PURETENSOR
INTELLIGENCE DISTILLED

Restore Drills a Fake Cannot Pass

148 agents worked on 27 waived drills, and a binary verdict that never converged

Paper: PT-R-2026-068 v1.0

Author: PureTensor Ltd

Date: 8 November 2026

Status: Published

Abstract

A restore drill is a scheduled job that restores a backup into scratch space, checks the content against live data, and exports pass or fail with a measured recovery time and recovery point. A drill that cannot fail is a false green, like an alert rule that cannot fire. On 8 November 2026 we put 148 agents to work on 28 backup jobs whose drills had been waived: 21 builders wrote drills, 67 adversaries tried to make each drill pass on a truncated, corrupt, empty, stale or swapped backup, 59 fixers closed what the adversaries found, and one agent landed the result through our review gate.

One drill met the bar and merged. The other 27 stay waived, and each waiver now names the branch that holds its drill and the damage that still passes. Under our starting rule, any passing attack makes a drill "vacuous", and it returned vacuous in all 48 recorded rounds. The pass fraction did not fall: 29.6%, 32.0% and 34.7% across rounds 1 to 3 for the 13 units that reached a third round. Of 650 individual attacks, 36.0% passed.

In a third workflow we split passes into material (a damaged backup passes without forging the drill's own reference) and residual (a coordinated forgery of backup and reference, or a cosmetic difference). Among 81 triaged passes, 44 were material and 37 residual. That made one verdict decidable, the Ceph configuration drill that landed, and left nine of ten triaged units with material holes at their last returned result. The triage did not demonstrably reduce holes; we say so plainly.

The drills also found a dashboard backup that had been archiving a dead instance (5 dashboards against 29 live), and a laptop backup whose newest completed snapshot was 63.3 hours old at the drill run, inside its 96 hour ceiling but heading for it. The landed drill's first production run was a false red, caused by the strict comparison we added to close adversary holes. This is one day, one campaign and one model family; Section 8 states what the numbers will carry.

1. A Registry Full of Waivers

Our registry of backup jobs holds one row per job, carrying either an expected restore drill or a dated waiver with a reason. At the parent of the closeout merge the registry had 66 rows: 35 expected drills and 31 waivers, of which 26 cited a pending drill.

A pending waiver is an honest statement that nothing tests the restore, and it is easy to leave alone indefinitely. The directive on 8 November was to convert about 28 of the 31 into real scheduled drills, leaving operator-only rows as they were. Each new drill had to be shown to fail on bad data before it landed: one agent builds it, a second tries to make it pass on a damaged backup, and it then goes through the review gate. The whole effort ran under a hard cutoff at 21:00 UTC with no new rounds after it.

A drill that passes against an empty database is worse than no drill, because it generates evidence of safety. The question was what an adversarial campaign against drills produces, and whether its verdict rule can converge.

2. Campaign Design

The unit of work is a family of related backup jobs that share one drill kit. We attacked 18 units: 17 whose drills stayed waived and one that landed. A nineteenth inventory entry was only retired. One of the 17 (a deduplicated-backup data unit) had no binary round, only triage attackers. Two units were built twice, in different workflows.

For each unit the loop was the same. A builder ported any unlanded draft onto current main, then proved the drill red on damaged copies and green on the real backup. Up to three attack rounds followed. In each, adversaries built damaged, empty, stale or swapped scratch copies and ran the drill against them; the real backup was never modified. Each round was followed by a fixer. Every attack ended in one of three outcomes: the drill went red as expected, the drill stayed green on damaged data (a vacuous pass), or the attack was skipped. A round was declared "vacuous" if any attack passed.

Three workflows ran. The first, with 89 agents (17 builders, 42 attackers, 30 fixers), ran from about 13:07 to 20:42 (first agent activity to last). The second, with 15 agents (4, 6 and 5), ran from about 14:09 to 21:00. The third was the triage workflow, with 44 agents (19 attackers, 24 fixers and the lander), from about 19:30 to 21:00:20. In it a pass was material if a damaged backup passed without forging the drill's independent reference, which blocked landing, and residual if it needed a coordinated forgery of backup and reference or was cosmetic, which was recorded without blocking.

The 148 agents divide into 21 builders, 67 attackers, 59 fixers and 1 lander, all non-interactive sessions of one hosted frontier model working on scratch copies. A Ceph health gate makes a drill skip unless the cluster reports HEALTH_OK; we return to that gate in Section 7.

3. The Binary Verdict Never Converged

The agent journals record 48 attack rounds, and all 48 returned "vacuous". Those rounds held 650 individual attacks.

Quantity	Value
Attack rounds recorded	48
Rounds with verdict vacuous	48 (100%)
Individual attacks	650
Failed as expected	390 (60.0%)
Passed	234 (36.0%)
Skipped	26 (4.0%)
Fewest passes in any round	1

Table 1. Binary rounds from workflows 1 and 2, from the agent journals.

The minimum number of passes in any round was one. No round ever had zero. A rule of this shape can only converge if the adversaries run out of ideas faster than the fixers close shapes, and the numbers say they did not. For the 13 units that reached a third round, the pass fraction did not fall (29.6%, 32.0%, 34.7%).

Unit (generic)	R1	R2	R3
Ceph config export (landed)	6/12	4/13	4/10
DNS zone export	5/11	2/11	3/10
Web-app backup (two rows)	6/18	7/16	4/10
Diary database replica	1/11	2/11	6/13
Drive sync family (two rows)	3/19	6/23	4/11
Git mirror sync	4/15	5/12	4/10
Clone sync	5/14	2/6	3/9
Photo library family (three rows)	6/24	5/24	8/31
Volume snapshot daily	2/10	9/16	3/13
Small-mirror rsync family	2/13	3/9	2/9
Task-queue local backup	5/14	1/10	4/17
CMS backup	7/13	7/14	6/11
Dashboard backup	3/12	3/10	7/13
Total	55/186 (29.6%)	56/175 (32.0%)	58/167 (34.7%)

Table 2. Passing attacks over attacks, by unit and round, for the 13 units that reached round 3.

Passes per unit moved up and down with no trend. The attack count per round fell from 186 to 167. A plausible reading, which these counts do not test, is that fixers closed the specific shapes found and the next attacker tried a neighbouring shape; the fraction also depends on which attacker agent drew each round.

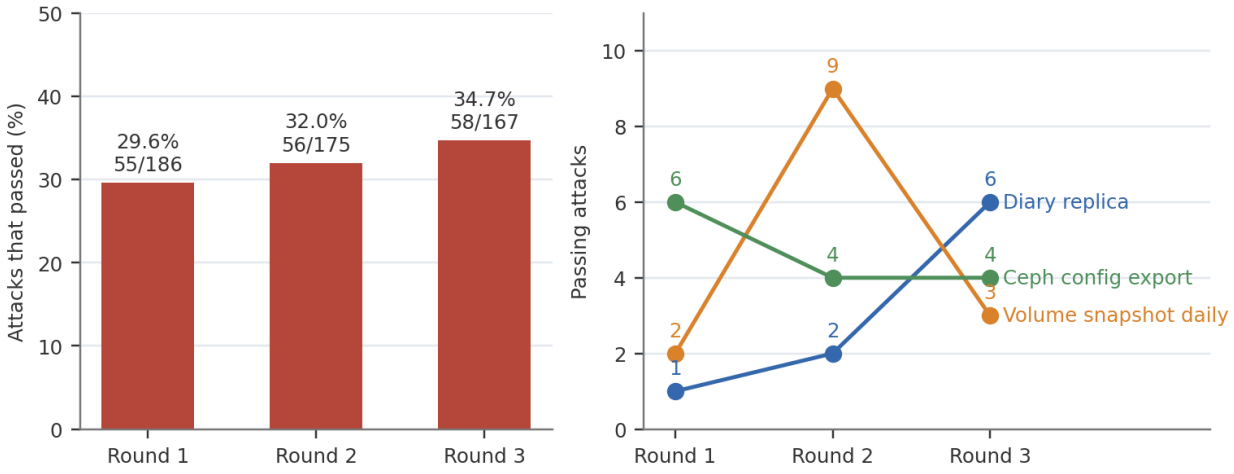


Figure 1. Left: pass fraction by round for the 13 units that reached round 3. Right: passes per round for three units, chosen to show counts that rose, fell or oscillated. Values are in Tables 2 and 3.

Unit	R1	R2	R3
Diary replica	1	2	6
Volume snapshot daily	2	9	3

Unit	R1	R2	R3
Ceph config export	6	4	4

Table 3. Passes by round for three units, chosen to show the absence of a trend.

Other units had different round counts. The dashboard-backup unit had a fourth round (5 of 9 attacks passed), the laptop-backup unit two (12 of 16, then 5 of 8), and three more units two in the journals (5 of 9, then 10 of 18; 10 of 14, then 11 of 22; 4 of 14, then 3 of 12). Of the 17 held units, 12 reached a third round, and all 16 with a binary round were vacuous at their last one; the seventeenth had triage attackers only.

The per-unit state records on disk hold 44 of the 48 rounds (589 attacks); the journals hold all 48. Headline figures use the journals; Table 2 uses the on-disk records, which match the journals for those rounds.

4. Triage: Material Versus Residual

After the binary rule kept returning vacuous, we introduced the material versus residual verdict in the third workflow. Nineteen triage attackers were launched; 12 returned results before the cutoff, covering 10 units.

Unit (generic)	Results	Material passes (field value)	Residual passes
Ceph config export (landed)	1	0 (0)	4
DNS zone export	2 (third not returned)	2 (3), then 1 (1)	2, then 4
Web-app backup	2	6 (7), then 7 (7)	1, then 5
Drive sync family	1	5 (4)	1
Git mirror sync	1 (second not returned)	1 (1)	7
Clone sync	1 (second not returned)	2 (2)	4
Volume snapshot daily	1 (second not returned)	5 (5)	2
Task-queue local backup	1	2 (2)	3
CMS backup	1	7 (7)	4
Deduplicated-backup data unit	1 (second not returned)	6 (6)	0

Table 4. Triage results by unit. The first number in the material column counts attacks labelled as material passes; the bracketed number is the attacker's own summary field.

Over the 12 results there were 157 attacks: 73 failed as expected, 44 passed as material, 37 passed as residual, and 3 were skipped. Residual passes were 37 of the 81 passes (45.7%). Under the binary rule all 81 count against the drill. The summary field and the outcome count disagree in three results (drive sync 4 against 5; DNS zone round 1, 3 against 2; web-app round 1, 7 against 6).

Scheme	Attacks	Failed	Passed	Material	Residual	Skipped
Binary, 48 rounds	650	390	234	not split	not split	26
Triage, 12 results	157	73	81	44	37	3

Table 5. Outcome mix under the two verdict rules.

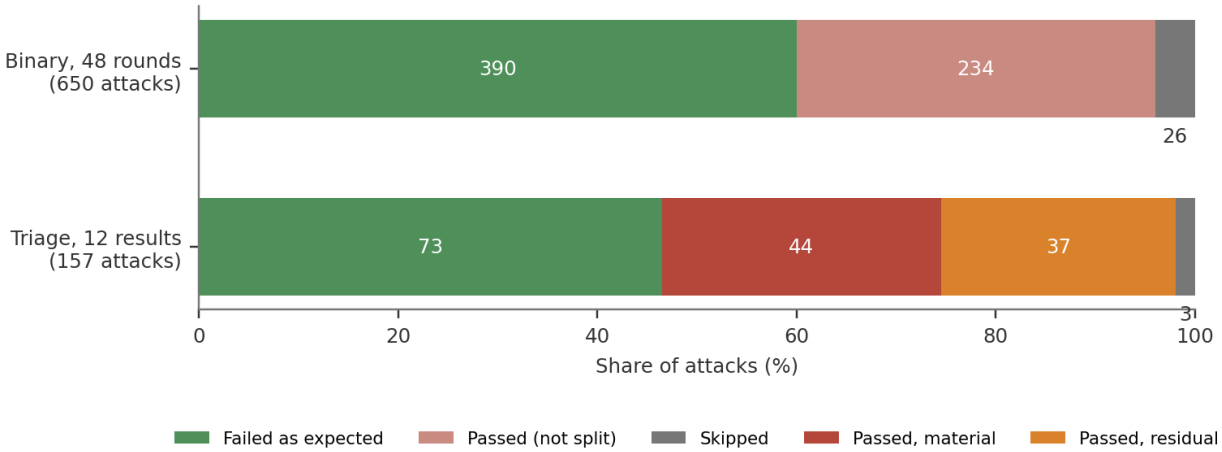


Figure 2. Attack outcomes under the binary rule (650 attacks) and the triage rule (157 attacks), as shares of each total. Values are in Table 5.

The claim that severity triage fixed the non-convergence is not supported. One unit of ten had zero material holes, and it was the one that landed. Nine still had material holes at their last returned result, up to seven in the CMS and web-app units. The only unit with two falling triage results was the DNS zone unit (3 material holes by the summary field, then 1); the web-app unit was flat by the attacker's summary field (7, then 7), though by outcome count it rose (6, then 7), and the DNS unit by outcome count fell from 2 to 1. The landed unit had already been through three binary rounds and several fixes, so its zero cannot be credited to the new rule. The rule made one verdict decidable.

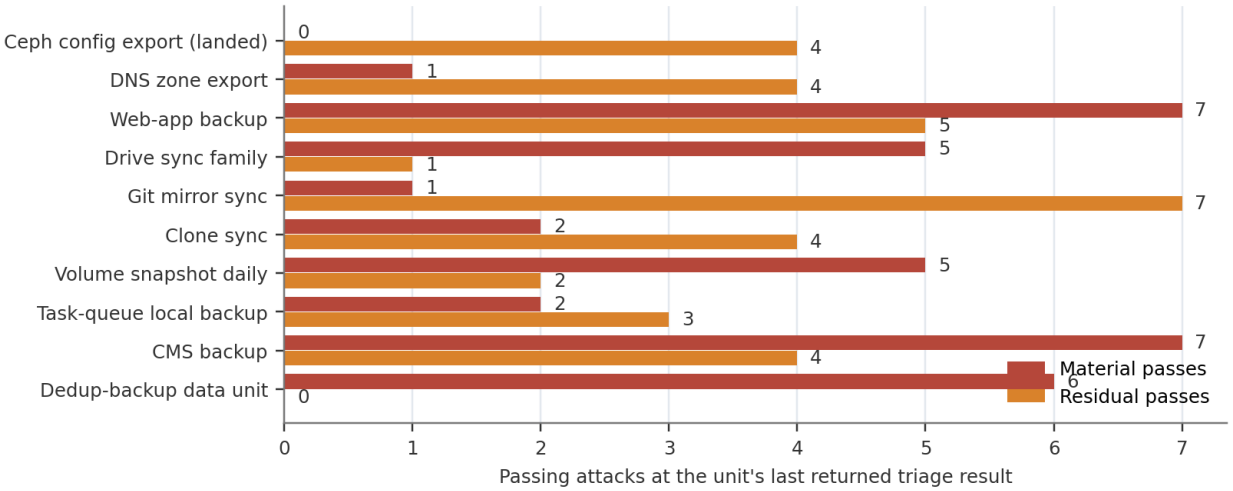


Figure 3. Material and residual passes at each unit's last triage result. Values are in Table 4.

5. The One Drill That Landed

The drill that met the bar checks the off-site copy of the Ceph configuration export: cluster maps, pool detail and configuration dumps, 8 files and 42,160 bytes in all. It runs daily, pulls the copy, restores it to scratch space and compares it with the live cluster. A pass requires 13 checks, with a floor of 10.

Its first live green came at 13:49 UTC with 13 of 13 checks, a recovery time of 13 s and a recovery point of 34,692 s (about 9.6 h, against a limit of 129,600 s). After later fixes a second live green at 19:53 gave 13 of 13, a recovery time of 10 s and a recovery point of 1,789 s. A green at 19:22 had gone red on genuine intraday drift in one dump (the export predated the

live cluster) and passed once the producer's own daily job had refreshed the export.

The red proofs grew with the adversary rounds, from 12 damage modes to 20: four added at 13:49 to 13:50 (storage daemon (OSD) weights rewritten, OSD addresses wiped, pool fields stripped, monitor header lines deleted) and four more at 19:53 (the config level column, the config read-only column, 16 trailing bytes after the placement map, junk in the pool file). Every red run returned a failing exit code, though two of the 13:50 runs were first skipped by the health gate and rerun with a wrapper that changed only the health report. Hermetic tests (needing no live services) grew from 69 to 100 to 115, and 116 after the later fix.

Round 1 passes (6 of 12) were valid-but-wrong placement maps, a truncated dump, a forged older epoch (a version counter) and an edited manifest. Round 2 (4 of 13, with 2 skipped) found edits to weights, addresses, pool fields and monitor headers; the fix was an exact line-for-line comparison against the map's plain text at the exported epoch. Round 3 found 4 more passes of 10.

The change was 20 files with 3,076 insertions and 105 deletions (kit 1,317 lines, tests 1,182). It merged at 22:18:32 UTC, after its first head failed its own registry tests because an expected-set pin had not been updated when a row flipped (from the landing record). The registry then stood at 65 rows: 36 expected and 29 waivers.

6. What Stayed Waived, and What the Drills Found

The 29 waivers left after the merge are 27 held build waivers and two operator-only rows that no agent can run. The arithmetic is 31 waivers before, minus the one that landed, minus one retired inventory row. The 27 sit on 17 unit branches, all pushed to our own forge.

Reason text	Rows
Green not re-verified at closeout	19
Red-proven on 8 November (no live run)	6
Final attack incomplete at the cutoff	5 (overlaps the first group)
Names open material holes	1 (web-app backup: 7)

Table 6. Reasons recorded on the 27 held waivers. Tiers are 17 tier 2, 6 tier 1, 3 tier 3 and 1 tier 0; review dates fall between 15 December 2026 and 12 February 2027.

Each waiver also names, at class level only, the damage that still passes. All the damage involved needs write access to scratch copies only.

Two findings surfaced along the way. The dashboard backup's nightly leg archived the volume of a dashboard instance on the monitoring host. That instance was dead: its database was last written on 28 March 2026 and held 5 dashboards, while the live instance in Kubernetes held 29. The old leg's only success test was that the archive and copy exited zero. The drill, comparing restored content with the live service, went red on 8 November. If the job ran nightly throughout, that is up to about 225 days of archives of the wrong instance; the dead instance's last write is dated by a comment in the script, and the start of the fault is not recorded. The fix archives the live instance (an integrity-checked database copy cross-checked against the API dashboard count, plus a full API export) with a per-file hash manifest. The drill is now green with 11 of 11 checks, 29 of 29 dashboards, a recovery time of 481 s and a recovery point of 74,257 s against a 36 h ceiling.

The laptop's Time Machine backup, read through a Ceph snapshot, showed a newest completed snapshot dated 02:13:49 UTC on 6 November: 63.3 h old at the 17:29 UTC drill run, and 68.3 h by calculation at 22:31 UTC (assuming none completed in between). Both are inside the drill's 96 h ceiling, which would be crossed at 02:13 UTC on 10 November. This is a stale backup, not

a confirmed breach. Six dead no-op backup jobs were also removed after re-verification, with copies archived.

7. A Stricter Drill, a False Red

The landed drill was enabled as a daily timer. Its first production run, about six minutes after the merge at roughly 22:24 UTC, failed the OSD-set check with "line order or blank lines differ". The export and a fresh dump of the same epoch held identical lines. Ceph had printed two blocklist entries (client addresses barred from the cluster) in swapped order; we believe because it renders them from an unordered map, which we did not confirm in its source, and the release is not recorded. The line-for-line comparison, added to close the OSD edit holes from round 2, treated order as content.

The fix, committed at 22:31:13 UTC, sets blocklist lines aside before comparing order. A swapped pair passes; any other change still fails. We checked it on 8 November: live green 13 of 13; red still returned for truncation, a short dump and pool-field stripping; all 116 Ceph tests passed. At 22:31 the fix was not yet merged. Why the earlier live greens did not show the ordering is not recorded.

This is a cost of the adversary loop that we met once: every hole closed makes the comparison more exact, and the more exact it is, the more ordinary variation reads as damage. The loop by design scored only false greens, not false reds on good data, and in practice we met two sources of false red: comparison order and intraday drift between an export and the live cluster. The health gate also made two of the red proofs skip until rerun.

8. What These Numbers Will Not Carry

This was one day, one campaign and one model family, with no repeats; a different attacker model might find different holes.

The verdict rule changed between workflows, not on the same units under both rules, so the comparison of binary and triage is between different units at different stages, not a controlled one.

The labels "material" and "residual" are the attacker agent's own; a human did not relabel them. The split exists for only 157 attacks on 10 units, so we cannot say how many of the 234 binary passes would count as material. A pass means the drill would accept that damaged scratch copy, not that production data was damaged, and passing attacks overlap rather than counting distinct defects. Green was not re-verified live for 19 of the 27 held drills.

What survives: the binary rule returned vacuous in every recorded round and the pass fraction did not fall across three rounds; the one drill that landed had zero material holes in its single triage result and still had four residual passes; the dashboard fault and the laptop snapshot age are measured facts; and the false red happened. What does not survive: any claim that triage reduces holes, any residual fraction outside the 157 triaged attacks, and any claim that the held drills would reach zero material holes with more rounds.

9. What We Changed

From the third workflow onward the verdict rule is material versus residual, and the next campaign will triage from round 1. Every one of the 27 held waivers now states its branch and the damage that still passes. One drill landed with a daily timer, one inventory row was retired, six no-op jobs were removed, and the dashboard backup now archives the live instance. The next experiment runs both rules on the same units, with a human relabelling a sample, and

adds a good-data variation check so exactness is scored against false reds too.

10. Methods and Artefacts

Counts and ratios only; no significance tests, no repeats. Live proofs ran on a single compute node. Round and attack figures come from the agent journals unless stated otherwise. The registry was parsed at the merge commit and its parent.

We retain the per-unit state records for 20 units, the agent journals for all three workflows, the registry at both commits, the landed drill's kit and tests, the dashboard-backup diff, the fix commit and one branch per held unit. Damage shapes here are paraphrased to class level.

PureTensor operates a sovereign AI infrastructure fleet (on-premises GPU compute, storage, and serving) run day-to-day by autonomous agents under human direction. This paper is PT-R-2026-068. The measurements were taken on 8 November 2026, the last at 22:31 UTC; per-unit records, agent journals and the registry snapshots are retained.