统: 实现 8 个可插拔工具 (网页搜索、计算器、天气查询、文件读写、SQL 查询、代码执行等), 统一 Function Calling 接口
- RAG 知识库: Chroma 向量库 + bge-large-zh-v1.5 嵌入, 支持 PDF/Markdown 文档分块索引检索, 实现带引用的问答
- 前后端: FastAPI 异步后端 + React 18 前端 (Ant Design), WebSocket 实时推送 Agent 执行过程

智慧楼宇能耗预测与优化 毕业设计 2025.10 – 2026.05

技术栈: PyTorch · CNN-LSTM · DQN · Flask · paddlepaddle · MQTTX · Vue.js

- 模型训练: 使用公开建筑能耗数据集, 构建 CNN(Conv1D)→BiLSTM→Attention 混合模型预测 24 小时能耗, AdamW 优化+EarlyStopping, 测试集 MAPE=8.2%, $R^2=0.52$
- DQN 优化: 将空调温度设定建模为 MDP, DQN 训练 5000 轮学习最优控制策略, 仿真环境实现节能约 15%
- 工程落地: MQTTX 模拟传感器数据→InfluxDB 存储→Flask 加载 TorchScript 模型推理→Vue.js+Node3D 可视化看板。

专业技能

LLM 部署: vLLM / Ollama / llama.cpp 部署与量化, FastAPI 封装 OpenAI 兼容接口, Prompt Engineering (Few-shot/CoT/ReAct)。

Agent 开发: LangChain / LangGraph 多智能体编排, Function Calling 工具链, RAG (Chroma/Milvus), ReAct 自主规划。

模型训练: PyTorch, CNN/LSTM/Attention 时序建模, DQN 强化学习, 数据清洗与特征工程, ONNX/TorchScript 模型导出

工程能力: Docker / Docker Compose, FastAPI/Flask, Redis/PostgreSQL/InfluxDB, Git, Linux(Ubuntu), MQTT 协议