苏哲

AI 应用开发工程师 / Agent 架构

电话：15367885946 · 邮箱：suzhe52@foxmail.com · GitHub：github.com/CodeAlex52

专业技能

Agent 架构与工作流编排：熟悉 Agent Runtime、Agent Loop、Function Calling 与 Tool Registry；做过 LangGraph 状态机、Checkpointer 状态持久化和 Human-in-the-Loop，能围绕工具权限、上下文预算和终止条件设计**确定性边界**。

检索增强生成 (RAG) 与向量检索： 做过复杂文档切片、Metadata 过滤与 RAG 检索链路；熟悉 HNSW 索引稠密检索、BM25 稀疏召回、混合检索、Cross-Encoder 重排及 pgvector 向量量化，能围绕召回质量、上下文噪声与延迟做**工程取舍**。

AI 基础设施 (Infra) 与质量安全护栏： 参与 LLM 统一网关建设，落地 API Key 鉴权、Trace ID 链路追踪与按项目 Token 配额治理，以逻辑模型名屏蔽底层供应商差异；熟悉 MCP 协议、Guardrails 护栏与模型 Fallback 降级，**兼顾成本可控与链路可观测**。

大模型微调与偏好对齐 (SFT+DPO)： 熟练使用 LLaMA-Factory 完成开源基座模型业务微调（LoRA / QLoRA / SFT）；基于真实交互语料执行 DPO 偏好对齐；结合 Pydantic Schema 约束结构化输出，搭建 **LLM-as-a-Judge 自动化评测**与数据闭环。

实时语音交互 (RTC) 与多模态工程： 熟悉 Audio In → RTC VAD → 流式 ASR → LLM → TTS → Audio Out 的低延迟语音链路；落地火山引擎 RTC 接入、CustomLLM 回调、Barge-in 语音打断与多模态检索，**免检索链路端到端延迟压至 800ms 内**。

大模型语料治理与数据闭环 (LLM Data Curation)： 搭建面向大模型微调与 RAG 的自动化语料治理管线，覆盖跨轮次去噪、口语冗余过滤与多轮会话结构化规整；落地“正则 + 轻量 NER”双层 PII 隐私脱敏，支持 ShareGPT / Alpaca / OpenAI JSONL 导出，**保障微调数据质量与隐私合规**。

工程基建、全栈研发与开源贡献： 熟悉 Python、FastAPI 与 asyncio 异步服务开发；使用 PostgreSQL / pgvector、Docker、Serverless（SAE / FC）、structlog 结构化日志与 Trace 完成项目落地；向 NVIDIA、Google、Dify、字节 DeerFlow 等 **主流 Agent 开源仓库贡献并合入 PR**。

项目经历

业务闭环 **业务应用层** 核心业务协同矩阵 社媒获客(03) · 私域导购(01) · 语音售后(02)

01 基于 Harness 的跨境高客单自进化导购 Agent

核心开发 / 导购 Agent 与策略闭环 Owner　　2023.10 — 2026.10

项目描述： 面向红光面罩、光疗腰带、光疗垫等海外高客单 PBM 光疗产品，历经三阶段架构演进：一阶段 RAG 被动问答 → 二阶段意图识别 + 状态机工作流促单 → 三阶段将价格 / 库存 / 禁忌下沉为原子 Tool，构建基于 Harness Agent Loop 的自主决策与话术策略自进化导购 Agent。

关键结果 KEY RESULTS	提升 28.6%	缩短 40%	40+ 收敛至 5
	咨询意向成单转化率	高客单成交决策周期	确定性 Tool 化与状态机瘦身

阶段一　RAG 知识库与被动问答基线

01 **知识库语义切片与元数据过滤** / 采用**按章节标题 + 自然段边界的语义切片**，规避定长切片的语义截断；构建 **Metadata 标签过滤**，按商品品类、规格与适用人群前置硬过滤，**从源头抑制无关召回与语义碎片化，提升检索精准度与召回率**。

02 **HNSW 混合检索与两阶段精准召回** / 基于 **HNSW 索引的稠密检索 + BM25 稀疏检索**做混合召回，按术语密度动态调节稠密 / 稀疏权重；召回后引入 **Cross-Encoder 重排模型**二次精排，**专有名词与长尾参数命中率显著提升**。

03 **提示词版本管理与被动问答瓶颈** / 对提示词做版本化管理，配多轮评测集自动评分与负样本重构；但纯 RAG 只能问什么答什么，多轮对话缺少需求追问、选型引导与促成动作，会话难以推进到成交，**由此推动向二阶段演进**。

阶段二　意图路由与工作流促单

04 **前置多意图识别与分流引擎** / 针对买家咨询行为构建多意图识别与前置路由，识别价格敏感、功效疑问、使用禁忌、竞品比价与售后报障等核心意图；**基于意图分类做前置路由分流，打通商品库与优惠券规则引擎**。

05 **状态机工作流编排导购推单流程** / 基于状态机工作流显式编排「探寻需求→筛查禁忌→产品定制推荐→阶梯发券→主动促单」销售路径；**以节点状态约束模型按标准销售 SOP 推进会话，将被动应答转为有目标的主动引导**。

06 **状态机分支膨胀与规则维护瓶颈** / 高客单多轮对话中插话、回跳与比价频繁，状态机分支随「体质禁忌 × 优惠叠加 × 库存状态」组合持续膨胀至 40+ 节点，偏离主流程时需依赖硬编码兜底分支修正路径，单条规则变更牵连全局；**复杂流程导致维护成本陡增且缺乏动态交互弹性，由此推动向三阶段自主决策演进**。

阶段三 **Harness Agent Loop 与策略自进化**

07 **统一 Agent Loop 与销售工具标准化解耦** / 应用层基于统一 Harness 抽象编程，以 **LLMProvider + Tool 双抽象与注册表**解耦导购业务逻辑与底层实现：新增模型只补一个 chat() 适配器，新增导购能力只登记一个 Tool，决策循环与业务代码无需改动；配置驱动在离线 Mock 与真实模型间切换，**tool_call / tool_result 与 Function Calling 协议对齐**。

08 **收敛 Loop Engine 自主决策确定性边界** / 构建「推理→选工具→调用→观察→再决策→终止」闭环，将 40+ 节点分支收敛为商品选型、价格、库存、禁忌、优惠 5 个确定性 Tool，意图识别隐式融入工具选择，省去规则分类器；单步多工具并行，**异常转 Observation 交回模型自纠，内置最大步数、超时、低置信转人工三层兜底**。

09 **业务话术自沉淀与案例库分层治理** / 业务主管给出「问题 + 标准话术 + 处置 SOP」做一次对话式教学，LLM 自动归纳话术原则注入提示词，审核发布后即时生效，**一线经验无需排期开发即可沉淀上线**；服务时向量 Top-K 召回相似案例复用经验，Prompt 按「红线→话术→案例」分层治理。

10 **上下文工程：信息预算与分层降载** / 上下文按系统提示、长期记忆、召回片段、工具返回四层分配 Token 配额，超限时按「摘要折叠 → 原文外存留引用 → 剔除低相关召回」三级降载；剥离工具噪声、会话档案分层压缩，关键事实外部存储双写兜底；**抑制 Context Rot，单 Agent 即达标，不引入子 Agent 以控制调度复杂度与 Token 成本**。

02 实时语音与多模态售后排障 Agent（RTC）

核心开发 / 实时语音与检索链路 Owner　　2024.08 — 2025.03

项目描述： 家用高客单 PBM 产品售后排障场景复杂、海外大件退货物流昂贵；项目融合火山引擎 RTC、豆包大模型、CustomLLM 回调桥接与 pgvector 单表多模态检索，打造低延迟实时语音售后排障链路。

关键结果 KEY RESULTS			
	压降至 800ms 内 免检索链路端到端延迟	**54% → 88%** 长尾故障精准召回率	**降低 34%** 非质量类大件售后退货率

01 **火山 RTC 与 CustomLLM 回调桥接** / 集成**火山引擎 RTC 与 CustomLLM 回调服务**，串联流式 ASR → LLM → TTS 低延迟实时语音管线；传输层复用 RTC SDK 的**弱网抗丢包与自适应抖动缓冲**，**弱网环境下保持语音交互连续**。

02 **Barge-in 语音打断与 VAD 静音检测** / RTC 云端 VAD 捕捉到用户开口瞬间，前端立即清空本地音频播放缓冲并停播，同时经 DataChannel 下发截断信令，中止后端 LLM / TTS 输出流；**用户插话即停播，语音交互无抢话卡顿**。

03 **多模态素材中心单表存算一体** / 基于 **PostgreSQL + pgvector** 构建 1024 维 HNSW 索引的统一底座 **rag_assets**，媒体资产与检索文本解耦单表建模，仅对 retrieval_text 建向量索引；一条 SQL 在同一事务内完成属性过滤与相似度打分，**免去跨库双写与同步链路，检索结果与业务状态强一致**。

04 **Int8 标量量化与两阶段重排** / 结合 **Int8 标量量化**将向量存储开销降低 75%；依托**“Int8 粗召 Top-20 + Float32 Cross-Encoder 重排 Top-5”**两阶段检索链路，对故障术语提高稀疏检索权重，**长尾故障精准召回率由 54% 提升至 88%，检索速度提升 4 倍**。

03 AIGC 社媒内容生产与运营 Agent

核心开发 / 内容工作流 Owner　　2024.01 — 2025.06

项目描述： 面向海外私域与社媒运营，将选题策划、脚本撰写、多语种本地化、配图生成与多平台分发拆解为可恢复的 LangGraph 内容工作流；通过 SSE、异步任务与模型分级调用提升内容产能和生成稳定性。

关键结果 KEY RESULTS			
	提升 7.5 倍 月度营销内容矩阵生成产能	**缩减 70%** 批量配图生成耗时	**降低 40%** 多模型分级调用成本

01 **LangGraph 多阶段内容生产工作流** / 基于 **LangGraph** 构建选题、写作、配图与本地化多阶段工作流，以 **StateGraph / SubGraph** 将内容生产拆为独立编排单元；通过 **AsyncPostgresSaver** 持久化状态，支持节点级回滚、重试与人工审核断点，**将长流程内容生产收敛为可恢复、可审核的 Agent 工作流**。

02 **SSE 流式输出与多模型配图异步化** / 将 **astream_events** 与前端 **ReadableStream** 对接，捕获 Token 级流式输出；批量配图通过**豆包 Seedream + nano banana** 多模型异步调用，结合 **asyncio.gather + run_in_executor** 并行执行，并引入指数退避重试应对模型抖动，**缩短感知首字延迟并提升批量生成吞吐**。

03 **MetricsContext 与多模型分级调用** / 通过 **MetricsContext** 采集各节点 Token 消耗、推理延迟与错误率，结合模型定价做成本归因；按任务复杂度分别调用轻量模型与旗舰模型，并以 **Pydantic Schema** 约束结构化输出，**在内容质量与推理成本之间做分级取舍**。

04 **LangSmith、Mock 与工程化保障** / 接入 **LangSmith + structlog + request_id** 完成 Graph 级链路追踪，配置 LLM / Image / Checkpointer 的 Mock 开关解耦外部依赖；结合健康检查与优雅关闭保障异步内容任务稳定运行，**将问题定位从小时级压缩到分钟级**。

数据飞轮 语料微调与数据中台

语料资产治理 · 模型自进化 | 数据治理中台(04) · 语料风格微调(05)

04 全链路 AI 语料数据中台

数据治理链路 Owner

2023.08 — 2024.10

项目描述： 面向跨境独立站、私域社媒与售后工单，构建从原始数据接入、会话结构化、质量清洗与脱敏、人工审核到 RAG / 微调 / 评测多下游供数的数据中台；以统一数据源驱动导购、售后与模型迭代，解决多源数据口径不一致与复用效率低的问题。

关键结果 KEY RESULTS			
高达 99.1%	提升 5 倍	累计 35 万+ 条
高危敏感 PII 隐私脱敏召回率	数据预标注效率	清洗脱敏真实业务对话语料

01 **Raw → Staging → Approved 数据分层** / 统一接入海外独立站、私域聊天与售后工单等多源数据，保留不可变 **raw_chats** 原始层；通过 ETL、5 分钟时间窗切分、同角色连续发言合并与去重生成标准化会话单元，**让 RAG、微调与质检共享同一份可追溯数据**。

02 **DataFilter 与双层 PII 隐私治理** / 构建 9 级优先级内容过滤与 **正则表达式 + 轻量 NER** 双层脱敏引擎，覆盖手机号、地址、支付账号与订单号；价格与内部折扣作为硬红线字段，**在清洗、写入与导出链路中防止敏感信息进入模型**。

03 **预标注、人工审核与可信数据源** / 自动生成意图分类、质量评分、敏感标记与候选问答，进入 **staging** 暂存区；由运营或销售执行 approve / reject / edit，只有审核通过的数据才能进入下游，**将 AI 预标注与人工确认组合为统一质量门禁**。

04 **多下游扇出与数据总线** / 以 **approved** 数据为唯一可信源，一键导出 **ShareGPT、Alpaca、OpenAI JSONL** 与 RAG CSV；同步生成 knowledge_chunks / knowledge_articles，供导购 RAG、微调与销售质检消费，**形成数据采集、治理、供数与反馈回流的中台闭环**。

05 **评测集构建与 Bad Case 回流** / 从 **approved** 会话中抽取高意向、转人工、客诉、人工纠错与失败样本，按业务场景构建回归评测集；线上失败样本与评测结果回流中台，驱动 Prompt、策略与微调迭代，**形成由数据中台驱动的 AI 应用优化闭环**。

05 销售风格 SFT / DPO 微调与评测

数据与评测链路负责人

2024.10 — 2025.03

项目描述： 在数据中台完成语料治理后，基于高质量销售对话开展风格微调与偏好对齐，使 AI 客服从通用表达转为业务销售话术，同时将价格、库存与政策保留在 RAG / Tool 链路。

关键结果 KEY RESULTS			
58 → 91 分	QLoRA → LoRA	SFT + DPO
销冠话术风格盲测对齐	本地验证到云端训练	风格模仿与偏好对齐

01 **“微调管风格、RAG 管知识、Tool 管实时”三级解耦** / 明确权重、检索与工具的职责边界：微调专注掌握专业光疗顾问的共情表达、答疑节奏与促单话术；产品规格、禁忌原理与政策条款由 RAG 检索注入；价格、库存与优惠资格由**确定性 Tool 按 SKU 实时取数**，**规避快照过期与越权召回，新产品上线无需重训**。

02 **本地 QLoRA 验证与云端 LoRA 微调** / 采用“本地快速验证 + 云端生产微调”策略：本地使用 **LLaMA-Factory + Qwen2.5-3B** 运行 QLoRA，验证语料质量与训练方向；方案通过后转为云端 LoRA 训练，**以较低试错成本验证数据与训练配置**。

03 **SFT → DPO 与五维评估闭环** / SFT 阶段学习销售话术风格，DPO 阶段基于销售主管标注的 chosen / rejected 偏好对优化回答质量；从 **Product Knowledge、Objection Handling、Closing、Policy Enforcement、Empathy** 五维构建评估集，**以盲评与裁判校准控制模型上线质量**。

工程底座 AI Infra 与工程平台

共享协议底座 · 统一治理平面 | MCP 服务共享 · 统一模型网关(06)

06 企业级 MCP 共享服务与统一模型网关

模块 Owner / 平台共建者

2025.04 — 2025.10

项目描述： 导购、售后、营销与数据治理等业务线重复实现库存、算价、RAG、Memory 与 Prompt 能力，且各自持有模型 Key、直连不同供应商；项目将通用能力下沉为 MCP 共享服务，并在其上收敛统一模型网关、鉴权、Trace 与 Token 配额。

关键结果 KEY RESULTS		
4 类	1 个入口	按项目
LLM / RAG / Memory / Prompt 共享服务	多供应商模型调用统一收敛	鉴权 · Trace · Token 配额

01 **四类 MCP 服务拆分与标准化暴露** / 将 **LLM 网关、RAG 检索、会话记忆、Prompt 管理**抽离为 4 个独立 MCP Service，经 **Streamable HTTP** 暴露标准化 Tool / Prompt 原语；新业务接入仅需实现 Agent 编排层，**通用能力由一套共享服务统一维护**。

02 **共享能力底座与解耦编排** / 基于 MCP 协议封装库存、运费、算价等共享 Tool 与 Prompt 模板；架构上坚持“能力下沉、编排解耦”，底座不绑死具体流程，无论是 **自主决策、实时语音还是复杂图工作流** 均可自由编排，并预置标准流水线供通用场景开箱即用；**业务逻辑留在各自工程，新业务配置租户 Key 即可直接复用全套底座，实现统一治理与平滑扩展**。

03 **跨 Agent 买家画像与健康禁忌共享** / 针对导购与售后 Agent 状态割裂、用户重复解释的问题，设计全局买家画像同步：导购识别出的肤质、敏感史与起搏器等禁忌写入 `user_facts`，按 `user_id + fact_key` 唯一约束；**售后 Agent 可按 user_id 直接读取，跨业务画像实现互通**。

04 **统一网关与调用治理** / 基于 FastAPI 收敛模型调用入口，落地 **API Key → project_id 映射鉴权**、Trace ID 链路追踪、按项目 Token 日配额与 429 前置拦截；以逻辑模型名屏蔽底层供应商差异，**实现多业务隔离、成本计量与模型切换解耦**。

05 **MCP 转发与 Serverless 部署** / 网关按 tool 类型转发至 LLM、RAG、Memory、Prompt 服务并做逐服务探活；各服务独立构建 **Docker 镜像**，按 MCP 服务 → 网关 → 业务项目的拓扑部署至 **阿里云 Serverless（SAE / FC）**，**以弹性伸缩、灰度回滚与 SIGTERM 优雅终止保障在途请求不丢**。

个人优势

01 **主流 AI 与 Agent 开源生态贡献** / 以 Contributor 身份向主流 AI / Agent 仓库提交并合入修复：**NVIDIA NeMo Agent Toolkit** #2275、#2236、#2308；**Google ADK-JS** #922；**Dify** #43875；**ByteDance DeerFlow** #6689；**Strands Agents** #4263，**覆盖 ReAct 解析、流式护栏、权限测试与结构化输出回归**。

02 **前沿追踪与主流大模型测评选型** / 持续跟进 OpenAI、Anthropic、Google、阿里通义、DeepSeek 等厂商新模型，**第一时间完成实测、Prompt 适配与业务选型**，沉淀模型对比与切换的最佳实践。

03 **业务价值导向与跨职能协同** / 长期与产品、运营、销售一线协作，**能将模糊的业务痛点拆解为可落地的 Agent 工作流与工程 Spec**；关注 GMV 转化、客诉压降与 Token 成本等指标闭环。

04 **复杂场景交付与生产稳定性** / 经历海外大促、架构重构与核心版本上线等关键节点，**具备从 0 到 1 架构设计与线上复杂问题排障能力**；推行自动化回归门禁与兜底降级机制，保障核心链路稳定交付。

工作经历

01 深圳市臻星瑞通科技有限公司 / AI 应用开发工程师 / Agent 架构

2023.04 — 2026.10

- **核心职责**：负责跨境 PBM 光疗产品海外独立站及私域生态的 AI 应用与 Agent 研发落地，覆盖智能导购、内容营销、售后客服与数据治理。
- **平台底座**：参与 Agent Runtime 底座、MCP 共享服务与 LLM 统一网关建设，落地模型统一调用、多层级 Memory、Token 计量与安全鉴权。
- **业务落地**：攻坚多 SKU 高客单导购、销售话术策略自进化、实时语音多模态售后排障与 AIGC 营销内容自动化流水线。
- **数据资产**：推动独立站、私域社媒、售后工单等多源数据统一治理，为 RAG、SFT / DPO 微调、销售质检与 Agent 评测提供数据底座。
- **工程规范**：推行 Spec-Driven 研发流程与自动化评测门禁，保障 AI Agent 与模型应用高效交付、生产稳定运行。

教育背景

张家界学院 / 软件工程 · 本科

2019.09 — 2023.06