崔维铁

18612247536 • nile.cui@gmail.com • 男
https://ncui-blog.netlify.app

求职意向

意向岗位：算法工程师	意向城市：北京
期望月薪：35000~45000 元/月	求职类型：社招

自我评价

1. **工作背景**：拥有 10 年以上算法、架构与嵌入式开发经验，现任中科雨辰科技算法工程师，负责大模型平台研发、RAG 系统集成、多模态建模与国产化适配，参与多个核心项目从 0 到 1 架构搭建与落地部署。
2. **个人优势**：在大模型、RAG、多 Agent 任务流构建等方向具有丰富实战经验，擅长从平台架构到模型部署的全链条任务拆解与落地，拥有产品化视角和跨模块协同能力，能在复杂异构系统中快速定位问题并高效迭代优化。
3. **专业技能**：熟练掌握 SGLang、vLLM、Ollama、Taskweaver、Dify、LangChain、LlamaIndex、langgraph 等大模型工具链，具备 BERT、LSTM、HDBSCAN、UIE、Qwen 等专业模型及大模型调优与蒸馏经验，精通 Python、gRPC、异步队列、分布式系统架构，擅长任务调度、内存优化与国产平台适配。
4. **综合素质**：具备强烈自驱力与持续学习能力，善于在跨行业与跨技术背景中寻找解决方案，乐于挑战前沿技术领域，具备良好沟通表达与技术文档能力，能够胜任大模型平台研发、算法架构设计与多系统融合等高复杂度岗位。

工作经历

2021.1-至今 **中科雨辰科技有限公司 | 算法工程师（人工智能）**

1. **平台架构与系统集成**：主导辰智标注平台、训练平台、事件系统、知识智能分析平台、Taskweaver BI 工具等 8 个核心产品的算法模块开发与接口标准制定，完成平台级接口规范文档 12 份，支持各产品间的数据联动与模块复用。
2. **RAG 与 Agent 工具研发**：参与 Agent 标书生成工具、翻译平台、细雨搭档（AiChat）等项目的 RAG 能力搭建与知识库集成，使用 Dify + RAGFlow 方案迭代上线版本 4 次，优化生成内容与查询效率，支撑文档处理任务日均 1200 次。
3. **多模态与事件抽取模型训练**：主导新闻与社交媒体事件系统的抽取模型训练，构建基于 BERT 和 LSTM 的双塔结构，累计处理样本数据 4500 万条；研发视频预警监测模型 2 个，结合图像帧提取、语音文本与事件联动实现自动检测方案。
4. **大模型推理与部署优化**：在 AIP 平台及辰悉平台中部署大模型推理服务，采用 SGLang、vLLM、Ollama 框架，实现 32B 与 70B 模型多基座运行，支持并发用户请求峰值 3000 次/分钟；完成模型微调任务 12 次，优化响应延迟 30 毫秒以上。
5. **国产化适配与系统迁移**：推动模型推理平台向国产硬件平台迁移，完成鲲鹏与飞腾芯片环境下的推理兼容性测试 40 项，优化内存与计算资源分配策略，实现部署成功率稳定在 98%以上，支撑长期业务稳定运行。
6. **系统调度与性能优化**：设计并实现分布式任务调度模块，支持任务分片与异步执行，服务日均处理新闻与社交数据 5000 万条；通过模型轻量化与异步缓存策略，整体系统资源利用率提升 25%，系统稳定性提升 30%。

2017.10-2020.12 **北京亿歆源科技有限公司 | Python 工程师 / 架构师**

1. **系统架构设计与业务建模**：主导公司进出口与电商交易平台架构设计工作，完成跨境贸易系统与供应链平台后端架构拆分方案 2 套，基于异步队列与分布式缓存搭建高可用方案，实现系统支持订单峰值并发量超每秒 500 条。
2. **数据处理与 NLP 任务开发**：负责商品、订单、清关记录等多源异构数据采集与清洗流程，月均处理数据量超过 3000 万条；开发实体识别模型并部署至生产环境，用于抽取产品参数、地址字段及税务条款，准确率达到 89%以上。
3. **模块解耦与微服务改造**：在系统重构中完成 8 个业务子模块的微服务化设计，使用消息队列完成服务通信与事件处理，接口调用成功率稳定在 99.8%；推动核心模块服务独立部署，支持业务快速迭代上线。
4. **全流程开发与团队协作管理**：带领 4 人开发小组完成 3 套业务系统从需求分析到技术选型、开发、上线全流程；主导制定接口规范、CI/CD 部署脚本与环境配置，缩短上线周期 20%，保障核心业务系统稳定运行。

2012.8-2017.7 **宇宙之讯科技有限公司 | 嵌入式/Python 开发工程师**

1. **高并发系统架构与协议实现**：设计并开发与三大运营商网关对接的短信协议接口 3 套，支持日均发送短彩信消息超 3 亿条，主导完成接口稳定性测试与异常重传机制设计，提升消息送达成功率至 99.7%。
2. **异步框架优化与吞吐能力提升**：基于 Twisted 异步事件驱动框架重构短彩信主系统核心模块，实现消息处理并发能力提升 3 倍以上，在多平台场景下完成单节点最大并发支持量达每秒 5 万条的优化改造，用于选票平台等实时业务。
3. **底层开发与跨平台应用移植**：参与嵌入式终端的系统定制与游戏类应用底层开发 6 项，完成代码移植至 3 个操作系统平台（包括 MTK、Android 原生、国产 RTOS），处理兼容性问题 20 余项，保障业务稳定运行。

2011.8-2012.7 **中国数码科技有限公司 / 中彩通 | 嵌入式开发工程师**

1. **通信协议开发与系统对接**：负责短彩信系统与三大运营商网关通信协议的设计与实现，完成接口对接方案 2 套，支持日均消息发

送量 8000 万条；协同测试团队完成压力验证 3 轮，确保系统在关键业务期间正常运行。

2. **底层开发与应用移植**：参与嵌入式设备上的音乐播放器与棋牌类游戏软件底层开发，完成应用在 3 种芯片平台上的适配与移植任务，处理内存管理与驱动兼容问题 10 项，提升跨平台应用上线效率与部署速度。

2010.8-2011.7 **泰圣思信息系统开发有限公司 | 嵌入式开发工程师**

1. **核心功能模块开发与协议实现**：参与中国移动 12530 音乐播放器、短信定制业务与动漫阅读器开发任务，完成音频播放、短信协议解析、图片渲染等模块设计与编码 10 项，处理接口对接问题 8 项，支撑平台日均活跃用户量 50 万以上。
2. **系统适配与性能优化**：完成产品在 4 种手机操作系统平台上的适配工作，针对内存占用、启动速度等指标优化音频处理与渲染逻辑，减少 CPU 资源占用率 20%，处理适配兼容问题 15 项，提升跨设备部署效率。

项目经历

2025.3-至今 **XXX 深度特征报告分析 | 算法工程师 / 架构师**

1. **多模态数据抽取与结构化转换**：负责 PDF 文档中的表格与图片内容解析，设计数据抽取流程 2 套，集成 OCR 与图表识别工具完成字段提取覆盖率提升 30%，支持后续结构化入库与分析任务对接使用。
2. **多 Agent 任务流与 Text2SQL 融合分析**：基于用户异构数据源构建多 Agent 任务链，配置数据清洗、分析指令、SQL 生成与修正模块，调用 Text2SQL 模型执行查询语句共计 1200 条，覆盖报告类分析任务类型 17 种。
3. **大模型选型与可视化联动开发**：对接并部署 XDeepSeek-32B、Qwen2.5-32B 与 Qwen2.5-Coder 等大模型，设计轻量化任务通过 Dify 平台任务流构建，完成 6 类分析结果与前端交互联动设计逻辑，开发图表展示逻辑并落地上线。

2025.1-至今 **AIP 一体化平台 | 算法工程师 / 架构师**

1. **平台架构设计与 Agent 集成**：基于 Taskweaver 构建多模块 Agent 任务流程，负责指令规划、数据调用、SQL 生成等 5 类子任务拆解逻辑，实现对接异构数据源 3 套，配置执行节点 12 个，支撑平台核心业务任务全流程处理。
2. **大模型选型与融合优化**：对接并部署 DeepSeek-32B、Qwen-32B、Uie-Lora-Qwen 等大模型，构建任务流调用策略，完成模型推理调度链 2 套；对模型执行效率进行评估与压测，调整权重与路由策略，减少平均响应时间 20%。
3. **智能化任务流与可视化联动**：使用 Dify 搭建业务流场景任务图 8 种，配置多跳决策路径与接口映射表，实现与前端交互模块联动的图表展示方案；对接 SQL 查询、摘要生成、文档比对等功能节点，完成前后端集成与上线发布。

2025.3-2025.5 **辰悉智能助手项目 | 算法高级工程师**

1. **智能体应用开发与 RAG 配置**：基于 MCP 工具完成 BI 对话、PDF 解析、企业知识库问答等智能体任务流程 4 类，对接 RAGFlow+Dify 知识库，构建知识图谱索引结构 3 套，完成提示词调试 30 轮，实现任务型智能体在 3 个业务场景上线。
2. **流程配置与数据治理执行**：设计智能体执行流程模板 5 种，配置上下文控制与插件路由规则 12 项，处理知识库数据治理问题 40 条，完成冗余文档清洗、文档重构及嵌入向量重新生成任务共计 5000 份。
3. **大模型部署与国产平台适配**：在 GPUStack 框架下完成 VLLM 大模型推理部署 2 次，分别在 A100 与 3090 GPU 环境中调试通过；完成昇腾 310I 与 910B 芯片下的模型编译、内存分配与接口映射调试任务 10 项，实现全链路适配并提交测试报告。

2024.10-2024.12 **XX 聚类主题领域事件分析项目 | 算法工程师**

1. **平台适配与推理优化**：负责事件聚类分析系统在麒麟操作系统与 ARM 架构下的部署与性能调优，设计轻量化文本向量化流程，集成 MNN、FastLLM 与 ONNX 三种推理框架，推理速度提升 40%，完成端侧推理能力测试与部署方案交付。
2. **算法优化与内存重构**：重构 HDBSCAN 算法中的内存管理模块，替换稠密结构为稀疏矩阵压缩策略，降低聚类过程峰值内存使用 35%，实现 5 万条/日事件数据的实时聚类分析，聚类结果用于主题发现与趋势建模。
3. **信息抽取与工程融合**：完成 UIE 模型蒸馏流程，融合正则引擎规则，部署在事件分析主线上，提取事件实体与要素字段 10 类，在 5 类样本集上测试准确率达到 85%，满足验收标准并通过项目交付评审。

2024.1-2024.9 **全球热点事件监控系统 | 算法工程师 / 架构师**

1. **多语言事件抽取与消歧建模**：主导中英文事件抽取模块设计与开发，英文采用 BLINK 自训练模型，中文通过大模型+UIE 微调策略，覆盖 5 类实体与事件类型，结合 Wiki 知识库构建跨语言实体消歧流程，事件抽取模块准确率达到 89.5%。
2. **分布式任务调度系统搭建**：设计并实现以 Celery、RabbitMQ、Kafka 为核心的分布式任务调度框架，打通数据爬取、清洗、翻译与分析链路，部署后日均处理新闻与社交数据约 120 万条，任务失败率控制在 0.3%以内。
3. **视频预警模块与多模态融合**：开发视频内容预警模块，集成关键帧提取、OCR 识别、ASR 音频分析与目标检测组件，完成预警模型触发逻辑 2 套，提升视频类敏感内容响应速度 40%，用于事件快速响应处理流程。
4. **模型部署优化与系统集成**：将多语言翻译模型迁移至 TensorRT 框架部署，结合 INT8 量化方案，推理延迟降低 35%；完成模型与 MongoDB、Milvus、NLP 模块的 Docker 容器化部署，集成入全链路处理系统并稳定运行。

教育经历

2018.3-2020.3 **中国人民大学 | 金融学 | 本科**

2012.3-2014.3 **北京大学 | 对外贸易 | 本科**

2006.9-2009.7 北京应用技术大学 | 电子信息工程 | 大专