

工作经验

甲丁甲科技 AI全栈工程师 2026/04 - 2026/07

- 负责 AI SaaS、抖音直播助手与数字人口播平台的算法方案、系统架构及全栈交付；
- 研究 Coze 工作流/智能体及 AI 应用，落地智能客服、微信小程序与飞书机器人集成；
- 基于 Next.js 开发 Google AI Studio 生图工具，完成 Vercel 部署并接入统一 SaaS 平台；
- 完成 LLM 服务接入、支付积分、OSS、多节点控制及 Windows/GPU 部署与稳定性优化。

小猿教育科技有限公司 AI算法工程师 2025/04 - 2026/03

- 负责 AI 教育产品的大模型应用框架设计与核心模块开发；
- 主导多轮对话 Agent、RAG 检索增强及多模态调度工程化落地。

秒优大数据科技有限公司 AI算法工程师 2023/11 - 2024/12

- 参与智能客服、推荐、翻译及企业智能化项目开发；
- 负责模型训练、数据处理与服务接口封装。

南京法奔信息科技有限公司 NLP算法工程师 2023/06 - 2023/09

- 参与需求分析、架构设计及法律 NLP 核心模块开发；
- 负责文本向量化训练与案例匹配系统实现。

项目经历

AI 工具调度与商业化 SaaS 平台 全栈负责人 2026/04 - 2026/07
<https://aibigtree.com>

项目概述：

面向企业 AI 应用统一运营场景，构建集用户、工具、权限、积分、支付、日志与素材管理于一体的全栈 SaaS 平台，统一接入 AiStudio、Coze、N8N 等 AI 工具与工作流。

- 全栈架构：基于 Next.js 14 App Router、React、TypeScript、Ant Design 与 Prisma/PostgreSQL，完成管理端、用户端、API 及数据模型设计，并部署至 Vercel/阿里云服务器。
- 工具与积分闭环：设计 launch/verify/consume 三阶段校验扣费流程，使用数据库事务记录积分与操作日志；统一封装 Coze、N8N、AiStudio 调用及 iframe/postMessage 交互。
- AI 生态落地：基于 Coze 搭建工作流与智能体，完成智能客服、微信小程序及飞书机器人对接；使用 Next.js 开发 Google AI Studio 生图工具，部署至 Vercel 并以链接/API 接入 SaaS。
- 商业化与工程化：实现短信验证、支付、积分订单与渠道控制；完成 OSS 存储、结果图保存、数据迁移、超时重试、结构化日志和统一异常处理。

抖音直播间智能自动回复与活跃助手 开发负责人 2026/04 - 2026/07

项目概述：

面向主播与直播运营团队，基于 FastAPI、Playwright 和大语言模型构建直播间智能助手，实现弹幕监听、规则/AI 自动回复、定时公告、冷场活跃及多账号多节点统一控制。

1. 自动化采集与发送：利用 Playwright + MutationObserver 监听 React 虚拟列表弹幕，过滤点赞/进场等系统消息；通过持久化 Profile 保存登录态，并模拟键盘输入完成消息发送。
2. 智能回复引擎：采用“关键词规则优先、LLM 兜底”策略，兼容 OpenAI、Gemini、DeepSeek；实现回复去重、20 字语义截断、黑名单、自身消息识别及全局/用户级冷却。
3. 多账号与多节点：为每个账号隔离浏览器 Profile、代理、规则与定时任务；通过 REST/WebSocket 完成主控与子电脑配置同步、状态回传、批量启停和远程发送。
4. 稳定性与交付：加入 asyncio 锁、发送限频、代理池检测、异常 Profile 修复、日志轮转与 Pydantic 校验；使用 PyInstaller 打包 Windows 桌面程序并完善单元测试与部署文档。

数字人口播与音色克隆视频生成平台 项目负责人 2026/04 - 2026/07

项目概述：

基于 HeyGem 与 IndexTTS2 构建 Linux GPU 数字人口播平台，输入文案、参考音频和图片/视频即可完成音色克隆、语音合成、人物口型驱动与成片输出。

1. 生成链路：串联文案优化/翻译、IndexTTS2 音色与情感克隆、图片转视频、HeyGem 人脸驱动和 FFmpeg 合成，支持图片/视频素材及批量生成。
2. 服务化架构：建设 FastAPI 任务接口与 WebUI，使用后台任务、GPU 互斥锁、状态轮询和安全取消控制推理并发；提供素材库、生成记录、预览、下载与删除。
3. 内容工作流：接入火山引擎流式润色/翻译，支持抖音链接解析、素材下载及 SiliconFlow ASR 回填文案，减少内容搬运与人工整理。
4. 部署优化：拆分 HeyGem Python 3.8 与 IndexTTS 独立环境，在 AutoDL Linux/CUDA/ONNX Runtime 上完成部署；完善显存清理、单 Worker、启动脚本与异常提示。

AI 智能教育辅助平台 AI模块负责人 2025/05 - 2026/03

项目概述：

本项目是基于Fast API架构与大模型（LLM）及向量检索能力深度融合的智能多模态交互式问答AI教育平台，系统以融LangGraph 作为对话流引擎核心，多Agent 调用，串联意图识别、上下文感知、RAG检索增强推荐。涵盖从多轮智能对话（SSE流式相应），图片理解识别分析，到利用FFmpeg 语音输入，教学视频生成等全业务线路。

1. 基于图结构对话工作流引擎设计与开发：基于LangGraph 设计高度模块化AI会话生命周期节点，包括语音输入，意图分类（Intent Classification）、闲聊判定（Chitchat Node），实现对话摘要压缩以解决长对话Token 限制；设计智能查询改写模块补全语义，并结合多规则判断意图延续。推荐资源召回层及兜底回复，采用Server-Sent Events（SSE）进行流式回复。
2. RAG检索增强与个性化动态推荐：采用BGE-M3加BGE-Reranker 二阶段检索，基于内容哈希实现增量计算；引入相似度阈值去重与多样性惩罚，实时提取偏好解决冷启动。基于用户交互历史或指定诉求智能推荐。
3. 多模态AI工具：① 识题解题：集成GLM-4.5V视觉模型，设计多层容错解析不稳定JSON；并开发答案一致性校验函数自动修正字段，支持符合题型递归映射与LaTeX规范化。② 大纲教案生成、PPT生成：提高老师备课效率；
4. LLM工程化：封装HTTP 客户端池管理多模态调度，处理SSE 流式输出异常保障实时体验，通过JSONMode 与自检指令约束缓解幻觉问题。

项目概述：

为教育平台课程、教师、场馆和群组提供基于用户行为与兴趣标签的动态推荐服务。

- 1. 构建 9 类兴趣标签及行为反馈链路，通过 API 实时更新用户画像。
- 2. 设计标签覆盖率与精确匹配加权算法，结合冷启动打散、地域和年龄等特征生成推荐结果。

项目概述：

本项目服务于GST项目中服装工序翻译任务,应用于中文-->英语、柬埔寨语、朝鲜语的翻译,节省人力成本。

- 1. 需求分析:调研并确定项目需求,确保翻译系统能够满足中文至英语、柬埔寨语、朝鲜语的翻译需求。
- 2. 数据预处理:实施高效的数据清洗流程,处理并优化约2万条翻译数据对,确保数据质量。
- 3. 专有词表构建:整理并构建服装工序领域的专有词表,提高翻译的准确性和专业性。
- 4. 基于 M2M100-418M 完成训练、调参与微调，并通过交叉验证保障翻译质量。
- 5. 完成模型服务集成与测试，提升专业术语翻译的准确性和稳定性。

项目概述：

面向工厂管理场景分析员工效率，实现低效率预警、沟通辅助与早会报告生成。

- 1. 对接 SQL Server 数据，构建员工效率指标计算与定时查询链路。
- 2. 集成大语言模型完成效率分析问答、改进建议与早会报告自动生成。
- 3. 使用 FastAPI 封装模块接口，完成系统集成和数据交互。

项目概述：

本项目是开发并集成一套成衣推荐系统至公司ERP系统中，旨在通过分析历史服装数据，智能推荐工厂及员工，从而减少时间及人力成本。

- 1. API 封装：使用FastAPI 框架封装图像处理和检索系统的API 接口，支持用户上传图片、链接或文件夹。
- 2. 特征提取：集成VGG16模型，对上传的图像进行特征提取，为图像检索提供数据支持。
- 3. 图像检索：开发功能，允许用户传入单张图片或图片链接，在图片库中查找相似图片。
- 4. 图片管理：实现单个或多个图片的删除功能，以及整个图片库的清空操作，维护图片库的整洁和高效。
- 5. 系统优化：优化FastAPI 应用性能，确保接口响应速度和系统稳定性。

项目概述：

为服装商城开发并嵌入一套系统,该系统基于商城信息、、供货商和设计师数据,提供问答和推荐服务。利用LangChain框架,结合大型语言模型和本地知识库,旨在为用户提供专业、准确的问题解答。

- 1. 数据预处理:负责商城信息、、供货商和设计师信息的数据清洗,确保数据质量和可用性。
- 2. 技术调研:调研并选择合适的模型和云服务器,为系统提供稳定的技术支持。

3. 模型应用:使用chatglm3-6B模型,编写Prompt以引导模型生成准确和相关的回答。
4. 数据存储:使用 Elasticsearch 完成数据存储、索引维护与检索，并结合本地知识库优化回答质量。

律呗法律助手

算法负责人

2023/06 - 2023/09

项目概述：

面向律师工作场景提供法律咨询、案例检索与相似案例推荐。

1. 完成法律数据清洗、增强、Elasticsearch 入库与 Query 检索。
2. 训练 Doc2Vec 向量模型，计算用户问题与案例相似度，并参与需求分析、架构设计与 NLP 核心模块开发。

专业技能

- 算法方向：NLP、文本分类与匹配、向量检索、推荐算法、Transformer、BERT、BGE
- 大模型应用：LangChain、LangGraph、skills、MCP、Agent、RAG、Prompt Engineering、Embedding/Reranker、SSE
- 多模态生成：GLM-4.5V、Google AI Studio、IndexTTS2、HeyGem、FFmpeg
- 模型框架：PyTorch、TensorFlow、Keras、Transformers、scikit-learn
- AI 生态集成：Coze 工作流/智能体、N8N、OpenAI/Gemini/DeepSeek、微信小程序、飞书机器人
- 工程开发：Python、FastAPI、Next.js、TypeScript、React、Prisma、Playwright
- 部署交付：Linux、Docker、Git、Vercel、AutoDL、CUDA、ONNX Runtime、PyInstaller
- 数据存储：PostgreSQL、MySQL、SQL Server、Elasticsearch、Neo4j

自我评价

- 具备从传统 NLP、文本向量化与推荐算法，到 LLM、Agent、RAG 和多模态应用的连续技术积累。
- 能够使用 Python/FastAPI 与 Next.js/TypeScript 完成算法服务、后台系统和前端产品的端到端交付。
- 熟悉 Coze、Google AI Studio、微信小程序与飞书机器人等 AI 生态的集成和商业化接入。
- 具备 Windows/Linux/Mac、GPU/CUDA、AutoDL、Vercel 等多环境部署、性能优化与稳定性交付经验。

教育经历

北京邮电大学

计算机科学与技术

2021/09 - 2026/07