How xChk measures — against the field, and against itself

Every framework, benchmark, and doctrine xChk has been scored on — including the control-plane-integrity axis nothing else measures, run adversarially against our own machinery and reported with the findings it surfaced.

THE CLAIM Prevention is not the bar; unfalsifiable records are. *You may be able to bypass the gate — you cannot bypass it without the evidence layer flagging the omission.* A gate that fails open quietly is theater; a gate that can be evaded but never evaded quietly is not, because the record still tells the truth. That second property is what an insurer or a court actually requires, and it is what these benchmarks measure.

0
Silent passes
Prevention axis: 23 techniques, 5 harness dialects + Hermes, 14 gate decisions.

0
Executed & undetected
Detection axis, 14 scenarios + 1 organic firing.

0
Silent authorizations
Authorize layer, 26 grant states through the real decision engine.

96.2%
AgentDojo blocked
25/26 attack sequences, 0.84 ceremonies per benign task.

0
Unsigned dispatches run
Dispatch axis: owner signature verified at the execution host.

94%
Injections noticed
Screens axis: 694/742 spikee payloads, 1/302 false positives, none on a gate screen.

EXTERNAL INSTRUMENTS — MEASURED, REPRODUCIBLE, SELF-REPORTED

INSTRUMENT	KIND	RESULT	WHAT IT SHOWS
AgentDojo v1.2.1 (ETH Zurich)	Empirical attack suite	**25/26** · 96.2%	Attack sequences blocked at low benign friction. The one slip is a read-only ingress — the STRIDE information-disclosure gap we name deliberately. Adaptive-LLM run pending.
spikee 0.9.2 (Reversec Labs)	Injection guardrail bench	**694/742** · 1/302 FP	742 document injections through every xChk screen as a guardrail target: 15% in the morning (authority claims caught, plain requests missed), 94% by evening after four families; the 48 misses are non-English, named. Murtaza OWASP LLM01: 6/6.
AgentGovBench (spec v0.2)	Governance benchmark	**47*/48** · 45/46 node	Seven of eight categories perfect; 45/46 as a node self-run (19 Sep), authority-ceiling 4/4. *The miss, fail_open_honored, rewards acting on with the authority unreachable; xChk holds and a signed person releases (RMF MEASURE 2.6; REPRODUCE.md; our comment to NIST, AI 200-2).
DEMM-Bench (arXiv 2606.20634)	Evidence rubric	**8/8**	A real anchored receipt reconstructs all eight decision properties; 4/8 → 8/8 in a day once the "why" joined the bundle. Their scorer still overclaimed a verifying bundle 12.5% — verifies ≠ explains.
AGB self-assessment (ExecLayer rubric)	Vendor rubric, 6 dims	**93.75** / 100	"Sovereign" tier — earned in code, not scoring (D1 hardened the same day). D4 containment & D5 jurisdiction held honestly at 75.
AARM v1.0 (CSA)	Industry standard	ALIGNED	Self-run of the 21-row test table (2 Oct): every Core technical row passes. Not a conformance claim.
AGSC v1.0.0 (COHUMAIN / SafeAlign)	Control catalog, 25 controls	**13/25** · 7/8 Tier 1	Self-assessed as a component (6 Nov): 13 met with verifiable evidence, 5 half ours, 7 model-lifecycle controls a deployer pairs it with; not a conformance claim.
STRIDE (agentic)	Threat model	**owns R, E**	Repudiation and Elevation — the accountability columns — are xChk's. Information-disclosure is a stated gap; Denial partial.

XCHK CONTROL-PLANE-INTEGRITY BENCHMARK — OUR OWN, RUN AGAINST OURSELVES

AXIS	COVERAGE	RESULT	THE FALSIFYING OUTCOME — AND WHAT WE FOUND
A · Prevention BYPASS	23 techniques · 14 decisions	**0** silent passes	Self-protection denies fire before any grant, in five dialects (Claude Code, Codex, Cursor, Antigravity, Copilot CLI) and Hermes; 15 of 16 gate-critical techniques still hold under an allow-all grant (F2 is the stated limit). Every unanswered, expired or declined hold and registry error refuses.
B · Detection RECONCILE	14 scenarios / 8 techniques	**0** exec & undetected	Omission, content-swap, fabrication, reference-reuse, approved-attest omission and unaccounted INPUT consumption all caught; examined reuses narrow to amber, never green. Self-found lamp-gap (D1) fixed; never-issued-and-never-recorded is the named limit. A second, ORGANIC firing (29 Aug): five real omissions, human-adjudicated same day.
C · Authorize GRANTS	26 grant states	**0** silent authorizations	The real decision engine over allow/attest/forbid, deny-by-default, amount caps, expiry, rate windows, cumulative spend, session context, behaviour drift and the dealer clock. No escalation returns an autonomous allow when it should hold.
D · Anchor ON-CHAIN	6 checks	**0** undetected tamper	The Ravencoin OP_RETURN commits the host-chain Merkle root; truncation, alteration, and reorder all diverge from the inscribed root. The bound is computable by anyone holding the heads and the txid — no trust in the registry.
E · Dispatch SIGNED	19 seeded + 2 live cases	**0** unsigned dispatches run	Workloads instantiate only on the owner's passkey signature, verified at the execution host against a provision-pinned key — the registry alone cannot start work. A command planted directly in the registry database was claimed and REFUSED; no pinned key → every deploy refuses.
F · Screens INJECTION	742 + 7 payloads · 5 screens	**0** FP · 48 named misses	One regex screen over everything an agent is about to read (skill files, errands, telemetry, tool inputs, compaction summaries), byte-identical across five copies — a parity test fails if one drifts. A signal to a person, never a block. English families; translations wait for the sidecar model.

WHY THIS AXIS, AND WHY SELF-SCORED

AgentDojo measures whether an attack *completes*; it cannot test an adversary going *around* the gate. The harnesses are reproducible and we report our own failures: each finding surfaced this way (F1, D1, the 630 spikee misses, three of our own checks reading green without running) was fixed that day.

Harnesses: bypass_bench.py · decisions_bench.py · configured_bench.py · canary_bench.py · detection_bench.cjs · authorize_bench.cjs · anchor_bench.cjs · runnerVerify.test.js · injection_bench.cjs · spikee/run.sh. Open work: *containment* of un-hooked execution; a mid-run fresh-anchor write. xChk · October 2026.