

# 苗静思

Agent/全栈 工程师

## About

📍 广州 CN

✉️ miaojsi@outlook.com

📞 +86 175 1209 0401

👤 Alenryuichi

## AI Agent & RAG

### Multi-Query RAG

Prompt Chaining CBR

RAGAS Langfuse

Self-Consistency

## AI 编程工具链

MCP 协议

Agentic Coding

Agent 自闭环

Cursor/Augment/Claude

Multi-Agent Spec

Swarm

## 编程语言

Go Swift

Python/PySpark

TypeScript C++/hikit

SQL/Gremlin

## 移动端 & CV

SwiftUI SwiftData

CoreML MobileNet V3

Detectron2 MediaPipe

## 后端 & 数据

Go/Gin GORM MySQL

Redis ClickHouse

图数据库 Hbase TDW

pkv

## DevOps & 监控

Docker Compose Nginx

Prometheus Grafana

Loki

腾讯云/阿里

云/Supabase/Vercel

## Interests

### 抱石

V3

磕线 App 独立开发者

### 健身

2023 第七届环球国际超级联赛

C 组第 4

## Summary

专注 Agent 与 RAG 系统设计，具备从 0 到 1 架构设计与落地能力。独立负责风控审核 AI Agent V1→V4 架构演进（准确率 70%→90%，日均 300+ 单自动结单），设计 Prompt Chaining + 策略短路机制，落地大规模反作弊系统（年化收益近 4000 万）。独立完成 iOS 应用全流程（产品/UI/前后端/AI 模型/运营），实践 Agentic Coding 与 VideoMind Agent 架构。

## Experience

### 全栈开发工程师

07/2022 - Present

负责 AI Agent 审核系统、10 亿级策略分析平台与反作弊治理系统的架构设计与落地，涵盖 RAG/图数据库/分布式系统

### Open Source Projects

#### OpenMemory Plus - AI Agent 双层记忆框架 · <https://github.com/Alenryuichi/openmemory-plus>

02/2026 - Present

为 AI Agent 设计的双层记忆管理系统，解决跨会话上下文丢失问题

MCP AI Agent Memory System Qdrant E2E Testing Multi-Agent

- 设计双层记忆架构，通过分离项目级（Git 版本控制）与用户级（MCP 跨项目）存储，实现跨会话上下文保持率 95%+
- 实现 ROT 智能过滤（Redundant/Obsolete/Trivial），通过 Qdrant 向量搜索（相似度阈值 0.85）+ Ebbinghaus 衰减模型，减少 40% 冗余记忆
- 构建 Agent 自闭环测试体系，通过分层测试策略（单元测试 + DiD 沙盒 + 完整 E2E）实现 15 项端到端测试全自动化验证
- 设计统一\_omp/ 目录结构与 YAML Schema，支持 4+ IDE Agent（Augment/Cursor/Windsurf/Cline）无缝共享项目记忆

### 磕线 (Betoline) - 独立开发全流程

09/2025 - Present

从 0 到 1 独立完成的攀岩 iOS 应用，涵盖产品设计、UI/UX、前后端开发、AI 模型训练、DevOps 与运营全流程

VideoMind Agent MobileNet V3 CoreML Go/Gin Docker Compose MCP Server Prometheus SwiftUI

- 独立完成产品/UI/前后端/AI/DevOps/运营全流程，前端 SwiftUI + SwiftData + Combine，后端 Go/Gin + GORM + MySQL + Redis，10 个容器微服务编排
- iOS 端实现攀岩训练（实时计时 + 阶段状态机）、数据分析（训练负荷 ACWR + 难度分布图表）、攀岩日记（AI 视频剪辑与识别 + 磕线项目追踪）三大核心模块
- 设计 VideoMind Chain-of-Roles 工作流（4000+ 行，6 角色），批量推理减少 80% API 调用，成本 -39%，准确率 92%；4 模型 VLM 统一接口 + 5 步 CoT Prompt，成本 ¥0.01/视频
- 训练 YOLO11m 岩点检测（mAP50 59%）+ MobileNet V3 分类器，VLM 伪标签自动化标注；CoreML 导出实现 300x 加速（30-60s→100ms），三层递进颜色识别准确率 88%
- 构建完整后端与运维体系：Go/Gin 三层架构 + JWT 双令牌轮换 + 8 层中间件 + 腾讯云全家桶（COS/CDN/MPS/APNs）；Docker Compose 编排 10 服务 + Prometheus/Grafana/Loki 监控 + acme.sh 泛域名证书自动续期；自研 MCP Server（25 个运维工具）实现 AI Agent 远程服务器控制

### 风控审核大模型 AI Agent

12/2025 - Present

独立设计开发的智能审核系统（V1→V4 四版本迭代），支撑企业微信日均 300+ 申诉自动处理

Prompt Chaining CBR Multi-Query RAG 策略短路 RAGAS 分布式异步 Go Langfuse

- 独立完成 V1→V4 架构演进：V1 素朴 RAG → V2 创新性引入 CBR 解决规则缺失 → V3 四步 Prompt Chaining 解耦提取/评估/决策/分类（Token ↓88%，准确率 70%→90%+）→ V4 Multi-Query + 策略短路
- 设计 V4 策略短路机制（规则 + LLM 协同），解决模型处理条件判断的幻觉问题，人脸解封场景 AI 识别率从 5-10% 提升至 94.89%，可复用架构可快速推广至其他场景
- 设计 V4 Multi-Query 并行检索策略（原始查询 + 核心提取 + 同义扩展），CBR 回调率提升 20%，整体覆盖率达 50-60%
- 构建 Langfuse 深度整合工具链（Trace/Session/Prompt 管理 + RAGAS 自动评估 + Badcase 七大失败点定位），实现质量监控闭环
- 实现分布式高可用审核系统（DB 原子抢占 + 30s 补偿循环），支撑 2000+ 并发，落地日均 300+ 单自动结单，节省 2+ 人力/天

### 策略全生命周期管理平台

06/2022 - 06/2025

整合画像中台、图能力、策略质量保障、特征系统的一站式策略管理平台，支撑反垃圾安全治理

ClickHouse 图数据库 策略质量 分布式优化 画像分析 OLAP/OLTP

- 架构设计 10 亿级画像中台（OLAP ClickHouse + OLTP TDW/MySQL），支持 200+ 标签 5 分钟复杂分析，刷单欺诈 ↓30%，贷款投诉 ↓23%
- 构建图数据库（20+ 实体，70+ 关系），支持 4 种扩散 + 4 种溯源分析，完成 17 项业务入图
- 主导策略质量保障平台（自动挖掘→分级→闭环），挖掘 228 个低质量策略，Review 时长 5 天→2 天，误封率 2.4%→0%
- 修复 PKV 分布式序列化缺陷，写入性能 90k/s→1200w/s（130 倍提升），解决百亿级数据写入瓶颈
- 设计客户端大包加速方案，特征拉取耗时减少 500-1000 倍（20ms 内），KM 文章热度 Top 5

### 自建应用反作弊治理系统

01/2023 - 06/2024

大规模智能检测与自动化治理系统，覆盖 311 万活跃用户

图算法 域名扩散 IP 聚类 社区发现 TDW/Spark

- 识别违规服务商 820 家、企业 5.2 万家，通过设计 10+ 维度检测特征（域名扩散、IP 聚类、社区算法）处理亿级实体数据
- 实现年化收益近 4000 万，通过多维度算法挖掘（域名 178 万/月 + IP 38 万/月 + Path 关键词 30 万/月）
- 将人工干预减少 80%，通过构建 30+ 模块端到端自动化决策系统（识别→灰度→迁移→反馈闭环）
- 实现零投诉运营，通过建立服务商违规条款并妥善处理百余家企业反馈

## Education

### 华南理工大学

09/2019 - 06/2022

硕士 计算机科学与技术

## Publications

### UAV Visual Navigation System based on Digital Twin

12/2022

IEEE 18th International Conference on Mobility, Sensing and Networking (MSN 2022)

CCF-C | 提出基于数字孪生的无人机视觉导航框架，集成深度神经网络实现物理空间与虚拟空间的实时映射