马晓晖

电话 13500038851 mason_joe
joey.gq joejoey.ma@gmail.com

技能

- Web开发： React、ReactNative(Expo)、Capacitor、Electron、TypeScript、Tailwind / CSS-in-JS、Redux、Zustand、Zod、Tanstack Query
  后端: Node.js、Next.js、Nestjs封装CRUD RESTful API、GraphQL、WebSocket，Middleware、Serverless Vercel、Cloudflare, AWS Lambda/ECS冷启动、Edge分布式微服务，API网关、Redis、Kafka、gRPC
- Web框架：Medusajs、Better Auth、GSAP Motion animations、Aceternity /shadcn UI AntD ，熟悉构建工具Webpack 、Vite、Turborepo、Monorepo、pnpm workspaces项目管理性能优化的实现例如代码优化、打包优化、资源优化，能结合实际业务场景进行优化
- DevOps：Prisma、Drizzle ORM、PostgreSQL/D1, R2/S3,并使用Docker镜像容器部署,Kubernetes集群，CDN缓存、负载均衡，Sentry.io事件抓捕、Elastic日志Kibana监控， CI/CD / MLOps、自动化Jest测试、灰度发布、A/B，熟悉 Azure 或其他云计算平台（如 AWS 、Google Cloud 、Vultr 、阿里云等）

工作

HKBN 香港宽频 Hong Kong Broadband Network（2023.8- NOW ）

AI ITBS Senior software Engineer

- 从0到1的基建部署上线AI Portal赋权BS-CS、API-Dev-向量知识库体系的AI驱动工具链。

广东力进物流股份有限公司（2021.2-2023.7）

工程化 全栈架构组 信息主管

- 三端（Web,APP,桌面端）ERP系统(框架重构)数据流并发高，跨部门协作集成，容器部署容灾挑战。

广州新中英跨境有限公司（2020.2-2021.2）

DevOps 前端开发

- 基于业务跨境电商平台研发,埋点，关联分析，风控数据运营，优化落地上线。

项目

AI Portal业务数据统筹Prompt

Next.js RAG Pipeline Data Lake Kubeflow PyTorch Airflow GSAP

- 从 0 到 1 企业级 AI 中台主导 **LLMOps** 、 **Microservices** 接入检索系统、数据管道、成本与 SLA 通过 API Gateway 聚合多模型与多检索链路 Prompt、检索命中策略、Token 成本、Re-rank 、延迟与模型参数**以数据驱动决策集成 Langfuse** 实现用户反馈并推动**全链路观测异步化**，Query / Vector / Prompt 多级缓存体系，引入 **A/B 测试与灰度发布机制迭代**具备可回溯性与可量化结果。 **整体吞吐能力提升 3 倍**、**P95 延迟下降 60%+**，同时将 **单请求 LLM Token 成本压缩 40%** 。
- 构建 **Data Lake + RAG Pipeline** 并由 **Kubeflow + Airflow** 编排实现多业务以文档**TB级**（25,000+ 份工单（表格）、手册（图文混合ppt/pdf）、手写/扫描案例和长合同）语音（销售电话客服记录，访谈例会总结）的自动化接入持续更新 **pgvector + OpenSearch** 分离存储并结合 **HNSW + BM25 + RRF Hybrid Retrieval 推动检索体系从“可用”升级为“可验证”**，**Top-K 命中率提升50%，上下文噪声降低 30%+**，并降低接入成本，覆盖运营、风控、数据分析与企业知识检索等基于 **1M Menu Agent 多 Agent 架构与可配置规则 40+ 核心业务场景** 直接将 AI 回答从反馈 → Q&A 循环以持续改进 **提升为>95%的 recall@k并附带证据链接响应一致性** 。使人工平均查询时间 **从10分钟至<30秒** 与重复沟通成本 **整体下降 60%+**，显著提升跨团队协作效率
- 平安钟**视频监控视觉 AI 自建**数据标注、关键帧抽取、**LoRA 微调与模型蒸馏**的训练流水线并推动模型端侧化部署，在真实场景中实现 **误报率降低约 45%（基于 30 天生产监控）** 检出率稳定标注评估集**上达到 92% 准确率**
- AI Portal 前端整体按Figma还原及UX动画参与以CSS-JS、动画GSAP完整交付用户体验与性能目标 **Next.js App Router** 设计多会话并发、流式响应与多模态（对话/绘画/视频生成）输出高复杂度交互系统，通过 **PPR / SSR、组件级 Cache与 Compiler** 优化在持续叠加 AI 功能的情况下保持 **首屏 FCP 秒级稳定、交互延迟降低 50%+**，并系统性解决 SSR Hydration 。

ERP/业务财务一体化/进销存物流管理SAAS

WebView PWA Service Worker monorepo Redux ant-design Apollo Cluster

- React + Nest.js + Capacitor + Electron 构建的企业级物流管理系统，Monorepo + 微服务**三端部署响应式适配**服务于省级/区域/分拨中心、手持终端、司机智控等解决**千万级吨数**运输业务动态调度订单性能痛点。
- 重构redux状态管理引入以请求缓存重试并发控制实现了实时**同步+Optimistic更新+支持弱网离线缓存**解决抖动中断**业务连续性提高至80%**，IndexedDB并封**装分页/排序/虚拟滚动、表单验证**等加入**事件速率优化**,解决初始渲染**从17秒降低至3秒**。
- 实现通信车辆,货车地图导航可视即时人工同步**秒级调整**、解析Excel、CSV、回单付水印章电子面单等组件结合业务逻辑实现从**路由成本控制审批流转**使**对账周期缩短**天 → 小时，**四级地址结算体系，末端落地配派**拼单抢单、网约车**智能合单**使空驶低装载率单票成本等**以可用绩效指标展现使运输效率整体提升两倍**。
- BFF层微服务分析聚合API并实施基于Redis Lua的**分级限流与熔断机制**，DataLoader**解决了 GraphQL N+1** 查询难题包含**DB Query数响应时间**显著下降、**JWT令牌动态判断权限、五层权限模型**，支持**1000+并发连接**，处理**600+车辆实时位置**推送**延迟<500ms**。
- 构建Kafka数据管道+Kubernetes**容器化部署**CI/CD自动化Jest/e2e**集成测试与持续交付**、Sentry**性能监控**、WebSocket**轮询卡顿心跳重连机制**，支持40+Topic**集群扩展系统**高可用架构。

基于node的独立站电商平台集POS线下一体购物系统

Express.js ORM react-i18next Stripe ChatWoot

- 全栈node线上线下一体**购物(POS)系统开发**，**RESTful API 和 MVC架构 ORM封装CRUD接**口结合react-scanner**扫码入库结账**,Express.js定制Stripe**支付网关**回调WebHook路由转跳。驱动层实现扫码枪小票打印机的**硬件级交互打通O2O交易闭环**。
- 搭建覆盖率超**90% 的用户行为追踪系统**，sendBeacon页面关闭时的数据发送，基于采集的各类指标及转化漏斗数据针对性优化LCP/CLS/PU/UV等指标给予运营决策。
- i18next实现多语言资源按需分块加载并定制开发基于ChatWoot的智能工单系统实现了**7x24小时自动化客服响应**转化率相比**提升28%**。

教育

华南理工大学-本科-计算机科学与技术-学士学位证书

全国英语等级考试 PETS-3证书 口语5/5