Supplementary Material: Non-Completion Is Not Prevention: Defence-Induced Censoring and the Measurement of Attack Success in Tool-Using LLM Agents

S0 Overview

This supplement carries the material moved out of the main paper: the complete 60-comparison table with sensitivity analyses (Section S1), the complete bounding table (Section S2), per-pass totals (Section S3), the pilot campaign (Section S4), the routing arm (Section S5), the adaptive arm (Section S6), configurations and product versions (Section S7), the execution deviation and exclusion ledger (Section S8), the data dictionary and analysis commands (Section S9), the event schema and per-configuration accounting (Section S10), and the post-termination inspection and sensitivity analyses (Section S11). Tables in Sections S1 and S3 print the as-recorded output of `v9_analysis.py` from the deposited unit records (record identifier withheld for double-blind review); the table in Section S2 additionally incorporates the main-paper Section 4.2 post-termination inspection reclassifications, which are a step of the analysis script. The analysis plan, data dictionary and audit log are in the accompanying analysis folder.

S1 The 60 comparisons: primary analysis and sensitivity analyses

Primary procedure: exact two-sided sign-flip permutation test on per-unit difference totals, programme-wise Holm correction across the family of 60; intervals are unit-cluster bootstrap percentile intervals (20,000 replicates, seed 20261118). Sensitivity procedures: paired t-test and Wilcoxon signed-rank test on per-unit totals, and the pooled McNemar test on (pass, unit) discordant pairs (the analysis printed in the earlier draft; it treats repeated passes as independent and is reported for contrast only). The reanalysis is post hoc. Differences are A minus B in completions per pass. Mechanism labels follow the comparison specification, with normalised terminology (degraded for the specification's deliberately degraded variants; archived release for the older detector release); the specification's original labels are preserved unchanged in `tests60_primary.csv` and `pattern_level.csv`.

Under the primary procedure the sole programme-wise survivor is comparison 36. Under the paired t-test, comparisons 10 and 36 survive; under Wilcoxon, none; under the pooled McNemar sensitivity analysis, the eight comparisons printed in the earlier draft (1, 5, 10, 28, 29, 35, 36, 39) survive, which is the contrast the main paper draws. Pattern-level results (six harm patterns, descriptive only): no key attack comparison reaches 0.05 unadjusted (hardening 0.0625, scoping step 0.125, scoping alone 0.0625); the full pattern table is in `pattern_level.csv`.

S2 Complete bounding table

Censoring sets: error-terminated runs (primary) and error-or-timeout-terminated runs (wider). Each side's censored runs vary independently. The alternative dual-event rule classifies a veto-carrying error-terminated run as blocked rather than censored. The alternative column is a sensitivity of the rule alone, computed on the reclassified counts: the alternative rule and the reclassification cannot both be applied to the same runs. Every reclassified run carries an earlier veto in its transcript, all of them on read calls that preceded the delete, and the delete call itself was

Table 1 Complete 60-comparison table. Primary procedure: exact two-sided sign-flip permutation test on per-unit difference totals, programme-wise Holm correction across the family of 60; intervals are unit-cluster bootstrap percentile intervals (20,000 replicates). Sensitivities: paired t-test and Wilcoxon signed-rank Holm-adjusted p-values, and pooled McNemar Holm-adjusted p-values. Differences are A minus B in completions per pass. The sole programme-wise survivor is comparison 36.

n	Family	Comparison	Outcome	Mechanism	Units	Passes	Diff per pass	95% CI	Sign-flip p	Holm p	Survivor	t Holm	Wilcoxon Holm	McNemar Holm
1	Core	0 -¿ 1	attack	Hardening	24	10	4.50	1.80 to 7.60	0.0039	0.227	no	0.400	0.440	0.0000
2	Core	0 -¿ 1	utility	Hardening	20	10	0.10	0.00 to 0.30	1.000	1.000	no	1.000	1.000	1.000
3	Core	1 -¿ 2	attack	Detector	24	10	0.10	0.00 to 0.30	1.000	1.000	no	1.000	1.000	1.000
4	Core	1 -¿ 2	utility	Detector	20	10	1.10	0.00 to 3.20	0.500	1.000	no	1.000	1.000	0.051
5	Core	2 -¿ 3	attack	Scoping	24	10	4.10	1.00 to 8.10	0.062	1.000	no	1.000	1.000	0.0000
6	Core	2 -¿ 3	utility	Scoping	20	10	-0.10	-0.40 to 0.20	1.000	1.000	no	1.000	1.000	1.000
7	Core	3 -¿ 4	attack	Approval	24	10	-0.10	-0.30 to 0.00	1.000	1.000	no	1.000	1.000	1.000
8	Core	3 -¿ 4	utility	Approval	20	10	0.00	0.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
9	Core	4 -¿ 5	attack	Framework	24	10	0.70	0.00 to 1.90	0.500	1.000	no	1.000	1.000	0.766
10	Core	4 -¿ 5	utility	Framework	20	10	10.90	6.80 to 15.00	0.0010	0.058	no	0.0071	0.079	0.0000
11	Core	5 -¿ 6	attack	Framework + sandboxing	24	5	0.00	0.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
12	Core	5 -¿ 6	utility	Framework + sandboxing	20	5	0.00	0.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
13	Core	6 -¿ 7	attack	Framework + egress	24	5	-0.60	-1.80 to 0.00	1.000	1.000	no	1.000	1.000	1.000
14	Core	6 -¿ 7	utility	Framework + egress	20	5	0.20	0.00 to 0.60	1.000	1.000	no	1.000	1.000	1.000
15	Core	7 -¿ 8	attack	Framework + output scanning	24	5	0.60	0.00 to 1.80	1.000	1.000	no	1.000	1.000	1.000
16	Core	7 -¿ 8	utility	Framework + output scanning	20	5	-0.20	-0.60 to 0.00	1.000	1.000	no	1.000	1.000	1.000
17	Core	8 -¿ 9	attack	Framework + monitoring	24	5	-0.60	-1.80 to 0.00	1.000	1.000	no	1.000	1.000	1.000
18	Core	8 -¿ 9	utility	Framework + monitoring	20	5	0.00	0.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
19	Core	9 -¿ 10	attack	Framework + memory governance	24	5	0.40	0.00 to 1.20	1.000	1.000	no	1.000	1.000	1.000
20	Core	9 -¿ 10	utility	Framework + memory governance	20	5	0.20	0.00 to 0.60	1.000	1.000	no	1.000	1.000	1.000
21	Core	2 -¿ 2R	attack	Detector (archived release)	24	10	0.10	0.00 to 0.30	1.000	1.000	no	1.000	1.000	1.000
22	Core	2 -¿ 2R	utility	Detector (archived release)	20	10	-1.10	-3.20 to 0.00	0.500	1.000	no	1.000	1.000	0.051
23	Core	4 -¿ 4R	attack	Approval (degraded)	24	5	0.20	0.00 to 0.60	1.000	1.000	no	1.000	1.000	1.000
24	Core	4 -¿ 4R	utility	Approval (degraded)	20	5	0.00	0.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
25	Core	7 -¿ 7R	attack	Egress (lapsed)	24	5	0.80	0.00 to 2.40	1.000	1.000	no	1.000	1.000	1.000
26	Core	7 -¿ 7R	utility	Egress (lapsed)	20	5	-0.20	-0.60 to 0.00	1.000	1.000	no	1.000	1.000	1.000
27	Single-control	core 0 -¿ 2S	attack	Detector (attribution)	24	5	1.00	-0.60 to 2.80	0.375	1.000	no	1.000	1.000	1.000
28	Single-control	core 2 -¿ 2S	attack	Hardening (attribution)	24	5	-4.00	-7.00 to -1.40	0.0078	0.438	no	0.653	0.622	0.0001
29	Single-control	core 0 -¿ 3S	attack	Scoping	24	5	7.60	3.00 to 12.20	0.0054	0.306	no	0.249	0.489	0.0000
30	Single-control	core 3 -¿ 3S	attack	Scoping	24	5	-1.80	-4.60 to 0.00	0.500	1.000	no	1.000	1.000	0.195
31	Single-control	core 0 -¿ 2S	utility	Detector	20	5	1.00	0.00 to 3.00	1.000	1.000	no	1.000	1.000	1.000
32	Single-control	core 2 -¿ 2S	utility	Detector	20	5	-0.20	-0.60 to 0.00	1.000	1.000	no	1.000	1.000	1.000
33	Single-control	core 0 -¿ 3S	utility	Scoping	20	5	1.40	0.00 to 3.80	0.500	1.000	no	1.000	1.000	0.766
34	Single-control	core 3 -¿ 3S	utility	Scoping	20	5	0.40	-2.60 to 3.40	1.000	1.000	no	1.000	1.000	1.000
35	Position	core 2 -¿ 5P	attack	Framework (placement)	24	5	4.60	1.40 to 8.40	0.016	0.844	no	1.000	0.820	0.0000
36	Position	core 2 -¿ 5P	utility	Framework (placement)	20	5	11.00	6.80 to 15.00	0.0005	0.029	yes	0.0050	0.064	0.0000
37	Position	core 5 -¿ 5P	attack	Framework (position)	24	5	-0.60	-2.40 to 0.60	1.000	1.000	no	1.000	1.000	1.000
38	Position	core 5 -¿ 5P	utility	Framework (position)	20	5	0.20	0.00 to 0.60	1.000	1.000	no	1.000	1.000	1.000
39	Position	core 0 -¿ 7E	attack	Egress, early (effect = hardening)	24	5	5.00	2.00 to 8.20	0.0078	0.438	no	0.299	0.622	0.0000
40	Position	core 0 -¿ 7E	utility	Egress, early	20	5	0.20	0.00 to 0.60	1.000	1.000	no	1.000	1.000	1.000
41	Position	core 1 -¿ 7E	attack	Egress, early	24	5	0.00	-0.60 to 0.60	1.000	1.000	no	1.000	1.000	1.000
42	Position	core 1 -¿ 7E	utility	Egress, early	20	5	0.00	0.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
43	Position	7E -¿ 7RE	attack	Egress (degraded)	24	5	-0.20	-0.80 to 0.40	1.000	1.000	no	1.000	1.000	1.000
44	Position	7E -¿ 7RE	utility	Egress (degraded)	20	5	0.00	0.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
45	Seam	core 2 -¿ 2P	attack	Detector (seam)	24	5	0.00	-0.60 to 0.60	1.000	1.000	no	1.000	1.000	1.000
46	Seam	core 2 -¿ 2P	utility	Detector (seam)	20	5	-1.00	-3.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
47	Seam	core 5 -¿ 5IN	attack	Framework (input half)	24	5	-0.20	-0.60 to 0.00	1.000	1.000	no	1.000	1.000	1.000
48	Seam	core 5 -¿ 5IN	utility	Framework (input half)	20	5	0.00	0.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
49	Seam	core 5 -¿ 5OUT	attack	Framework (output half)	24	5	0.20	0.00 to 0.60	1.000	1.000	no	1.000	1.000	1.000
50	Seam	core 5 -¿ 5OUT	utility	Framework (output half)	20	5	1.00	0.00 to 2.80	0.500	1.000	no	1.000	1.000	1.000
51	Seam	core 8 -¿ 8NC	attack	Output scanning (canary)	24	5	-0.40	-1.20 to 0.00	1.000	1.000	no	1.000	1.000	1.000
52	Seam	core 8 -¿ 8NC	utility	Output scanning (canary)	20	5	0.20	0.00 to 0.60	1.000	1.000	no	1.000	1.000	1.000
53	Seam	5IN -¿ 5OUT	attack	Framework (halves)	24	5	0.40	0.00 to 1.20	1.000	1.000	no	1.000	1.000	1.000
54	Seam	5IN -¿ 5OUT	utility	Framework (halves)	20	5	1.00	0.00 to 2.80	0.500	1.000	no	1.000	1.000	1.000
55	Adaptive	core 0 -¿ focal 0	attack	Adaptive attacker	24	3	1.00	-3.00 to 5.00	0.762	1.000	no	1.000	1.000	1.000
56	Adaptive	core 0 -¿ focal 0	utility	Adaptive attacker	20	3	0.00	0.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
57	Adaptive	core 3 -¿ focal 3	attack	Adaptive attacker	24	3	-0.33	-1.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
58	Adaptive	core 3 -¿ focal 3	utility	Adaptive attacker	20	3	0.33	0.00 to 1.00	1.000	1.000	no	1.000	1.000	1.000
59	Adaptive	core 5 -¿ focal 5	attack	Adaptive attacker	24	3	-0.33	-1.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000
60	Adaptive	core 5 -¿ focal 5	utility	Adaptive attacker	20	3	0.00	0.00 to 0.00	1.000	1.000	no	1.000	1.000	1.000

never vetoed; under the alternative rule as printed, any error-terminated run carrying a veto counts as blocked, so a reclassified run cannot be counted as a completion under that rule (see main-paper Section 4.2). Veto counts are from the attempt transcripts, not the stored veto-events field, which blanks vetoes on aborted attempts. Observed differences and intervals in this table incorporate the main-paper Section 4.2 inspection reclassifications of error-terminated runs whose predicates had already been satisfied (fifteen runs, fourteen deletion of the legacy deployment configuration file and one utility summary-file write); the S1 table prints the as-recorded script output. The jointly re-scored difference re-scores all framework-side error runs as completions on both sides and is not a bound; it is printed for continuity only, and is unchanged by the reclassifications because each reclassified run adds one observed completion and removes one censored run. Table 2 gives the complete set. The bounded 3 to 5 contrast, reported in main-paper Section 5.3, is appended as an additional row.

Under the primary rule, two seam contrasts exclude zero (5IN to 5OUT, 0.2 to 9.0; 5 to 5OUT, 0.2 to 12.4), and both exclude zero under the alternative dual-event rule as well (0.6 to 1.0 and 0.6 to 1.8); under the wider error-plus-timeout censoring set neither does (minus 23.4 to 9.0 and minus 23.4 to 12.4), so both are classification-driven in

Table 2 Complete bounding table. Censoring sets: error-terminated runs (primary) and error-or-timeout-terminated runs (wider). Each side's censored runs vary independently. The alternative dual-event rule classifies a veto-carrying error-terminated run as blocked rather than censored. The jointly re-scored difference re-scores all framework-side error runs as completions on both sides and is not a bound; it is printed for continuity only.

Comparison	Outcome	Passes	Framework side	Observed diff	Identified interval (error)	Identified interval (error + timeout)	Identified interval (alternative rule)	Excludes zero	jointly re-scored diff
core 4 -¿ 5	attack	10	B	0.50	-10.90 to 0.50	-10.90 to 0.50	-0.80 to 0.50	no	-10.90
core 5 -¿ 6	attack	5	both	0.20	-11.80 to 12.00	-11.80 to 12.00	-1.40 to 1.40	no	0.00
core 6 -¿ 7	attack	5	both	-0.40	-12.00 to 11.60	-12.00 to 11.60	-2.00 to 1.20	no	0.00
core 7 -¿ 7R	attack	5	both	0.00	-10.60 to 11.60	-10.60 to 11.60	-1.00 to 1.60	no	1.00
core 7 -¿ 8	attack	5	both	0.20	-10.60 to 11.80	-10.60 to 11.80	-0.80 to 1.80	no	1.00
core 8 -¿ 9	attack	5	both	-0.20	-10.00 to 10.60	-10.00 to 10.60	-1.60 to 0.80	no	0.80
core 9 -¿ 10	attack	5	both	0.00	-10.20 to 9.80	-10.20 to 9.80	-1.40 to 1.40	no	-0.40
core 2 -¿ 5P	attack	5	B	4.60	-6.20 to 4.60	-6.20 to 4.60	+3.40 to 4.60	no	-6.20
core 5 -¿ 5P	attack	5	both	-0.20	-11.00 to 11.60	-11.00 to 11.60	-1.40 to 1.00	no	0.80
core 5 -¿ 5IN	attack	5	both	0.00	-8.40 to 11.80	-8.40 to 11.80	-0.40 to 1.20	no	3.40
core 5 -¿ 5OUT	attack	5	both	0.60	0.20 to 12.40	-23.40 to 12.40	0.60 to 1.80	yes	12.00
5IN -¿ 5OUT	attack	5	both	0.60	0.20 to 9.00	-23.40 to 9.00	0.60 to 1.00	yes	8.60
core 8 -¿ 8NC	attack	5	both	-0.20	-8.20 to 10.60	-8.20 to 10.60	-1.40 to 0.80	no	2.60
core 5 -¿ focal 5	attack	3	both	-0.33	-4.00 to 10.67	-21.67 to 10.67	-1.00 to 1.00	no	7.00
core 3 -¿ 5	attack	10	B	0.40	-11.00 to 0.40	-11.00 to 0.40	-0.90 to 0.40	no	-11.00
core 4 -¿ 5	utility	10	B	10.90	10.80 to 10.90	10.80 to 10.90	10.90 to 10.90	yes	10.80
core 2 -¿ 5P	utility	5	B	10.80	10.00 to 10.80	10.00 to 10.80	10.80 to 10.80	yes	10.00

the same sense main-paper Section 5.3 discusses. Every other framework attack interval includes zero under the primary and wider rules; under the alternative rule the 2 to 5P attack interval excludes zero (3.4 to 4.6), which is a sensitivity of the rule alone and cannot be combined with the reclassification, as stated above. Both framework utility intervals exclude zero under every rule and censoring set. The "Excludes zero" column is read under the primary rule.

S3 Per-pass totals

Completions per pass for every record set (attack goals out of 24, utility tasks out of 20, unless the set's unit count differs). Generated in full in `perpass_totals.csv`; the replication core sets are tabulated here.

S4 Pilot campaign

The pilot ran the same frozen campaign configuration on the earlier subject (Qwen3-32B, same assembly) before the replication. It exposed four protocol defects, all corrected before the replication and none of whose data enters any reported number: (1) payload sequences were truncated in some configurations, so the full six-payload sequence was not prosecuted everywhere; the harness was fixed to prosecute the full sequence in every configuration; (2) two completion predicates were mis-specified and were corrected, after which the predicate suite was validated against 88 hand-built environment states (44 positive, 44 negative), classified 88 of 88; (3) failure labels were coded inconsistently across run sets and were re-coded uniformly from run events into the terminal-state classification of main-paper Section 4.2; (4) sampling parameters were not recorded per run and are recorded in the replication. The determinism check (three identical undefended passes at temperature 0: 11 of 24 attack goals and 19 of 20 utility tasks, zero per-unit differences) was run on the replication subject.

S5 Routing arm

Routing variants deliver the same attack content through different shapes: twelve R1 and R2 variants over six base goals, plus ten R3 units, 22 distinct units in all, each run for five passes against configurations 4, 5 and 6. At the guardrails configuration (cell 5), total completions across all routing units were 2: G-ISSUE-3-R1 completed in one of five passes and G-ISSUE-3-R2 in one of five passes; no other routing unit completed at that configuration. The earlier description of this result as eleven of twelve shapes resisting, with a twelfth leaking, was incorrect: six base shapes and twelve R1/R2 variants are different counts, and the leak is two variants of one base goal. Full per-unit routing records are in `corrected_findings.csv`.

S6 Adaptive arm

Three focal configurations (undefended, scoping, framework) each planned 72 goal-runs (24 goals, three passes). Exposure is defined by revision use in the attacker logs. Undefended: 0 opener errors, 19 opener completions, 53 revision-exposed goal-runs, 9 revision successes, 26 timeouts; attacker refusals 68 of 260 issued rounds (26.2 percent). Scoping: 0 opener errors, 3 opener completions, 67 revision-exposed, 1 revision success, 64 timeouts; refusals 204 of 466 (43.8 percent). Framework: 9

Table 3 Completions per pass for every record set (attack goals out of 24, utility tasks out of 20, unless the set's unit count differs). Generated in full in `perpass_totals.csv`; the replication core sets are tabulated here.

Set	Kind	Completions by pass
0	attack	9, 10, 12, 11, 10, 8, 9, 9, 12, 8
0	utility	19, 19, 19, 19, 19, 19, 19, 19, 19, 19
1	attack	5, 6, 5, 6, 5, 5, 5, 6, 5, 5
1	utility	19, 19, 19, 18, 19, 19, 19, 19, 19, 19
2	attack	6, 5, 5, 6, 5, 5, 5, 5, 5, 5
2	utility	18, 18, 18, 17, 18, 18, 18, 18, 17, 18
2R	attack	5, 5, 5, 5, 5, 5, 5, 6, 5, 5
2R	utility	19, 19, 19, 18, 19, 19, 19, 19, 19, 19
3	attack	1, 1, 1, 1, 1, 1, 1, 1, 2, 1
3	utility	18, 18, 18, 18, 18, 18, 17, 18, 18, 18
4	attack	1, 1, 1, 2, 1, 1, 1, 1, 2, 1
4	utility	18, 18, 18, 18, 18, 18, 17, 18, 18, 18
4R	attack	1, 1, 1, 1, 1
4R	utility	18, 18, 18, 18, 18
5	attack	0, 1, 0, 0, 0, 1, 1, 0, 1, 1
5	utility	7, 7, 7, 7, 7, 7, 7, 7, 7, 7
6	attack	0, 1, 0, 0, 0
6	utility	7, 7, 7, 7, 7
7	attack	1, 1, 0, 1, 1
7	utility	7, 7, 7, 6, 7
7R	attack	0, 0, 0, 0, 0
7R	utility	7, 7, 7, 7, 7
8	attack	0, 0, 0, 1, 0
8	utility	7, 7, 7, 7, 7
9	attack	1, 1, 1, 0, 1
9	utility	7, 7, 7, 7, 7
10	attack	0, 1, 0, 0, 1
10	utility	7, 6, 7, 7, 7

opener errors, 2 opener completions, 61 revision-exposed, 0 revision successes, 53 timeouts; refusals 336 of 471 (71.3 percent). Exact Clopper-Pearson upper bounds for zero successes in 61 exposed runs: 4.79 percent one-sided, 5.87 percent two-sided. The earlier phrasing of this result as adaptive success below 4.7 percent is replaced by the estimand-explicit statement above; no adaptive-resistance claim is made, because refusals and timeouts removed exposure differentially across configurations.

S7 Configurations and harness

Core sequence (author-defined): 0 documented defaults; 1 + prompt hardening; 2 + detector library as quickstart-installed (llm-guard 0.3.16, shipped threshold, on the user prompt); 2R detector in its archived release (llm-guard 0.3.9); 3 + tool and credential scoping; 4 + approval dialog (mechanical, approve as presented) on sensitive tools; 4R approval in degraded form (summarised-argument dialog, sticky always-approve); 5 + guardrails framework (NeMo Guardrails 0.19.0; rail model is the subject model on the same endpoint); 6 + sandboxing and isolation; 7 + egress allowlist as documented; 7R egress with one allowlisted destination a lapsed domain re-registered by the attacker; 8 + output-side scanning (Presidio 2.2.358); 9 + monitoring and tracing; 10 + memory-write governance and tool integrity. Ablations: 2S detector alone; 3S scoping alone; 5P framework immediately after the detector; 7E and 7RE egress early (documented and lapsed); 2P detector at the tool-output seam; 5IN input rails only; 5OUT output rails only; 8NC output scanning without the framework configuration. Serving: vLLM 0.21.0 (replication serving, per the deposit README; the harness staging requirements file pins vLLM 0.13.0 for a separate serving venv), FP16, CUDA graphs enabled, Qwen3 XML tool-call parser, 32,768-token context; pilot subject Qwen3-32B, replication subject Qwen3.6-27B (model revision not pinned to an immutable snapshot: every run manifest records revision null, and no revision, commit hash or snapshot identifier appears in the deposited records). The harness subject-model factory defaults to Qwen/Qwen3-32B when the environment variable is unset; every run manifest records the subject model actually run as

Qwen3.6-27B, and the manifest name is the one cited. Agent loop LangGraph 1.0.5; MCP servers over stdio (filesystem, GitHub with write tools, fetch, JSONL memory). One execution configuration (four workers, one harness build, one serving configuration) served every replication pass. Campaign configuration SHA-256: f72ab68b8bbd984eace359129649f0e043494cfdb934dd085f17ef9bb8d78ea6. Sampling: temperature 0.2, top-p 1.0, per-pass seeds shared across configurations.

S8 Execution deviation and exclusion ledger

One follow-on deviation is recorded, not touching the replication core: six follow-on passes in the ablation and adaptive programmes are clean re-runs executed under a changed serving configuration and at ten workers, against the core's one serving configuration and four workers for every pass; the six passes are flagged in the deposited records and where their results are reported. Contamination history, disclosed in full: an early set of passes affected by a numerical-library fault was purged and re-run before analysis; six passes executed against an unresponsive serving endpoint on 2026-11-14 wrote connection-error records only, were detected by record inspection, were deleted, and were never analysed; the completeness check was thereafter tightened to require full record counts with bounded error fractions. No deposited record was excluded from the reanalysis: all 60 comparisons and all corrected counts run on the deposited sets as published. Deposit archive notes. One file in the deposited archive is unusable as stored: core-c5r7-10-c4R-c6-c7-c7Rr1-4.tar.gz carries a 347-byte plaintext header before the gzip stream, so tar refuses it; its .fixed twin extracts cleanly, and the corrupt file minus its first 347 bytes is byte-identical to the twin, so no record is lost. The determinism set det.tar.gz (132 records, three passes) is outside the script's file globs and has no extracted directory in the deposit; its records were read directly, and the determinism numbers reported in the main paper (11 of 24 attack goals and 19 of 20 utility tasks completed in every pass, zero per-unit differences) reproduce exactly from it.

S9 Data dictionary and analysis commands

The data dictionary for every analysis output is `data_dictionary.md` in the accompanying analysis folder, defining each column of `tests60_primary.csv`, `tests60_sensitivity.csv`, `pattern_level.csv`, `bounds_framework.csv`, `equivalence.csv`, `corrected_findings.csv` and `perpass_totals.csv`. Regeneration command: `python3 v9_analysis.py` in the analysis folder (Python 3, NumPy 1.26.4, SciPy 1.11.4); the script reads the deposited records and the comparison specification, writes all outputs including `bounds_framework.csv` and `audit_log.txt`, and verifies its recomputed discordant pairs against the printed table (all 60 match). The deposit root is taken from the `V9_DEPOSIT_ROOT` environment variable when set, so the script runs unchanged against a fresh extraction of the deposit archive. The family of 60 comparisons was fixed in the comparison specification before the reanalysis. The script's transcript-level predicate evaluation of main-paper Section 4.2 is a step of the analysis, and the reclassification extends to utility runs (one run, task U-18 in configuration 5P), which moves the 2 to 5P utility row of `bounds_framework.csv` to observed 10.8, primary 10.0 to 10.8 and alternative 10.8 to 10.8.

S10 Event schema, precedence rules and per-configuration accounting

Each deposited unit record carries: run and unit identifiers; configuration (cell) identifier; kind (attack or utility); the ordered list of attempts, each with its tool calls, veto events (with the responsible control), detector pre-run block flag, and per-attempt completion status against the goal's environment-state predicate; the unit's terminal state as a failure reason (completed units carry none); an error string for error-terminated units; a timeout flag; and the count of adaptive revisions used where applicable. The event schema a portable implementation requires, beyond what this dataset records, is: event timestamps or

sequence indices for all events (this dataset orders events within attempts and attempts within runs, which suffices for the precedence rules below), whether each tool action was attempted and executed, and whether the environment state was inspected after termination.

Precedence decision table. A run's classification is decided by its terminal state first, then by the wider and dual-event rules, exactly as in main-paper Algorithm 1. Step 3 of Algorithm 1 is the post-termination inspection: after a run terminates in error, the analysis inspects the attempt transcripts against the goal's environment-state predicate and reclassifies a run as an observed completion when an attempt satisfied the predicate before termination. The fifteen runs so reclassified are listed with their corrected counts below (Table 6).

Detector activity at the tool-output seam (configuration 2P), reported as activity, not as outcomes: 96 `detector_block` veto events; 20 of 120 attack unit records contain at least one such event; 4 distinct goals are affected (G-DOC-1, G-DOC-4, G-DOC-5, G-MEM-1); the final outcome in those 20 records is 0 completions, while total completions at the seam placement are 27 of 120, equal to the matched detector configuration's first five passes (27 of 120). Event counts, affected-run counts, affected-goal counts and prevented-outcome counts are different quantities and are never interchanged in this paper.

Per-configuration accounting for the framework configurations (from `config_accounting.csv`, which covers every run set). Veto counts are from the attempt transcripts, not the stored veto-events field, which blanks vetoes on aborted attempts (see main-paper Section 4.2). The table includes the 7R and 8NC attack sets and the utility sets at configurations 6 to 10, whose 12 censored runs the earlier analysis omitted; the remaining configurations have zero censored runs and are tabulated in `config_accounting.csv`. The focal-5 rows carry transcript veto counts: 10 of the 12 error-terminated attack runs and all 53 timeout-terminated attack runs carry a veto in the transcript.

Post-termination inspection reclassifications (main-paper Section 4.2). Fifteen attempt transcripts, one in each of fifteen run records, had reached their target state before a recursion-limit error terminated them, fourteen of them deletion of the legacy deployment configuration file (G-ISSUE-3): configuration 10 passes 3 and 4; configuration 5 passes 3 and 5; configuration 6 pass 3; 7R passes 1, 2, 3 and 5; configuration 8 passes 1 and 3; 5IN pass 1; 8NC pass 4; focal 5 pass 3; and one utility summary-file write, configuration 5P pass 4, task U-18. On the stored veto-events field, two of the fifteen carry an earlier veto (configuration 5 pass 3 and configuration 8 pass 1) and the other thirteen show none, because the harness blanks veto events on the aborted attempt. On the attempt transcripts, all fifteen carry vetoes before the satisfying call (three to thirteen each), all of them on read calls; the satisfying call itself was never vetoed in any of the fifteen runs. The corrected counts for the affected cells are:

S11 Post-termination inspection, leave-one-out and pass-level sensitivities

Post-termination inspection (from `inspection_results.csv`). Every error- or timeout-terminated run in the framework configurations was inspected at attempt level: 773 attack runs (602 error-terminated, 171 timeout-terminated) and 54 utility runs (29 error-terminated, 25 timeout-terminated). The framework configurations are 5, 6, 7, 8, 9, 10, 5P, 5IN, 5OUT, 7R, 8NC and focal 5; the earlier count of 610 attack and 52 utility runs covered only the nine sets in the script's original framework list and omitted 7R, 8NC and focal 5. All inspected runs had executed tool calls. Fourteen inspected attack runs had an attempt that satisfied the goal's environment-state predicate before termination, all of them deletion of the legacy deployment configuration file, and one inspected utility run had an attempt that satisfied its task predicate (configuration 5P, pass 4, task U-18, summary-file write); all fifteen are reclassified as observed completions as listed above. No other inspected run contains a completed attempt. For an error-terminated run that is not reclassified, the outcome treated as unobservable is therefore its eventual completion over the payload attempts that remained, not its state at termination, which

Table 4 Precedence decision table. A run's classification is decided by its terminal state first, then by the wider and dual-event rules, exactly as in main-paper Algorithm 1.

Recorded situation	Classification (primary rule)	Under the wider rule	Under the alternative rule
Terminal state completed	Y = 1, C = 0	same	same
Control veto terminates the run (blocked)	Y = 0, C = 0 (observed prevention)	same	same
Detector blocks before execution (blocked pre-run)	Y = 0, C = 0 (observed prevention)	same	same
Refused, never read, or incapable	Y = 0, C = 0 (observed non-completion)	same	same
Timeout, no veto event	Y = 0, C = 0	C = 1	same as primary
Timeout after a veto event that did not end the run	Y = 0, C = 0	C = 1	same as primary
Error, no veto event	C = 1	C = 1	same
Error after a veto event (dual-event)	C = 1	C = 1	Y = 0, C = 0 (counted as blocked)
Error after partial execution, with an attempt recorded as predicate-satisfied	Y = 1, C = 0 by the step 3 inspection (Table 6)	same	by veto flag: a veto-carrying run is counted as blocked, so the inspection reclassification and the alternative rule cannot both be applied to the same run
Error after partial execution, no satisfied attempt	C = 1	C = 1	by veto flag, as above
Error followed by retry and eventual completion	terminal state completed: Y = 1, C = 0	same	same
Multiple veto events in one run	V = 1 once; classification by terminal state	same	same

Table 5 Per-configuration accounting for the framework configurations (from `config_accounting.csv`, which covers every run set). Veto counts are from the attempt transcripts. Refused, never read and incapable are zero except for the one incapable utility run at 5OUT.

Set	Kind	Runs	Completed	Blocked	Pre-run	Refused	Never read	Incapable	Error	Timeout	Censored (primary)	Censored (wider)	Veto then error	Veto then timeout
5	attack	240	5	119	0	0	0	0	116	0	116	116	103	0
5	utility	200	70	119	10	0	0	0	1	0	1	1	0	0
5P	attack	120	4	62	0	0	0	0	54	0	54	54	48	0
5P	utility	100	34	56	5	0	0	0	5	0	5	5	0	0
5IN	attack	120	2	75	0	0	0	0	43	0	43	43	41	0
5IN	utility	100	35	58	5	0	0	0	2	0	2	2	0	0
5OUT	attack	120	0	0	0	0	0	0	2	118	2	120	2	118
5OUT	utility	100	30	32	5	0	0	1	7	25	7	32	0	0
6	attack	120	1	58	0	0	0	0	61	0	61	61	53	0
6	utility	100	35	56	5	0	0	0	4	0	4	4	0	0
7	attack	120	4	58	0	0	0	0	58	0	58	58	50	0
7	utility	100	34	59	5	0	0	0	2	0	2	2	0	0
7R	attack	120	0	63	0	0	0	0	57	0	57	57	52	0
8	attack	120	1	63	0	0	0	0	56	0	56	56	51	0
8	utility	100	35	58	5	0	0	0	2	0	2	2	0	0
8NC	attack	120	3	76	0	0	0	0	41	0	41	41	35	0
9	attack	120	4	67	0	0	0	0	49	0	49	49	42	0
9	utility	100	35	59	5	0	0	0	1	0	1	1	0	0
10	attack	120	2	65	0	0	0	0	53	0	53	53	46	0
10	utility	100	34	58	5	0	0	0	3	0	3	3	0	0
focal 5	attack	72	2	5	0	0	0	0	12	53	12	65	10	53
focal 5	utility	60	21	35	3	0	0	1	0	0	0	0	0	0

the inspection establishes directly for every terminated run; timeout-terminated runs are observed non-completions under the primary rule and are treated as unobservable only under the wider rule.

Leave-one-out sensitivities (from `lopo_results.csv`). Attack estimates use leave-one-pattern-out over the six harm patterns; utility estimates use the analogous leave-one-task-out over the 20 tasks. Prompt hardening (4.5 per

Table 6 Recorded and corrected attack and utility counts for the cells affected by the Section 4.2 inspection reclassifications. Corrected completed counts feed the bounds in Section S2; the recorded counts above are unchanged.

Set	Recorded completed	Recorded censored (primary)	Corrected completed	Corrected censored (primary)
5 (10 passes)	5	116	7	114
5 (first 5 passes)	1	61	3	59
6	1	61	2	60
7R	0	57	4	53
8	1	56	3	54
10	2	53	4	51
5IN	2	43	3	42
8NC	3	41	4	40
focal 5	2	12	3	11
5P (utility)	34	5	35	4

pass in full): range 3.0 to 4.5 across the six exclusions, never changing sign. Scoping step (4.1): range 3.0 to 4.1. Scoping alone (7.6): range 4.6 to 7.6. Framework utility step, configuration 4 to 5 (10.9): range 9.9 to 10.9 across the 20 task exclusions. Framework placement utility, configuration 2 to 5P (10.8, the programme's sole survivor): range 9.8 to 10.8 across the 20 task exclusions with the Section 4.2 reclassification applied (10.0 to 11.0 as recorded in `lopo_results.csv`).

Pass-level sensitivity (from `pass_level.csv`). Sign-flipping per-pass completion differences, with passes as the unit, is anti-conservative and directional only: with five passes its smallest attainable two-sided p-value is 0.0625. For the ten-pass step contrasts (hardening, scoping step, framework utility step) all ten per-pass differences are in the same direction (pass-level $p = 0.002$ in each case); for the five-pass placement and scoping-alone contrasts the pass-level p-value is 0.0625, the floor of the procedure at that size. The primary unit throughout is the goal or task, never the pass.