

张文飞

AI 应用开发工程师 | Python • RAG • Agent workflow

电话: 19534845035 | 邮箱: 2730793998@qq.com | 2026 届本科 | 已取得双证 | 可随时到岗 | 接受全国城市
个人作品集 | 项目界面、架构与脱敏静态展示: <http://47.111.231.30:8888/>

核心优势

- **AI 应用落地与业务提效:** 具备公司 AI 实习与企业项目制交付经历, 能识别资料整理、知识检索、数据分析和业务录入等重复环节, 将知识库/RAG、Text-to-SQL、OCR/语音和 Agent 工具调用接入现有系统, 减少重复操作、提高处理效率并提升系统智能化程度。
- **RAG 全流程与知识工程:** 完成过多格式解析、清洗分块、Embedding、BM25/向量/结构化混合索引、RRF 融合、来源引用、权限过滤和索引版本治理, 并通过检索参数、标准答案、失败降级及调用日志提升稳定性与可追溯性。
- **AI 框架与工程交付:** 实际使用 LangChain、LangGraph、LlamaIndex、Vanna 完成固定链、状态编排、工具调用与结构化输出, 能以 Django/FastAPI + Vue/React 开发业务系统, 并用 Docker、Linux、Git 完成测试、部署和维护。

工作经历

洛阳矩阵软件有限公司 | 大模型应用工程师 (实习)

2026.02–2026.09

- 在团队中负责系统全栈与 AI 链路开发、联调、部署和维护, 将 AI 应用于学习前的课件解析/备课、学习中的讲解/问答/测验和学习后的掌握度分析/学习报告, 系统已在实际教学场景使用。
- 面向 PPTX/PDF 课件递归提取文本框、组合元素和表格, 构建页级知识库; 从章节知识点抽取 prerequisite、contains、application 等 **6 类关系**, 经节点白名单、去自环等校验后生成可编辑知识图谱。
- 构建课程材料与章节课件范围过滤的向量查询, 并结合知识图谱关系查询形成混合查询能力; 检索由 **Top 15 初筛至 Top 5**, 通过 SSE 返回答案与来源页码, 以关键词、相似度和 LLM 判定拦截离题问题。
- 统一记录模型、action_type、Token、耗时、课程/章节上下文及 requestId, 支撑管理端按模型和业务场景统计; 对出题与学习报告采用结构化校验和失败降级, 并完成 **129 项接口集成测试**。

项目经历

工业设备智能售后与知识服务平台 | AI 应用/全栈独立开发 (企业项目制)

2026.04–2026.09

技术栈: Django REST Framework、Vue 3、MySQL、Vanna、LangChain LCEL、BGE-M3、Qdrant、BM25、Redis、SSE、WebSocket、Docker

- 与驻场需求人员协作梳理业务, 独立负责架构、前后端、AI 链路、测试与部署, 覆盖产品版本、客户设备、售后工单、知识问答、工作日志和分析报表等核心业务; 完成测试环境验收与一期试运行。
- **混合检索与权限:** 针对型号、版本号、故障码和自然语言并存的查询, 构建 BM25、BGE-M3/Qdrant、结构化数据 **3 路召回 + RRF 融合**; 高频确定性问题优先命中审核答案, 并在候选阶段复核角色、客户、资产及 active 索引成员权限。
- **Text-to-SQL 分析:** 在售后分析模块接入 Vanna, 将自然语言问题转换为受控数据查询; 执行前由后端复核用户权限和可查询范围, 异常或不可执行结果回退到固定指标与筛选查询, 将 AI 查询能力安全接入业务平台。
- **索引与流式问答:** 以 **5 态索引生命周期** 支持失败分块、断点续建、完整性校验和验证后原子激活; LCEL 固定链路通过 **6 类 SSE 事件** 返回回答、引用、心跳及错误, 构建失败时保持旧 active 版本可用。
- **多模态业务接入:** 使用 PP-StructureV3 识别送货单, 先生成私有草稿、人工校准后落库; ASR 转写工作日志并保留原文和结构化字段, 支持空转写补录、人工确认与失败降级, 音频不落盘。
- **可信分析与交付:** 确定性查询先生成带证据 ID 的指标, 再由 Writer/Reviewer 组织分析并程序化复核数字; 持久化任务、租约和心跳支持中断恢复, 模型不可用时返回基础数据报告; 完成半双工 ASR-RAG-TTS 及 Docker/Nginx 部署、迁移、备份和回滚。

项目经历

洛阳智能旅游系统（本科毕业设计） | 独立开发 2026.02–2026.05

技术栈：FastAPI、LangGraph、MCP、LlamaIndex、ChromaDB、DashScope、SSE、React 19、高德地图 API

- 独立完成需求、架构和前后端开发，基于 LangGraph 构建单协调器、多节点旅行规划 workflow，以 **8 个图节点、26 个 State 字段、3 条条件路由** 管理需求分析、工具调用、规划与校验状态。
- 将高德 POI、天气、路线和本地文旅 RAG 能力封装为 MCP Server 工具，统一参数与返回结构，供规划 workflow 调用；模型输出结构化决策，应用层负责工具执行、结果观察和异常降级。
- 使用 **16 个校验方法、21 个约束编号** 检查开放时间、时间冲突与路线合理性，将失败原因回传规划节点，默认最多修正 2 次，形成规则校验与有限重规划闭环。
- 使用 FastAPI、SSE 与 React 19 展示 workflow 状态和最终行程，完成高德地图与 DashScope 真实链路联调。

专业技能

- **后端与前端开发**：使用 Python、Django REST Framework、FastAPI、SQLAlchemy、MySQL、Redis 开发业务 API、权限与异步任务；使用 Vue 3、React、TypeScript 完成管理端与交互页面，处理 JWT/RBAC、事务、文件上传和第三方服务接入。
- **知识库构建与 RAG 优化**：覆盖 PPTX/PDF/OCR 解析、清洗分块、Embedding、索引、召回、融合、生成和来源回传全流程；使用 BM25、BGE-M3/Qdrant、ChromaDB、结构化数据构建混合索引与查询，以 RRF、Top-K、权限校验、标准答案、版本切换和失败回退优化效果。
- **AI 框架与 Agent**：使用 LangChain/LCEL、LangGraph、LlamaIndex、Vanna；具备 Prompt 设计、结构化输出、MCP 工具封装与调用实践，理解 Function/Tool Calling、ReAct、CoT 等基础方法，可实现状态路由、结果观察、规则校验和有限重规划。
- **知识图谱与多模态**：使用 Neo4j 建模与展示业务关系，完成知识点关系抽取、校验和图谱查询；使用 PP-StructureV3、ASR/TTS、SSE、WebSocket/Channels 接入 OCR、语音转写与合成，并通过人工确认、状态机和降级控制结果风险。
- **部署维护与研发提效**：使用 Linux、Docker Compose、Nginx、Git 完成部署、迁移、备份和回滚，以 pytest、Playwright、requestId 日志及调用统计验证主链路；熟练使用 Claude Code、Codex 辅助需求分析、编码、重构、测试和排错，关键结果由本人审查验证。

教育经历

洛阳理工学院 | 计算机科学与技术 | 本科 | 2022.09–2026.06
GPA 3.87/5.0（专业前 5%） | 国家励志奖学金 | 校级二等奖学金 | 优秀学生干部

认证与荣誉

阿里云大模型高级工程师 ACP 认证 | 人工智能训练师四级/中级工 | 蓝桥杯河南省二等奖 | 超级智能体 Campathon 三等奖 | iCAN 河南赛区三等奖