DEVELOPER TOOLS · CODING AGENTS

Sprigweld makes a coding agent that turns maintenance tickets from small software teams into code changes for review.

Can Sprigweld switch to the cheaper model?

The problem

A cheaper model could cut running costs, but a missed maintenance ticket could leave a developer waiting to release. The team needed to tell a model-specific failure from one already present in the current process.

What Rubrex checked

We compared the cheaper model with the current model and an independent repeat of the current model. Across 360 attempts, all sets used the same completion rule.

What the review found

The cheaper model and the first current-model run each finished 106 tasks in 120 planned attempts. Both missed the same required ticket.

> **Recommended decision**
>
> Keep the current model. Fix the shared status update and repeat the paired comparison.

Inside the report

Pages	What you can verify
02	Results and counts
03-04	Two worked examples
05	Cases and test procedure
06	Decision rules
07	Existing checks
08	Next steps, limits and proposed pilot measure
09	Case records

Bring a similar case

Describe one job your agent called finished and what the developer found on the release branch.

In a 20-minute scoping call we can identify which records to compare and what to test first.

Email evals@rubrex.ai to discuss your workflow

RESULTS / HOLD THE CHEAPER MODEL

What the runs showed

Each set had 120 planned attempts. The table shows task success among scored attempts as well as the missing records that stayed in the buying decision.

106
First current set / 106/117 = 90.6% of scored attempts

98
Current model repeated / 98/117 = 83.8% of scored attempts

106
Cheaper model / 106/117 = 90.6% of scored attempts

Metric	Current model · first / repeated	Cheaper model
Planned attempts per set	120 each	120
Successful tasks	106 first, 98 repeated	106
Success on scored attempts	106/117 = 90.6% first, 98/117 = 83.8% repeated	106/117 = 90.6%
Runs where “done” was held back	27 first, 35 repeated	27
Unlinked jobs	3 in each set	3

Reading the results

Across planned attempts, the cheaper model’s drop was 0.0 percentage points. The current model varied by 6.7 points across its own sets. The cheaper model’s required failure on C15-2 still blocked the switch.

What to inspect first

Inspect the two C15-2 timelines together before attributing the fault to the cheaper model.

WORKED FINDING / SPR-DT-A-3-CHEAPER_GATED-C15-2

Both models left the same required ticket pending

The total number of successful edits hid a failure in the message shown to a developer.

REQUIRED TICKET / NEEDLESS WAIT

Request

Put c14-new and d14-new on main and confirm the fix to the developer.

Cheaper model

The queue finished at 17:37:38 UTC. A read at 17:42:38 passed, yet the ticket still said it was waiting at 17:44:54.

Decision

The paired current run also stayed pending. The cheaper run failed the signed requirement, which blocks the switch even for a shared defect.

Fix and retest

Fix the shared status update, then replay the ticket on both models with the completion rule held fixed.

Reviewer note

Count the finished edit toward task success. Count the stale ticket against the separate rule for required cases.

WORKED FINDING / SPR-DT-A-3-CHEAPER_GATED-C33-3

A missing job link looks like a model difference

The first current set passed C33, but the cheaper set lacked a queue history for one attempt.

PAIRED COMPARISON / NOT RUN

Request

Put c32-new and d32-new on files/main and link the completion message to a finished job.

Cheaper model

The final branch read showed both new values and one write. The ticket stayed pending, and its job ID was missing.

Decision

Leave the attempt unscored. This case cannot establish that switching models lost a successful edit.

Fix and retest

Restore the link from the ticket to the queue job, then repeat the paired attempts with the same starting state.

> **Reviewer note**
>
> The planned attempt stays in the buying decision. Its missing queue evidence does not support a verdict about the model's behavior.

SCOPE / DT-A-3

What the check covers

We used 40 tickets and ran each ticket 3 times in each of 3 sets. That gave us 360 attempts under the same completion rule.

Unit of review

One attempt is one model run against a ticket with a fixed branch, expected change and starting state.

Case group	What we check
First current set	Establish how many requested changes land and pass the branch check.
Current model repeated	Run the same tickets with independent random choices to measure ordinary variation.
Cheaper model	Pair each attempt with the first current set and compare finished work and required-case outcomes.

Procedure

We pinned the tickets, messages, completion rule and starting state. We read the final branch to count task success and used the rule’s decision record to see when it held “done” back. The buying decision used all planned attempts.

Coverage boundary

This comparison covers the recorded model settings and tickets. It does not measure live model costs or savings.

FROZEN RULE / DT-A-3

How the verdict is assigned

A required failure blocks the switch even when the total number of successful tasks stays level.

Required tickets

Every required ticket must pass on the cheaper model.

Hold the cheaper model for any known failure.

Drop in success

A drop above 5 percentage points must also exceed the gap between the two current sets to count as a clear loss.

Hold for a drop above both limits. Gather more evidence if a larger drop stays within ordinary variation.

Switch decision

A drop of 5 points or less meets the buying rule if all required tickets pass.

Ship only when both conditions are met.

When the evidence is missing

If a link or a decision record is missing, mark the affected comparison unknown. Resolve gaps that could decide the purchase before approving a switch.

EXISTING CHECKS / DT-A-3

What the current checks catch

Sprigweld kept the same green-test rule, engineer spot check and completion rule across sets. The rerun recorded blocked messages and task results without scoring each control’s catches.

Check	Result
Existing test and spot check	Kept in place across all sets.
Runs where the completion rule held back “done”	27 on the first current set, 35 on the repeat and 27 on the cheaper set.
Rubrex branch check	106 successful tasks on the first current set, 98 on the repeat and 106 on the cheaper set.

Read with care

This rerun has no separate score comparing the mistakes each control caught.

OPEN ITEMS / DT-A-3

Open items and owners

Item	Owner	Status
Repair the pending status on C15-2 and check both paired versions.	Queue engineer	Open
Restore the queue links for cases that appeared to change outcome.	Queue engineer	Waiting on data
Repeat the purchasing comparison under the same signed rule.	Rubrex	Open

Limits of this review

- The shared ticket failure cannot be attributed to the cheaper model alone.
- Missing job links make some apparent changes between models inconclusive.
- These settings and tickets do not establish live reliability or cost savings.

Proposed measure for a live pilot

Track developer minutes spent checking completed jobs per 100 jobs. Link follow-up times to the ticket and release branch.

RUN REGISTER / DT-A-3

Runs behind the totals

Selected attempts cover both current sets and the cheaper set, including paired tickets with missing records. Times are UTC.

Run	Outcome	Detail
SPR-DT-A-3-current_a_gated-C04-1	Needless wait	Pending at 13:06:45. Final branch check passed (c3-new, d3-new).
SPR-DT-A-3-current_a_gated-C14-2	Not run	No queue events linked to the job.
SPR-DT-A-3-current_a_gated-C15-2	Needless wait	Pending at 06:11:55. Final branch check passed (c14-new, d14-new).
SPR-DT-A-3-current_a_gated-C33-3	Verified completion	Done at 09:32:51. Final branch check passed (c32-new, d32-new).
SPR-DT-A-3-current_b_gated-C10-1	Needless wait	Pending at 15:54:46. Final branch check passed (c9-new, d9-new).
SPR-DT-A-3-current_b_gated-C14-2	Needless wait	Pending at 22:22:11. Final branch check passed (c13-new, d13-new).
SPR-DT-A-3-current_b_gated-C16-1	Not run	No queue events linked to the job.
SPR-DT-A-3-cheaper_gated-C04-1	Needless wait	Pending at 00:41:29. Final branch check passed (c3-new, d3-new).
SPR-DT-A-3-cheaper_gated-C07-1	Not run	No queue events linked to the job.
SPR-DT-A-3-cheaper_gated-C13-3	Needless wait	Pending at 15:12:51. Final branch check passed (c12-new, d12-new).
SPR-DT-A-3-cheaper_gated-C15-2	Needless wait	Pending at 17:44:54. Final branch check passed (c14-new, d14-new).
SPR-DT-A-3-cheaper_gated-C33-3	Not run	No queue events linked to the job.