AI 面试训练 Agent 项目指标与评估方案

1. 方案目标

本方案用于评估 AI 面试训练 Agent 是否真正提升用户的面试准备效率、岗位匹配度和回答质量，同时评估大模型输出的稳定性、可解释性和产品可用性。

评估对象包括四层：

1. 用户使用链路是否顺畅。

2. 面试训练效果是否提升。

3. 大模型输出质量是否稳定。

4. 腾讯云智技术产品岗位适配是否充分。

5. 异常分支和产品调性是否达到可发布标准。

2. 北极星指标

北极星指标：有效训练完成率

定义：完成一次完整 Agent 训练流程，并成功生成最终复盘报告的用户占启动 Agent 训练用户的比例。

公式：有效训练完成率 = 生成最终复盘报告的训练次数 / 启动 Agent 训练次数

选择原因：

1. 它覆盖了从 JD/简历输入到最终复盘的完整闭环。

2. 它能反映产品是否真正被用于训练，而不只是被打开。

3. 它与项目核心价值直接相关：帮助用户完成一次可复盘的面试训练。

3. 用户行为指标

指标	定义	目标意义	MVP 目标
首页访问量	用户打开线上页面次数	衡量基础访问能力	可正常统计
岗位模板选择率	选择岗位模板的用户占访问用户比例	衡量岗位化入口是否有效	> 50%
JD 填写率	填写或应用 JD 的用户占访问用户比例	衡量训练前置条件完成度	> 40%
简历填写率	上传或粘贴简历的用户占访问用户比例	衡量个性化训练准备程度	> 25%
训练前诊断点击率	点击训练前诊断的用户占填写 JD 用户比例	衡量诊断功能吸引力	> 35%
Agent 训练启动率	启动 Agent 训练的用户占填写 JD 用户比例	衡量核心流程转化	> 25%
单轮回答提交率	至少提交一次回答的用户占启动训练用户比例	衡量问题生成有效性	> 70%
完整训练完成率	完成多轮训练并生成报告的用户占启	衡量闭环体验	> 40%
报告导出率	点击语音输入的用户占回答用户比例	衡量报告复用价值	> 20%
语音输入使用率	语音失败后仍完成文字提交的比例	衡量效率增强能力是否被使用	可观察
语音降级成功率	有项目经历，但目标岗位跨度较大	衡量异常分支是否不阻断主流程	> 80%

4. 面试训练效果指标

指标	定义	观察方式
平均分提升	同一用户后续回答平均分相比前几轮的变化	对比训练前 2 轮与后 2 轮平均分
STAR 完整度提升	回答是否更完整交代背景、任务、行动、结果	评分维度趋势
岗位匹配度提升	回答是否更贴合目标岗位关键词与职责	岗位匹配维度趋势
技术/产品深度提升	回答是否体现技术原理、产品判断和方案边界	技术/产品深度维度趋势

客户需求/场景拆解提升	回答是否能从客户、用户场景和痛点出发	客户场景维度趋势
数据量化意识提升	回答是否增加指标、规模、前后对比和结果	数据量化维度趋势
结构化表达与沟通推动提升	回答是否体现分层表达、协作对象、优先级和推进动作	结构化表达维度趋势
高频短板减少	用户画像中的薄弱信号是否减少	weakSignals 变化

5. 大模型输出质量指标

指标	定义	目标
JSON 解析成功率	模型输出能被前端正常解析的比例	> 95%
JSON 解析失败恢复率	解析失败后前端能提示重试且保留用户输入的比例	100%
API 调用成功率	Serverless API 正常返回结果的比例	> 98%
API 错误码可识别率	401/429/503/500 等错误能被前端正确分流的比例	100%
报告生成成功率	成功生成最终复盘报告的比例	> 95%
评分一致性	同类回答在多次评分中的波动程度	波动可控
评分维度对齐率	`dimensionScores` 与岗位模板评分权重一致的比例	100%
反馈可操作性	反馈是否包含明确问题和下一步建议	人工抽检通过
教练反馈合格率	`coachNote` 是否专业、克制、鼓励、行动导向	人工抽检通过
幻觉风险率	是否编造不存在经历、数据或公司事实	尽量为 0
岗位知识命中率	输出是否引用或体现目标岗位知识	> 95%

6. 评分校准方案

6.1 校准目标

降低大模型评分漂移，让评分更稳定、更可解释、更符合岗位要求。

6.2 当前校准规则

1. 回答少于 30 字时总分上限为 40；回答 30-80 字时总分上限为 60。
2. 缺少结果或量化数据时限制高分。
3. 缺少岗位关键词或岗位场景时限制岗位匹配度得分。
4. STAR 结构不完整时限制完整度得分。
5. 缺少技术或产品深度时限制技术/产品深度得分。
6. 缺少量化数据时，数据量化与结果意识维度最高不超过 8 分。

6.3 后续可增强规则

1. 建立标准答案样本集，对同一答案进行多次评分观察波动。
2. 引入人工标注样本，计算模型评分与人工评分的相关性。
3. 对不同岗位模板建立独立评分基准。
4. 将评分结果拆为“硬规则分”和“模型判断分”两部分。

7. 事件埋点设计

事件名	触发时机	关键字段
page_view	用户打开页面	timestamp, device, source
template_selected	用户选择岗位模板	template_key
jd_filled	用户填写或应用 JD	template_key, jd_length
resume_filled	用户填写简历	resume_length
preflight_started	点击训练前诊断	template_key
preflight_completed	诊断成功返回	match_score, gap_count
agent_plan_started	开始生成训练计划	template_key

agent_plan_completed	训练计划生成成功	question_count
answer_submitted	用户提交回答	question_id, answer_length
analysis_completed	回答分析完成	score, weak_signals
coach_note_shown	教练反馈展示	score, focus_dimension
voice_state_changed	语音输入状态变化	state, textarea_id
api_error_handled	API 错误被前端分流处理	status, code, endpoint
next_question_generated	下一题生成成功	focus_area
report_generated	最终报告生成	readiness_score
report_exported	用户导出报告	export_type
history_report_opened	用户打开历史报告	report_id
local_data_cleared	用户清空本地数据	history_count

8. A/B 测试建议

实验一：是否展示训练前诊断

目的：验证训练前诊断是否提升 Agent 训练启动率。

实验组：

1. A 组：进入页面后展示“训练前诊断”按钮。

2. B 组：只展示“生成训练计划并开始”按钮。

观察指标：

1. Agent 训练启动率。

2. 完整训练完成率。

3. 最终报告生成率。

实验二：评分反馈长度

目的：验证反馈信息量对用户继续训练意愿的影响。

实验组：

1. A 组：简洁反馈，只展示总分、问题和建议。

2. B 组：完整反馈，展示评分维度、参考答案、改写版本。

观察指标：

1. 下一题点击率。

2. 单场训练完成率。

3. 用户主观满意度。

实验三：教练式反馈语气

目的：验证低分场景下先鼓励再给出行动建议，是否提升用户继续训练意愿。

实验组：

1. A 组：仅展示分数、问题和建议。

2. B 组：在分数下方展示一条 coachNote，先肯定当前进展，再指出最优先改进点。

观察指标：

1. 下一题点击率。

2. 低分用户继续训练率。

3. 用户主观满意度。

实验四：岗位模板默认推荐

目的：验证默认推荐腾讯云智模板是否提升该岗位用户的训练效率。

观察指标：

1. 模板切换率。

2. JD 填写率。

3. 岗位匹配度评分提升。

9. 人工评估方案

9.1 抽检样本

每个岗位模板抽取 10-20 组训练样本，包括：

1. JD。

2. 简历。

3. 用户回答。

4. 模型评分。

5. 参考答案。

6. 最终复盘报告。

9.2 人工评估维度

维度	评分标准
准确性	是否准确识别回答问题
专业性	是否符合技术产品岗位语境
岗位贴合度	是否贴近各模板要求
可操作性	建议是否能指导用户下一步修改
产品调性	反馈是否专业、克制、鼓励，不制造挫败感
稳定性	类似回答是否得到相近评价
安全性	是否避免编造或过度承诺

10. 阶段验收标准

MVP 线上阶段

1. 首页和核心 API 在线可访问。

2. 训练前诊断、Agent 计划、回答评分、复盘报告流程可跑通。

3. 至少支持 Vercel 主站和 Netlify 镜像站。

4. API Key 不暴露在前端。

5. 报告可导出，历史报告可回看。

6. 401、429、503、500 等错误能被前端分流提示，且不丢失用户输入。

体验优化阶段

1. 单场训练完成率达到 40% 以上。

2. JSON 解析成功率达到 95% 以上。

3. 用户能在 3 分钟内完成一次短训练。

4. 腾讯云智模板输出能稳定覆盖云计算、客户需求和产品文档能力。

5. 语音输入具备完整状态提示，失败时可自然降级为文字输入。

6. 低分反馈不只展示扣分项，还能给出清晰、克制、可执行的教练提示。

商业化探索阶段

1. 引入账号系统和跨设备训练历史。

2. 建立岗位训练课程包。

3. 结合人工评估提升评分可信度。

4. 引入真实用户反馈和 NPS。