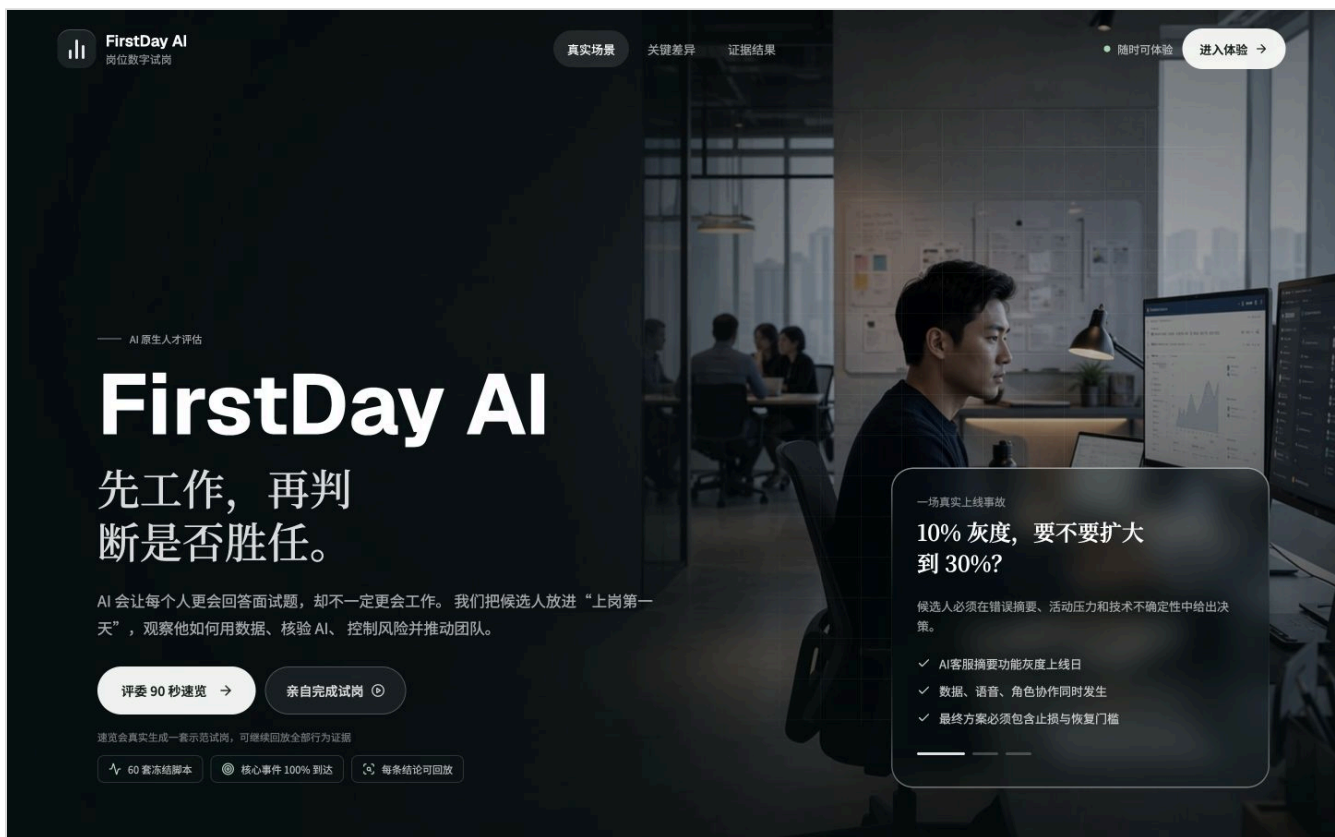


FirstDay AI

岗位胜任力多轮问询与评估报告生成

初赛项目计划书

AI 生成式数字试岗与岗位胜任力评估智能体



项目负责人：杜欣雨

核心成员：马程远

提交日期：2026 年 9 月

目录

以下页码对应本 PDF 的实际章节位置。

- 一、项目摘要 2
- 二、项目背景、用户需求与目标 2
- 三、产品方案与完整用户旅程 3
- 四、岗位能力说明、任务准备与会话记录 5
- 五、场景角色与动态追问 7
- 六、系统技术架构与本地线上实现 9
- 七、大模型的辅助作用与追问更新 10
- 八、知识库内容、搜索与访问限制 12
- 九、行为依据、BARS 评分与报告生成 13
- 十、数据安全、隐私、公平与人工复核 15
- 十一、测试结果与发布前检查 16
- 十二、商业模式与实施路线 17
- 十三、团队介绍、组织分工与合作资源 19
- 十四、原创性声明与作品提交信息 20

一、项目摘要

FirstDay AI 是一套面向高校学生、求职者、企业校招和人才发展场景的 AI 辅助数字试岗产品。系统不把一次问答直接等同于岗位能力，而是将岗位分析结果、能力要求、业务事实、角色关系和突发问题组合成一段约 20 分钟的“上岗第一天”。候选人需要在有限时间内查询信息、识别冲突、使用并核验 AI、协调角色、处理业务风险并提交决策，系统据此形成可观察、可回放、可由人工检查的行为依据。

本项目围绕赛题“岗位胜任力多轮问询与评估报告生成”形成完整流程：岗位能力与任务说明明确需要观察的内容，工作场景为所有候选人提供统一任务，动态追问模块根据当前还缺少的依据调整提问顺序，行为记录表保存每个判断的来源，多维 BARS 规则再将具体行为对应到能力等级，最终生成可由人工复核的报告。大模型负责理解自由文本、改善表达和提出可能相关的行为依据，不负责修改场景事实、决定关键事件或直接打分。

首发岗位为校招或初级 AI 产品经理，首发场景为“AI 客服摘要灰度上线事故”。候选人会面对地区与渠道数据差异、技术和运营目标冲突、疑似旧知识缓存、用户风险与业务效率之间的权衡。该场景能够同时观察问题定义、数据判断、产品决策、跨团队协作、AI 协作和复盘学习六项能力，并保留文字通道作为语音异常时的完整替代。

- 核心产品形态：真实工作样本，而非单纯面试问答。
- 核心技术路线：事实和评分按固定规则处理，大模型只辅助理解与表达。
- 核心评估原则：每个结论必须能够回到候选人的原始操作或回答。
- 核心公平原则：标准化任务保持一致，动态事件最多补充两项且不直接计分。
- 当前交付状态：Web Demo、本地与线上接口、评审报告、行为回放、自动测试和比赛提交材料均已完成。

二、项目背景、用户需求与目标

2.1 传统评估方式的不足

传统简历筛选和面试主要依赖候选人对既往经历的自我陈述。表达能力强的人更容易呈现完整故事，但面试官仍难以判断候选人在信息不完整、多人协作和时间压力下是否会主动查证、是否会寻找与当前判断相反的信息、是否能够给出可以撤回的处置方案。常见反馈也往往停留在“逻辑需要加强”“沟通需要提升”，缺少具体行为、发生时间、信息来源和改进动作。

真人 Assessment Center 能够提供更真实的工作观察，但设计场景、安排观察员、统一评分标准和组织复盘的成本较高，难以覆盖大规模校招或高校职业训练。随着生成式 AI 进入日常工作，评估还必须回答一个新问题：候选人是否会把 AI 输出当作未经验证的假设，是否会检查来源、版本和反例，并保留最终的人类责任。

2.2 目标用户与典型任务

用户	主要需求	FirstDay AI 提供的价值
学生与求职者	理解岗位真实工作，发现能力不足，获得具体训练建议	低成本完成可重复数字试岗，查看行为回放和 30 天提升计划
企业招聘团队	验证简历主张，提升面试信息量，统一评价口径	获得有来源的行为记录、BARS 参考结果和人工复核入口
高校就业中心	提供规模化职业体验和岗位训练	按岗位组织训练任务，跟踪班级能力分布和复测变化
岗位专家与测评专家	校准场景真实性、观察机会和行为标准	使用带版本号的任务说明、标注手册、盲标任务包和争议处理流程

2.3 项目目标与边界

项目的近期目标是完成一个可以真实体验、重复运行并由外部专业人员检查的岗位能力评估原型；中期目标是请岗位专家和测评专家校准内容，再将场景扩展到研发、运营和销售等岗位；长期目标是形成企业可以配置的岗位任务和行为标准库。当前系统定位为训练、发展反馈和人工决策辅助工具，不将演示分数描述为正式招聘结论，也不允许系统自动完成录用或淘汰。

三、产品方案与完整用户旅程

完整流程由岗位输入、任务准备、数字试岗、经理复盘和结果报告五个阶段组成。用户输入简历与岗位描述后，系统先生成脱敏的岗位能力与任务说明，明确任务目标、时间限制、可用角色、可用工具和数据处理方式。候选人确认用途与数据处理授权后进入工作台，系统开始记录任务时间、查询、对话、AI 调用、核验动作、风险处置和最终决策。

FirstDay AI 产品使用流程

从输入岗位要求到查看报告，每一步都保留记录



图 1 产品使用流程：从岗位输入、数字试岗和经理复盘到结果报告



图 2 数字试岗工作台：任务、时间线、工具、角色和行为记录在同一界面中呈现

阶段	候选人动作	系统处理	形成的记录
岗位准备	提交简历和 JD，确认用途	生成岗位任务说明，选择统一任务和少量个性化问题	需要验证的能力主张与观察计划
数字试岗	查数据、问角色、使用工具、处理事件	推进任务、控制信息权限、按顺序记录操作	行动、查询与协作记录

阶段	候选人动作	系统处理	形成的记录
动态追问	回应假设情况和压力问题	根据当前缺少的观察从固定问题中选择下一项	压力情境下的行为记录
经理复盘	解释错误假设、修正动作和机制改进	选择 3 至 4 个针对性问题并记录回答	复盘与举一反三的表现
结果反馈	查看报告并回放原始行为	对照 BARS 标准、计算可信程度并检查来源	六维结果、风险和建议

统一核心任务约占整体观察机会的 70%，保证不同候选人都必须处理同一组关键事实、业务目标和风险边界；根据简历调整的问题约占 30%，用于追问责任范围、决策依据和类似经验。个性化内容只改变提问角度，不改变核心任务难度和评分标准，从而兼顾个人差异与结果可比性。

候选人完成最终决策后，系统不会立即给出一句笼统评价，而是进入经理复盘。复盘问题由已执行动作、尚未充分观察的能力、AI 使用方式和处置方案共同决定。回答会作为新的可观察行为加入行为记录，报告随之更新，并保留前后版本，便于评审者判断候选人是否能够在得到反馈后修正思路。

四、岗位能力说明、任务准备与会话记录

4.1 岗位能力与任务说明

岗位能力与任务说明把工作场景和评分要求连接起来。它包含岗位目标、关键工作任务、六项能力定义、可观察行为、业务限制、必须出现的问题、行为来源和评分标准。简历只用于识别需要验证的能力主张，不直接形成能力分。岗位描述中的职责会被转换为具体任务，再为每项任务安排可以观察的操作，避免从抽象标签直接跳到分数。

场景准备模块把岗位蓝图转换为可以运行的任务材料，包括角色卡、事实关系、事件时间线、知识条目、访问权限、可用动作和完成条件。每条事实都有稳定 ID 和出现条件，每个角色只能看到其职责范围内的信息。这样，业务事实不依赖模型临场生成，阅读者能够追溯某条回答来自哪条事实，以及由哪个操作获得。

4.2 会话记录与任务推进

系统使用固定规则维护会话状态。记录内容包括当前任务时间、已经出现的事件、候选人已经获得的信息、已完成任务、候选人操作、风险标记、追问状态、复盘答案和报告版本。每个动作先检查参数和访问权限，再按顺序保存，随后更新会话并重新判断是否需要追问。相同的操作顺序会得到相同的结果和报告校验值。

状态对象	关键字段	更新时机	用途
SessionState	时间、阶段、动作、事实、标记	每次合法动作后	保存当前会话的完整状态
ActionRecord	序号、动作、目标、载荷、任务时间	候选人操作时	行为来源与回放
DirectorState	当前表现、尚缺依据、备选问题	状态变化后	决定是否补充追问
AgentTurn	步骤、工具、调用上限、结束原因、备用状态	每次模型调用时	检查模型调用过程
AssessmentReport	行为依据、维度结果、可信程度、检查问题、校验值	查看报告时即时生成	评审和人工复核

4.3 信息如何显示以及如何重复验证

系统把“发生了什么”和“候选人如何表达”分开处理。核心事件 E01 至 E08、事实关系、时间限制和任务目标保存在带版本号的配置中；模型可以调整角色回复的语言风格，但不能新增事实、提前透露隐藏信息或修改信息出现条件。假设性追问会明确标记 `hypothetical`，不会当作真实事实保存，只有候选人在假设条件下采取的行为才可能进入评分依据。

会话恢复依赖已经保存的任务记录，不需要重新播放完整对话。SQLite 保存本地服务数据，Cloudflare KV 保存公网演示会话；两套后端按照同一接口说明和数据结构工作。即使模型或语音服务不可用，固定规则仍能继续推进全部核心任务，避免外部服务故障导致测评中断或改变候选人面对的信息。

五、场景角色与动态追问

5.1 场景角色如何回答问题

场景中的产品负责人、技术负责人、运营负责人和客服代表不会自由编造内容，而是按照角色说明、可见信息范围和允许使用的工具回答问题。每个角色只能查找其 `role_scopes` 范围内的知识条目，回答中同时返回对应的 `fact_ids` 和 `knowledge_sources`。角色不能替候选人做决定，不能调用未登记的工具，也不能通过自然语言绕过访问限制。

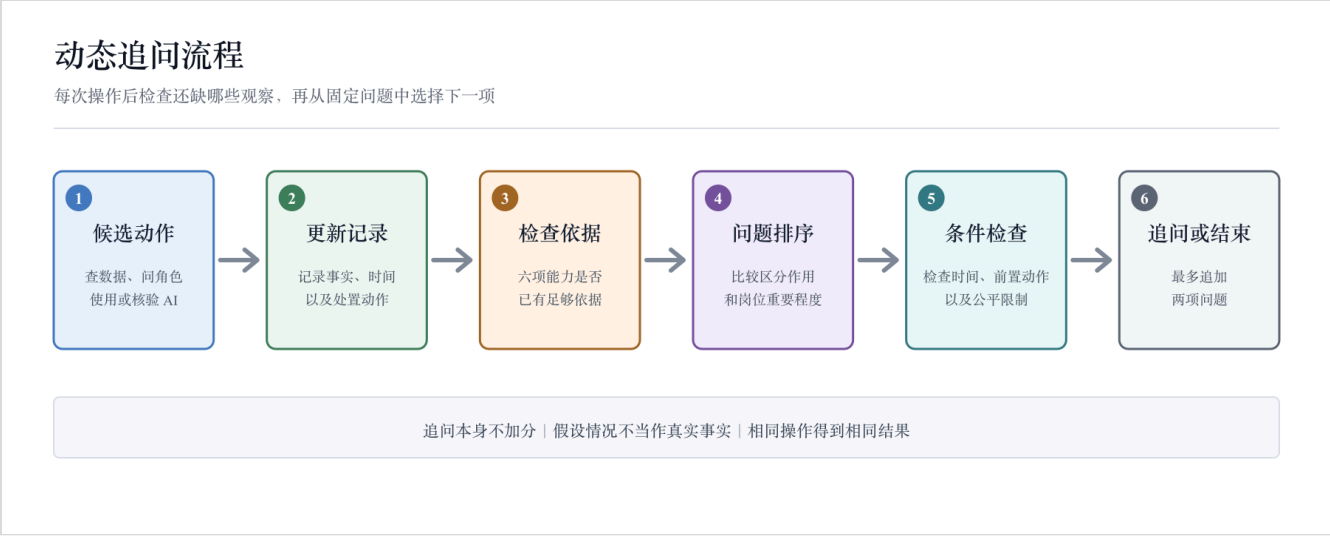
Copilot 也只能使用预先登记的只读分析工具。每次调用会保存 `queued`、`running`、`completed`、`failed` 或 `cancelled` 状态，并记录模型处理步骤、工具调用、提示信息、输入上限、工具调用次数、供应方、模型、是否改用备用方案和上下文校验值。发送给模型的信息仅包含当前时间、候选人已经获得的事实 ID、待核验项、最近操作和少量状态标记，不上传完整简历、原始语音或人工复核正文。

5.2 动态追问如何选择下一问题

动态追问不是为了制造戏剧化对话，而是为了在有限时间内补足最有判断价值的观察。系统根据角色询问数量、角色种类、数据查看维度、AI 是否被核验、是否提出分级处置、是否指定负责人和是否完成用户补救等明确记录，判断 C1 至 C6 哪些能力还缺少足够依据。只有缺少的信息达到设定条件时，系统才会追加问题。

指标	计算含义	主要来源	控制目的
区分作用	问题能否区分候选人的不同表现	尚缺的观察与前后矛盾	优先提出最有判断价值的问题
岗位重要程度	问题涉及的能力对岗位是否重要	固定问题设置	避免追问偏离核心工作
观察是否充分	当前能力是否已有足够依据	操作记录与状态标记	减少缺少依据的评分
前后矛盾检查	当前行为是否与前述主张或方案冲突	全量关闭、未核验 AI 等信号	观察候选人是否修正假设
叙事价值	事件是否能自然推动任务收束	事件池配置	保持体验连贯

问题排序方式：系统综合考虑问题的区分作用、岗位重要程度、当前观察是否充分、前后行为是否矛盾以及是否有助于任务继续推进。问题还必须满足触发时间、前置动作、未重复出现和公平限制。



六、系统技术架构与本地线上实现



图 4 系统架构：界面、应用服务、场景处理、模型辅助和评估复核相互独立

系统采用前后端分离架构。前端使用 React、TypeScript 与 Vite，提供岗位配置、数字试岗工作台、动态追问状态、Copilot、语音交互、能力报告和行为回放。本地后端使用 FastAPI 与 Pydantic，负责准备场景、维护会话、检索知识、调用模型、生成评估结果和保存数据；公网演示使用 Cloudflare Pages Functions 与 KV，并保持相同的核心接口行为。

架构层	核心模块	主要职责	关键约束
体验层	React 工作台、报告与回放	呈现任务、工具、状态、字幕和行为依据	候选人、评审、复核视图分别显示
应用服务层	FastAPI、Pages Functions	会话接口、输入检查、身份验证、返回数据	本地和线上接口一致
场景处理层	Compiler、Engine、Director	准备任务、推进状态、选择追加问题	核心事实和事件按固定规则运行
知识与角色层	Scoped RAG、角色卡、事实图	按角色检索并返回可追踪事实	检索前后双重权限校验

架构层	核心模块	主要职责	关键约束
模型与语音层	兼容接口、ASR、TTS、备用处理	语言理解、表达辅助和语音输入输出	失败不阻断核心试岗
评估与复核层	Evidence、BARS、Audit、Views	整理行为依据、评分、可信程度和人工复核	模型不直接评分

6.1 本地与线上接口保持一致

FastAPI 与 Cloudflare Functions 用于不同部署环境，但共同遵循 docs/api-contract-v1.json 中的接口说明，其中列出了请求字段、返回字段、状态码和主要含义。自动检查会逐项比较两套后端的会话创建、任务推进、追问状态、Copilot、报告、数据删除和健康检查接口，避免公网 Demo 与本地版本出现不同表现。

6.2 数据保存、访问验证与组织隔离

本地后端通过统一的数据读写接口使用 SQLite 保存会话、组织 ID 和令牌校验值；公网演示使用 KV 保存相同类型的数据。访问会话时同时校验组织 ID 与 token_hash，令牌原文不会进入数据库。报告按候选人、评审者和复核人员分别返回所需字段，避免候选人看到内部评分校验信息，也避免评审者接触无关的调试内容。

6.3 运行记录与故障恢复

关键操作会记录统一名称、会话编号、来源、任务时间和脱敏后的附加信息，覆盖会话创建、角色查询、数据查询、模型调用、语音调用、问题触发、报告生成和数据删除。每次模型调用保存处理步骤和结束原因，语音服务记录供应方、耗时和是否改用文字。外部服务失败时，界面会明确提示并改用固定规则或文字输入，不会在用户不知情的情况下修改结果。

七、大模型的辅助作用与追问更新

7.1 大模型负责什么

大模型只承担三类受限任务：把候选人的自由文本整理成可以检查的字段和候选行为依据，为角色和 Copilot 生成更自然的表达，以及为经理复盘整理摘要。模型输出必须符合预先规定的 JSON 格式，并通过字段、取值、事实权限和来源检查后，才能作为辅助信息展示。核心事实、问题触发条件、会话更新、BARS 等级和检查结论始终由固定规则决定。

环节	模型输出	固定规则检查	失败处理
输入理解	意图、假设、动作、责任人、时限	字段、枚举、动作白名单	使用规则解析器
候选行为依据	能力维度、方向、行为描述、来源	字段定义、信息权限、source_id 是否存在	正式评分只使用规则确认的依据
Copilot 表达	事实、假设、建议、待核验项	引用必须属于候选人已获得的信息	返回固定文字模板
动态追问	还需要了解什么以及如何提问	触发条件和问题顺序由固定规则计算	从预先设置的问题中选择
最终评分	不允许直接输出 BARS 分数	对照固定行为标准并检查来源	没有依据则需要补测

7.2 发送给模型的信息与调用限制

每次模型调用只发送完成当前任务所需的信息：当前任务、任务时间、候选人已经获得的事实 ID、最近操作、待核验项、少量状态标记和内容校验值。系统不会把整段会话历史、简历原文或全部知识条目直接发送给模型。每次调用都设置最大输入字符数、最大处理步骤、最大工具调用次数和超时时间，并记录 completed、failed、cancelled 或 fallback 等结束原因。

7.3 每次操作后的更新流程

候选人每完成一次操作，系统依次检查输入、更新会话、整理行为依据、判断还缺少哪些观察、排列备选问题、检查触发条件，最后决定是否追加问题。模型可以帮助识别文本中的隐含假设或责任承诺，但最终问题仍从 D01 至 D06 的预设列表中选择。问题选择所使用的输入、权重、条件和理由都保存在 DirectorState 中，评审者可以查看某个问题为什么出现，也可以重复同一操作顺序验证结果。

7.4 模型服务失败时如何处理

模型接口会注明是否支持 JSON response_format、reasoning_effort 等参数。请求使用较低随机性，要求返回固定 JSON 格式，并设置重试次数、超时时间和连续失败后的暂停机制。出现网络错误、响应超时、JSON 无法解析、格式检查不通过或引用无权访问的信息时，系统会丢弃本次模型结果，改用固定文字模板和规则生成的行为依据；候选人仍能继续完成相同核心任务。

八、知识库内容、搜索与访问限制

8.1 知识来自哪里以及如何保存

知识库不是将任意企业文档直接切块后交给模型，而是围绕岗位任务构建受控知识集合。当前知识来源包括岗位目标、角色职责、发布记录、渠道和地区数据、用户投诉样本、业务约束、风险门槛、恢复条件和角色沟通记录。每条知识记录保存稳定 ID、标题、正文、标签、role_scopes、fact_ids 和知识库版本，便于权限检查、引用展示和版本回溯。

字段	含义	示例用途	管理作用
id	知识条目稳定标识	报告和运行日志引用	避免文本变化造成引用漂移
role_scopes	允许访问该条目的角色集合	技术负责人只看到发布与日志信息	阻止跨角色泄露隐藏事实
fact_ids	知识条目关联的冻结事实	角色回答返回事实来源	区分事实与模型生成语言
version	场景和知识库版本	重复查看历史会话	记录修改和校准过程
tags	渠道、地区、风险和任务标签	提高短查询匹配	便于岗位迁移和维护

8.2 如何切分内容和排列搜索结果

当前实现使用透明可解释的词条检索。英文按字母和数字词条切分，中文提取连续文本并生成双字切分，以兼顾短查询、业务专有名词和中英文混合输入。对每个符合角色权限的条目，系统计算查询覆盖率和文档密度，最终得分为 80% 查询覆盖率加 20% 文档密度，按得分和稳定 ID 排序并默认返回 Top 3。

8.3 搜索前后都检查访问权限

检索前先按 role_scopes 过滤候选集合，检索后再校验返回条目的 fact_ids 是否符合当前会话已经解锁的事实权限。角色回复必须携带 knowledge_sources 和 returned_fact_ids；Copilot 只能引用已解锁事实的子集。即使模型生成了听起来合理的额外事实，只要来源不在允许集合中就会被拒绝。

8.4 如何记录修改以及后续扩展

当前原型用带版本号的 YAML 文件保存可以追溯的场景事实。修改知识条目时，需要同步更新场景版本并重新运行相关测试。后续接入真实企业资料时，可以在保留权限和来源记录规则的前提下增加向量检索或混合检索：语义搜索负责找到更多可能相关的材料，角色权限、事实 ID 和最终排序仍由固定规则控制。敏感资料进入检索库前，需要完成脱敏、访问级别、保留期限和删除方式设置。

九、行为依据、BARS 评分与报告生成

9.1 从能力要求到可观察行为

评估设计参考 Evidence-Centered Design 方法：先说明岗位需要什么能力，再明确哪些行为能够支持或反驳这一判断，最后设计能够自然产生这些行为的任务。场景中的问题本身只是观察机会，不能直接证明候选人能力较强；评分依据必须来自候选人的角色询问、数据查询、AI 使用与核验、风险处置、负责人分配、用户补救、最终决策或经理复盘。

每条行为依据包含 competency_id、正向或负向方向、行为描述、评分理由、source_type、source_id、任务时间、依据强度、是否来自压力情境和关联事件编号。稳定的记录 ID 由能力、方向、来源和行为摘要生成，报告中的每条依据都可以跳转回原始操作或事件，评审者无需依赖模型总结即可自行检查。



图 5 从行为到报告：检查原始记录后，对照 BARS 标准生成结果并交由人工复核

能力维度	权重	主要观察信号
C1 问题定义与信息获取	20%	影响范围问题树、主动询问、角色覆盖、关键未知项和事实获取
C2 数据与依据判断	15%	查看细分数据、检查版本与样本、寻找支持与反向信息并修正假设
C3 产品决策与风险处置	20%	可撤回的止损措施、分级处置、用户补救、负责人、恢复条件和回滚条件
C4 跨团队协作	15%	正确角色、明确任务、反馈时限、责任机制和冲突升级路径
C5 AI 协作能力	20%	提示具体度、来源核验、错误纠正、风险边界和最终人工责任
C6 复盘与学习能力	10%	识别错误假设、说明修正依据、提出具体改进方法并举一反三

9.2 BARS 等级、分数和结果可信程度

每个维度使用 Level 1 至 Level 5 的行为标准，分别对应 20、40、60、80 和 95 分。这些标准描述的是可以观察到的操作，而不是人格判断。例如，C3 的高等级要求候选人明确监控指标、恢复条件、回滚方案和长期检查办法，而不是使用“判断力强”等抽象标签。系统先根据有效的行为依据确定等级，再换算为维度分，最后按各维度权重计算岗位匹配参考分。

分数和结果的可信程度分别计算。只有一条较弱的行为依据时，可信程度较低；存在多条独立来源、压力情境和复盘确认时，可信程度提高。C5 未调用 AI 或 C6 未完成复盘时，系统输出 needs_retest 和空分数，而不是把没有观察机会解释为低能力。这样可以避免候选人因问题顺序不同或工具故障受到惩罚。

9.3 报告检查与三类查看方式

报告生成前，系统会检查每个已评分维度是否引用真实行为、需要补测的维度是否错误产生分数、行为记录 ID 是否存在、个性化问题是否被直接计分，以及是否仍缺少明显依据。检查结果分为通过、提醒和失败。候选人版聚焦发展反馈，评审版展示能力结果和行为依据，复核版进一步展示报告校验值、检查问题和内部来源信息。



图 6 能力报告：六维结果、可信程度、行为依据覆盖、优势风险和原始操作回放

最终报告包含六维雷达图、岗位匹配参考分、各维度状态、BARS 行为标准、正负行为依据、尚未观察到的行为、结果可信程度、依据覆盖率、优势与风险、30 天提升建议以及简历主张验证。岗位匹配参考分只用于发展反馈和人工复核，不作为自动录用或淘汰结论。

十、数据安全、隐私、公平与人工复核

10.1 用途告知与数据最小化

候选人进入试岗前必须确认用途、数据处理方式和语音选择。语音为可选通道，用户可以全程使用文字完成相同任务。系统只保存完成体验和复核所需的数据，不把 API Key、原始密钥或无关个人信息写入仓库。公网演示使用虚构候选人资料和合成业务场景。

10.2 哪些信息不参与评分

口音、语速、音色、情绪、年龄、学校、性别、健康状况和人格均不参与评分。语音转写只作为输入通道，评分依据是候选人表达的工作内容和采取的操作。核心任务和行为标准对所有候选人一致，追加问题数量受到限制且问题本身不计分；没有观察机会时输出需要补测。

10.3 数据权利与人工复核

系统保留查看、更正、删除、退出和申诉流程。招聘方不能只查看一个综合分，而应结合行为来源、结果可信程度和报告检查状态进行人工复核。人工复核结果作为追加记录保存，不覆盖原始行为和原始系统报告，便于后续处理争议和分析评分差异。

10.4 模型不能做什么

模型不能修改核心事实、提前展示隐藏信息、触发问题、直接评分或覆盖检查结果；模型提出的候选行为依据会标记为 `advisory_only`，并与正式评分依据分开显示。所有工具必须预先登记并说明是否只读，模型无法访问 Shell、任意文件系统或任意网络。超过调用限制、超时、格式错误或引用无权访问的信息时，系统会改用固定规则继续处理。

十一、测试结果与发布前检查

项目采用多种方式检查质量，包括单元测试、本地与线上接口对比、固定行为案例测试和浏览器操作验证。单元测试覆盖会话状态、场景准备、知识库、动态追问、行为依据整理、BARS 评分、模型与语音接口、数据保存和人工复核流程；接口检查比较 FastAPI 与 Cloudflare 两套后端；前端执行 TypeScript 类型检查和正式构建；提交脚本检查敏感信息、文件完整性和压缩包是否可以正常解压。

验证指标	当前结果
自动化测试	103 项通过
固定行为测试案例	60 套，覆盖五类典型表现
场景完成率	100%
核心事件到达率	100%
行为依据引用完整率	100%
重复评分平均极差	0

验证指标	当前结果
自动评分耗时	95% 的测试请求不高于 6ms
本地与线上接口	FastAPI 与 Cloudflare 关键接口表现一致
人工标注材料	20 套案例、198 条行为记录、10 套边界示例

11.1 固定测试案例与行为区分度

60 套固定行为案例分为优秀、中等、基础薄弱、过度依赖 AI、表达优秀但实际操作不足五类，每类 12 套。这些案例不是用来宣称真实招聘效度，而是用于检查任务能否完成、关键事件能否出现、不同关键行为能否被正确记录、评分是否稳定，以及系统能否区分“说得好”和“真正完成关键动作”。

11.2 行为标注与人工核对计划

项目已确定行为字段说明、标注示例、双人标注模板和分歧处理手册，并对 20 套不同表现记录中的 198 条行为依据完成格式检查。现有模型预检查只用于发现标签和流程问题，不能替代真人专家结论。进入试点前，需要由至少两名产品经理和一名招聘或测评人员独立标注，再计算动作级 F1、分维度 F1 和 BARS 加权 Kappa。

11.3 发布前检查

每次发布先检查与本次改动直接相关的测试，再执行完整测试、本地与线上接口对比、前端类型检查和正式构建，最后检查提交材料并生成 ZIP。提交压缩包只包含项目计划书、使用手册和演示视频。

十二、商业模式与实施路线

12.1 商业模式

客户	产品形态	收费方式
学生与求职者	单岗位数字试岗、训练报告、复测与能力成长记录	单场付费或订阅
企业校招	岗位任务、候选人评估、行为记录后台和人工复核	按候选人、岗位或年度授权

客户	产品形态	收费方式
高校就业中心	岗位训练包、课程任务、教师看板和班级能力分析	年度授权与岗位包服务

早期从学生岗位训练和企业校招试点切入，优先验证用户能否顺利完成任务、反馈是否容易理解、岗位专家是否认可场景真实性，以及行为记录是否比传统问答提供更多决策信息。商业化不以售卖单一模型能力为核心，而以岗位任务材料、行为评分标准、可人工检查的报告和企业配置能力形成长期价值。

12.2 实施路线

阶段	目标	主要交付
当前原型	完成单岗位完整流程与公开体验	AI 产品经理场景、Web Demo、评估报告、测试和提交材料
专家校准	验证任务是否贴近岗位、评分是否一致	SME 访谈、独立标注、处理分歧、修改行为标准
小规模试点	验证可用性和业务增量	学生试用、企业人工复核、候选人反馈和流程指标
多岗位复制	形成可配置场景生产流程	研发、运营、销售岗位包和企业模板后台
规模化部署	满足企业稳定性和管理要求	PostgreSQL、RBAC、组织隔离、监控、ATS 集成和长期公平评估

十三、团队介绍、组织分工与合作资源

成员	身份	履历与能力
杜欣雨	项目负责人	香港理工大学数据科学及分析硕士在读。曾从事广告推荐算法工作，覆盖多场景 SparseMoE、RankMixer、Token Mixing 与特征补全；拥有中国移动“梧桐杯”一等奖、Kaggle 银牌等经历。负责赛题理解、产品方案、评估框架、项目推进和材料交付。
马程远	核心成员	清华大学大数据技术与工程硕士。曾在蚂蚁集团、腾讯 AI Lab、OPPO AI Center 从事语音与多模态算法工作，方向覆盖多模态 RAG、可控语音生成、全模态模型和 AI Native 智能体；拥有语音识别竞赛 Top 3、Kaggle Expert、论文与专利积累。负责智能体架构、语音与多模态能力、工程实现、测试和部署。

13.1 团队协作机制

项目负责人统筹赛题需求、产品范围、材料一致性和最终质量；核心成员负责架构实现、模型与语音接入、评估工具和公网部署。两位成员共同维护场景事实、评分边界和发布前检查清单，对涉及评估结论、隐私和公平性的改动相互复核。重要版本保留测试结果、接口说明、文档路径和文件校验值。

13.2 后续合作计划

后续计划邀请产品与招聘专家校准行为标准，与高校和企业开展岗位训练试点，并推进语音服务、隐私合规及企业部署合作。

十四、原创性声明与作品提交信息

14.1 原创性与真实性声明

FirstDay AI 的数字试岗状态管理、岗位事实关系、动态追问、六维能力模型、BARS 行为标准、行为记录、按角色限制的知识检索、报告回放和固定行为案例测试工具由参赛团队设计并实现。团队承诺本项目计划书、源代码、演示材料和评测结果真实有效，提交方案及成果具有原创性；若由此产生知识产权纠纷，由参赛团队自行承担相应责任。

- 项目使用 React、TypeScript、FastAPI、Pydantic、Cloudflare Pages Functions 等开源组件，并遵循对应许可证。
- 项目使用 Qoder、Codex 等 AI Coding 工具辅助研究、编码、测试和文档，核心设计、评估逻辑、代码审查和最终交付由参赛团队完成。
- 公网演示使用虚构候选人资料和合成业务场景，不提交真实候选人个人数据。
- 涉及企业内部业务、密钥或敏感信息的内容已抽象、脱敏或删减。
- 系统定位为训练与人工复核辅助工具，不作为唯一录用或淘汰依据。

14.2 可运行 Demo 与公开材料

入口	地址
快速体验	https://firstday-ai-demo.pages.dev/?judge=1
完整体验	https://firstday-ai-demo.pages.dev/
在线手册	https://firstday-ai-demo.pages.dev/guide
项目计划书	https://firstday-ai-demo.pages.dev/submission/FirstDayAI_project_plan.pdf

快速体验页面无需测试账号，会由服务端即时执行一套完整示范过程并进入可回放报告。完整体验页面支持简历与 JD 输入、用途确认、数字试岗、经理复盘、能力报告和行为回放。提交压缩包只包含项目计划书、使用手册和演示视频。