袁晨杰

北京/西雅图 | chenjy4@uw.edu | 186-1258-7299 | antaresyuan.site | 领英

University of Washington, Seattle（QS Top 100） | B.S. Human Centered Design & Engineering **2023.9 - 2027.6**

实习经验

字节跳动 - 管理研究院 | Hiring Agent | AI 产品经理 **2026.6 - 至今**

阶段性目标: 在 Hiring Agent 的 1->N 扩量阶段，通过策略让 Agent 推荐的候选人的简历评估通过率趋近资深招聘官人工推荐的水平（深度）；不断丰富 Hiring Agent 的原子能力，把原本依赖 ATS 的招聘流程，升级为基于 Hiring Agent 打造的 AI 原生招聘工作台。（宽度）
核心矛盾: 扩量阶段召回标准陷入放宽通过率下滑 vs 收紧漏放的权衡；产品不能只是一个新的对话式 AI 工具，而是一套 AI 招聘工作体系。

- **KR1 | 岗位分类知识治理，修复经验度误判:** 归因 200+ 条 badcase，根因定位为模型缺乏相近职能的岗位分类知识，深层症结是 JD 与 Title 不一致导致匹配度被系统性低估；基于 Title、JD 与原子技能治理上万条产研运岗位，沉淀 13 大类 96 子类分类体系，完成离线回测后推动灰度扩量，作为召回与精排的结构化知识输入；对" 评估通过但两方分类无交集"case 人工复核，交付算法全量打标。
- **KR2 | 迭代优缺点标签 prompt，建标签质量评测体系:** 针对呈现给用户的候选人优缺点标签存在内容不精准的问题，用 1k 条离线数据完成 5 轮迭代、修复 12+ 个共性问题，将标签可用率从 **25.4%** 提升至 **99.5%**，通过合规检测后上线；进一步建标签评测体系，设计准确度 /必备条件覆盖度 / 可读性 / 合规四维下限 + 用户意愿覆盖上限的评分 rubric，沉淀出 sota 模型打分 prompt。
- **KR3 | 指令上下文的稳定段/变化段分离:** 针对用户不会写 prompt 的核心负担，将指令拆为固定骨架与可变槽位，按用户所处阶段动态下发预制入口，在降低使用门槛的同时提升输出稳定性；与测评同学在模拟线上环境注入 Ground Truth 逐个回测，通过后上线。

AI 项目经历 &AI 论文

Coze | 字节跳动 AI 全栈开发挑战赛 | 多 Agent 协作编排平台 | AI 产品经理 **2026.5 - 2026.6**

目标: 通过面向不同开发任务的多 Agent Workflow 编排，降低 AI Coding 新手完成复杂任务的门槛，提升熟练用户的编排与复用效率。
核心矛盾: 多 Agent 协作依赖用户自行拆解和编排；固定 Workflow 灵活性不足，开放编排对新手门槛过高，且有效流程难以沉淀复用。

- **KR1 | 规划入口推荐准确性:** 多 Agent Workflow 选错会浪费后续全部节点的执行成本，因此以" 准确识别适用边界并区分易混淆场景" 为成功标准；定义任务类型、适用场景、Agent 角色、输入输出、流程节点及限制条件等元信息，构建含易混淆项的 30 条评测集；通过 Bad Case 归因定位模型判断知识不足，在工程测引入带正反例及原因解释的 5-shot，采用 ICE 结构提示词与 Schema 结构化输出，避免模型产生幻觉；将 Seed 2.0 Pro 的准确识别率由 16/30 提升至 28/30 **（+ 40pp）**。
- **KR2 | 编排资产复用:** 上线基于任务元信息的 Workflow 推荐、可视化编辑器与模板市场；**项目获 AI 全栈开发挑战赛 Top 3**。

飞书 | 字节跳动 AI 校园挑战赛 | 基于 IM 的项目长程协作 Agent | AI 产品经理 **2026.4 - 2026.5**

目标: 将需求澄清、任务拆解、内容生产、交付评审与复盘等项目协作环节沉淀为预设 Skill，支持团队在统一对话中推进项目全链路。
核心矛盾: 传统 IM 团队协作只能沉淀沟通信息，团队需自行判断任务并重复整理上下文，导致信息衔接断裂、协作推进与交付成本高。

- **KR1 | 记忆召回精度与上下文组装:** 将历史消息召回由语义相似度 top-20 检索重构为基于 Seed 1.5 Lite 的 Memory 结构化存储 + 特定关键词与任务类型双路定向提取，以精确匹配收敛语义漂移、以任务类型限定召回范围；在基于任务完成维度和模型思考链路沉淀的 60 条测评用例上，通过率由 65% 提升至 80%。
- **KR2 | 按子任务复杂度分层调用模型:** 将上下文抽取、结构化等确定性处理下沉到 Seed 1.5 Lite，Seed 2.0 Pro 只承担核心推理与生成，减少高成本模型需要处理和输出的 token；人工抽检对比未见明显质量差异，**将单任务平均响应时延由 95s 降至 50s（- 47pp）**。

字节跳动内部开源项目 Kaboo | AI-Native 社区共建与产品化探索 | AI 产品经理 **2026.6 - 至今**

- 参与 Seed 内部开源项目 Kaboo（Token 消耗统计平台）共建，基于 Mew 探索" 需求方即 Builder" 的 AI Native 协作方式；主导 Token 效率榜与效率分析模块，已覆盖数千名内部用户。
- 现有统计只能回答" 消耗了多少"，无法回答" 哪些消耗是浪费"；定义无效重试、工具定义冗余、上下文冗余、缓存未命中、模型选型过重五类浪费的判定标准，设计仅采集结构化元数据（不含任务内容）的诊断方案，采用 Effective Tokens 口径归一化跨模型成本。

SusBench:An Online Benchmark for Evaluating Dark Pattern Susceptibility of Computer-Use Agents | IUI'26 二作

DevPost | Agentic RAG 引用可信度评测基准 | Honorable Mention | AI 评测产品设计 **2026.5 - 2026.6**

- 针对 RAG 系统中 "答案正确但引用不支撑结论" 的问题，设计引用可信度评测框架，基于 HotpotQA 构建 500 题、5,000 条干扰引用证据集，以 "引用是否真实支撑答案" 为核心标准。
- 将引用质量拆解为虚构引用、伪装相关、合理推导与证据缺失等指标，完成 Claude 与 GPT 模型评测，引用可信度分别为 42.5% 与 26.7%

技能

AI 产品能力：LLM Eval、评测集与指标口径设计、Badcase 归因、Agent 上下文与记忆策略、RAG Evaluation、LLM-as-Judge、HITL
数据与工具：Python / BOE 评测脚本 / Pandas / NumPy、Figma、FigJam、Miro、HTML / CSS / JS、LLM API