居浪

求职意向：大模型评测工程师（Coding / Agent 方向）



男丨25岁丨硕士　17314833596　langju123456@gmail.com
github.com/langju123456

个人优势

1. 圣路易斯华盛顿大学信息系统管理硕士（2025.12）；AI软件公司创始人兼技术负责人，独立设计并开源Agent评测平台EvidenceLoop，覆盖出题、判分标准、失败归因与评测工具全链路。
2. 评测出题与评测集构建：按多维分桶程序化出题并配独立oracle，建成550道冻结评测集（含OOD划分、查重与哈希可复现）；主动发现并堵住"只保留ok行即可答对"的出题捷径。
3. 评测标准与判分体系：五维验证器（数值/单位/来源/状态/权限）加9类错误路径表，检查最终回复与系统状态是否一致；修复验证器放过编造数值的漏洞，杜绝虚高成功率；在Qwen3三个尺寸上采集776条真实轨迹做失败归因。
4. Coding Agent体感与工程实践：以Claude Code、Codex为主力完成6个产品研发，熟悉其在真实工程中的典型失误；Python为主、Java为辅，配套pytest、CI、Docker、结构化日志与健康/就绪探针等运维实践。

工作经历

Lang Horizon Technologies llc　创始人 · AI应用全栈工程师　2026.01-至今

自营AI软件项目研发与交付。创立并经营公司，主导EvidenceLoop、ForgeFlow、SoloScale AI OS、BuildLog、美业服务Copilot、X Post Lab六项自研产品从0到1的需求、研发与交付，近期重点建设大模型后训练数据与评测体系。

1. 需求与范围：与真实店主做需求访谈和原型测试，把业务场景拆成可验收的流程与数据边界；按证据成熟度决定各产品的范围和下一步，未完成项在交付材料中明确列出。
2. 研发与集成：用Codex、Claude Code推进AI辅助开发，本人负责架构取舍、关键执行路径和代码审阅；集成多家云端与本地模型、OAuth及外部API，统一结构化输出、超时与错误处理。
3. 数据与实验规范：关键决策写入决策记录，规则先写后跑；题目划分与环境、验证器版本冻结并记录哈希；用真实模型运行反查评测与产品本身的缺陷；每个产品配自动化测试、类型检查、CI与独立审阅。
4. 交付与部署：按客户实际设备与使用条件选择部署方式，完成门店本地轻量化部署、macOS桌面打包与云端GPU实验环境配置；BuildLog完成LinkedIn与X各1次受控发布并保留HTTP 201回执。

南京康尼机电股份有限公司　海外技术销售工程师　2023.12-2024.05

参与2-3个海外轨道交通投标及技术项目，连接海外客户与研发、生产、销售团队。

1. 将客户规格、交付要求和问题清单转化为技术方案、需求澄清材料及执行事项，支持英文客户沟通，协调跨部门事项推进投标。
2. 结合CRM/ERP业务流程识别人工操作、信息断点和数据管理问题，输出流程分析及改进需求。

项目经历

EvidenceLoop｜Agent后训练数据飞轮与评测平台（开源）　2026.10-至今

项目定位：面向工具调用Agent的后训练数据与评测平台，以实验测量数据分析为科学任务场景（版本化数据集与分析规程、质量标记过滤、单位换算、科学计数法数值、统计计算）。独立负责架构、数据管线、验证器、训练与评测全链路。

1. 题目生成与难度控制：按单位档×题型×版本数×瞬时故障×数值格式多维分桶程序化出题，配独立oracle；建成550道冻结评测集（train_mining 200 / validation 50 / test_id 150 / test_ood 150），划分查重、OOD模板隔离、全量哈希可复现。
2. 程序化验证与失败归因：五维验证器（数值/单位/来源/状态/权限）加9类错误路径表，将失败精确归因；独立审计中15类错误策略、804条错误轨迹全部被拒绝并正确归因；主动发现并堵住"只保留ok行"即可答对的出题捷径。

3. 多尺寸评测与失败挖掘：在Qwen3-0.6B/1.7B/4B上累计采集776条真实轨迹，输出排行榜与失败分布；定位1.7B失败主因是嵌套工具参数（259次格式错误），重构接口后单位/状态/来源通过率由38%/43%/17%提升至78%/85%/65%；修复验证器放过编造数值的漏洞，杜绝虚高成功率。
4. 失败驱动的数据合成与配比：以参考解法生成示范轨迹，经验证门筛选、去重与谱系追踪，产出816条SFT样本；按6组失败率为数据配比加权，构建均匀/失败定向/含错误示范三组消融数据，预热剂量分档并事先写定达标标准。
5. 训练与回归：在Kaggle T4上搭建Qwen3-1.7B LoRA训练与vLLM评测流水线，支撑预热、配比消融训练与回归评测。
6. 工程化：91项自动化测试、11条决策记录，题目、环境与验证器全部版本化；代码开源于github.com/langju123456/evidenceloop。
技术栈：Python / Qwen3 / MLX / vLLM / Transformers / PEFT LoRA / SFT / pytest / Kaggle

ForgeFlow｜面向企业的Agent受控执行与审计平台 2026.10-至今

1. 以PostgreSQL完整记录Agent的run、step、event、artifact、approval与receipt，形成可审计的执行轨迹；lease/epoch与幂等保证崩溃后可恢复、可回放，验证SIGKILL回滚与跨进程同键回放；基于MCP SDK设定sandbox工具边界，模型提议、服务端判定、审批与回执分离。
2. Supervisor、PM多角色接入本地Qwen3:8B，完成5个业务场景共20次模型调用；构建离线重评工具，结构与证据选择5/5；经一轮系统性审阅修复，932项隔离测试通过。

SoloScale AI OS｜LLM应用与工程证据工作台 2026.01-至今

1. 对授权的Git项目与文档重叠分块、增量同步，保留稳定分块ID、来源与内容哈希；SQLite FTS5/BM25、RRF融合排序；BoundedEvidenceAgent按查询规划→检索→覆盖检查→有依据草稿执行，核对引用ID/哈希。
2. PostgreSQL持久化任务，幂等键、SKIP LOCKED事务领取与租约，双worker完成20/20合成模板任务、无重复完成；720项测试通过、10项跳过。源码：github.com/langju123456/soloscale-ai-os

美业服务 Copilot｜需求访谈与原型试点 2026.09-至今

1. 原音频、首次转写、人工修订与最终采用文字分离保存，人工纠错全程可追溯，可作为高质量反馈数据；写回前检查来源版本与租约，原话修改后旧建议自动失效；按门店现有电脑完成轻量化本地部署。

Ardent Mills｜供应链数字化转型校企Capstone（项目负责人） 2025.09-2025.12

带领6人团队，经14场客户、业务与技术相关方会议，澄清供应链B2B集成、主数据与数据治理问题；比较EDI、API与混合集成方案，向客户管理层及技术团队汇报分阶段路线图。

教育经历

圣路易斯华盛顿大学 信息系统管理（MISM） 硕士 2024.08-2025.12

佛罗里达大学 信息系统（B.S.B.A.） 本科 2019.09-2023.05

专业技能

评测：Coding / Agent评测集构建、评测出题与难度控制、判分标准设计、程序化验证、失败归因、OOD划分与防污染
评测工具：自动化评测流水线、vLLM批量推理评测、排行榜与失败分布报告、版本化题库
后训练数据：SFT数据构建、指令微调数据生成、数据合成与筛选、去重 / 谱系 / 版本管理、数据配比与消融
训练与推理：LoRA / PEFT / Transformers、SFT、vLLM / Ollama、Qwen3 / Llama、MLX、Kaggle GPU
大模型应用：Prompt工程、Function Calling / Tool Calling、RAG、LangChain / LangGraph / LlamaIndex、MCP
工程：Python（主力）、Java、SQL、pytest、Docker / Git / GitHub Actions CI/CD、AI Coding（Claude Code / Codex）

证书与语言

Hugging Face：Agents Course 卓越证书（涵盖 smolagents / LlamaIndex / LangGraph） | MCP Course | Deep RL Course | Robotics Course
DeepLearning.AI / Coursera：Machine Learning Specialization | Machine Learning in Production
语言：英语精通（听说读写） | 西班牙语熟练