洪思栋



手机号码：13434767116 |**邮箱：**tremann023@gmail.com | **意向：AI产品经理**

作品集 ： https://www.yuque.com/yuqueyonghusboarc/ca7q13/ql3hvwv5becobd2v?singleDoc

GitHub：https://github.com/GemHsid/Gym-RAG-System

教育背景

广州工商学院（本科） **软件工程 AI工程师学院** **2022.09 -- 2026.07**

主修课程：软件设计模式，软件工程导论，Python程序设计，机器学习基础，人工智能，数据结构与算法。

实习经历

模力方舟（Gitee AI） **AI 产品经理** **2025.10 -- 2026.06**

项目1：平台文档与国产算力应用认证

背景： 原平台文档结构不完整、内容口语化；缺少官方平台与社区主流模型的应用 showcase；开发者对算力、模型与资源消耗的理解门槛高，影响算力与 MaaS 业务的用户转化。

结果： **支撑 500+ 用户完成模型工作流实践，新增 3500+ 算力订单、7400+ 卡时租赁**；认证模块**累计答题 1000+ 次，120+ 用户通过考核并申请技能证书**。

职责：

- **重构平台文档**：优化文档对象与内容的结构，覆盖了平台产品功能、供应商与社区生态。规范功能的理论说明，精简操作指南。
- **实操教程体系**：产出LLM、Image、OCR、ASR、TTS、RAG (Embedding、Rerank) 的模型部署SOP，覆盖主流模态，最大化的满足用户需求、上手难度，加速平台用户的转化速度。
- **国产 GPU 技能认证**：
  - **从 0 到 1** 负责壁仞的 L1 技能认证，**主导设计方案，深度对接算力供应商**，解决认证方案中出现的适配性卡点，快速推动项目上线。
  - 深度参与沐曦、燧原、天数智芯的 L1 技能认证的**上线测试；担任平台社群的技术支持**，协助用户解决技能认证中的难点。

项目2：算力容器公网与后台的公网风控管理

背景： 算力容器服务无法部署到私有项目；公网能力商业化需要平衡开放性与通道暴露风险。

结果： **客诉从 20+ 降至 0**；公网能力从无法安全开放到可开放、可管控、可运营。

职责：

- **差异化开放策略：开放节奏与客户价值、风险等级匹配。**划分平台用户群体，通过申请表功能，包周期用户需提交可量化数据，如：业务场景、端口数量、服务时长等，申请并经飞书机器人同步至算力市场群人工审核。算力专属部署客户直接开放功能。
- **主导公网设计方案：**
  - **用户端：用户秒级拿到公网地址，平台在开放前就完成风控。**入口设实名认证与实例状态拦截，风险实例直接冻结；端口配置即时校验、重启前二次确认，关机自动切断通道。
  - **后台：每个公网通道全程监控。**统一面板收录全部公网域名，规则量化与模型辅助持续检测，高风险自动封禁、处置全程留痕。

项目3：客诉汇总与分析 Agent

背景： 保险行业客诉信息来源分散，存在采集分散、分析口径不统一、跨渠道汇总效率低的问题，导致客诉洞察产出慢、复用弱，也难以为产品、运营和管理决策提供稳定支持。

结果： 试点期间单份报告撰写时长从人工 120 分钟压缩至 8 分钟（**缩短 93.3%**），**累计自动生成分析报告 30 份**。

职责：

- **口径与任务拆解**：拆解采集、分析、分发三类任务，基于重疾险公开投诉数据定义公司、时间、产品、投诉类型与报告结构口径，保证结果可比。
- **链路编排**：基于 PocketClaw 编排**投诉采集、文档分析与飞书分发**能力，完成从**投诉收集、结构化分析到报告生成与分发**的链路设计。

项目经历

24小时健身房 AI 客服系统 **独立负责人** **2026.02 – 2026.04**

背景： 24 小时健身房，夜间无人值守，存在服务真空期、前台响应不稳定，会员在营业时间、会员卡价格、退款规则、器械用法等高频且标准化咨询上无法获得及时的解答。

结果： 相较基础 RAG 方案，**召回率提升 22%、忠实度提升 30%、回答相关度提升 6%**（RAGAS 离线评估，四层指标体系 + 固定评测集驱动）

职责：

- **知识库预处理**：整理所有业务线的现有文档结构，协同管理员补齐缺失的业务文档，采集用户的高频问题。
- **RAG 问答搭建**：提升长文档场景下的召回覆盖与上下文完整性。基于 LlamaIndex ，完成数据清洗、层级切块、检索优化 、层级回溯与精排。
- **效果评估与兜底**：针对低置信度、多轮未解决和高风险对话设置拦截与转人工策略，转接前结构化沉淀用户意图、历史对话和已执行动作。
- **性能与算力成本：已在算力专属部署用户场景完成验证。**显存测算容量；对比精度量化前后的 RAGAS评分 与并发吞吐，锁定成本-质量平衡档；评估夜间无人值守的并发、延迟与算力成本。

技能与证书

- **技能工具**：Office、Figma、Axure、XMind、Linux、Git、Docker、Python、Codex、ComfyUI、N8N、Transformers、Diffusers、 vLLM、Prompt Engineering、RAG、模型本地部署与测试、熟练使用海内外 GPU。
- **证书**：沐曦 AI 技能认证、燧原 AI 技能认证、CET-4

自我评价

- **AI 应用与边界：具备丰富的AI 应用经验。**理解AI 在业务流程中的使用场景与局限，曾在公网的请求路径检测结合大模型完成检测方案的落地。
- **团队沟通协作：具备良好的沟通协作和项目推动能力。**在算力平台推动公网风控从策略到上线，打通**产品、研发、运营与客服**；与燧原、天数智芯等**芯片厂商深度对接**认证适配，解决上线卡点；长期在社区承接开发者问题，把外部反馈带回产品。
- **数据指标与迭代：具备较强的数据监控与量化的意识。** 在算力公网的大模型自动化中，以申请表的数据采集，完成自动化的正确性的强化与校准、数据闭环。