

杨爽



出生年月：1992 年 4 月

职位：架构师

电话：+86 13520143313

邮箱：13520143313@163.com

住址：上海市闵行区

教育背景：本科

自我评价

11年Java开发经验，5年以上大型系统架构设计经验，3年以上AI产品架构设计经验
目前负责架构设计与研发的AI产品有多家世界500强的客户使用，包含家乐福、林德集团等

技术能力

深度参与产品需求评审，擅长从技术可行性、用户体验、长期维护成本等维度分析问题
具备优秀的逻辑思维能力，对咀嚼业界全新前沿技术，思考问题实践充满热情。
精通java，熟悉python编程
精通RAG设计以及调优策略
精通Milvus向量数据库的分布式部署以及调优
熟悉RAG+RAGAS构建知识库检索和评估机制
熟悉Dify、LangChain(UnstructuredLoader)、LangGraph(LLM,Chat,Agent,Chains)等大模型应用的工程框架
熟悉基于vllm进行大模型私有化部署
熟悉基于PEFT的常用微调手段(QLoRA、LoRA、IA3等)及微调优化(Accelerate多GPU微调)等
熟悉RASA开发框架,自定义开发实现韩语管道，支持韩文的会话
理解Transformer架构以及相关变体应用，深入理解LLama3大模型架构
精通JVM原理(内存模型、GC机制、性能调优)、多线程编程、集合框架、IO/NIO。
精通Redis(分布式锁、集群、持久化、缓存穿透/雪崩解决方案)
精通RabbitMQ, RocketMQ, Kafka等消息中间件，阅读过Kafka源码并基于拦截器做自定义处理
熟练使用Spring全家桶及Spring Cloud Alibaba(Spring Boot、Spring MVC、Spring Cloud、nacos、sentinel)
掌握分布式事务(如TCC、Saga、Seata)、分库分表(ShardingSphere)、分布式锁等解决方案。
精通MySQL(索引优化、事务隔离、锁机制、慢查询分析)，熟悉PostgreSQL等关系型数据库。
熟悉Linux系统及Shell脚本，掌握Docker、Kubernetes容器化技术。
熟悉CI/CD流程(Jenkins、GitLab CI)，了解Prometheus、Grafana等监控工具。
熟悉云原生部署与管理
Milvus开源社区贡献者

工作经历

2022.02 - 至今 上海微福思软件科技有限公司(原惠普软件部) 架构师

工作内容：担任架构师，主要负责系统的技术选型、架构设计、业务分析、平台搭建、性能优化等任务，构建公平、可解释性、真实性和稳健性、隐私与安全、安全、透明度、可控性、治理的AI应用

项目名称：IT Operations Aviator（生成式AI运维助手）

项目简介：基于LLM+RAG架构的智能运维对话系统，实现企业级IT运维问答自动化。

核心挑战与解决方案：

➤ RAG架构设计

完成5种主流向量数据库（Milvus/Weaviate/Pinecone/Qdrant/PGVector）的对比测试，最终选用Milvus实现毫秒级768维向量检索

索引优化：通过HNSW与FLAT索引的AB测试，在百万级数据量下实现召回率98%+延迟<50ms

分块策略：基于LangChain的文本分割算法，实现上下文连贯性提升40%

测试验证：基于RAG + RAGAS实现知识库的检索和评估

➤ 大模型工程化

模型部署：

初期采用AWS SageMaker+DJL部署Llama2-7B，生成延迟达10s+

升级为TGI推理框架，通过连续批处理（continuous batching）、语言特性(Rust vs Java)、KV Cache 优化等实现吞吐量提升60%，P99延迟降至7-8秒

推理优化:

系统化调优temperature(0.7→0.3)、top-k(50→20)、max_new_tokens(512→256)等参数，使回答相关性提升35%

实现动态参数配置体系，支持根据query类型自动调整生成策略

➤ 高性能架构

设计分级缓存体系：Caffeine本地缓存(命中率95%)+Redis分布式缓存,吞吐量提升5倍，同时缓解Redis带宽压力

➤ 构建CI/CD矩阵

建立30+ Pipeline, 涵盖部署, 变更测试, 自动化测试, 性能测试, 安全扫描, 发布等, 效率提升90%

一天之内完成质量测试并发版

➤ 安全扫描

集成 Trivy / Fortify 实现自动化漏洞扫描

➤ 可观测性体系

基于OpenTelemetry + fluent bit+ Open Search+ Prometheus + Grafana 构建可观测系统, 保障系统 SLA 99.95% 以上

项目名称：RAG系统优化

项目简介：重构企业级RAG框架，支持跨源异构数据融合与智能检索增强。

技术成果：

➤ 数据工程：

基于IDOL web connector实现支持Share point、Confluence、在线文档等多渠道数据的导入

构建Kafka流式处理管道，实现PDF/PPT/HTML/Text等多格式文档的异步解析

语义相似度切块算法，基于spaCy、Numpy、向量相似度动态调整切块，上下文完整度提升55%

➤ 检索增强：

实现混合检索策略：BM25+稠密检索+语义检索的多路召回，召回率达92%

采用FlashRank进行结果重排序，文档相关性0.68提升至0.83

➤ 系统设计：

设计可视化调试平台：支持实时修改分块策略/检索参数并预览效果，调试效率提升70%

实现最终一致性保障：通过客户端定时调度任务，解决数据不一致问题

技术创新:探索graph RAG的生产落地

项目名称：大模型部署优化

项目简介：使用vLLM + AWS EKS替代AWS SageMaker+TGI部署Llama3-8B

技术成果：

➤ 提升吞吐量：提升6倍

减少部署成本：减少25%

➤ 丰富观测性：基于vLLM提供的几十个metrics指标构建更详细的大模型运行监测,例如TTFT，TPOT等

项目名称：大模型微调及推理优化

项目简介：构建多agent及微调大模型垂直领域知识

技术成果：

➤ 多版本智能体 (Agents) 设计：提供配置页面，可以配置不同的Prompt，不同的RAG策略，满足情感分析，文档生成，文档总结，风险评估，方案推荐，异常检测等使用案例

➤ 微调方式：PEFT常用算法，QLoRA、LoRA、IA3、Prompt-Tuning、Prefix-Tuning、BitFit等

➤ 微调优化：采用Accelerate进行多GPU微调，4*24 GPU微调Llama3大模型

技术创新：

- **LangGraph**: 基于LangGraph实现Agentic AI应用
- **MCP**: Model Context Protocol
- **A2A**: Agent2Agent Protocol

项目名称: 虚拟客服平台

项目简介: 平台基于RASA和botkit构建, 根据客户输入信息自动识别意图并匹配相对应的执行操作。减少60% IT支持工作量。

工作职责:

- 主导业务需求到技术方案的转化
- 主导基于RASA的对话系统架构设计
- 主导LLM (Llama) 与RASA的混合架构设计
- 构建高效的CI/CD

工作成果:

- 动态按需加载NLU模型, 减少70%内存需求
- 自定义实现韩语管道, 实现韩语的高精度匹配
- 主导意图识别准确率 (目标≥92%) 和对话完成率 (目标≥85%) 的持续优化
- 自动化测试覆盖率≥80% (包括单元测试、对话流测试)

- 技术勋章**
- ❖ 2023年最佳优秀员工奖
 - ❖ 2024年最佳优秀员工奖

2017.04 - 2022.02 泛微网络科技股份有限公司 高级软件工程师


工作内容: 为提升企业流程审批效率及适应复杂业务场景, 对泛微OA系统的核心流程引擎进行功能扩展、性能优化及二次开发, 支持包括财务报销、采购审批、人事调动等企业核心业务流程的自动化处理。


工作职责: 流程引擎设计与优化、集成开发与系统对接、复杂业务场景解决方案、技术规范

2016.06 - 2017.03	北京久其软件股份有限公司	高级软件开发工程师
2015.04 - 2016.05	亚联荣博信科技有限公司	中间业务组组长
2014.08 - 2015.04	高伟达软件股份有限公司	开发工程师

 **教育背景**

2010.09-2014.06	鞍山师范学院	信息管理与信息系统 本科
-----------------	--------	----------------

 **资质证书**

 SAFe® 5 Practitioner