

林芮瑭

☎ 15307128607 | ✉ lofisuchat@gmail.com | 🗣️ susususuyuyu66

👤 20 years old · Full-Stack AI Agent Developer · Open to opportunities

🏠 github.com/LofiSu (Apache Fory Committer | Maintainer of ByteDance/DeerFlow)

🎓 中南民族大学 (中央部委院校) 威斯康星州立大学合办软件工程创新人才实验班 | 2023年-2026年 (本科跳级毕业)

实习经历

实习期间多篇论文解析及技术文章产出到内部论坛，多次登上公司《LLM技术周刊》与部门团队号优秀征文榜单，荣登ByteTech首页文章周报13次，ByteTech作者等级LV8，2025年字节跳动优秀技术创作者榜第13名。

github.com/bytedance/deer-flow ★ stars 76k · forks 10.3k



- 项目简介：字节排名第一的开源Long-Horizon Agent Harness项目，能够研究、编码与创作，内置AIOSandBox沙箱化执行环境、可插拔 Skill 体系、跨会话长期记忆、动态 Sub-agent 调度与系统性 Context Engineering，可处理多层级 long-horizon 任务。
- 我的职责：作为项目owner核心作者之一，主导1.0到2.0的架构升级与迭代
 - 1.0 (Multi-Agent + ReAct风格子 Agent)：Coordinator (意图识别) → Planner → Research Team (Researcher / Coder 支持 MCP 动态扩展外部工具) → Reporter (汇总上下文生成最终报告)；Prompt 由 OpenAI Meta Prompt 生成。
 - 2.0 (Agent Teams Harness + Middleware Chain + Skill System + Context Engineering)：
 - Agent Teams Harness：相比 1.0 五节点固定流水线，2.0 改为单一 Lead Agent 统一决策；system_prompt 动态写入——按当前启用的 Skill 列表、跨会话持久化记忆、Sub-agent 并发规则组装；Lead Agent 通过 task 工具触发 SubagentExecutor 新起独立 Agent 实例；调度与执行双线程池分离，最多 3 并发 / 单任务超时 900s；Sub-agent继承父线程目录 + 复用父沙箱，每步实时推流至前端，完成后压缩摘要回传 Lead Agent
 - Middleware Chain：1.0 上下文压缩 / 沙箱 / 记忆逻辑散落各处，改一处要动多处；2.0 抽成 11 层有序流水线，新增或修改能力只动对应层，Agent 推理代码零改动 (ThreadData → Uploads → Sandbox → DanglingToolCall → Summarization → Todo → Title → Memory → ViewImage → SubagentLimit → Clarification)
 - Skill System：每个 Skill 是一个目录，内含 SKILL.md；system_prompt 只注入索引，Agent 按需 read_file 渐进加载，避免全量注入 token 爆炸；支持自定义 Skill，custom/ 下新增目录即可扩展
 - Context Engineering：Write——中间结果写入文件系统，索引 / 规则每轮注入 system_prompt；Select——按 token 预算从 memory.json 裁剪注入记忆，Skill 按需读全文；Compress——超限调轻量 LLM 压缩历史 + 异步提炼持久化记忆；Isolate——Sub-agent 仅见任务描述，结果摘要回传，沙箱按 thread_id 隔离

🎵 抖音AI基建-AI4SE基座算法&Infra --- 2025.12 - 2026.4

AgentScaling for SFT & RL | MultiModel Relay + Multi-Scaffold Diverse Trajectory Synthesis Pipeline

- 背景与目标：
 - 单一模型轨迹风格单一易过拟合——借鉴「Mix Diverse GPTs」，引入多种 scaffold 提供风格不同的轨迹，artifacts 生成任务上混合训练已验证有显著收益；目标：规模化合成覆盖多 scaffold 的高质量多样化轨迹
 - 针对 Claude Opus 4.6 连续失败 3 次以上的难题，用接力和定位策略提升通过率，降低重复推理成本；目标：提升难题通过率，降低重复推理成本
- 我的职责：项目owner，完全负责从 0 到 1 设计并落地以下系统架构与技术链路 (CLI: relay run → export)：
 - 多 scaffold 统一适配：统一 Agent 接口接入 10+ scaffold (codex / claude_code / cline 等)，新增 scaffold 实现统一 run / get_answer 接口即可接入
 - relay run 执行引擎：同一 Docker 容器内多 LLM 顺序接力，共享代码现场。切换策略为三阶段交接：规则压缩历史为 ~12k 结构化文本 → LLM 以 gold patch 为隐形顾问生成 RFC (复盘) + Plan A/B/C → 裁剪对话为 [首条任务描述 + handoff 消息] 作为下一模型输入；可选每轮 Test-in-Loop 提取 patch 评测，通过则提前退出，失败则注入失败报告给下一模型。续跑时从 checkpoint 恢复仓库 patch 与对话，从上一模型继续。针对反复失败的难题，复现 SWE-Replay 论文并提升策略：维护历史轨迹 Archive，从关键中间步骤分叉、apply git diff 恢复现场后重试，避免全量重跑。Benchmark 指标：SWE-Bench 系列难题通过率较全量重跑提升约 25%+，单任务推理成本 (API 与时长) 降低约 30%+，多 scaffold 混合轨迹覆盖 10+ 种风格
 - 上下文压缩 + Checkpoint 续跑：超 180k 触发 Micro compact，超 250k 触发 LLM 兜底；每轮 checkpoint，中断后可从上次模型续跑

Agent Tracer Evaluation for SFT & RL | Trajectory Diagnosis + Rubric Eval and Wash Pipeline

- 背景与目标：
 - 多模型 MultiSWEBench 完成率接近，但 Agentic 能力 (规划 / 工具链效率 / 错误恢复 / patch质量) 参差不齐；目标：用 Rubric 评判标准逐 check 打分，筛选高质量轨迹片段用于 SFT / RL 训练
- 我的职责：项目owner，完全负责从 0 到 1 设计并落地以下系统架构与技术链路 (CLI: evaluate run → wash → export)：
 - LLM-as-Judge 评估引擎：按 check 类型分组 batch 减少 API 调用，规则类 check 零 LLM 直接判断，失败时回退逐条补全；支持多模型投票；遇 429 自动退避重试，并发数与请求频率双重限流
 - Failure Onset 定位：Agent 失败根因是能定位相关代码证据但无法正确解读并转化为行动 (Evidence-to-Action Gap)；通过将 Rubric check 逐步打分定位到具体轮次，自动识别 failure onset (首次出现错误决策的步骤)，过滤 onset 后的级联失败片段，只保留 onset 前的高质量决策轨迹用于训练
 - 轨迹压缩：逐轮按内容类型差异化截断；超 180k token 触发 micro_compact；超 250k 触发 LLM 总结兜底，保留首条 USER 消息与失败报告
 - 增量续跑 + wash / export：自动增量补全未完成 check；按 CSR / ISR / Score / Effective Action Ratio (有效工具调用占比) 多维 Top-K 筛选 + 早期质量评估，导出 pkl 格式。Benchmark 指标：Agentic Rubric 模型区分度 35%，Failure Onset 定位后高质量决策片段占比提升，用于 SFT/RL 的轨迹通过率与 patch 质量显著优于未筛选基线

SWE等系列Benchmark 评测服务化改造 | 接入字节云 AIPaaS 容器平台

- 我的职责：项目 owner，将分散的本地评测接入 AIPaaS 云容器统一入口：spec 配置 → CreateSession 申请隔离容器 → 还原仓库 → Agent 修复 → diff 送测 → 异步轮询 → 返回结构化结果，供全业务按需调用。



EC-Monica 电商产运Agent

- **背景：**为产品、运营、研发提供电商业务Agent，支持数据分析与报告生成、业务归因与需求文档解析等任务，提高业务理解效率与文档编写效率。
- **指标与效果：**上线后服务产运、产品、研发等多角色；调用量月均 600+ 次，报告类任务占比约 55%，需求/归因解析类约 45%；单次分析报告生成从人工 1-2 小时缩短至 5-15 分钟；业务范文 Few-shot 注入后，输出风格与内部模板对齐率提升，复购与主动推荐率持续增长。
- **我的职责：**项目owner，完全负责产品从 0 到 1 设计并落地以下内容
 - Agentic Workflow：意图识别 → 条件路由（需工具 / 直接生成）→ cfg.json 动态注册 FunctionSchema 的 Tool Registry 选工具循环执行 → 数据预处理计算同比 / 环比指标后注入 Context → LLM 生成子结论 → 循环直至完成后汇总报告；工具层封装超时重试与错误分类，失败时 Fallback 直接生成
 - Context Engineering：注入业务范文与运营分析模板作为 Few-shot 样例强化风格对齐；多轮推理中对 Observation 做摘要压缩 + 滑动窗口截断，携带压缩后的前轮结论滚动输入



直播 AI 智能搭建平台 — Agentic Workflow 搭建

- **背景：**基于 GenUI 平台，为直播搭建团队打造 AI 驱动的低代码智能建站能力，实现“自然语言→页面/流程”的端到端自动化搭建，显著降低页面搭建门槛与人力成本。**使用率与效果：**上线后 AI 搭建能力在直播搭建场景中的使用率占新建/改版页面的约 40%+，单页搭建耗时从人工 2-4 小时降至 15-30 分钟，团队采纳率与复购率持续提升。
- **我的职责：**项目owner，完全负责产品从 0 到 1 设计并落地以下内容
 - 设计并实现多条 Agentic Workflow 链路 与 Prompt 工程设计：
 - Prompt 工程：覆盖意图分析、任务拆分（new_page / new_section / update_section / new_flow / bind 等 8 种动作）、数据源缺失检测等完整推理链路
 - 主流水线：负责意图分析，将用户输入与页面 UIDL 传入 Claude3.5_v2，输出结构化 taskList（含 UI / Flow / Bind 三类任务）
 - Controller：按优先级并行调度各子任务，通过 Promise.allSettled 实现多任务并发执行；封装 XML Patch 服务实现 UIDL 增量更新
 - UI 生成链路：支持文生/图生双模式，经 JSX→XML→UIDL 多级 IR 转换渲染页面
 - Flow 生成链路：将 AI 输出的 IRFlow 转换为可执行的流程结构，完成节点与边的构建
 - Bind 链路：负责将 UI 组件与 Flow 按事件类型（onClick / onMount 等）绑定到 UIDL flowMap
 - 一期向 AI 喂入 54 个基础组件的 Schema/DSL 定义完成模型训练，为后续接入 Semi / Arco Design 组件库奠定基础



AI Startup 兼职&开源项目&AI 黑客松比赛：github.com/LofiSu/Awards

Thumbnail AI · Mountex

- **背景：**基于 MkSaaS、Next.js 搭建的多平台封面图生成 AI 工具 (thumbnail-ai.com)，面向 YouTube、Instagram、Facebook、X、LinkedIn、TikTok 等多平台内容创作者，支持选择平台尺寸、风格参考与主体上传，三步内完成专业封面图生成，帮助创作者提升点击率。
- **产品与链路：**前端：Next.js + 响应式多端适配，支持平台预设尺寸一键选择、本地上传/URL 主体、风格参考图；AI 链路：主体与风格图输入 → 提示词编排（平台场景 + 风格关键词）→ 调用图像生成 API 生成多尺寸封面 → 预览与下载；支持免费额度 + 订阅付费，付费用户可解锁高清导出与批量生成。
- **数据与指标：**累计注册用户 2000+，付费用户 200+；月活 500+，月生成封面 1500+ 张；付费转化率约 38%；用户从上传到出图平均时长 < 2 分钟。
- **我的职责：**项目 owner，从 0 到 1 负责前端工程（Next.js、支付与订阅流程、多平台尺寸与风格配置）、AI 封面生成链路设计与对接（提示词工程、多尺寸批量生成、降本与体验优化）。

Krene-Art

- **背景：**面向游戏/二次元角色设计师的 AI 创作平台 (krene.com)，通过 Agent 完成意图解析、参考图检索、视觉特征提取与 Prompt 生成，输出带设计说明的角色卡片（含核心设定、概念整合与角色立绘），缩短从脑洞到成图的设计周期。
- **ReAct Workflow 与链路：**
 - search：根据用户意图生成检索关键词 → 爬取 Bing 图搜解析高清链接 → LLM 按游戏角色维度（造型、配色、气质等）打分筛选，单次聚合约 300 张参考
 - design：GPT-4o Vision 对参考图做多维视觉特征提取 → 与用户设定拼接成结构化 Prompt → NovelAI nai-diffusion-3 生成约 10 张角色图
 - evaluate：按设计师设定维度打分，低分时 ReAct 触发重新 design（调整 Prompt 或参考集），直至达标后输出结构化角色卡片
- **数据与指标：**累计用户 1200+，付费用户 150+；月生成角色卡 400+ 张；付费转化率约 42%；单角色从输入意图到出卡平均 6-8 分钟。
- **我的职责：**项目 owner，从 0 到 1 负责前端工程与 AI 角色设计 ReAct Workflow（意图解析、search/design/evaluate 三阶段编排、Vision 特征提取与 NovelAI 对接、角色卡片结构化输出与展示）。
- **开源：**持续半年 Apache 社区贡献，[18 岁成为全球最年轻 Apache Committer](#)；深度参与 ByteDance/DeerFlow、Apache Fory、Spring AI Alibaba、Ant Design X 等社区开发与维护；2024 阿里天池云原生全球编程挑战赛 issue 解决数第一，技术文章荣获天池赛最佳质量奖；2024 腾讯犀牛鸟开源大赛课题实战奖 & Issue 实战奖；2024 OSPP 中科院开源之夏（ByteDance/VisActor 组件开发）
- **AdventureX AI 黑客松：**开发全球首个 Apple Vision Pro Agent，斩获 Kimi For Vibe Coding 奖、空间智能赛道第二名及 Injective \$1000。
- **AI Startup 好朋友：**帮多家真格、红杉、蓝驰、奇绩等知名 VC 被投 AI Startup 实现 Agent 业务落地；与千问、Kimi、智谱 GLM、MiniMax、豆包、Trae、灵光、阿福等多家大模型 / AI 产品深度合作，参与官方评测与 Demo 制作，产出[技术文章](#)、[评测笔记](#)与[Demo 视频](#)，为产品宣传与迭代提供反馈。
- 高中时期做过 Live2D/3D 数字人，二次元平台产品经理，对二次元文化有深入了解。