万彬宇

电话：18809164801 | 邮箱：wanbinyu18@163.com

GitHub：github.com/Wanbinyu | 个人主页：portfolio.wanbinxy.icu



熟悉 Agent 工具调用与多轮运行机制，独立开发并上线 Resume Fit AI 求职助手，完成模型适配、Agent 运行与业务应用的分层重构。具备 RAG 检索评测、企业客服知识库与商业服务端开发经验，能够围绕实际需求实现工具、处理异常并验证结果；参与 Spring Boot、Gradle、OpenRewrite 等开源项目贡献。

专业技能

- **Agent 架构：**熟悉模型适配层、通用 Agent 运行层与业务应用层的职责拆分；理解上下文组织、Tool Calling 多轮循环及结构化结果校验。
- **工具注册与执行：**使用工具目录注册名称、说明、JSON Schema 和执行函数；按任务装配白名单，处理参数校验、调用 ID 关联、输出限制与业务权限。
- **运行控制与兜底：**熟悉轮数 / 工具调用次数 / Token 预算、总超时与单次超时；区分错误类型实施有限重试，处理取消、迟到结果及任务状态回传。
- **上下文与记忆：**结合项目实践实现分层记忆、来源追踪、时态事实与纠错；使用遗忘曲线、巩固与固定锚点控制回忆，区分短期情绪信号和角色回应策略。
- **RAG 与模型接入：**使用 Spring AI 接入 DeepSeek / OpenAI 兼容接口；掌握内容哈希增量索引、Vector / BM25、RRF、Cross-Encoder 重排、引用与证据不足拒答，使用 Recall@K / MRR 评测。
- **后端与数据：**Java 21、Spring Boot 3、Spring Security、JWT、MyBatis-Plus、H2 / MySQL；熟悉 REST API、异步任务、SQL 与事务 / 索引；使用过 Python / FastAPI、Node.js / Express、Redis、Kafka。
- **工程实践：**Git、日志排查、接口联调与回归测试，使用 pytest、node:test、GitHub Actions；具备 Linux、Docker Compose、Nginx 环境下的项目部署实践，使用 AI 工具辅助编码与审查。

核心项目

Resume Fit AI - Agent 求职助手

作用与用法：面向求职者的简历与岗位分析工具。用户导入简历、选择版本并填写岗位 JD，即可检查简历表达、对照岗位要求与原文证据、保存修改版本，并整理公司资料、记录投递和导出 PDF。

jianli.wanbinxy.icu | github.com/Wanbinyu/resume-fit-ai

技术栈：Java 21 / Spring Boot 3.4.4 / Spring AI / DeepSeek / Spring Security + JWT / MyBatis-Plus / H2 / Tika / PDFBox

- **分层重构：**将模型 SDK 与消息转换下沉到 AI 层，通用 Agent 层集中处理运行循环与工具执行，应用层负责求职业务和报告校验；真实模型与离线 Demo 复用工具协议。
- **工具注册：**实现 ToolCatalog / ToolExecutor，拒绝重复工具名并在运行前冻结目录；校验 JSON 参数、输出大小和简历版本归属，限制模型只能调用本次任务登记的工具。
- **多轮编排：**实现“模型决策 → 工具执行 → 按原调用 ID 回传 → 下一轮”的循环，保留上下文并校验最终输出；设置轮次、工具次数、Token 预算及总时长，防止任务持续空转。
- **失败处理：**通过 @Async 与持久化状态管理任务；对尚未执行工具的临时模型错误最多重试一次，失败返回明确状态；支持取消、输入快照与迟到结果隔离，记录执行轨迹、耗时和用量。
- **岗位对比：**预留搜索来源适配接口，线上提供招聘平台搜索入口；录入完整 JD 后收藏并带入选定版本分析，区分表达评分与关键词覆盖，展示原文依据以辅助核对岗位要求。
- **质量验证：**后端 79 项测试通过，3 项真实模型测试按设计跳过；完成核心流程浏览器回归与前端边界检查，覆盖工具参数、输出限制、输入快照、取消、超时与账号隔离。

万彬宇

RAG LangChain Lab - 知识库检索与调查 Agent

作用与用法：面向本地文档的知识库问答与调查工具。用户导入资料后，通过 CLI / Web 提问、核对引用和检索过程，也可运行固定 RAG 与多步检索 Agent 的评测对照。

github.com/Wanbinyu/rag-langchain-lab

技术栈：Python / LangChain / ChromaDB / BM25 / Cross-Encoder / OpenAI-compatible API / pytest / GitHub Actions

- **增量索引与召回：**以内容哈希与稳定 Chunk ID 跟踪文档变更，仅更新受影响片段；使用 ChromaDB 管理向量，组合 Vector、精确词与 BM25 召回，经 RRF 融合、Cross-Encoder 重排并补充相邻证据。
- **多步检索：**构建"选择检索工具 → 读取原文 → 判断补充检索 → 生成引用回答"的单 Agent 循环，校验参数和工具权限，记录中间结果；证据不足时拒答，控制台展示来源、工具调用、证据状态与耗时。
- **运行约束与恢复：**设置最大步数、超时、Token / 成本预算，区分可重试与不可重试错误并限制次数；持久化步骤、工具结果和检查点，支持任务中断及服务重启后恢复。
- **评测与回归：**分离开发 / 校准集与冻结集，原基线冻结集首次证据决策正确 24/30；比较固定 RAG 与 Agent 的 Recall@K、MRR、质量、耗时、成本及失败类型，将回归接入 GitHub Actions。
- **外部挑战集：**在 609 篇文档、75 个可回答题的 MultiHop-RAG 外部挑战集中，固定检索基线至少命中一条证据 72 题、命中全部证据 33 题；单独分析检索命中但证据不完整的失败样本。
- **性能取舍：**在第二领域 5 篇文档、22 题的本机受控评测中，重排候选从 18 条缩至 6 条，检索 P50 降低约 24%，Recall@3、MRR 与证据决策不变；计时不含答案生成，包含冷启动样本。

工作经历

陕西森豪齐创 | 服务端开发

2026.03 - 至今

- **开发与交付：**使用 Lua 与 Excel 配置参与战令、商城、任务奖励和首领掉落等版本开发，完成 FGUI / Lua 联调；维护源表与生成配置，核对物品 ID、任务进度、变量、怪物映射和奖励格式，处理跨端配置差异。
- **AI 客服与知识库：**接入大模型实现游戏内自动应答，维护问答知识与提示词，异常时兜底并转人工；为公司客服部署 AnythingLLM，导入培训手册、注意事项和常见问题。
- **排障与回归：**维护游戏内人机玩家的行为与配置，处理与真人玩家的交互及跨端联调；结合日志与配置比对定位持久化、怪物刷新、编码和资源引用问题，在文档不完整时还原业务链路，完成发布前回归。

开源贡献

- **主流开源贡献：**Spring Boot 3 项、Gradle 1 项、OpenRewrite 相关仓库 6 项 PR 已合并；涉及配置解析、健康端点路径、构建追踪、XML 缩进与许可证识别，补充对应回归测试。
- **DeepSeek Harness 启动器：**用于便捷安装、启动和管理 Windows 上的 DeepSeek Harness；实现依赖检查、后台启动、托盘管理、脱敏诊断与重复启动请求合并，并维护配套插件。
- **个人陪伴 Agent（在研，私有仓库）：**面向日常聊天与记忆管理，使用 TypeScript / React / Fastify / SQLite 开发可运行的 PC 原型；实现四层记忆、遗忘与巩固、来源追踪和事实纠正，结合短期情绪及"先倾听"等偏好调整回应。编写 DeepSeek 兼容的提取与回复适配器，已有 34 项自动化测试及离线对照，真实模型联调持续推进。

Spring Boot PR | Gradle PR | OpenRewrite PR | dsh-launcher

教育背景

咸阳师范学院 | 计算机科学与技术 · 本科

2021.09 - 2025.06

专业前 10%；"解忧维修铺"微信小程序主要开发，项目获校级铜奖。

墨香阁社区：支持文章发布、检索、延迟发布与异步通知；使用 Spring Cloud、MyBatis-Plus、JWT、Redis ZSet、Elasticsearch 和 Kafka。