张欣耕

工作地点不限 | zhangxingeng970221@gmail.com | +86 18652158890
shanechang.com | github.com/zhangxingeng | https://www.linkedin.com/in/zhangxingeng

求职意向

全栈 AI 工程师（大模型方向）—期望从事大语言模型（LLM）预训练、微调及强化学习（RLHF）相关工作

专业概述

拥有 5 年以上生产级 LLM 系统开发经验的**全栈 AI 工程师**。专注于**智能体架构**（工具调用、规划、记忆）、**RAG 检索增强生成**及**分布式工作流编排**。主导构建的系统服务 2,000+ 用户，日均 API 调用量超 10 万次。

专业技能

生成式 AI: 智能体系统（Agent）、工具调用（Tool Use）、规划与记忆、提示工程（Chain-of-Thought, Few-Shot）、RAG 检索增强生成、多智能体编排、LLM 评估、结构化输出
编程语言: Python, TypeScript, JavaScript, SQL, C++ **框架:** FastAPI, Litestar, SvelteKit, React, Node.js, Temporal.io
机器学习: PyTorch（分布式训练）, Transformers, HuggingFace, LangChain, Pydantic-AI, LiteLLM
数据存储: PostgreSQL, Redis, 向量数据库（Chroma, pgvector）, Elasticsearch, Neo4j
DevOps: Docker, Kubernetes, OpenTelemetry, GitHub Actions, AWS (S3, SageMaker, Bedrock), Azure

工作经历

Empath Legal (empathlegal.com) 美国新泽西州
创始全栈工程师（GenAI） *2024 年 12 月 - 至今*

- 独立构建端到端**智能体 AI 系统**：**Python/Litestar** 后端 + **SvelteKit/TypeScript** 前端 + **PostgreSQL** 数据库，编排 78+ 条 GenAI 流水线，实现陪审员背景调研自动化
- 设计基于 **Temporal.io** 的**多智能体工作流编排**：支持容错处理、人机协作反馈、**Chain-of-Thought** 推理、步骤级暂停/恢复
- 架构**多供应商 LLM 接入层**（**LiteLLM/Pydantic-AI**）：无缝切换 **GPT-4/5**、**Claude 3/Opus**、**Gemini**、**Mistral**、**Llama 3**，统一结构化输出与工具调用接口
- 实现**类型安全架构**：OpenAPI 自动生成 TypeScript 类型、Pydantic 模型驱动 API 契约；基于 Redis Pub/Sub 的实时 **SSE** 推送
- 搭建生产级基础设施：**Docker Compose**（10+ 服务）、**OpenTelemetry** 可观测性、OAuth 2.0 多平台认证（Google/Microsoft/Apple）、**353 条**自动化测试

花旗集团 (Citigroup Inc.) 美国新泽西州
AI/机器学习工程师 *2024 年 8 月 - 2024 年 12 月*

- 构建**智能体 RAG 系统**：结合知识图谱与**向量数据库**（Elasticsearch, pgvector）进行合规性分析，F1 分数从 0.32 提升至 0.71，幻觉率降低 **30%**
- 设计 **FastAPI** 后端服务：日均处理 **10 万 + 次 API 调用**，将单文档人工审核时间从 **4 小时缩短至分钟级**；连续 120 天零宕机
- 构建数据处理流水线：文档处理耗时从 10 秒降至 **1 秒以内**，服务 **2,000+ 用户**；基于 Terraform 部署至 **AWS**（SageMaker, Bedrock）

罗伯特・伍德・约翰逊大学医院 (RWJUH) 美国新泽西州
AI/机器学习工程师 *2023 年 1 月 - 2024 年 8 月*

- 构建基于 **Chroma** 的 **RAG 系统**：自动摘要 **5,000+ 篇**研究论文，大幅提升神经科学研究人员的文献检索效率
- 训练定制化**多模态 Transformer 模型**：处理神经科学数据（图像 + 时序信号），研究成果已发表
- 与 **15 位研究人员**紧密协作，整合 **200+ 条**反馈意见；开发 **React** 前端，将新用户上手时间缩短 **50%**

罗格斯新泽西州立大学 - 研究实验室 美国新泽西州
研究生助研 *2021 年 9 月 - 2022 年 12 月*

- 研究基于 **Transformer** 的 3D 点云重建模型；实现 **PyTorch 分布式训练**，利用多 GPU 集群将训练时间缩短 **40%**

Fiskkit Inc. 美国旧金山
机器学习工程师 *2020 年 1 月 - 2021 年 7 月*

- 将 **NLP 功能**集成至 **Node.js** 后端，基于 **PyTorch**（CUDA）实现实时文本生成与摘要，响应时间降低 **70%**
- 构建 **PySpark** + **PyTorch 分布式**数据流水线，模型训练效果较基线提升 **20%**

教育背景

罗格斯新泽西州立大学 (Rutgers University), 美国新泽西州 2020 年 12 月 - 2022 年 12 月
计算机科学硕士（机器学习方向） GPA: **3.71**
罗格斯新泽西州立大学 (Rutgers University), 美国新泽西州 2017 年 9 月 - 2020 年 9 月
计算机科学学士

其他信息

认证: AWS 机器学习专项认证 (MLS-C01) 2024 年 6 月
教学经验: 罗格斯新泽西州立大学助教，指导 **400+ 名**学生学习计算机科学与机器学习课程；具备工程师入职培训与团队带领经验
语言能力: 英语（流利）、普通话（母语）