匠人学院
JR ACADEMY

2026 技术背景版

把 AI 做进企业业务

FDE 实战入门

面向有技术基础的开发者，从岗位认知到第一个企业 AI 交付项目。跟做知识助手，留下方案、代码、评估与交接材料。

岗位认知　跟做项目　工程交付

从一个业务问题，做出可用的原型



jiangren.com.au · 12 章指南 / 本地 AI 实战 / 完成示例

AI 来匠人 ✦ FDE FIELD GUIDE ✦ JIANGREN.COM.AU ✦ 学 AI 来匠人 ✦

CONTENTS

先认识岗位，再完成一个项目

01 认识 FDE 与企业需求 3

02 岗位与薪资：澳 → 美 → 中 6

03 技能准备与项目地图 9

04 客户调查与范围 11

05 方案选择与预算 13

06 架构、数据与工具 15

07 跟做本地 AI 知识助手 17

08 评估与失败分析 21

09 生产准备与发布 23

10 成本与用户采用 25

11 变更、验收与交接 27

12 作品集与继续学习 29

附录 A 文件导航与术语 31

附录 B 情境练习 33

附录 C 来源与学习资源 34

讨论你的 FDE 学习路线

扫码咨询课程与项目训练，说明你已有的开发经验和想补齐的交付能力。

01 / 认识 FDE

FDE：把技术带进客户的工作现场

Forward Deployed Engineer，常译为前沿部署工程师或前线部署工程师。

你要解决什么问题?

客户已经有模型或产品，但还需要有人理解业务、接通数据和系统、验证结果、推动使用。FDE 在这些环节中动手实现，也与客户一起决定做什么、做到什么程度。

常见岗位	通常更关注什么
软件工程师	产品功能、系统设计、工程质量与稳定性。
AI Engineer	模型驱动的应用、检索、工具调用、评估与成本。
解决方案工程师 / 架构师	技术可行性、方案设计与验证；部分岗位也承担生产交付。
FDE	客户场景中的问题定义、代码与集成、上线和使用反馈，并把共性问题反馈给产品团队。

例子：客户想做一个知识助手

先问员工怎样找政策、谁维护资料、哪些人能看，再决定是否采用检索增强生成（RAG）：先找出相关资料，再让模型依据资料回答。之后还要证明答案能支持实际工作，并交给有人维护的团队。

> 这些是工作重心，不是统一的职位标准。本书聚焦企业 AI 应用交付，读者需要已有代码、API 和基础部署经验。

来源：S1 OpenAI | What FDE means at OpenAI · S13 OpenAI | Sydney FDE

01 / 企业为什么需要 FDE

演示能回答，业务却还没用起来

沿着同一个企业政策查询场景，看 FDE 的工作怎样出现。

企业遇到的情况	FDE 要推动的具体决定
演示用十份干净文件，真实资料混有旧版	确认资料负责人、版本规则与更新流程。
演示账号能看全部资料，员工权限不同	接入可信身份，把访问控制放进检索路径。
回答很流畅，业务方仍不敢使用	一起定义正确答案、失败情形与人工核对方式。
原型完成了，没人负责故障与更新	明确运行责任、上线条件、交接与支持范围。

一个公开案例提供的启发

OpenAI 发布的 Morgan Stanley 案例介绍了专家参与评价与持续检查。这里值得借鉴的是让业务专家参与判断，而不是直接照搬别人的系统。本书用虚构的 Northstar 项目练习这件事。

怎样理解岗位前景

当企业需要连接旧系统、处理数据权限、衡量任务效果时，这类交付能力有用。标准化工具也会吸收部分配置工作，因此长期要积累工程判断、行业理解和协作能力。这是基于工作需求的分析；具体机会再看下一章的澳洲、美国和中国招聘样本。

来源：S1 OpenAI | What FDE means at OpenAI · S2 OpenAI | Morgan Stanley 案例

01 / 职责拆解

一项需求，要经过哪些人和决定?

FDE 需要把工程工作和客户决策接起来，也要说清责任边界。

工作阶段	FDE 的动作	留下什么
需求不清	观察用户流程，核对假设与限制	需求记录、范围与验收人。
选方案	比较替代方案，说明代价与依赖	方案建议书、架构决定。
实现与验证	编写集成代码，调查失败	可运行原型、测试与失败记录。
上线与使用	组织发布、培训和交接	运行手册、试点反馈、产品建议。

谁做决定?

业务负责人批准规则、范围和收益口径；资料负责人确认内容与版本；IT 与安全负责人批准接入；项目负责人安排资源。FDE 提出方案、实现控制并报告风险，不能替这些角色默默签字。

看 JD 时，确认实际重心

有的岗位偏系统部署，有的偏 Agent 配置和工作流，有的要求承担客户架构与技术领导。面试时问清：是否负责生产代码、同时支持几个客户、怎样验收、谁决定追加需求。

> 本书练习：面对“我们需要一个 AI 助手”，先写出用户、资料负责人、批准人和三个待验证问题，再开始选模型。

来源：S10 天云融创 | 南开大学就业网招聘信息 · S11 凡泰极客 | FDE 前沿部署工程师 · S15 Anthropic | Staff FDE

02 / 澳洲

澳洲：从真实招聘看工作要求

以下按“岗位、工作、经验、薪资”阅读。金额保留澳元，核查于 2026–11–18。

招聘样本	工作与经验要求	公开薪资
OpenAI · FDE 悉尼	负责客户系统交付；5 年以上工程或技术部署经验，包含面向客户的工作。	未披露。
PROFOUND 招聘机构 · FDE 悉尼	连接数据与系统、开发可用界面；约 3 年以上生产软件经验。	**AUD 200,000–300,000 / 年**；另提股权，super 是否包含待确认。

读懂岗位条件，再看数字

OpenAI 岗位写明每周 3 天到办公室，出差最高可达 50%。PROFOUND 广告页头标 Remote，但正文要求现场工作；这类信息冲突需要在投递前确认。招聘机构的客户名称没有公开，不把它写成某家企业的已核实岗位。

怎样准备澳洲岗位的材料

建议用一个项目说明：你怎样理解客户流程、写了什么代码、怎样比较方案、上线后如何处理问题。把个人贡献和团队成果分开，再核对具体岗位的工作地点与经验要求。

> 三地统一阅读口径：都是特定岗位的招聘样本。薪资不代表行业平均；基本工资、奖金、股权与 super 分开核实，招聘状态以原页为准。

来源：S13 OpenAI | Sydney FDE · S16 PROFOUND | 悉尼 FDE 招聘（SEEK）

02 / 美国

美国：薪资与职级一起看

以下两个样本都要求较强的工程与客户经验，金额保留美元。

招聘样本	工作与经验要求	公开薪资
OpenAI · FDSWE 旧金山	面向客户开发全栈系统、制定交付计划；7 年以上全栈经验。	**USD 185,000–325,000 / 年**基本工资；股权另列。
Anthropic · Staff FDE 纽约 / 旧金山 / 西雅图	客户技术策略、架构、生产交付与技术指导；Staff 职级。	**USD 280,000–320,000 / 年**；保留原标签 Annual Salary。

同样是 FDE，不一定是同一级别

Staff 样本包含跨工作流技术判断和客户高层沟通。OpenAI 的 FDSWE 更强调亲自开发全栈实现。比较机会时，先看岗位责任和自己已有的证据，再看工资区间。

一份可比较的 Offer 记录

分别填基本工资、保证奖金、浮动奖金、股权、福利、地点和出差安排。OpenAI 样本提到最高 50% 出差，Anthropic Staff 样本提到 25%；这些也影响实际工作体验。

> 本页 Annual Salary 不改写成“含股权总包”。这些样本也不能用于推断澳洲或中国同名岗位的薪资。

来源：S14 OpenAI | San Francisco FDSWE · S15 Anthropic | Staff FDE

02 / 中国

中国：同名岗位也要看实际交付

金额保留人民币。企业官网与招聘平台样本采用同一套阅读字段。

招聘样本	工作与经验要求	公开薪资
中兴新云 · FDE 北京 / 上海 / 成都等	2027 届校招；业务场景、AI 方案与跨团队交付。	未披露。
帆软软件 · AI FDE 杭州	平台展示 Agent 项目交付；经验不限，本科。	**20—40K·14 薪。**
阿里云 · 大模型 FDE 杭州	平台展示客户大模型应用落地；3—5 年，本科。	**40—70K·16 薪。**

看懂月薪与薪数

K 表示千元。20—40K × 14 的名义年薪是 28—56 万元；需要确认额外薪数是否保底、绩效条件和固定工资范围。这个乘法不能直接表示实际到手收入。

职责还有哪些差异?

凡泰极客的深圳岗位偏部署、集成与运维；天云融创的招聘材料包含 Agent 配置、知识组织与评测。找工作时可以扩大职位关键词，但仍要逐项核对自己会承担什么。

帆软和阿里云数字来自 BOSS 列表；详情页触发访问验证，招聘人身份及薪资兑现条件未核实。天云融创页面的金额未标周期，未纳入本页薪资比较。

来源：S9 中兴新云 | 2027 届 FDE 校园招聘 · S12 BOSS 直聘 | 杭州岗位展示 · S11 凡泰极客 | FDE 前沿部署工程师 · S10 天云融创 | 南开大学就业网招聘信息

03 / 技能准备

从已有经验出发，补齐交付短板

下面是学习建议。选择真实目标岗位后，再决定需要补到什么程度。

你的基础	可以迁移的优势	优先补齐
后端 / 全栈	接口、数据模型、业务系统开发。	需求发现、AI 评估、范围和客户验收。
前端	交互、用户体验、工作流界面。	服务端、身份权限、数据接入与发布。
数据 / 算法	数据质量、实验与指标分析。	应用工程、业务接口、可靠性与使用反馈。
云 / 运维	环境、部署、排障和监控。	应用开发、模型行为评估、业务方案表达。

先做一次小规模交付演练

选一个自己理解的流程，找使用者验证问题；实现一条最小链路；保留五种失败情形；找另一位工程师按文档运行。最后讲清楚为什么这样做、哪里还不能上线。这比堆满工具名更有助于展示能力。

职业发展可以怎样延伸?

喜欢系统深度，可继续做高级 FDE 或架构工作；喜欢行业问题，可发展行业解决方案；喜欢跨项目协作，可尝试交付负责人；喜欢提炼共性需求，也可以转向产品工程。这些是能力延伸方向，并非所有公司都有固定晋升路径。

接下来进入工程实战：第 4–6 章形成方案，第 7 章运行离线实验，第 8–11 章补评估、发布和交接。每读完一部分，留下能被别人检查的材料。

03 / 项目地图

跟着 Northstar，完成一轮交付练习

Northstar 是虚构企业；角色、预算和试点计划是教学设定。代码与实跑记录单独标明。

业务目标与第一版边界

计划让运营和财务员工查询现行政策，核对原文，并知道何时需要人工帮助。首期只读，不自动批准报销。未来试点设为 40 人；本地程序只使用 4 个测试身份和少量教学文档。

章节	你会得到的成果
4–5 · 需求与方案	填好的需求记录 D01、方案与预算 D02。
6–7 · 架构与实现	架构记录 D03、权限实验、实际调用本地模型的问答程序。
8–10 · 验证与试点	实跑报告、验收矩阵 D04、发布记录 D05、测量计划 D06。
11–12 · 交接与展示	变更单 D07、接手步骤 D08、作品集说明 D09。

先学会读这几个词

PoC 是验证想法是否可行的小原型；ACL 是谁可以访问哪些资料的规则；Agent 会按任务调用工具或决定下一步；评估（Eval）用一组问题检查实际行为。其余术语随用随讲，附录提供速查。

配套包中 templates/ 是空白模板，worked-example/ 是填好的教学示例，reports/ 是实际运行记录。教学计划中的客户审批、用户收益与生产上线均未发生。

04 / 客户调查

先观察工作，再整理需求



一次访谈的成果应该能改变方案，而不只是增加会议纪要。

找谁	问什么	记录什么
一线用户	最近一次查政策，从哪里开始？最后找了谁？	操作步骤、等待时间、无法继续的原因
业务负责人	什么结果值得继续投入？哪些错误不可接受？	业务目标、基线、验收负责人
资料负责人	现行版本在哪里？旧版如何撤回？	资料目录、版本、更新与撤回责任
IT / 安全	身份从哪里来？接口和数据访问由谁批准？	接入条件、环境、权限、审批依赖

具体问法

把“你们查资料是不是很慢”改成“请演示昨天那次差旅政策查询”。记录用户在搜索、阅读、询问同事、核对版本上分别用了多久。等待别人回复的时间，和用户实际操作的时间分开记。

事实与假设分栏

事实：资料负责人确认有两个现行目录。假设：目录里的所有文件都适合入库。待验证：嵌入的表格能否正确解析。每条待验证项必须有负责人和完成日期，不能在方案建议书（Proposal）中悄悄写成已经具备。

本页交付：需求调研摘要（Discovery Brief）。至少写清一个具体流程、一位最终用户、一位批准人和三个可能改变方案的未知项。

04 / 范围与基线

把“提高效率”写成可检查的目标

先测现在如何工作，再提出目标；不要用演示速度代替业务收益。

一份可执行的基线采集方案

教学建议：选择正常、复杂、无答案三类政策问题，邀请运营与财务用户完成代表性任务。记录任务类型、开始与结束时间、答案是否正确、是否求助、重做时间。测试样本尽量覆盖不同熟练度；保留原始记录，不只报平均值。

范围项	首期决定	批准条件
现行政策	纳入已批准目录	资料负责人确认版本和权限
历史档案	暂不纳入	访问获批、预算与测试更新后再讨论
自动审批	不纳入	未来独立评估业务权限和写入风险
两个部门	分批启用	各部门有负责人和验收记录

写一条能验收的需求

“运营用户询问差旅政策时，系统返回现行版本的依据和引用；找不到可用依据时不猜测，提示联系负责人。”把“回答准确”拆为事实支持、版本正确、权限正确和下一步清楚，才能知道问题出在哪。

本章成果 · D01 需求记录

打开 worked-example/01-discovery.txt。已写出运营与财务的不同权限、三条本地需求，以及仍待客户验证的耗时和资料责任。练习时把其中一个假设换成你观察到的事实。

05 / 方案选择

不是每个问题都需要 Agent



先用最小的方案验证业务价值，再为确有必要的能力增加复杂度。

路线	适用前提	本案例判断
整理目录与传统搜索	主要问题是资料混乱或版本不可辨	必须先做，可作为比较基线
现成企业 AI 工具	现有权限、数据与功能条件满足需求	先验证接入与治理约束，再决定是否采购
定制 RAG 工作流	回答需要企业文档、引用与权限控制	首期候选：流程可预测、以只读为主
自主 Agent	任务路径变化大，需要动态选择工具	首期不需要；若扩展动作再单独评估

用相同问题比较候选方案

让相同身份在相同文档版本下完成相同任务。比较质量、总处理时间、维护工作、权限实现和完整成本。某个模型在通用榜单上的表现，不能替代你的业务验收。

一条架构决定记录

决定：首期只读 RAG。理由：主要任务是查找现行政策，不需要修改业务系统。替代方案：先改进目录检索。重新评估条件：用户确实需要跨系统执行动作，且动作权限与责任已明确。

> Anthropic 的公开文章区分预定义工作流与动态 Agent，并建议从简单方案起步。本页选型表是针对教学案例的应用。

参考：S3 Anthropic | Building effective agents。

05 / Proposal 与预算

让客户批准一个具体范围

方案要让业务、IT 和预算负责人都知道自己批准了什么。

Northstar 首期 Proposal 摘要 | 教学设定

目标：验证政策查询是否能降低查找与核对负担。范围：两个部门、已批准的现行资料、只读问答与引用。交付：原型、权限与评估记录、受控试点、运行手册。依赖：身份接入、资料批准、业务专家参与。接口延期时先完成离线验证，不承诺按原计划上线。

教学工作量	小时	按 A$120 / h 计算
调查与方案	16	A$1,920
开发与集成	48	A$5,760
评估与修复	24	A$2,880
试点与交接	16	A$1,920
小计 / 20% 准备金	104	A$12,480 / A$2,496
合计	—	A$14,976

上述单价和工时仅用于演算，不是市场报价；税费、采购费用与持续运行费用未含。资料清理超出假设、增加写入能力或审批延期，都应重新估算。

本章成果 · D02 方案摘要

打开 worked-example/02-proposal.txt。它把本地原型与未来企业试点分开，列出替代方案和批准条件。A$14,976 是试点预算演算，不是几个脚本的报价；运行与采购费用需另列。

06 / 系统架构

让权限成为数据路径的一部分



先确认谁能看什么，再决定把哪些内容交给模型。

已验证的用户会话
- → 服务端解析租户、用户、组与权限版本
- → 检索前限制允许访问的现行文档
- → 排序与选择证据
- → 生成带引用的草稿（或无依据转人工）
- → 检查引用范围与答案支持程度
- → 返回结果，并记录最少必要的运行信息

三条需要画在架构图上的边界

身份服务负责证明用户是谁；检索服务负责限制能取出的内容；模型只处理已经允许进入上下文的资料。不要把前端传来的 department 当成可信权限，也不要指望提示词替代访问控制。

引用链接也要验证权限

回答显示“见文档第 3 页”并不意味着文档可公开下载。打开原文件时仍应校验用户身份和权限。日志、缓存、错误信息与历史会话同样可能泄露内容，必须纳入设计。

本页交付：画出数据流，并在每条跨系统连线上写明身份来源、传输内容、授权方式与失败行为。微软公开样例的 ACL 文档可作为深入阅读。

参考：S6 Microsoft | 登录与文档访问控制。

接入之前，先约定失败时怎么做

接口契约既要定义成功结果，也要定义重试、撤回和不确定状态。

字段 / 约束	用途
document_id / version / status	识别资料、区分现行与已撤回版本
tenant_id / allowed_groups	限定可访问范围；缺失授权信息时默认不进入检索
owner / source_url / updated_at	追溯负责人、原文与更新状态
request_id / timeout / error_code	定位一次调用，区分超时、无权限与服务故障

MCP 与工具调用放在哪里

MCP 是连接模型应用与外部工具的一种协议。它和工具调用可以帮助接通系统，业务权限仍需自行设计。工具入口仍需校验当前身份、允许的参数和操作范围。首期提供只读查询；将来的写入能力，应具备审批、审计和重复执行保护。

写操作超时，不要盲目再执行

例如创建采购申请后响应超时，服务器可能已经成功创建。先用幂等键（标记同一业务操作，避免重复执行）或业务标识查询状态，再决定重试。只有重复执行不会新增副作用的操作，才适合自动重试；重试次数与总时长都要有限制。

本章成果 · D03 架构记录

打开 worked-example/03-architecture.txt：身份过滤发生在检索前，模型只接收一份获准资料。引用 ID 与原文一并返回；接下来用程序观察这条路径。

07 / 第一步：权限与版本



先验证哪些资料可以进入模型

配套实验用确定性检索隔离问题，让你先验证权限与版本逻辑。

解压 fde-tech-companion.zip，在 companion 目录打开终端。需要 Python 3.9 或更新；只使用标准库，无需 API 密钥或云资源。

```
python3 lab.py --user ops_alice --question "差旅申请"
python3 lab.py --user ops_alice --question "关账"
python3 lab.py --user fin_bob --question "关账"
python3 evaluate.py
```

请求	预期行为
运营用户查差旅	answered；引用 travel-v2
运营用户查关账	no_evidence；不返回财务内容或文档标识
财务用户查关账	answered；引用 finance-close
运行评估	passed = 12，total = 12；退出码 0

这一步确认了什么

实验让每条结果可以重复，便于观察权限是否真的在检索前生效。它不做语义理解或生成，不能代表完整 RAG 的效果。后续接入模型时，仍需要保留这些边界检查，并增加答案事实支持、提示注入和模型失败测试。

> 重要：--user 是可信本地测试身份。生产应用必须从服务端验证后的会话读取身份，不能让浏览器任意指定用户名。

07 / 第二步：启动真实模型

把小模型放在本机运行

Python 程序只用标准库；模型推理由另行安装的 llama.cpp 提供。

准备环境

需要 Python 3.9+、llama.cpp 的 llama-server 命令，以及 Qwen2.5-0.5B-Instruct 的 Q4_K_M 文件。它是较早的小模型，选它是为了降低教学资源需求。模型文件约数百 MB，配套 ZIP 不含权重。安装入口、固定版本和下载校验步骤见 README.txt。

```
python3 download_model.py
llama-server -m models/qwen2.5-0.5b-instruct-q4_k_m.gguf \
  --host 127.0.0.1 --port 8089 \
  --alias northstar-qwen -c 2048 -ngl 0 -t 4
```

命令里的参数是什么意思？

-m 指向模型文件，--alias 给客户端一个固定名称，-c 是上下文长度，-ngl 0 用 CPU 推理。服务只监听本机地址。保留这个终端，在另一个终端进入配套包的 companion 目录运行 Python。

> 首次准备需要下载公开模型；准备完成后推理在本机进行，不需要付费 API 密钥。本书的实际机器、版本与哈希以 reports/runtime.json 为准。

来源：S17 llama.cpp | 本地模型服务与发布版本 · S18 Qwen | Qwen2.5-0.5B-Instruct-GGUF

07 / 第三步：生成与核对

从一条问题走到可核对的草稿

ai_app.py 已实现模型请求、超时处理、输出解析与引用检查，可以直接运行。

```
python3 ai_app.py --user ops_alice \
  --question "差旅申请需要先做什么？"
python3 ai_app.py --user ops_alice \
  --question "关账检查码是什么？"
python3 evaluate_live.py > my-live-report.json
```

程序步骤	你可以检查的结果
读取本地教学资料	使用 data.json；先根据可信测试身份过滤权限和版本。
调用实际模型	仅把命中且获准的资料送到本机模型服务。
检查模型输出	JSON 格式、完成状态和引用 ID 必须符合规则。
返回待核对的草稿	status=draft；citations 中包含 ID、版本和原文，逐句核对。

读报告时先看两条路径

运营问差旅应得到现行资料的草稿；运营问关账应得到 no_evidence，且 model_called=false。第二条请求在调用模型之前就被挡住，不能靠提示词让模型自行判断权限。

本次实跑：2026-11-18

差旅回答：“差旅申请应在预订前提交给直属经理审批。”引用 travel-v2。6 个教学用例符合预期，其中 3 次实际生成；租户 B 的回答出现了重复措辞，事实仍与原文一致。完整结果见 reports/live-model.json。

> 引用检查证明来源在允许范围内，不证明每句话正确。draft 表示尚需人工核对；模型生成内容以你当次运行结果为准。

07 / 第四步：让失败也可观察

模型关掉以后，系统应该怎么回答?

用真实停机和可重复的故障注入，观察两种不同的失败来源。

真实停机演练

在模型服务终端按 Ctrl+C，再运行上一页的差旅查询。应得到 service_unavailable，并提示稍后重试或联系负责人。重新启动服务后再问一次，确认恢复。完整记录保存在 reports/failure-drill.json。

```
python3 -m unittest -v test_ai_app.py
python3 evaluate.py
python3 evaluate_live.py > my-live-report.json
```

输出状态	表示什么
no_evidence	没有可用依据；本例不会调用模型。
needs_review	资料多份匹配，或生成格式、引用、完成状态未通过。
service_unavailable	模型服务连接失败、超时或服务响应不可解析。
draft	已获得生成结果，仍需对照原文核查事实。

本章成果：一条实际调用模型的链路、真实回答报告和失败记录。test_ai_app.py 使用模拟响应测试控制逻辑；evaluate_live.py 才调用实际模型，两者的结果分开记录。

08 / 评估设计

把“看起来不错”拆成可判定的问题



不同的错误需要不同的测试；一个总分解释不了所有风险。

维度	测试方法	教学验收示例
访问边界	跨部门、跨租户、撤权、缺失 ACL	关键检查全部通过；出现泄露立即停止
答案依据	逐条核对事实与引用是否对应	业务负责人按约定量表评分
版本	旧版撤回后再问相同问题	只使用现行资料；冲突转人工
拒答与转人工	无资料、模糊问题、证据冲突	返回正确状态与后续路径
成本与延迟	记录每次调用及复核工作	在项目约定预算和体验范围内

测试集要与开发材料分开

开发集用于改提示词和调检索；验收集由业务方确认并保留。看到验收失败后反复调到通过，相当于把它变成开发集，需要新增独立样本。对重复运行不稳定的任务做多次试验，不只展示最好的一次。

黄金答案未必是一句话

可以记录必须包含的事实、禁止出现的内容、允许的引用、预期工具状态和应转人工的条件。用代码检查权限与结构，用业务专家判断内容是否支持实际工作；模型评分也要抽检。

参考：S4 Anthropic | Demystifying evals for AI agents。

08 / 失败分析

先找到坏在哪一层，再改系统

下面是教学故障清单，不是任何真实客户的事故报告。

观察到的失败	优先查哪里	修复后补什么测试
政策明明存在却找不到	解析、索引、关键词与权限范围	该文档的代表性查询及相似负例
回答引用了旧政策	索引刷新、状态过滤、缓存版本	撤回后下一次请求与缓存命中路径
引用存在但结论不支持	上下文截断、生成、事实核对	逐事实评分，不只验证链接有效
换部门后仍看到旧答案	权限缓存、会话与历史消息	撤权后的检索、缓存与历史访问
多次请求得到重复操作	超时与幂等控制	响应丢失后重试与状态查询

实验中的 12 项检查意味着什么

配套 evaluate.py 的 12 项检查覆盖身份、权限、租户、版本、缺失依据、歧义与撤权。evaluate_live.py 另行记录实际模型回答；不能把两组结果合并成生产质量得分。离线检查通过只说明这些确定性样例满足预期；本次模型报告也仅覆盖少量教学问题。生产可用性和业务收益需另行验证。

一份失败记录至少写这些

输入与身份、期望结果、实际结果、相关版本、最小复现步骤、影响范围、原因假设、修复方式和回归样例。运行日志里的“成功返回 200”不能代替任务是否成功。

本章成果 · D04 验收矩阵

打开 worked-example/04-acceptance.csv。N01 对应现行差旅，N02 对应越权拦截，N04 对应模型停机。每项都指向报告或测试；真实身份和业务收益保留为待完成。

09 / 生产准备

上线前，验证系统失效时的行为



先让故障可见、可控、可恢复，再扩大用户范围。

检查点	你应该展示的证据
身份与密钥	服务端验证会话；最小权限账号；无密钥写入代码或浏览器
可观测性	request_id 串起检索、模型和工具；日志按需脱敏并限制访问
超时与限额	单步超时、总时长、并发与预算上限；明确限流提示
恢复	旧版本可回退；索引快照与配置一致；有演练记录
数据生命周期	新增、更新、撤回、删除可传播到索引、缓存和派生数据
人工接管	问题能到达有人处理的队列，责任人与响应方式明确

设计一个故障演练

教学场景：检索服务不可用。用户应看到暂不可用及求助方式，不能把超时伪装成“没有这项政策”。日志记录故障类型，但不把内部堆栈直接发给用户。恢复后验证资料版本与权限仍正确。

本页交付：上线清单与证据位置。每项标记已通过、未通过或不适用，并写明理由、负责人和复核日期。没有证据的勾选不算完成。

09 / 发布与回滚

发布的是一组版本，不只是代码

代码、模型、提示词、索引和权限配置共同决定系统行为。

建立发布清单

为一次试点保留代码提交号、模型标识、提示词版本、资料清单、索引版本、权限配置和评估报告。变更其中任何一项，都要判断哪些测试需要重跑。模型升级不应直接等同于可以扩大范围。

阶段	进入条件	退出 / 停止条件
隔离验证	模拟资料与测试身份可用	边界检查失败则修复后重测
小范围试点	业务、IT 与资料负责人确认	出现越权或关键错误即停用相关路径
扩大使用	质量、成本、维护与使用反馈达标	收益无证据或维护负担过高则缩减
版本回退	已验证旧版与数据仍兼容	回退也需重查权限与现行资料

回滚不能把已撤回资料重新放出来

恢复旧代码时仍应执行当前有效权限和资料状态。代码回退、数据恢复和安全撤权是不同动作，运行手册（Runbook）要分别说明。不要让“退到昨天”成为重新暴露敏感信息的操作。

本章成果 · D05 发布记录

打开 worked-example/05-release.txt。模型、运行器和文件哈希记录在 reports/runtime.json；按同一组版本重新启动，再验证差旅回答和拒绝访问用例。

10 / 成本与效果

先算每次成功任务，再谈节省



模型账单只是成本的一部分，人工复核与维护也需要计入。

```
模型成本 = 输入量 × 输入单价 + 输出量 × 输出单价
月运行成本 = 模型 + 检索/存储 + 监控 + 维护 + 复核
每次成功任务成本 = 当期总成本 ÷ 成功完成的任务数
净释放工时 = 基线操作工时 − 新流程操作工时 − 新增复核工时
```

一个可复算的教学例子

假设每月 1,200 次查询，平均每次少用 3 分钟：释放 60 小时。若每次另需 1 分钟复核，净释放 40 小时。按 A$50 / h 的内部估值计算，产能价值为 A$2,000。若月运行成本假设为 A$900，价值扣运行成本为 A$1,100；这不是实际现金利润。

什么情况下不能说省钱

工资没有减少、工时没有转向有价值工作，就不能把所有释放时间写成现金节省。还需考虑一次性实施成本、试点样本偏差、返工和未被覆盖的任务。真实采购时，用当前供应商报价替换教学假设。

> 不要用一个快速成功请求代表全部体验。记录中位数、较慢请求表现、错误率与人工升级率，按任务类别比较。预算模板提供公式和待填字段。

让试点回答三个业务问题

用户会不会用、能不能完成任务、值得不值得继续投入。

要回答的问题	采集方式	避免的误读
用户为什么不用	观察真实工作、访谈未使用者	登录人数不能解释使用障碍
是否帮助完成任务	记录完成率、复核时间与改写量	生成一段文字不等于任务完成
是否值得扩大	结合成本、质量、维护与反馈	次数增长不等于业务价值增长

培训围绕具体工作

让运营用户完成一次差旅查询：查看引用、核对版本、识别没有依据的结果、提交反馈。让财务用户体验权限边界和人工升级。培训时直接展示已知限制，不用一串顺利问答代替操作训练。

试点结论允许“暂缓”

如果使用者愿意用，但资料版本持续混乱，先修资料治理；如果效果仅在少数任务成立，缩小范围；如果成本和返工抵消收益，提出停止或换路线。停止条件应在试点前约定，不能等结果不好才改口径。

本章成果 · D06 试点测量计划

打开 worked-example/06-pilot-plan.csv。运营流程记录搜索与复核时间，财务流程还检查权限。该文件是待执行计划，尚无真实客户采集值；本页金额也只是演算。

11 / 变更与沟通

预算削减时，先改承诺再改代码



客户变化本身并不可怕，未记录的变化才会让验收失去共同标准。

教学情境 | 客户希望把开发预算降到 A$10,000

原预算 A$14,976 包含验证与交接。不能简单删掉测试。可讨论先只覆盖一个部门、减少资料范围、分阶段接入身份系统，或先改善现有搜索。每个选项都应重新说明交付能力、依赖和不能做的事。

变更单字段	本案例要填什么
触发原因	预算变化的时间、提出人与批准人
影响	范围、工时、验收、运行费用、风险与时间
选项	缩小首期 / 分阶段 / 暂停；各自代价
决定	批准的方案、明确排除项、负责人及日期
跟进	更新 Proposal、计划、测试与用户说明

一段可直接改写的客户表达

“按新预算，我们建议首期只覆盖运营政策，保留权限验证和交接。财务资料下一阶段评估。这样可以验证查询价值，但本期无法证明跨部门推广效果。请确认是否接受这个范围，以及谁负责资料批准。”

本页交付：一份变更单。若客户没有批准新范围，不能把删掉的能力默认变成双方共识。

让接手的人独立完成一次恢复

把代码交出去只是交接的一部分，运行责任也需要有人接住。

材料	接手者需要完成的动作
系统与配置说明	找到服务、环境、权限入口与配置来源
资料更新流程	新增一份获批政策，撤回旧版并验证查询
评估与发布记录	运行验收检查，解释失败项并决定是否发布
Runbook	定位故障、暂停功能、恢复服务、核对权限
支持与责任表	知道谁负责资料、代码、费用、审批与事故沟通

交接演练 | 研发人员暂时不操作

接手者按文档处理一次“旧政策仍被引用”的故障。观察他能否识别索引与缓存状态、找到当前资料负责人、恢复正确版本并留下记录。卡住的位置就是文档或工具需要补充的地方。

明确已知限制与剩余责任

列出未接入资料、尚未验证的高风险任务、依赖的外部服务、维护工作和支持时间。验收矩阵逐项关联最初承诺与实际证据；尚未通过的项目要明确处置，不能藏在代码注释里。

本章成果 · D07 / D08 变更与交接

示例文件分别写出缩小范围的建议和接手命令。把模型服务关掉再恢复，让另一位工程师按 D08 操作；记录实际卡住的位置后再确认交接完成。

12 / 复用与作品集

把项目经验留下来，下一次少重做



复用前先区分通用能力与客户特有规则。

适合复用	需要重新确认
连接器接口与错误处理	客户 API、限额、身份与允许操作
评估运行器与报告结构	客户问题、业务规则与验收阈值
发布清单与 Runbook 骨架	实际环境、责任人、恢复步骤
权限过滤测试模式	租户边界、部门关系、撤权机制

把现场问题反馈给产品

例如：多个客户都需要验证撤权是否传播到缓存。整理最小复现、预期行为、影响范围、现有绕行方式和共同需求，提出可配置的能力。避免把某个客户的特殊流程硬编码成所有客户的默认值。

作品集按证据组织

展示客户问题、你的职责、方案取舍、架构、关键实现、失败与修复、测试结果及未验证范围。模拟项目明确标注模拟；团队项目解释个人贡献。截图之外保留可运行材料，让别人能够复核。

> 一句可信的项目描述：“完成教学知识助手的权限过滤、本地模型接入与引用检查，保留离线测试和真实生成记录；尚未验证生产身份与真实用户收益。”

用四轮练习形成一个完整项目

以下是自学练习安排，不承诺固定时间内达到某个岗位要求。

练习轮次	完成什么	自查问题
第 1 轮：定义问题	访谈模拟、流程图、试点范围、Proposal	为什么值得做？谁批准？哪些仍是假设？
第 2 轮：实现关键链路	运行权限实验，新增部门与测试	身份来自哪里？无权限和无依据如何处理？
第 3 轮：接近真实环境	运行 ai_app.py 与实际模型，再扩展一条企业接口	哪些已实测？费用如何记录？失败怎样恢复？
第 4 轮：评审与交接	演示、变更单、验收矩阵、交接演练	换个人能否运行？什么条件下应该停止？

工程评审的六个追问

为什么选这个范围？为什么需要这个架构？如何证明没有越权？模型或接口失败时会怎样？完整成本怎么算？你还没有验证什么？用材料回答，不用“应该没问题”结束讨论。

何时继续读官方培训

你能解释一个端到端项目后，再按缺口选 API、模型、评估、治理或推广课程。完成课程与通过业务验收是两件事；学习记录可以展示，实际效果必须来自项目证据。

本章成果 · D09 作品集说明

打开 worked-example/09-portfolio.txt。先写清跟做部分，再列自己的改动、失败和新增测试。作者的模型实跑报告可用来对照，不能替代你自己的运行记录。

附录 A / 配套文件

空白模板、完成示例和实跑记录

解压 fde-tech-companion.zip，先读 companion/README.txt。

文件 / 目录	用途
lab.py / data.json	确定性权限实验与教学资料。
ai_app.py	读取获准资料，调用本地模型并检查引用。
evaluate.py	12 项离线边界检查。
test_ai_app.py	模拟失败响应，检查错误处理与输出控制。
evaluate_live.py	实际模型的少量教学用例，输出当次回答。
worked-example/	D01–D09 填好的项目材料。
templates/	5 份可编辑空白模板。
reports/	作者实跑报告、版本与哈希；不代替你的测试。

完成示例怎么用?

先对照书中决定，再看 D01–D09 的完整文件。把预算、范围与责任替换成自己的项目条件；客户审批、试点测量和交接确认需要实际执行，不能因为示例填好了就视为已经完成。

CLI 中的 --user 是可信教学身份。它不能原样变成 Web API 的登录参数；生产身份必须由服务端验证后的会话提供。

附录 A / 术语速查

读到陌生词，先理解它在项目里做什么

用同一个知识助手场景解释，避免只记缩写。

术语	在本项目中的含义
RAG · 检索增强生成	先查获准资料，再让模型根据资料回答；检索不一定必须用向量库。
Agent · 智能体	根据任务调用工具或决定下一步；本书的首期问答采用固定流程。
PoC · 概念验证	小范围验证技术或业务假设，成功后还需要试点和验收。
ACL · 访问控制列表	说明谁能访问哪份资料，本例用租户与组来表达。
Eval · 评估	用约定的问题、预期与评分方法检查系统行为。
MCP · 模型上下文协议	连接模型应用与工具或资料服务的协议，不替代业务授权。
幂等	同一业务操作重复请求时，不产生重复扣款、重复申请等额外效果。
Runbook · 运行手册	让接手者知道怎样检查、暂停、恢复和升级处理。
Token	模型处理文本的计量单位，不能简单当成中文字数。
Super	澳洲招聘中常见的退休金缴款项目；核对是否在报价之外。

> 不要为了展示工具而增加流程。比如本项目不执行报销审批，就无需为了使用 Agent 而加入自动写入动作。

附录 B / 评审练习

先自己做决定，再看参考思路

以下没有唯一答案，但答案必须交代依据和代价。

情境一：资料负责人说历史档案下周才获批

参考思路：保留现行政策试点，移除历史档案承诺；重新说明覆盖率与验收范围。接口未批准前，不能用个人账号或导出文件绕过审批。补充记录：谁批准、何时重新评估、增加多少验证工作。

情境二：多数答案正确，但出现一次跨部门泄露

参考思路：暂停相关路径，按事件流程评估影响；检查检索、缓存、日志和历史消息，修复后做负例回归。平均正确率不能抵消权限失败。上线决定应依据事先约定的阻断条件。

情境三：大家说很好用，却没人能说清节省多少时间

参考思路：保留定性反馈，但不写收益结论。补采任务基线、实际操作与复核时间，检查样本差异。若测量成本高，可先缩小到一种高频任务；不能直接把好评数量换算成现金。

情境四：客户要自动批准报销

参考思路：将只读问答升级为写入业务系统，是独立范围变化。确认规则、金额权限、审批、幂等、撤销与审计；先提出人工确认版本，不能靠加一段提示词就上线。

附录 C / 来源导航

方法与公开案例

核查日期：2026–11–18。招聘和技术页面可能持续更新。

S1 · OpenAI | What FDE means at OpenAI

打开原始资料 ↗

官方工作方法与项目叙述；不代表公开完整客户实施材料。

S2 · OpenAI | Morgan Stanley 案例

打开原始资料 ↗

供应商发布的客户案例；用于解释评估与推广，不作为独立效果审计。

S3 · Anthropic | Building effective agents

打开原始资料 ↗

工作流与 Agent 的架构原则；2024 年文章，工具细节需查当前文档。

S4 · Anthropic | Demystifying evals for AI agents

打开原始资料 ↗

任务、运行记录与评分方法；本书练习和阈值为自行设计。

事实依据、教学设定与作者判断分开标明。公开客户案例没有披露的预算、代码与内部架构，本书不补写。

附录 C / 来源导航

工程参考与继续学习

核查日期：2026–11–18。招聘和技术页面可能持续更新。

S5 · Microsoft | Azure Search OpenAI Demo

打开原始资料 ↗

官方公开样例；已检查 README，未在本次制作中部署 Azure。

S6 · Microsoft | 登录与文档访问控制

打开原始资料 ↗

用于深入阅读身份与 ACL；接入前核对当时版本。

S7 · OpenAI | 人工参与的客服助手 Demo

打开原始资料 ↗

已检查 README；业务操作为占位实现，本次未运行此仓库。

S8 · JR Academy | FDE Tech 课程

打开原始资料 ↗

用于对齐读者与训练目标；本书预算、实验和企业背景独立设计。

公开仓库用于延伸学习，本次未部署 Azure 或运行 OpenAI 客服样例。PartnerU 与 Claude 登录后培训未逐门核验，访问条件以官方为准。

附录 C / 来源导航

岗位来源：澳洲 → 美国

核查日期：2026-11-18。招聘和技术页面可能持续更新。

S13 · OpenAI | Sydney FDE

打开原始资料 ↗

官方招聘；5 年以上相关经验；未披露薪资，不能套用美国报价。

S16 · PROFOUND | 悉尼 FDE 招聘（SEEK）

打开原始资料 ↗

招聘机构发布的单个岗位；AUD 200k–300k/年。页头 Remote 与正文 fully on-site 不一致，工作方式需招聘方确认。

S14 · OpenAI | San Francisco FDSWE

打开原始资料 ↗

官方招聘；7 年以上全栈经验；美元年基本工资，股权另列。

S15 · Anthropic | Staff FDE

打开原始资料 ↗

官方招聘平台；美国 Staff 职级；保留 Annual Salary 标签，不改写为总包。

附录 C / 来源导航

岗位来源：中国

核查日期：2026–11–18。招聘和技术页面可能持续更新。

S9 · 中兴新云 | 2027 届 FDE 校园招聘

打开原始资料 ↗

企业官网；北京、上海、成都等城市；岗位页未披露薪资。

S10 · 天云融创 | 南开大学就业网招聘信息

打开原始资料 ↗

企业招聘信息，发布于 2026–11–12；薪资原文未标周期，不能直接年化。

S11 · 凡泰极客 | FDE 前沿部署工程师

打开原始资料 ↗

企业官网；深圳，3 年以上部署或运维经验；薪酬面议。

S12 · BOSS 直聘 | 杭州岗位展示

打开原始资料 ↗

平台列表展示样本；详情页触发访问验证，未核实招聘人身份或薪资兑现条件。不是企业官网确认或市场统计。

本地实战：运行器与模型

核查日期：2026-11-18。招聘和技术页面可能持续更新。

S17 · llama.cpp | 本地模型服务与发布版本

打开原始资料 ↗

官方服务文档；本书配套使用本机回环地址与 chat/completions 接口。固定运行器版本见 reports/runtime.json。

S18 · Qwen | Qwen2.5-0.5B-Instruct-GGUF

打开原始资料 ↗

官方模型仓库；本书选 Q4_K_M 小模型用于本地链路教学，不作为模型质量或企业选型推荐。模型按原仓库许可另行下载。

本地实验的验证范围

运行环境、模型哈希和程序版本见配套报告。成功跑通少量教学问题，不等于完成生产质量、安全或用户收益验收。

匠人学院
JR ACADEMY

NEXT / 把项目带进评审

让你的工程判断
有材料可以验证

准备一份 Proposal、一条可运行链路和一份失败记录。再找人追问你的范围、权限、成本和交接方式。

FDE Tech 课程围绕企业背景、方案展示、技术评审和交付展开训练。具体安排以课程页面为准。

查看 FDE Tech 课程 ↗

想根据已有经验选择训练重点，可以联系 Rain：

Rain

Senior IT Career Consultant

在澳洲近20年，Master of HRM，15年企业HR及猎头招聘工作经验，熟悉中国及澳洲就业市场情况。尤其在IT领域中拥有超过6年的咨询及招聘经验。

是你拿下offer的最佳辅助！

毕业生求职做规划　辅助在职人士跳槽晋升

已帮助数千余人成功就业　免费简历诊断

3000+ 3000+咨询案例

1200+ 1200+辅助成功斩获offer

1000+ 1000+简历诊断

http://linkedin.com/in/rainqiu

匠人学院 JR ACADEMY



jiangren.com.au

© JR Academy · 本书为独立编写的实践指南，非厂商官方教材。