杨嘉焕



AI 应用开发工程师 • Agent 设计 • 对话记忆系统 • LLM 评测工程

13609668934 · 1665526933@qq.com · 珠海（可接受远程） · flowwater011223.top

Pluto：pluto-mem.top（已上线） · FTA：easyforeigntrade.top（已上线） · GitHub：flowing-water1 · Linux.do：4 篇实践长文

> 全栈开发背景，近两年专注 AI 应用落地与上线。2024.12–2026.08 在星越实业独立承接公司全部内部 AI 需求，从需求、架构到前后端与部署运维一人完成；2026.08 起独立开发，个人对话记忆 Agent「Pluto」已上线公网（pluto-mem.top），此前交付的外贸获客工作台亦上线运行（easyforeigntrade.top）。两个项目均处于早期用户阶段，我不把规模说大，把重点放在系统正确性与可验证性上。
>
> 工程上明确划分模型自主决策与确定性约束的边界：记忆内容、证据引用、检索完备性等关键状态均以可机械校验的规则判定，不依赖模型自述，使故障可归因至模型、链路或评测环节。

工作经历

独立开发者（个人项目 Pluto） 2026.08 - 至今

全职推进对话记忆 Agent「Pluto」从架构设计到上线与持续运维（pluto-mem.top），负责产品定义、全栈开发、评测体系、部署与迭代。

星越实业（珠海）有限公司（约 20 人） • AI Agent 训练师 2024.12 - 2026.08

公司未设专职研发岗，AI 相关需求均由本人从 0 承接并独立交付，覆盖外贸获客、销售通话分析、企微会话追踪、多语种名片识别、装柜优化等场景，全部上线投入内部使用。岗位职责实际等同于 AI 应用全栈开发：产品定义、架构设计、前后端、部署与运维。

核心项目

Pluto • 对话记忆 Agent（个人项目，从 0 到 1） 已上线：pluto-mem.top

独立完成架构设计、全栈开发、评测体系、部署与上线 2025.11 业余启动，2026.08 后全职推进

让 AI 真的记得住用户。设计前对 Gemini、GPT 的记忆机制做了逆向分析：Gemini 记忆强依赖近期上下文，长对话间隔后早期信息衰减失效；GPT 的多层记忆结构可作参考，但用户对记忆内容的可见性与控制权有限。Pluto 围绕三个问题设计：记忆可追溯、冲突可裁决、检索与生成表达分离——对应三层记忆分离、检索/生成两阶段、检索完备性的确定性判定（由系统记录检索过程，不依赖模型自述是否找全）。

三层记忆分离与版本化裁决（解决冲突记忆的权威问题）

- 记忆分三层独立存储：会话摘要层（L1，负责定位与会话级压缩）、跨会话用户画像与表达偏好层（L1.5）、客观事实层（L2）。三层采用不同更新频率与可信度阈值：偏好允许单轮对话内覆写；事实仅在系统已检索到的候选范围内判断新旧关系（重复 / 被取代 / 共存 / 存疑），由事务层核验归属与版本后提交，不交给模型自由判定。
- **原始消息为唯一权威来源**，摘要仅承担定位与压缩；每条记忆引用绑定具体候选证据，可回链原始会话与原始消息，摘要不得单独作为证据。
- 用户画像采用版本化存储加「当前有效指针」与乐观锁，避免并发写入相互覆盖；PostgreSQL 为原文与版本化记忆的唯一权威，语义索引仅作可重建的候选召回，命中后回库核验。

检索 / 生成两阶段分离

- 检索阶段（Core）负责任务理解、来源选择与强类型工具调用，产出结构化交接包（发现、证据、冲突、未决项），不直接成文；生成阶段（Answer）仅依据交接包组织用户可见回答。
- **设计动机**：保持检索上下文干净、把「查到什么」与「怎么说」解耦——生成阶段接触不到原始记忆与工具，检索故障与表达故障可独立定位，更换回答模型时记忆范围保持不变。
- **检索完备性由运行时记录机械判定**，不依赖模型自述：所有候选对话下钻读取、命中消息全部处理、无未决错误，三项同时满足时「未找到」结论方成立；并区分未查完、预算耗尽、供应商失败、有界范围内未找到四种状态。

评测与上线前验收

- 采用「开发集迭代 + 冻结验收集」两阶段：开发期先后用过多套评测集——V8（20 个案例、约 47 万字，其中 2 个为超长对话压力测试：一段对话分 16 批写入约 23 万字业务记录，每批夹带一次事实变更，最后只问 2 个问题，考查记忆能否扛住大量输入）、公开基准子集 29 条（LongMemEval / LoCoMo / HaluMem / MemoryAgentBench / MemoryBench），首轮 3/29；按失败归因修根因后，把挂过的题固化为 24 例日常开发集（硬检查 + 证据留痕），回归全过。
- 发布门禁用**每轮新写、运行前先定死案例与评分标准**的 8 个产品案例（24 段对话、54 条消息，跨度约两周）做冻结验收，覆盖偏好记忆、事实记忆、版本更新、边界抑制四类能力。该门禁从 6/8 起步：两个案例 Judge 满分、但证据绑定硬检查失败，定位到候选采用未绑定当前有效版本后修改产品代码，最终 8/8 通过机械校验与双模型独立评分。
- 失败按环境 / 产品 / 评测自身三类归因（评分误判、读取旧字段等不计为产品缺陷），避免把评测问题误判成产品质量。

并发一致性与上线运行

- 会话锁由进程内锁改为数据库咨询锁（以会话 ID 哈希为锁键，数据库不可达即拒绝放行）；双 API 实例并发下持锁区间重叠由 2.844 秒降为零，第二实例等待 2.847 秒正常获锁。
- 上线前通过 12 项发布门禁（配置校验、健康检查、模型策略、追踪、成本、部署、用户流程、故障恢复、多实例、正式评测等），正式评测通过后才执行发布；上线后以 **Docker Compose 生产编排 + 健康检查 + 域名与 HTTPS** 持续运行，并保留回滚版本。
- 评测环境与线上账号严格隔离：一次性评测资源用后即删并逐个验证 HTTP 404，避免测试数据污染生产。

技术：Python、FastAPI、LangGraph、Next.js、React、TypeScript、PostgreSQL、pgvector、Redis、Socket.IO、Langfuse、Docker

AI 外贸获客工作台 ForeignTradeAgent（公司项目，从 0 到 1）

已上线：easyforeigntrade.top

产品定义、架构设计、全栈开发、上线运维 2026.01 - 2026.08 | 服务 5 人外贸销售团队

系统由执行 Agent、三类工具层（客户搜索 / 内容抓取 / 公司调研）、生成与人工审核队列、事务计费账本构成主闭环：Agent 按业务需求调用工具完成客户检索与公司调研，生成开发信草稿，经「客户范围确认、草稿审核」两道人工节点后进入发送队列；Agent 不持有自主发信权限。

工具分层降级（主供应商 → 备用，禁止跨能力）

- 「客户搜索」「内容抓取」「公司调研」设计为独立能力层，各层内部配置多供应商降级链：主搜索 xAI Responses → OpenAI 兼容接口，网页搜索 智谱 → Tavily → Firecrawl，正文抓取层另设备用供应商。某层供应商全部失效时返回该能力缺失，不以其他能力的结果冒充；配套专项测试锁死边界，保证进入生成环节的信息来源类型明确。

并发计费：事务型积分账本

- 基于 PostgreSQL 的事务型积分账本，支持并发安全的多阶段扣费（预占 → 结算 → 释放）、业务幂等键防重扣，以及**价格快照**机制——任务执行期间调整计费标准时，在途任务仍按创建时价格结算。

生成质量验收

- 8 个定向案例覆盖「不主动报价」「必须提及起订量」「历史异议处理」等约束，每封邮件从正文质量、结构合规、协议遵守、可发送条件四维度独立评分，任一维度不达标整封退回；最终 8/8 通过，四维度各 8/8、平均 0.94。

自动化边界与部署

- 两道人工确认控制自动化边界：客户范围确认后进入起草，草稿审核通过后进入发送队列。
- 部署于京东云，Docker Compose + Nginx 同源代理 + 健康检查，GitHub Actions 发布。初版由 Runner 推送完整镜像 58 分钟未完成，改为服务器按同一 commit 本地构建后，**验收约 20 分钟完成**构建、备份、部署与健康检查。

技术：Python、FastAPI、LangChain、DeepAgents、PostgreSQL、pgvector、React、TypeScript、SSE、Docker、GitHub Actions

自研工程工具与公开输出

早期即系统探索人机协作开发，在 Linux.do 社区发布 4 篇实践长文，**累计近 2 万浏览**并形成讨论。核心判断：**AI 协作的瓶颈是证据是否落盘**——不落盘则跨会话丢失状态、跨模型丢失约束，故障也无法归因。为此自研三套工具并集成到实际项目：

- **任务契约与可恢复状态机（mission 框架）**：长任务拆为带验收标准的原子动作清单，落盘为可恢复状态文件，支持开发 → 初验 → 回归 → 提交闭环与断点续跑，已用于 Pluto 与 FTA 的开发流程。
- **评测编排与证据链管理**：固化权威来源、阶段、通过判据与证据快照，区分确定性检查、模型评分、运行级指标、人工标注四类判定，避免「分数好看」掩盖「证据缺失」，已用于 Pluto 多阶段评测与 FTA 生成质量验收。
- **架构决策记录生命周期管理（lite-arch）**：按「难回退、会使既有决策失效、存在真实方案权衡」三条判据决定是否记录，每条写明选择项、否决项与否决理由；两个项目累计沉淀 44 份决策记录，配套按标签与路径召回。

三套工具在 Pluto 与 FTA 的真实开发中持续使用，**实际拦截过「看起来通过、实则无效」的假结论**：评测器未调用模型、仅凭硬编码规则放行，导致「全量通过但结果为空」；双模型对照中模型标识未被运行时变量覆盖，实验条件无效；单测只覆盖终态、未覆盖真实控制点。此类问题均定位后修复并复验通过——工具的实际价值，是让「分数好看」无法掩盖「证据缺失」，效果已在两个项目中得到验证。

公开文章：用 mission + goal 让 codex 跑一天都不累 · 结合 OpenSpec 让 codex 做 8 小时班 · 如何让 codex 任劳任怨跑几个小时 · 放弃 MCP：Claude Code / Codex / Gemini 并行协同

其他产品交付（公司内部，均已上线使用）

销售通话自动分析系统 2025.02 - 2025.07 · 覆盖 4 名销售，月处理数百通录音 · 结果纳入 KPI

打通录音上传、多文件并发转写、大模型分析与 Excel 报告，每月自动汇总推送工作群，分析结果作为销售考核依据。对比 Whisper、火山引擎、讯飞的中文识别、说话人分离、速度与成本后接入火山引擎；持续迭代多音频格式兼容、截图存证与去重校验，保证结果可审计。

企业微信 AI 秘书 • 会话分析与项目追踪 2025.05 - 2025.08 · 覆盖 2 个群 · 从 0 到 1 全栈

面向多项目并行时的信息分散问题，研读企微会话存档接口拉取授权账号文本会话；按群组清洗合并消息，用大模型区分项目进展与闲聊，将摘要、状态、历史结构化入库，交付仪表盘、消息分析、项目追踪与系统设置。修复同步逻辑每次固定读取前 1000 条导致的遗漏，改为按新增消息增量更新。

PDF 名片识别与内贸客户管理系统 2025.09 / 2026.04

处理上万张多语种名片（阿尔及利亚、泰国、俄罗斯、尼日利亚、哥伦比亚等），针对背景色与文字相近的布局定向优化识别策略，用 Gemini 2.5 Pro 做视觉提取，输出供人工确认；另两周内完成内贸客户管理系统（销售录入、权限、统计看板、测试与正式数据隔离），替代原有付费工具。

装柜优化算法与计算工具 2024.12 - 2025.02 · 已稳定运行于业务系统

在大模型编程助手普及前，与 GPT 深度协作设计实现多约束装柜优化算法（含锦标赛选择机制），长期稳定运行于业务系统。

技能

Agent 与 LLM 应用：Agent 工作流、工具调用与编排、记忆系统、RAG、提示词工程、LLM 评测、LangChain、LangGraph、DeepAgents、Langfuse

后端与数据：Python、FastAPI、SQLAlchemy、PostgreSQL、pgvector、Redis、异步任务、事务与幂等

前端与交互：React、TypeScript、Next.js、Vite、SSE、WebSocket、Playwright、Vitest

工程与交付：Docker、Docker Compose、GitHub Actions、Linux、Nginx、Git、京东云、域名与 HTTPS、健康检查与回滚

教育背景与学习经历

广州大学 • 电气工程及其自动化 • 本科 2019.09 - 2023.06

GPA 3.56/4.0，专业排名 24/138 · 全国计算机二级（C 语言）、三级（网络技术）

语言：雅思 6.5（阅读 7.0）· 英语六级 497、四级 545

2023.06 - 2024.12 自主学习 Python 与 LangChain / LLM 应用开发（官方文档与视频自学，为入职 AI 应用方向做准备）