

邓思宇 Siyu Deng

AI Agent 应用工程师

178 2607 8582 | ciyudeng@163.com | 南京 | [个人主页 ciyufish.cn](http://个人主页.ciyufish.cn) | [GitHub](#) | npm@moyu-build/council

4 年软件研发经验（前 3 年前端 / 全栈，近 1 年聚焦 AI 应用） | 江苏大学 · 计算机科学与技术 · 本科

个人简介

- 具备从需求分析、系统设计、Agent / RAG 链路开发、前后端集成到测试验收的端到端交付能力，覆盖教育 AI、企业研发自动化、企业智能客服、长期 Context / Memory、推荐匹配和医疗资料整理等场景。
- 擅长把非确定性模型接入可验收的业务流程：以 Tool Calling、结构化输出、证据约束、权限过滤、Trace / Eval、受控降级和 Human-in-the-loop 管理结果质量与业务风险。
- 兼具 React / Vue / TypeScript / Node.js 全栈工程基础与 Python AI 服务开发能力，能够跨前后端定位 Trace、拆解失败组件，并通过自动化测试与 Eval 验证优化结果。
- 习惯按 **Prompt** → **Context** → **Harness** → **Loop** 拆解 Agent 系统，以“推理 → 执行 → 观察 → 校验 → 修复 / 人工接管 → 停止”的有界反馈回路替代不可控的单次生成。

核心技能

- Python / AI 服务**：Python、FastAPI、Pydantic、NumPy、scikit-learn、pytest、Ruff；具备推荐排序、特征处理、行为数据及 API 服务开发实践。
- Agent 架构与编排**：Mastra、Prompt Engineering、Function / Tool Calling、JSON Schema、结构化输出、任务状态机、有界 Agent Loop、SSE、Human-in-the-loop、Critic / Repair Loop。
- RAG / Context / Memory**：文档解析与分块、Embedding、向量与全文混合检索、RRF、Rerank、元数据及权限硬过滤、上下文预算、证据门控、引用校验与拒答；具备长期 Context 采集、蒸馏、持久化和检索实践。
- MCP / Skills / LLMops**：MCP Server、工程 Skills、多 Provider 接入、Trace、Benchmark、自动化 Eval Gate、错误归因、回归验证、Token / 延迟记录、风险分级、人工复核与审计留痕。
- 全栈与数据基础设施**：TypeScript、React、Next.js、Vue、Node.js、Vite、Tailwind CSS、PostgreSQL、Prisma、Supabase、SQLite / LibSQL、Monorepo、RESTful API、SSE、阿里云云效。

代表项目

课灵 AI 教育系统 | 核心 AI 链路研发

技术栈：React、TypeScript、Node.js、Mastra、SSE、Zod、Java、PostgreSQL、Doc2X、MinerU

- 基于 Mastra 组织试卷解析、智能组卷与命题、教师 AI 工具和数学题图等核心能力，将不同模型及业务接口统一抽象为 Agents、Workflows 与 Tools。
- 设计“试卷上传 → 多引擎解析 → 题目切分 → Schema 结构化 → Eval Gate → 受限 Repair → 人工复核 → 导出 / 入库”链路，处理题号、选项、公式及题图归属等复杂问题；全程保留 Provider、耗时、降级原因和质量结论。
- 建立题号覆盖、选项完整、公式风险三类确定性评估器及 High Issue 否决规则；在 **596 份试卷、10,420 题**回归中，规则免人工复核率 **86.8% (9,041 题)**，其余 1,379 题显式进入复核队列，任务超时 0、空结果 0、DOCX 审计 596 / 596 通过。
- 构建对话式组卷工作流，将教师自然语言转换为学段、学科、知识点、题型、难度和配额等结构化约束，支持真实题库召回、草稿隔离、换题、增删题、改分、7 类变题及 Revision 管理。

AI 研发效能与质量闭环中台 | 企业内部项目·方案与开发

技术栈：TypeScript、Python、Claude Code、Codex、阿里云云效 Projex / Codeup / Flow、Figma Plugin API、Vitest、Playwright

- 以阿里云云效为研发事实源，串联缺陷读取、仓库与分支定位、代码修复、质量验证、合并评审和发布跟踪，形成从 Bug 到交付结果的统一工作台。
- 设计 待修复 → 修复中 → 已修复 → 待验证 → 已验证 → 待发布 → 已发布 业务状态机；通过定时任务中台控制任务领取、循环执行、失败重试、超时终止和人工接管，避免 Agent 内部重试污染业务状态。
- 将 PROJECT_CONTEXT、开发规范、排障方法及测试策略沉淀为可复用 Skills，使不同 Coding Agent 按统一步骤完成上下文聚合、影响分析、代码修改和验收检查。
- 开发 Figma 上下文提取插件，采用“截图语义理解+目标化局部结构检索”方案，仅提取目标元素的尺寸、间距、层级、背景及组件关系；典型整页数据由 884 KB 压缩至 60 KB，减少原始节点树噪声。
- 以类型检查、静态检查、单元测试和 Playwright E2E 作为确定性质量门禁，将测试结果、执行日志和代码差异作为修复任务的验收证据。

咖啡客服运营 Agent

技术栈：TypeScript、Mastra、Next.js、PostgreSQL / pgvector、BM25、Chatwoot、Medusa、OpenTelemetry

- 构建“知识问答+业务操作”一体化客服 Agent：Agent 负责意图理解与 Tool 选择，Workflow 承载订单查询、取消、退款申请、支付链接生成及异常转人工等确定性流程。
- 构建企业政策、商品及售后规则的 Hybrid RAG 链路，结合元数据过滤、向量与 BM25 混合召回、RRF / Rerank 和引用生成；价格、库存、订单及支付状态均通过业务 API 实时获取。
- 将 Chatwoot、Medusa 及支付能力封装为类型安全 Tools，在写操作前完成身份鉴权、参数补全、规则校验、二次确认、幂等控制和结果回查，并通过 Trace / Evals 记录完整任务轨迹。

Council 长期上下文记忆与多角色协作系统 | 个人开源项目·架构与开发 | 2026

技术栈：Node.js、TypeScript、MCP、React、Claude Desktop、Cursor

- 针对对话框有限、个人信息难以长期沉淀及不同 AI 客户端记忆相互隔离的问题，设计个人长期 Context 与多角色协作系统。
- 设计 Capture → Distill → Convene → Refine 工作流，将原始对话提炼为带来源证据的结构化记忆，并通过反馈持续修正身份、偏好、关系和长期目标。
- 实现 MCP Server，统一暴露身份、记忆读写、角色 Persona 及多角色协作能力；提供 Web 与 CLI 双入口，支持本地持久化、相似记忆去重、证据回溯和人工编辑。
- 发布开源 npm 包 @moyu-build/council，支持 Claude Desktop、Cursor 等 MCP 客户端接入，形成可独立部署、重视隐私的个人 AI 记忆基础设施。

其他项目

- 直达医路管家 | 医疗资料智能整理与人工审核系统 | 2026：基于 Next.js、Mastra、PostgreSQL、Prisma 构建“资料上传 → 解析 → 结构化提取 → Schema 校验 → 人工审核 → 归档”流程；以 PENDING / APPROVED 状态门禁限制未经确认的信息进入正式档案，并在 Prompt、Workflow、Tool 三层禁止输出诊断和用药建议。
- Bonjour 双边推荐与离线评测实验室 | 2026.07：基于 Python、FastAPI、scikit-learn、SQLite 实现用户画像、冷启动召回、多目标混合排序和多样性重排，并以 Precision@K、Recall@K、NDCG@K 对推荐策略进行可重复的离线评测。

工作经历

课灵 Corling (万卷互联) | AI 应用研发工程师 | 2026.05 - 2026.07

- 独立负责教师端核心 AI 产品链路，覆盖试卷资料资产化、智能组卷与命题、教师 AI 工具及数学题图，推动存量 PDF、扫描卷和拍照试卷进入可解析、可复用、可复核的教研工作流。
- 建立“自动处理+异常分流+人工复核”的质量机制，以确定性规则筛出可直接交付结果，并把低置信度题目显式交给教师确认，兼顾处理效率与教学内容准确性。
- 将教师组卷从多页面手工操作升级为对话式草稿工作流，支持持续修改、版本管理和确认后入库；同时完成 AI 服务与既有 Java 题库、用户、权限及导出体系的协同接入。

江苏南数大数据产业集团 | 高级前端工程师 | 2025.08 - 2026.04

- 负责场馆服务平台管理端、运营端和用户端等多角色产品的工程交付，统一路由、权限、组件及接口规范；主导官网 SPA → Nuxt SSR 迁移，改善首屏体验与 SEO，核心页面加载时间由 **3.8 秒降至 1.2 秒**。
- 发起企业内部 AI 研发自动化中台建设，以阿里云云效连接 Bug、代码、验证及发布流程，用状态机和定时调度管理任务生命周期，推动 AI 从个人编码工具转化为可追踪、可复核的研发流程。
- 沉淀 `PROJECT_CONTEXT`、工程 Skills、测试规范和验收标准，统一团队使用 Claude Code、Codex 等工具时的上下文与执行方式；开发 Figma 上下文插件，降低设计稿进入 AI 开发流程时的噪声和沟通成本。

独立交付 / Freelance | 全栈开发工程师 | 2024.11 - 2025.07

- 独立交付多租户 SaaS、企业官网及运营后台，负责需求澄清、方案拆解、前后端开发、第三方服务接入、部署上线和交付验收。
- 覆盖订阅计费、Google OAuth、邮件 / 短信、Webhook、即时通讯、账号权限和租户隔离等常见业务能力，使用 Next.js、Node.js、Supabase / PostgreSQL、Vercel 完成从需求到上线的 MVP 闭环。

南京溪水网络科技股份有限公司 | 前端工程师 | 2023.05 - 2024.10

- 参与商家工作台及政企业务系统建设，负责权限、流程、复杂表单和数据看板等核心模块，持续跟进需求澄清、联调、上线与迭代维护。
- 参与 Vue2 → Vue3 技术升级，推动 TypeScript、Vite 与组合式 API 落地；维护企业级 CLI、组件库和 VitePress 文档，沉淀 10+ 通用业务组件，降低重复开发成本。

中国移动紫金院 | 前端开发 | 2022.06 - 2023.02

- 参与内部 CRM / Admin 管理后台及碳中和方向数字孪生可视化项目，完成多个业务模块从需求拆解、页面开发、接口联调到交付上线。

教育背景

江苏大学 | 计算机科学与技术 | 本科 | 2018 - 2022 | CET-6

- 高中开始接触算法并参加 NOIP；具备数据结构与算法基础，持续关注 LLM、Agent、RAG、推荐系统和 AI for Science。