

任德耀

大模型应用工程师 / AI Agent 开发工程师

2 年制造业 agent 业务经验 + 全栈 AI 工程能力 —— 把「模型能跑」推进到「线上能扛」



电话 | 13395478541 (微信同号)

邮箱 | rdy1131635252@gmail.com

现居 | 杭州 · 25 岁

籍贯 | 山东济宁

学历 | 统招本科 · 泰山科技学院

主页 | <https://de-yao-ai.com>

专业技能

大模型 / Agent	自研 Agent 编排引擎 (Plan-Execute + ReAct、Supervisor 模式)、OpenAI 兼容 Function Calling / Tool Use、上下文工程 (token 感知截断、滚动摘要压缩、多轮会话持久化)、Prompt 工程、SSE 流式思维链; 了解 MCP、LangGraph 多智能体编排
算法基础	机器学习 (监督 / 无监督学习、特征工程、模型评估 Precision / Recall / F1)、深度学习 (Transformer / Attention、损失函数、梯度下降与反向传播、正则化与过拟合)、LoRA / 量化原理、解码策略 (温度 / Top-p)
微调 / 训练	LoRA / QLoRA (PEFT)、SFT / DPO、CoT 知识蒸馏 (Teacher-Student 数据合成)、LLaMA-Factory; 自建评测体系 (独立测试集 + recall / F1 + 消融对照), 实测驱动选型
推理 / 部署	vLLM (offline batch / 在线服务)、Ollama + GGUF 量化、模型分层路由 (小模型抽取 + 大模型判别, 推理成本省约 50%)
RAG / 检索	BGE-M3 多语言嵌入、Chroma (HNSW) 向量召回、企业知识库构建; 了解 Milvus、混合检索 (BM25 + 向量) 与 Rerank 重排
后端	Python (FastAPI、SQLAlchemy 2.0 async、Pydantic v2、asyncio、Pandas、PyTorch)、JWT + RBAC 鉴权体系、Node.js
前端	React 18 / 19、TypeScript、Vite、Tailwind CSS、Recharts、TanStack Query、Zustand
数据 / 存储	MySQL、MariaDB、Redis、Milvus / Chroma / FAISS (向量库)、Hadoop 生态 (HDFS / Hive / MapReduce)
工程化	Docker / Docker Compose、GitHub Actions CI/CD、结构化日志与健康检查、frp / Cloudflare Tunnel
开发工具	IDE: VS Code、Cursor、PyCharm; AI 代码辅助: Claude Code、Codex、OpenClaw
系统环境	Linux (Ubuntu, 主力开发与部署环境)、Windows; Bash / Shell 脚本
低代码平台	Dify、Coze (扣子) —— 快速搭建原型, 用于与自研方案做能力边界对比
业务系统	ERP (ERPNext v15 / Frappe、鼎捷 T100)、CRM (缤纷销客)、OA —— 懂销售到财务的全域业务数据模型

工作经历

宁波巨隆机械有限公司 大模型应用工程师

2 年

- 主导推动 CRM / ERP / OA 接入智能体，梳理跨部门数据流并沉淀标准 SOP
- 用 Python + Pandas 对历史订单做客户分群，输出高价值客户优先级清单，直接支撑销售资源分配
- 深度参与 ERP 落地，积累销售 / 采购 / 仓储 / 生产 / 质检 / 财务全业务域的数据模型与流程理解

杭州绿兴环保有限公司 数据分析师

实习

- 用 Python 完成实验数据整理与可视化分析，编写脚本提升实验记录录入效率

项目经验

项目一 | 企业 ERP 智能问数 Agent 平台

2026 · 独立负责全栈交付 · 持续迭代

线上演示: app.de-yao-ai.com | ERP 后台: erp.de-yao-ai.com

业务背景

制造企业 ERP 取数高度依赖 IT 和熟练员工——业务侧要一张报表平均等半天到两天，月底对账、客诉追溯要人工跨 3-4 张表核对。本项目把销售 / 采购 / 仓储 / 生产 / 质检 / 财务 / 导出 **7 大业务域** 的高频取数场景重写为自然语言问答：一句话提问，Agent 自动规划、调用参数化 SQL 工具直连 ERP 生产库，秒级返回结构化结果与可下载图表 / 报表，全程流式展示推理过程。

职责

独立完成需求分析、架构设计、Agent 引擎、前后端开发、ERP 数据接入与上线部署运维。

核心技术方案

- 自研 Agent 编排引擎**：Plan-Execute 主循环（max 20 turns 防失控）+ OpenAI 兼容 Function Calling，LLM 基于语义自主选工具；Provider 无关封装，DeepSeek / OpenAI / Qwen 配置级热切换
- 43 个参数化 SQL 工具 × 7 业务域**：100% 参数化绑定 + Agent 专用只读数据库账号 + scope 行级数据隔离，三重防线杜绝 SQL 注入与提示词注入拖库
- 三层 RBAC 权限体系**：REST 路由 require_role → 导出域白名单 → 工具级 allowed_roles（按角色过滤工具后注入系统提示，LLM 对越权工具不可见、调不动），7 角色 / 21 个端点权限自测全量通过，越权零泄露
- SSE 流式思维链**：自研 start / thinking / tool_call / tool_result / answer 5 种事件协议，前端实时渲染 Agent 推理全过程
- 上下文工程**：MySQL 三表持久化（sessions / messages / context_summaries）+ token 感知截断 + 长对话自动摘要压缩，多轮成本可控
- RAG 知识问答**：BGE-M3（1024 维）+ Chroma HNSW 向量召回，覆盖企业制度 / 质检 SOP 等非结构化文档，新员工首日即可自助答疑
- 双数据源架构**：本地 MySQL（22 张业务表）与 ERPNext v15 MariaDB 生产库之间表名 / 字段自动映射，配置一键切换，演示 / 生产环境无缝迁移
- 工程化交付**：Docker Compose 编排、JWT + bcrypt 鉴权、结构化日志、三端健康检查、frp + Cloudflare Tunnel 公网发布

成果

- 高频报表 / 对账 / 跨表追溯场景，从人工 30 分钟-2 小时压缩到**约 7 秒**端到端返回
- 权限合规：三层 RBAC，**21/21** 端点自测通过，不同角色同一问题返回不同数据（行级隔离）
- 全链路公网部署，所有功能可在线实测验证

FastAPI

Function Calling

SQLAlchemy 2.0

React 18

TypeScript

Tailwind

Docker Compose

项目二 | 合同评审大模型微调工程

2026 • 独立负责算法全链路

业务背景

制造企业销售 / 采购 / 加工承揽 / OEM 合同评审依赖人工 —— 一份 20 页合同识别风险 + 抽取要素需 30-60 分钟，经验不足的评审漏检率高，规则 / 关键词方案换个说法就漏。本项目训练两个任务专用模型：**要素抽取**（甲乙方 / 金额 / 期限 / 关键义务 / 生效与解除条件）+ **12 类风险条款识别**（R01-R12，含等级 + 原文证据 + 判定依据），整份合同 **1-2 分钟**出评审报告，推理全程合同数据不出内网。

核心技术方案

- **数据生产管线（解决中文合同无公开数据集）**：DeepSeek-Reasoner 作 teacher 蒸馏合成中文样本（强制 CoT 150-300 字 + 严格 JSON schema），英文复用 CUAD 公开数据集并将其 41 类标注映射到自建 R01-R12 风险体系；共 2000 条双语样本，约 20% 对抗 / 边界样本，全量 pydantic v2 强校验，按合同 ID 切分 train / test 防数据泄漏
- **横向选型基准**：5 个候选（DeepSeek-chat / GLM-4.5-Air / Qwen3-4B / 8B / 14B）统一 prompt + schema + 独立测试集同条件 zero-shot 实测，本地推理走 vLLM offline batch 保证同口径
- **LoRA 微调**：LLaMA-Factory， $r=16 / \alpha=32 / \text{target=all}$ （注意力 + MLP），bf16 + gradient checkpointing，A800 80G 单卡，抽取任务 12 分钟 / 风险任务 30 分钟完成训练
- **关键决策（实测推翻预设）**：thinking vs no-CoT 严格对照实验 —— no-CoT recall 0.9494 vs 0.9207、延迟降 2.65 倍、输出 token 省 2.41 倍，精度成本双优，最终上线 no-CoT 版本
- **分层部署（model cascade）**：4B 抽取模型走本地 Ollama（GGUF 量化，约 8GB 显存即可跑），14B 风险模型走云端 vLLM，按任务复杂度分配算力，推理成本省约 50%

成果

- 风险识别 recall **0.658 → 0.949 (+28pt)**，超 DeepSeek-chat zero-shot 24pt；要素抽取 macro_acc **0.623 → 0.821 (+19.7pt)**
- 微调后 JSON 解析率 **100%**（pydantic 一次通过）；本地推理单条 0.05-0.2s，较 API 方案快 50-100 倍
- v2 路线：真实评审意见回流再训、on-policy 蒸馏、LangGraph 多智能体评审流水线

LLaMA-Factory

LoRA

Qwen3-4B/8B/14B

vLLM

Ollama / GGUF

DeepSeek-Reasoner

CUAD

pydantic v2

TensorBoard

AutoDL A800

自我评价

- **工程化思维**：擅长把「模型能跑」推进到「线上能扛」——鉴权、权限隔离、可观测、成本控制全链路落地
- **实测驱动**：不迷信直觉与论文结论，选型 / 架构决策全部用自建评测集对照实验说话（曾用实测数据推翻自己 PRD 的 thinking 上线预设）
- **业务理解**：2 年制造业 agent 经验，能把技术方案讲给非技术同事听，清楚哪些流程适合接 AI、哪些应保留人工判断