王杨

AI质量工程 / AI驱动测试 / 高级测试开发 SDET

男 | 北京 | 10年工作经验 | cowboy231bb@gmail.com

wangmouren.online | github.com/cowboy231

核心技能

AI自动化测试：将测试目标、业务规则和数据约束整理为 Agent 可执行任务，覆盖接口、数据库核对、文档巡检及复杂业务 UI 测试；结合人工接管、关键路径复测和证据审查，形成可追溯的测试结论。

AI测评与Agent测试：中、日、英 ASR Benchmark，分语种 CER/WER、LLM 语义诊断与人工抽检。

自动化与协议测试：Python、pytest、Playwright、Robot Framework；HTTP、WebSocket、Mock、契约与集成测试；具备 Selenium、Appium 使用经验。

性能与数据质量：Locust、JMeter；吞吐、首包首块、实时率、P95、容量与 GPU 负载分析；MySQL、Redis、SQLite，数据对账、自动巡检与异常定位。

AI Coding与工程化：临床标本库使用 Codex + MiniMax 辅助全栈开发，覆盖 Vue 3、FastAPI、SQLite、RabbitMQ 与 Windows 内网交付；具备 Jenkins、GitLab CI、Docker 工程经验。

技术教学与团队协作：设计并迭代 17 周自动化测试课程，完成 3 次改版，累计覆盖 300+ 学员；具备录播交付、知识沉淀、新人培养及测试团队协作经验。

工作经历

全国采伐审批系统测试 · AI 自动化测试
AI 测试工程师（顾问合作） | 2026-10 - 至今

临床标本库自动化系统 · AI Coding 全栈实践
个人项目实践 | 2026-07 - 至今

商汤科技 AI研究院
AI质量工程师 / AI驱动测试工程师 | 2025-12 - 至今

积家科技-外贸自研摄像头项目
高级测试工程师 | 2024-07 - 2025-10

九坤私募基金管理系统
高级测试工程师 | 2023-08 - 2024-06

小米 CRM / RMS / Dynamics 365
高级测试工程师 | 2022-08 - 2023-08

天眼查 OpenAPI / 专业版
高级测试开发工程师 / 测试组长 | 2021-02 - 2022-08

自动化测试讲师（码同学）
自动化测试讲师（后期兼职收尾） | 2020-01 - 2021-03

马上消费金融
测试工程师 -> 测试组长 | 2016-11 - 2019-08

搜狗输入法
安卓测试工程师 | 2015-06 - 2016-08

2021-02 入职天眼查后，以兼职方式完成码同学剩余授课至 2021-03。

项目经历

全国采伐审批系统 · AI 自动化测试

顾问合作 · 复杂业务流程与人机协作测试 | 2026-10 - 至今

- 业务探索：接触系统三四天后，与 AI 结对梳理审批流程、字段依赖和业务规则，在测试执行中持续核对、建立业务理解。
- AI 自动化执行：将测试目标、前置条件和预期结果拆成 Agent 可执行任务，由 AI 准备数据、填写表单、复核字段和重开业务记录；本人处理关键审批、异常接管和验收。
- 业务规则验证：排除身份校验等前置干扰，对目标阈值执行测试，并通过独立重开同一记录，补充问题值持久化的缺陷证据。
- 跨环节结果核对：人工完成审批，AI 查询并计算前后额度差异，验证该对账子项；将业务规则、样本、操作记录和证据沉淀为后续可复用的测试资产。

临床标本库自动化系统 AI Coding 全栈实践

个人项目 · 需求、开发、集成与分层验证 | 2026-07 - 至今

- 从现场操作、既有业务系统与设备约束中梳理实体、状态、不变量和异常路径，明确每条业务链路的验收条件。
- 使用 Codex + MiniMax 辅助实现 Vue 3 前端、FastAPI 服务、SQLite 数据模型与 RabbitMQ 适配；连接上游业务系统及喷码、整板扫描、自动化冰箱，形成可运行的内网系统。
- 将上游业务记录、本地绑定与过程记录、设备物理结果分层保存；长任务依据设备终态与实际结果判断完成，并设计异常恢复与库存回读。
- 本人把控业务判断、证据可信度、变更粒度与真实写操作；AI 辅助实现、回归和排障。扫码器现场记录覆盖四个接口、95 管回包，发现六个协议漏写字段。
- 2026-08-24 离线快照：后端 1051 passed、7 skipped，前端 92 passed，Vite 构建成功；现场链路按设备和版本另行核验。

多语言 ASR 准确性 Benchmark

商汤科技 · AI 模型效果评测 | 2025-12 - 至今

- 将人工听测整理为数据制备、文本归一化、指标计算、语义诊断与人工抽检流程，复现公开评测框架并扩展中、日、英测试数据。
- 按语种设计指标：英文 casefolded CER/WER，中文 normalized CER，日文结合字符 CER 与 LLM 语义审核，识别表面差异与实际语义偏差。
- 将错误类型与 none/minor/medium/severe 四级严重度交叉归因；报告支持样本筛选、排序与定位，人工优先复核高风险样本。
- 2026-07-10 nightly 报告覆盖三语各 200 条样本，共 600 条；支持模型发布日生成报告。历史批次不外推为当前线上准确率。

AI 测试资产与 Agent 执行体系

商汤科技 · AI 驱动测试 | 2025-12 - 至今

- 在开发使用 AI 编码的快速迭代中，将测试目标、API/数据库约束、预期结果和证据要求整理为可执行任务。
- 接口测试和数据库核对以 Agent 执行为主；使用 AI 完成首轮冒烟，本人复测异常与关键路径，审查请求、响应、数据及日志后形成业务结论。
- 将复核过的方法、用例、边界及执行记录沉淀到测试资产库，分别记录当前规则、历史观察与待裁决提案，支持后续复跑和追溯。

LLM 容量矩阵与成本评估

商汤科技 · 性能与容量分析 | 2026-05 历史项目

- 围绕上下文长度 × 并发构建测试矩阵，在一致输入输出和流式口径下分析 E2E P95、TTFT P95、TPOT P95 与吞吐。
- 按 SLA 判断可用容量，区分请求成功与延迟达标；整理不同档位的厂商对比，保留测试配置、失败和高风险档位。
- 将容量数据组织为混合流量估算工具，展示输入占比、等效资源消耗和成本口径；用于方案评估与假设对比，按 SLA 和实测覆盖标注估算边界。

企业席位测试建模与证据审计

商汤科技 · 身份、权限与权益一致性 | 2026-06 历史项目

- 梳理手机号、用户身份、企业、批次、席位、套餐与权益的关系，依据实际身份核对权限和权益归属，明确状态转换与异常边界。
- 以测试矩阵关联 API、数据库和页面证据；历史报告记录 111 条用例：81 PASS、18 WARN、8 FAIL、4 BLOCKED，按状态保留未通过原因。
- 对有状态操作记录执行前后快照及清理结果；无待执行项不代表全部通过，历史状态也不替代当前线上结论。

三账本计费顺序与流水核对

商汤科技 · 计费链路验证 | 2026-04 历史项目

- 围绕积分、代金券、现金三个账本核对扣费顺序与跨账本切换，结合 usage 与交易流水验证实际落账。
- 2026-04-29 nightly 报告包含 9 个有效计费样本、3 个无效或异常样本，覆盖积分优先、积分不足转券、券不足转现金等组合。
- 区分余额快照和消费流水：现金消费在 BALANCE / CONFIRMED 流水中可见，账户余额字段仍为充值快照；逐笔对账以交易流水为准。

项目经历 续

17 周自动化测试课程体系

码同学 · 课程设计与技术教学 | 2020-01 - 2021-03

- 自主设计自动化测试课程，从 Python 基础、接口与 UI 自动化推进到持续集成，累计完成 3 次改版并覆盖 300+ 学员。
- 用业务规则设计编程练习，将条件分支、循环和用例设计联系起来；授课、答疑与录播内容并行，沉淀 Robot Framework、Airtest、Git 等专题。
- 2021-02 入职天眼查后，以兼职方式完成剩余授课周期。

九坤私募基金测试体系

九坤私募 · 复杂业务建模与数据验证 | 2023-08 - 2024-06

- 在真实生产数据不可用的保密约束下，梳理上下游业务规则，建立场景化 Mock 数据和 SQL 数据集。
- 按实体、字段和规则关系拆解基金业务，重构 Robot Framework 参数化与 BDD 脚本，推动上下游统一字段定义和取值口径。
- 形成可执行的基金全生命周期验证方案，将测试范围从页面功能延伸至业务规则与数据一致性。

积家科技 外贸自研摄像头项目

高级测试工程师 · 海外智能硬件 | 2024-07 - 2025-10

- 参与面向海外市场的自研电池摄像头项目测试；公司同时提供面向合作方的 SaaS 系统。
- 项目主要测试约束包括高频迭代、文档不完整、系统耦合与基础设施薄弱，复杂场景构造和问题定位成本较高。

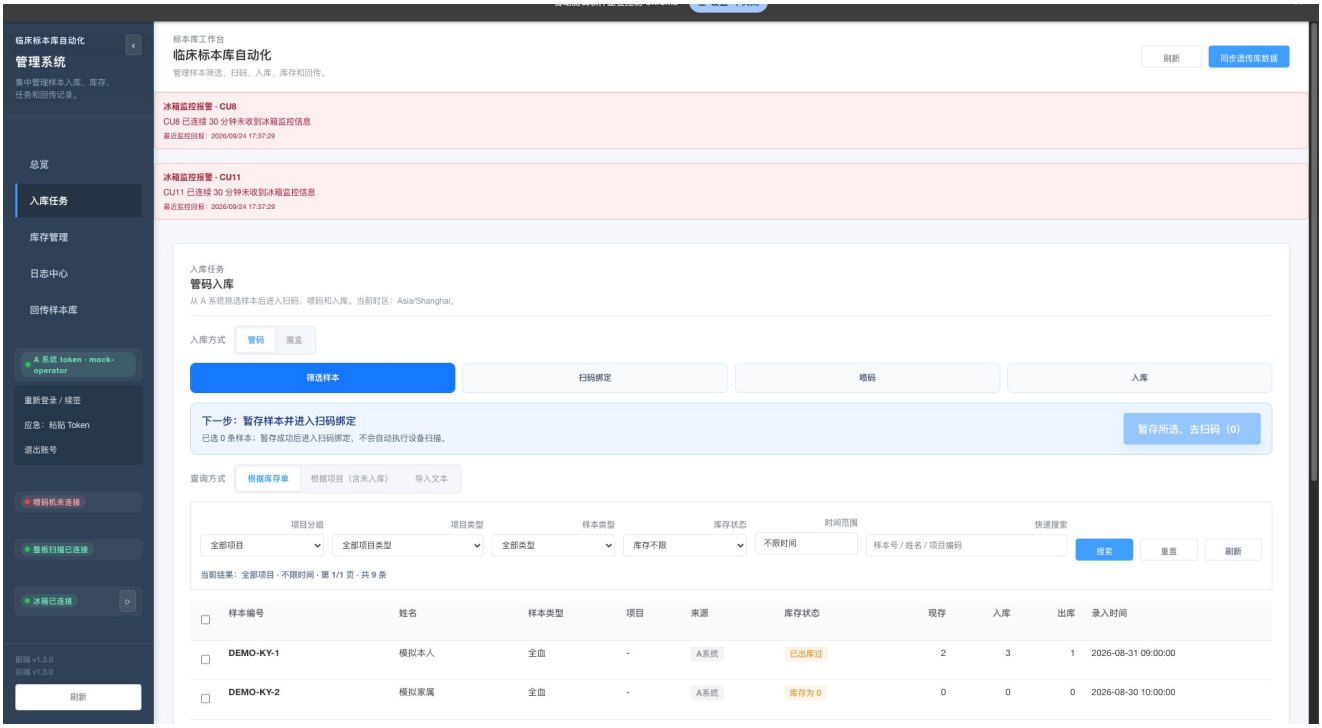
天眼查数据质量治理

高级测试开发工程师 / 测试组长 | 2021-02 - 2022-08

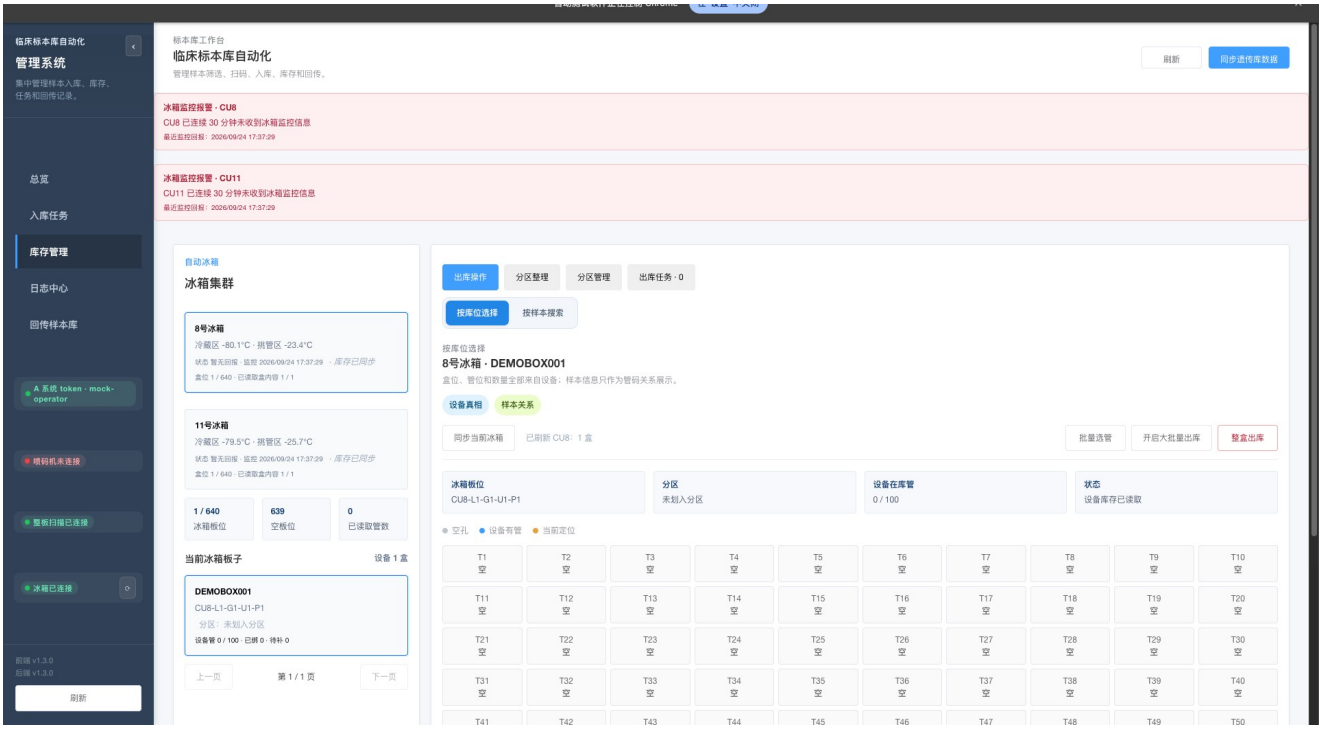
- 面向多实体 ToB 数据产品与 OpenAPI 梳理测试边界、异常分类及数据准确性风险，关注接口成功但业务数据不正确的问题。
- 建设自动巡检、数据准备与飞书告警套件，沉淀异常根因分类方法；承担新人培养、任务协调与测试团队协作。

作品集 临床标本库业务工作台

个人 AI Coding 实践。截图使用演示数据，展示业务流程与库存视图；不作为全系统设备验收证明。



入库工作台：以业务阶段组织筛选、暂存、扫码绑定、喷码与入库操作。



库存工作台：展示位置与状态。界面监控回报时间不等于回归测试快照日期。

作品集 ASR 三语评测与样本定位

2026-07-10 nightly 历史报告。三语各 200 条样本；调用成功和转写准确性分别统计。

Generated 2026-07-10

ASR Lite Nightly 三语精度测试综合结论

1. 测试结论

本轮在 nightly 环境完成 `senseaudio-asr-lite-1.5-260319` 的中/日/英 ASR 精度测试，每个语种 200 条样本，三语接口调用均为 `200/200` 成功。综合来看：

- 英文表现最好：按英文 casefolded 归一化后，CER 为 `0.5713%`，WER 为 `1.6128%`。
- 中文短音频表现较好：normalized CER 为 `4.1360%`。
- 日文 Reazon 子集字符级 CER 为 `18.8758%`；进一步用 `senseaudio-s2-lite` 做语义审核后，none/minor 语义可接受率为 `76.50%`，允许 medium 局部偏差时语义可用率为 `79.00%`。
- 本轮结果只代表 nightly 环境，不代表 staging 或 online；staging HTTP ASR 入口本轮曾返回 HTTP 500，不能作为精度结论环境。

3. 指标汇总

语种	样本数	成功数	推荐解读指标	CER	WER	平均延迟
英文	200	200	casefolded CER/WER	`0.5713%`	`1.6128%`	`909.3ms`
日文	200	200	normalized CER + 模型语义审核	`18.8758%`；语义可接受 `76.50%`	不建议按空格 WER 解读	`804.5ms`
中文	200	200	normalized CER	`4.1360%`	不建议按空格 WER 解读	`667.9ms`

结论与指标：按语种选择 CER/WER 等解读口径，并声明环境边界。



样本明细：通过语义诊断、错误类型及 CER 筛选样本，支撑人工抽检。

作品集 ASR 运行诊断与错误归因

同一历史批次的执行概览与错误分布。每个样本可有多个错误标签，占比不能简单相加。

运行概览

任务名	asr_lite_nightly_ja_200_20260710
运行 ID	20260710_152529_845_c77276
样本总数	200
成功样本数	200
成功率	100.00%
空转写样本数	0
空结果率	0.00%
协议失败数	0
打分失败数	0
平均原始CER	22.41%
平均规范化CER	18.88%
无吐Prompt样本平均规范化CER	18.88%
疑似吐Prompt样本数	0
疑似吐Prompt占比	0.00%
简体对齐平均规范化CER	18.88%
简繁转换后CER变化样本数	0
平均延迟(ms)	804.51

运行概览：调用成功、空结果、协议或打分失败与准确性指标分别呈现。

错误类型统计

错误类型说明

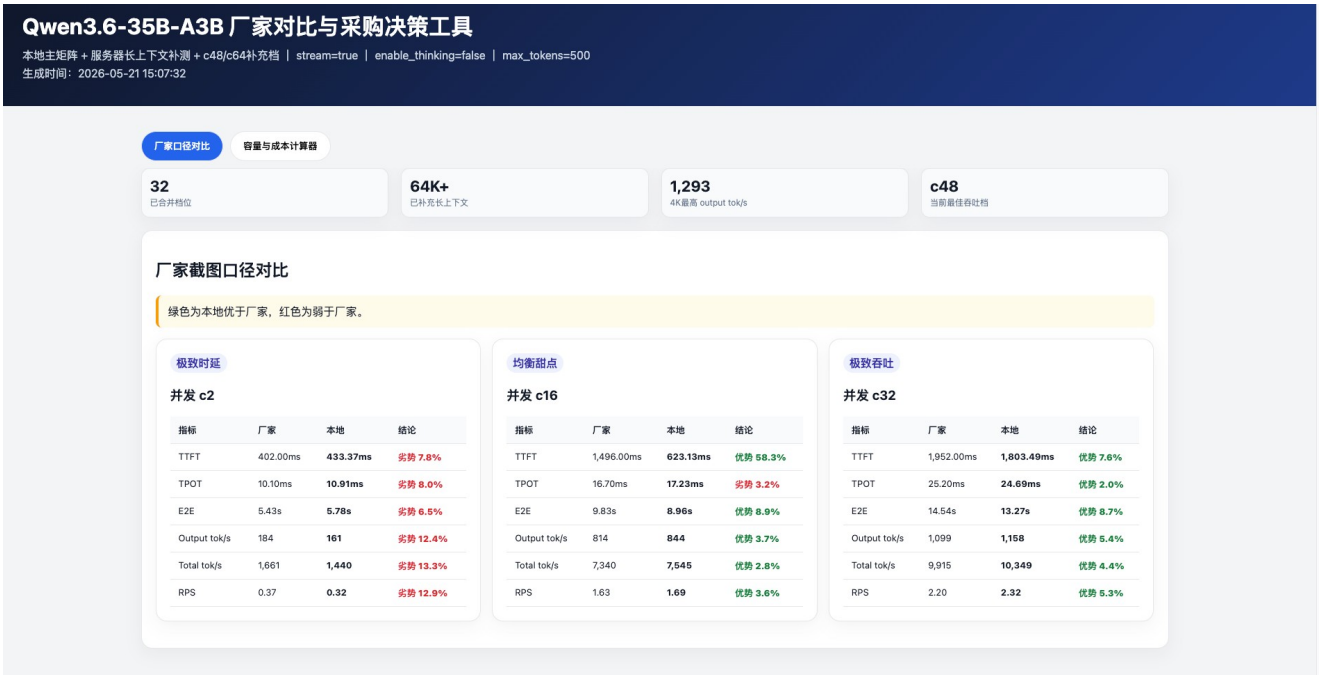
按大模型语义诊断输出的错误类型统计；一个样本可能包含多个错误类型，所以占比相加可能超过 100%。分母为已成功诊断且非 judge_error 的样本数。

错误类型	中文描述	样本数	占比
surface_form_difference	表记/标点/形式差异	75	37.88%
substitution	替换错误	34	17.17%
insertion	插入多余内容	31	15.66%
exact_or_punctuation_only	exact_or_punctuation_only	27	13.64%
omission	漏识别	26	13.13%
named_entity_error	专有名词错误	26	13.13%
homophone_error	同音/近音错误	17	8.59%
hallucination	杜撰/幻觉输出	6	3.03%
word_order_difference	语序差异	4	2.02%
number_error	数字错误	3	1.52%
empty_output	空结果	2	1.01%

错误分布：图中是当轮出现的类别，不能解释为系统永久只有 11 类错误。

作品集 LLM 容量分析与估算方法

2026-05-21 历史产物。按相同测试口径比较延迟与吞吐，成本工具仅作为假设下的估算参考。



厂商比较：按并发档位呈现时延、吞吐及差异，保留优势与不足。

估算环节	方法与检查点
输入约束	上下文长度、输入输出量、流量占比、流式与 thinking 配置
实测基线	各上下文与并发档位的 TTFT、TPOT、E2E P95、吞吐和失败
流量折算	以等效资源消耗估算混合流量容量，标记缺乏实测的区间
成本评估	结合假设运行时间与费用估算单位成本；不展示未核定价格
使用边界	先核验公式与 SLA 风险，再用于方案讨论，不宣称已产生采购收益

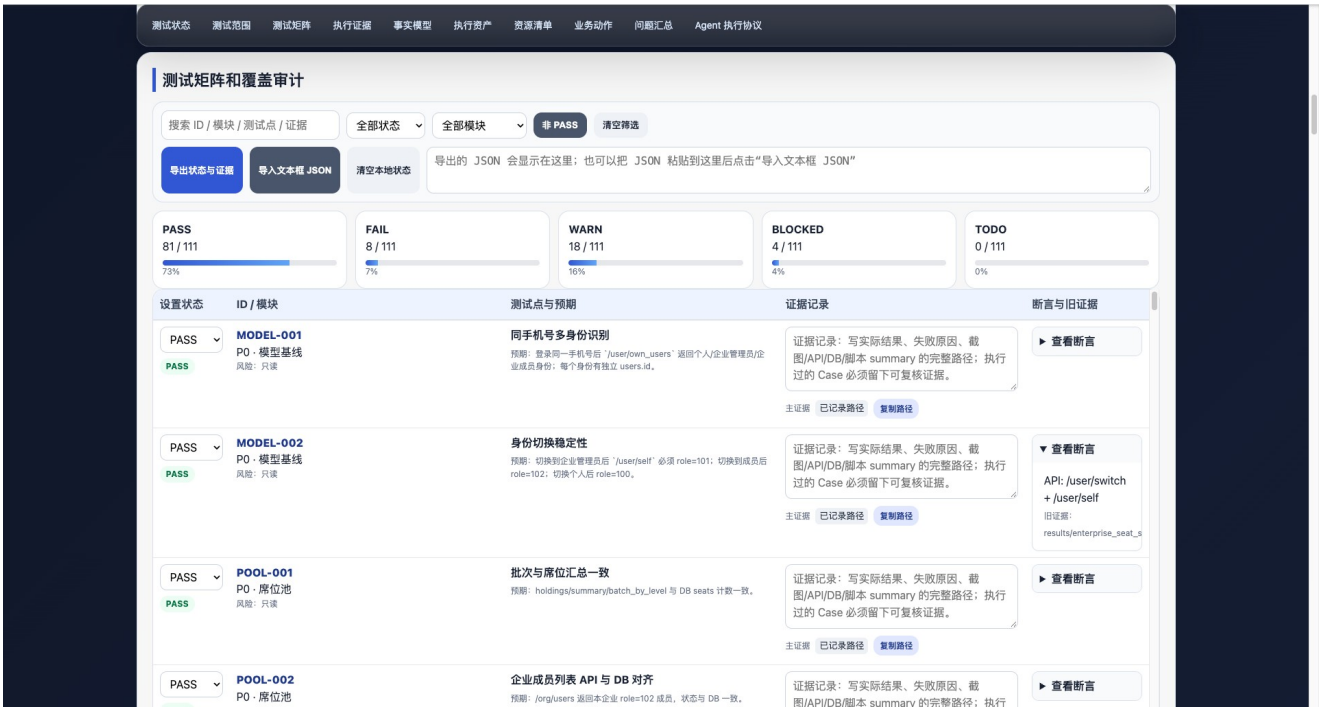
本页方法表由报告逻辑重述，不包含原截图的价格输入与账户数据。

作品集 企业席位建模与覆盖审计

2026-06 历史证据中枢。先明确实体、状态与权限，再按用例关联证据。



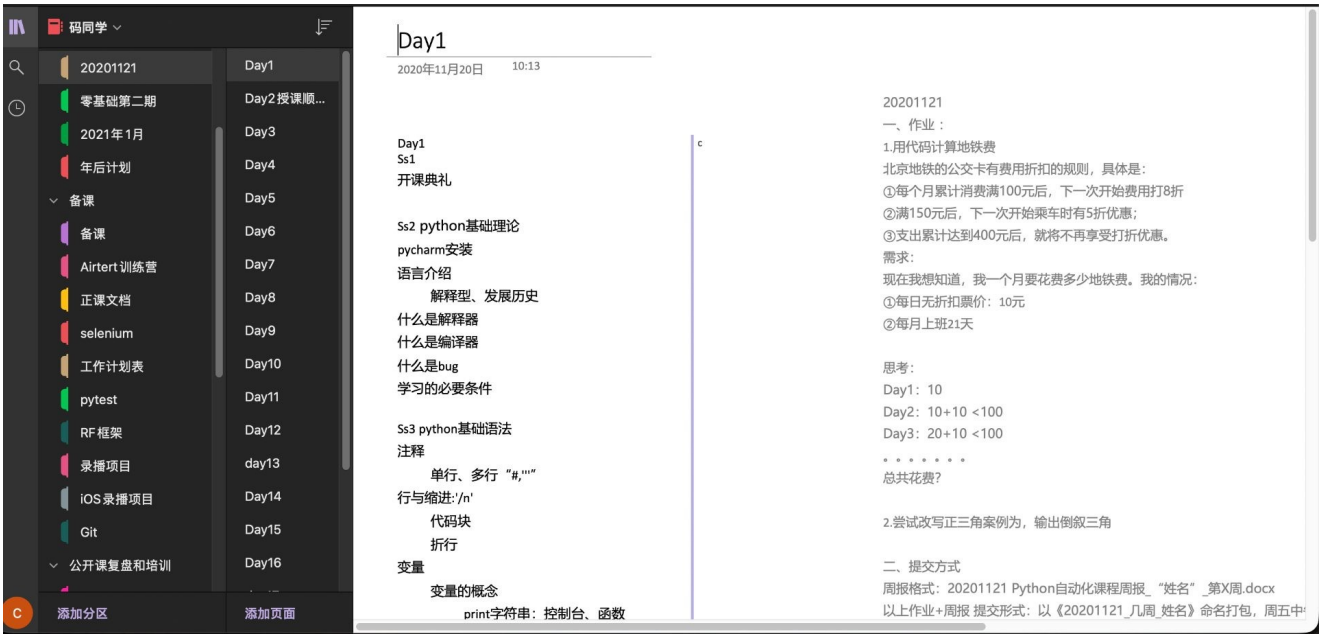
事实模型：区分手机号与用户身份，整理企业、席位、套餐与权益关系。



覆盖审计：111 条中 81 PASS、18 WARN、8 FAIL、4 BLOCKED；TODO 为零不代表全通过。

作品集 技术教学与计费验证方法

展示技术表达和业务核对能力。计费方法表依据 2026-04-29 历史报告重述，不含账号或请求标识。



课程资料：按授课阶段组织 Python 与自动化专题，以业务规则设计练习。

计费场景	核对重点
只有代金券	请求使用记录与代金券交易流水对应
积分优先	确认积分账优先消费，未触发的账本不扣费
积分不足转券	校验积分耗尽后的剩余费用进入代金券账
代金券不足转现金	分别对账代金券与现金流水，核对跨账本分配
券耗尽后现金扣费	以 BALANCE / CONFIRMED 流水核对现金消费
异常样本	与有效计费样本分开统计，不把样本数写成路径数

历史报告：9 个有效计费样本、3 个无效或异常样本。余额充值快照不替代消费流水。

作品集 需求指标与测试证据对照

2026-07-03 历史报告示例。将需求、测试结论、证据及风险放在同一矩阵，保留 PARTIAL 与待复核项。



官方要求满足矩阵

根据征集文档，语音与音频类核心要求包括首包时延、单账号 QPM、中文/英文、热词/VAD/语义修正、语音翻译准确度与端到端时延。下表先给商务可用结论，再链接证据页。

官方关注指标	状态	当前结论	证据页	边界/风险
ASR 首包时延 语音识别：首包时延≤2s	PASS	Pro 流式首服务端事件约 103ms，首转写约 1885ms	JIRA-ASR-STREAM-001	首转写口径接近 2s，建议按评审脚本复测确认
ASR 单账号吞吐 语音识别：单账号≥300 QPM	PASS	同一 online API key 下，Pro 1.1s/10.56s 两档 300 QPM 均 300/300 成功	JIRA-ASR-PERF-001	当前为短窗口压测；正式提交建议延长窗口
中英文识别与音频时长 支持中文（含方言）/英文，覆盖短音频与 10 秒级音频	PASS	Pro 已覆盖中文 1.1s/10.56s/15s 在线转写；英文样本在翻译与多语种包中保留	JIRA-ASR-PERF-001	方言专项样本需按评审语料补充
热词/VAD/语义修正 具备热词/VAD/语义修正能力	PARTIAL	已完成 20 条运营商术语链路验证，19/20 命中	JIRA-HOTWORD-001	当前是术语识别链路，不是热词开关 A/B 增益证明
语音翻译准确度 语音翻译：准确度≥90%	PASS	20 条小样本链路成功，语义初审 19/20，可支撑初版 95% 口径	JIRA-TRANS-ACC-001	正式外发前需要人工 final review
语音翻译端到端时延 端到端延迟≤4s（流式处理），单账号≥50 QPM	PARTIAL	已按正常语速补 50 QPM 专项压测，60/60 成功；平均约 6.28s，时延仍需优化/口径确认	JIRA-TRANS-ACC-001	当前不能对外承诺 50 QPM 下端到端≤4s 全量满足
接口/SLA/账号交付 标准公网 HTTPS、Token 鉴权、API 调用账号/密钥、流量管控	PASS	本轮所有 Pro 证据均通过 online HTTPS + 单个 API key 执行，可证明单账号调用链路	JIRA-CAPACITY-001	生产 SLA 仍需商务/平台承诺文件补充

推荐评审路径

从 ASR 单账号 300 QPM 与稳定性开始，再看流式首包、翻译和热词。Lite HTTP 页面仅作为备选模型附录，不放在商务主结论里。

生成信息

生成时间：2026-07-03 11:22:01 CST
主线模型：senseaudio-asr-pro-1.5-260319
执行口径：online HTTPS 接口、Token 鉴权、同一 API 调用账号/密钥。

指标与证据对应：首包、吞吐、识别与翻译分开评价；短窗口结果不升级为长期 SLA。

报告表达先给出适用范围，再说明执行配置和证据；未达项保留限制条件，避免仅凭请求成功判断验收。

这组作品展示历史方法与产出，不代表当前线上表现、商务承诺或已完成正式采购。