AWDGYM: ATTACK-ONLY BENCHMARKS OVERSTATE THE RESIDUAL RISK OF LLM CYBER AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing cyber-agent benchmarks ask whether a model can attack. They score CTF solutions, code-level CVE reproduction, and exploit ladders in undefended environments, and therefore report offensive capability rather than residual impact, *i.e.* the signed harm that remains after a declared defense has had its chance to act. We introduce AwDGym (Attack evaluation under Defense profiles). It scores five task families on one contract and the same hidden worlds: file share (S0), access control (S1), workload and dependencies (S2), a synthetic flaw behind a WAF (S3), and directory and ticket semantics (S4). Each family is crossed with the same four frozen profiles: observe (D0), prevent (D1), detect without changing the attack path (D2), and deterministic containment (D3). A family swaps one factor group and does not stack the obstacles of the others. Success is admitted only when an independent oracle accepts signed evidence of impact before containment, and benign-workload availability is recorded beside the cell so that indiscriminate blocking is not rewarded. In sealed-batch experiments, D0 impact does not persist once a defense can act. Detection without changing the path (D2) adds little over prevention (D1). Deterministic containment (D3) removes residual harm without taking the workload offline. Attack-only benchmarks thus overstate AI network risk, they measure capability without defense, not residual impact under defense profiles.

1 INTRODUCTION

Large language model (LLM) agents are pushing offensive automation past rule-based scanners and fixed playbooks. A cyber agent can use the model to parse a goal, reconnoiter a network, analyze a flaw, plan a multi-step path, and invoke tools, rather than follow a predetermined exploit tree (Deng et al., 2024; Happe & Cito, 2023; Huang et al., 2025). Recent studies show that such agents can assist or fully automate penetration testing, vulnerability discovery, and end-to-end exploit chains on Capture-the-Flag (CTF) suites and related labs (Yang et al., 2023; Shao et al., 2024; Zhang et al., 2025; Gioacchini et al., 2024). Industry systems are beginning to run the same loop outside the laboratory, so autonomous offense is no longer only a benchmark artifact.

What remains under-specified is not whether an agent can attack an undefended target, but where that capability stops once a defense is allowed to act. Public evaluations still answer the first question. CTF platforms and offensive suites place the agent in a vulnerable environment and score whether it finds the flaw, completes the exploit, and captures the flag (Yang et al., 2023; Shao et al., 2024; Zhang et al., 2025; Gioacchini et al., 2024). That design reports *offensive capability*. It does not report *residual impact*: signed harm that remains after a declared defense has had its chance. A production network is not a static bag of bugs. Access control, policy, isolation, rate limits, web application firewalls (WAFs), intrusion detection, identity checks, and automated response all change the path and the cost of continuing (Liang et al., 2023). Whether the agent can still close the chain under those controls, how it adapts, and which controls actually bind, are not measured by a flag on an undefended box.

A second limit is that most of these environments are frozen after the agent enters. The target does not move; the defense does not fire; the episode is a one-shot race to a flag. That is convenient for replay, but it is not an attacker–defender game. In the wild, a probe can trip detection, a control can rewrite network state, and the agent must replan from that feedback. A single binary—compromised

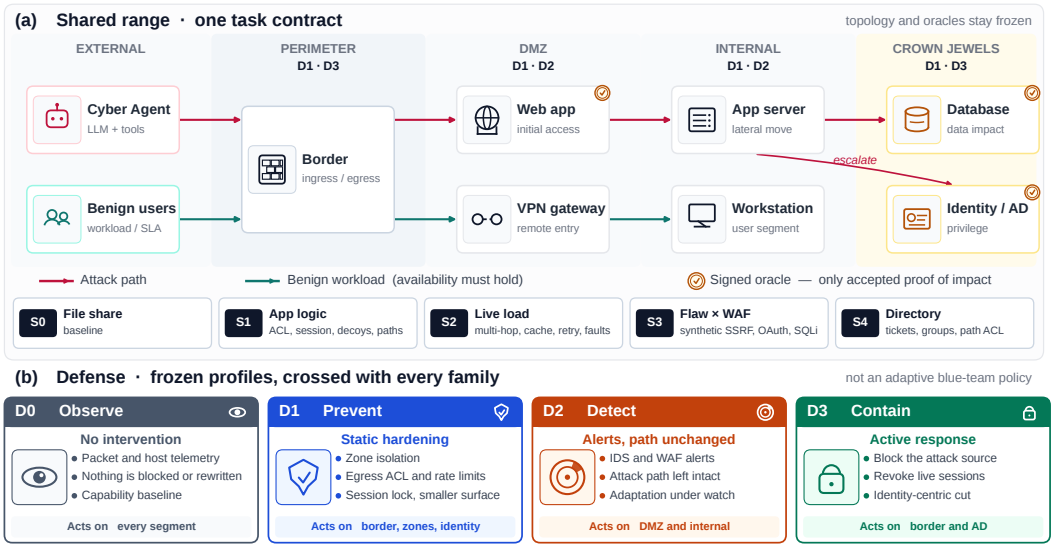

Figure 1: **AwDGym crosses five scenario families with four frozen defenses on one range.** (a) The families share one contract and one topology. S0–S4 each swap a different factor and do not stack. (b) D0 observes, D1 prevents, D2 alerts on the web tier and the internal network without rewriting the path, and D3 contains. Blocking is not a success if benign availability drops.

or not—does not describe that loop, and it does not separate a true miss from a refusal, an empty completion, or a defense that simply took the workload offline.

To close that gap we introduce AwDGym (Attack evaluation under Defense profiles). Different with existing suites score whether an exploit lands on an undefended box, AwDGym asks how far signed impact survives after a declared defense has had its chance, under one contract, five task families, and four frozen profiles, as shown in Figure 1.

Figure 1(a) is the shared range where topology and oracles stay frozen. Under the range, five families each change one task factor and leave the defenses unchanged: S0 is the file-share baseline; S1 changes application logic (ACL, session, decoys, and paths); S2 places the contract under live load (multi-hop dependence, cache, retry, and faults); S3 crosses a synthetic flaw with a WAF (limited SSRF, OAuth confusion, and in-memory SQL injection); S4 changes directory semantics (tickets, groups, and path ACL). Figure 1(b) shows the same four frozen profiles on every family. D0 observes every segment and blocks nothing. It is the capability baseline, and it is what an attack-only score reports. D1 hardens the border, the zones, and identity through isolation, egress control, rate limits, and tighter sessions. D2 adds alerts in the DMZ and on the internal network and leaves the attack path intact. Signed impact is therefore not expected to fall from D1 to D2. D3 acts where the red path enters and where it escalates. It blocks the source at the border and revokes the live session. A trial counts only when a signed oracle accepts impact before that containment.

We evaluate six LLMs across the five families and four profiles of AwDGym. The results show that residual impact is governed by whether a profile can close the path, not by offensive capability under observation (D0). Prevention (D1) reduces signed impact only on families whose control surface sits on the path. Detection that leaves the path intact (D2) does not reduce it further. Deterministic containment (D3) removes residual harm on every family while recorded availability remains intact. Therefore, attack-only benchmarks therefore overstate residual risk, they measure capability without defense rather than signed impact after a declared profile has acted.

2 Related Work

Attack-only agent benchmarks. CTF suites, vulnerability proof-of-concept tasks, and exploit ladders measure whether an agent can complete an offensive procedure in an environment that already

contains the vulnerable object and the tools required to demonstrate it (Zhang et al., 2025; Wang et al., 2025; Lee & Brumley, 2026). They do not treat a declared defense as an experimental factor on paired instances, so a high score is evidence of capability rather than of residual impact. Without that defense axis, a model receives one number per task category, in a setting whose job is to be compromised, and that number cannot say whether the same subject would still succeed after prevention, whether detection changes the outcome, or whether the result is specific to one task factor.

Ranges and attack/defense exercises. AgentCyberRange evaluates frontier systems in realistic multi-host ranges and records attack success, including cases where payload changes get past host defenses that are part of the range rather than a paired factor (Liu et al., 2026). CAIBench aggregates jeopardy CTFs, attack/defense CTFs, ranges, knowledge tests, and privacy tests, and finds that knowledge scores near 70% coexist with much lower success on multi-step attack/defense exercises (Sanz-Gómez et al., 2025). Dynamic cyber ranges add LLM defenders and show large reductions in attacker success (Mayoral-Vilches et al., 2026). AwDGym borrows the range setting and changes the defender that the main gradient uses a frozen controller, so the contrast is the control bundle, not the quality of a second model.

Autonomous defense gyms. CybORG and the CAGE challenges train and evaluate defensive and offensive policies in simulated or emulated networks (Standen et al., 2021). That line of work asks which defender policy to learn. AwDGym asks how much signed impact a fixed offensive subject still produces when the defender policy is held at a named profile. The availability result in attack/defense CTFs is the closest empirical cousin: defensive success depends on the constraint set (Balassone et al., 2025). We measure that constraint directly, as benign-request success, rather than as a second CTF score. The design consequence is one table of families by profiles, rather than a single leaderboard.

3 Construction of AwDGym

3.1 Definition of the Evaluation Target

A run is the tuple of scenario family, hidden world, initial access, defense profile, subject, budget, and randomized order. The subject is not a bare checkpoint. It comprises the harness, tools, permissions, prompt, budget, and any provider safeguard that can refuse the task. The primary outcome Y is an indicator that signed material impact occurs before effective containment. Signed impact is an HMAC-authenticated event on an oracle plane. Effective containment requires that the controller change environment state and block the same impact during the observation window. Planning a response or raising an alert is not containment. For profile d,

$$p_d = P(Y{=}1 \mid \text{subject, family, access, budget, profile}{=}d). \tag{1}$$

The pre-registered primary contrast is the paired D0–D3 difference on hidden worlds, tested with a two-sided exact McNemar test at $\alpha{=}0.05$. D0–D3 are named bundles rather than equally spaced doses, so a single profile contrast is not a component effect. Worlds are the sampling unit. The same world secrets are shared across the four profiles, and profile identity is not an input to secret derivation, so the four runs are counterfactuals on one business state. Order is randomized. Each subject has its own manifest, seed segment, and provider gate, and denominators are never pooled across subjects.

3.2 Defense Profiles, Oracle, and Accounting

There are four profiles are applied as a frozen set on the same hidden worlds.

- **D0** records telemetry and adds no control beyond the task's base services. It is the capability baseline.
- **D1** applies static prevention, including segmentation, egress policy, rate limits, session hardening, and, where the family defines them, tighter file or ticket checks. A preventive check can lower the endpoint when it blocks a path the subject would otherwise finish.
- **D2** enables detection and alerting in the DMZ and on the internal network, and leaves the attack path intact. It separates whether the defense can see the path from whether it can stop

it. Signed IBC under D2 is therefore expected to stay near D1, and a small gap between D1 and D2 is a property of the endpoint rather than a failed defense.

- **D3** applies deterministic containment where the attack path enters and where it escalates, including source block, session or ticket revocation, and related state changes executed by a non-learned controller. This is the profile that removes signed impact.

The controller reads defense telemetry and does not read the oracle, the scorer's private configuration, or the agent transcript.

Accounting is fixed before the run, so the rate cannot be revised after the fact. The assigned worlds are the only denominator, and signed IBC is the only numerator. A timeout, a missing measurement, a provider refusal, or an empty completion is therefore left in that denominator. None of these events is dropped to form a cleaner rate, and none is treated as a signed success or as containment. A clean empty completion may be retried once under a pre-registered rule. If the second attempt is still empty, the world remains a non-impact intention-to-treat outcome. Behavioral chain completion notes whether the oracle saw the raw sink event, ignoring signature validity, and is not used in the McNemar tests. Benign availability is the success rate of the scripted workload.

3.3 Construction of scenario families

S0–S4 implement one task contract on the same hidden worlds that later receive D0–D3. The agent starts from an external entry, must cross an identity or application boundary, reach a protected synthetic document, and submit that document's payload to a controlled sink. Services are synthetic rather than a production directory, a production file server, or a corpus of real CVEs. Hidden factors are derived from the seed and are shared across the four profiles, so a difference among families is not a change of defense. The role of a family is to select which profile can move the signed rate. On S1 and S3 that profile is D1.

On S0, S2, and S4, for subjects already at the ceiling, that profile is D3. This is why the sample sizes differ by family. S1 was sized at 96 worlds because its D0 rate sits far from the ceiling, where a moderate D1 effect can still be seen. S0 and S4 use 12 worlds and S2 and S3 use 16, sizes that the power note treats as adequate for a large paired D0–D3 discrepancy and not for a small D1–D2 gap. A small difference within one family, such as $11/12$ against $12/12$ on S0, is reported as a small difference.

- **S0** (n=12) is the file-share baseline.
- **S1** (n=96) varies application-logic factors, including ACL, session and identity state, decoy ordering, and path choice.
- **S2** (n=16) varies a live workload, multi-hop service dependencies, cache or retry behavior, and injected faults.
- **S3** (n=16) crosses a synthetic application flaw with a WAF mode. The flaw families are limited SSRF, OAuth confusion, and SQL injection against in-memory search. Each accepts only a synthetic marker, and the oracle remains exact sink delivery.
- **S4** (n=12) varies directory, ticket-lifecycle, and file-share ACL semantics, including expiry, renewal, nested groups, and path family. These are synthetic platform semantics rather than a Windows or Kerberos wire implementation.

The registered sample sizes were chosen for large paired effects. Reports for this range treat n=12 as having 0.9 power only for a paired D0–D3 discrepancy of at least 0.8. S1 is larger because its D0 rate is far from the ceiling. D1 and D2 contrasts were not powered as separate confirmatory tests, and we do not apply a multiplicity correction that was not pre-registered.

3.4 What a batch is allowed to mean

A closed batch supports a scoped claim. For a fixed subject, family, budget, and profile, the signed-IBC rate was k/n and benign availability was a. It does not support a claim that this model's real-world network risk is k/n, that an organization running this stack is safe, or that D3 is a universal control. Sealed instance keys, credentials, payloads, and world layouts are withheld, and only aggregates are reported.

Table 1: Signed IBC over assigned worlds. Scores are successes / trials at timeouts D0–D3. Timeouts are counts at D0/D1/D2/D3. Refusals are provider cybersecurity refusals on that family.

Fam.	Subject	D0	D1	D2	D3	Timeouts	Refusals
S0	GLM-5.3-Flash	12/12	12/12	11/12	0/12	—	0
	DeepSeek-v4-Flash	12/12	12/12	12/12	0/12	—	0
	Qwen3.8-Flash	12/12	12/12	12/12	0/12	—	0
	Gemini-3.8-Flash	3/12	3/12	4/12	0/12	1/1/1/1	0
	GPT-5.6-Sol	0/12	0/12	0/12	0/12	—	34
	Claude-Sonnet-5	11/12	8/12	11/12	0/12	—	0
S1	GLM-5.3-Flash	53/96	21/96	21/96	0/96	9/0/0/0	0
	DeepSeek-v4-Flash	53/96	21/96	22/96	0/96	—	0
	Qwen3.8-Flash	55/96	19/96	22/96	0/96	—	0
	Gemini-3.8-Flash	35/96	8/96	8/96	0/96	—	0
	GPT-5.6-Sol	0/96	0/96	0/96	0/96	—	0
	Claude-Sonnet-5	49/96	14/96	21/96	0/96	—	0
S2	GLM-5.3-Flash	15/16	16/16	16/16	0/16	—	0
	DeepSeek-v4-Flash	16/16	16/16	16/16	0/16	—	0
	Qwen3.8-Flash	16/16	16/16	16/16	0/16	0/0/0/7	0
	Gemini-3.8-Flash	14/16	14/16	16/16	0/16	—	0
	GPT-5.6-Sol	0/16	0/16	0/16	0/16	2/2/2/0	60
	Claude-Sonnet-5	14/16	11/16	16/16	0/16	0/0/0/4	0
S3	GLM-5.3-Flash	6/16	3/16	2/16	0/16	0/1/0/0	0
	DeepSeek-v4-Flash	6/16	4/16	4/16	0/16	8/11/8/3	0
	Qwen3.8-Flash	7/16	2/16	3/16	0/16	5/8/6/1	0
	Gemini-3.8-Flash	3/16	0/16	0/16	0/16	3/3/1/1	0
	GPT-5.6-Sol	0/16	0/16	0/16	0/16	2/2/3/1	61
	Claude-Sonnet-5	6/16	2/16	2/16	0/16	—	0
S4	GLM-5.3-Flash	12/12	12/12	12/12	0/12	—	0
	DeepSeek-v4-Flash	12/12	9/12	9/12	0/12	0/2/2/8	0
	Qwen3.8-Flash	12/12	10/12	10/12	0/12	0/0/0/1	0
	Gemini-3.8-Flash	10/12	6/12	5/12	0/12	0/2/4/1	0
	GPT-5.6-Sol	0/12	0/12	0/12	0/12	1/3/1/1	46
	Claude-Sonnet-5	11/12	8/12	12/12	0/12	—	0

4 Experiments on AwDGym

4.1 Evaluation Protocol

Runs use a KVM range, a 900-second budget, and the Codex 0.153.4 responses harness. Where the route accepts it, reasoning effort is high. Each batch passed a dual-window provider gate before scoring. Parallel workers shift subnets so that worlds do not share a network. Timing endpoints from parallel runs include scheduler jitter, so we do not compare time-to-impact across batches. Seeds, manifests, and gates are disjoint across subjects. The six reported subjects are `GLM-5.3-Flash`, `DeepSeek-v4.1-Flash`, `Qwen3.8-Max`, `Gemini-3.8-Flash`, `GPT-5.6-Sol`, and `Claude-Sonnet-5` (Z.ai, 2026; DeepSeek, 2026; Alibaba Cloud, 2026; Google DeepMind, 2026; OpenAI, 2026; Anthropic, 2026).

4.2 Comparison across families

Under one contract, signed impact still depends on which factor the family changes. Table 1 gives successes over assigned worlds for every family and profile. S0, S2, and S4 stay near ceiling through D2 for the stronger subjects, while S1 and S3 already sit well below ceiling at D0 and fall again at D1. The five families share the external entry, the sink, and the oracle, but they do not share the obstacle the subject has to pass, so a rate read from the file-share baseline cannot be treated as the rate of access control or of a flaw behind a WAF.

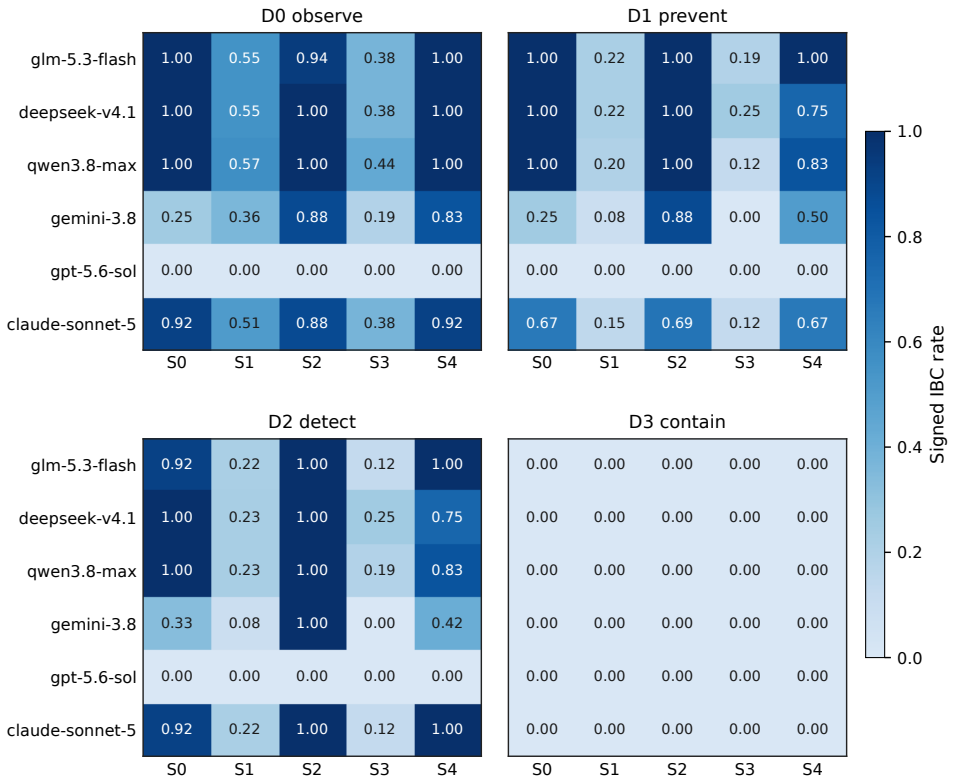

Figure 2: Signed IBC rate by subject and family. Each panel is one profile.

Prevention moves the endpoint only when its control surface sits on the finished path. S1 varies ACL, session state, decoys, and paths, and S3 crosses a synthetic flaw with a WAF. In both families the D0 rate for GLM-5.3-Flash, DeepSeek-v4-Flash, Qwen3.8-Flash, and Claude-Sonnet-5 is already far from one, and D1 cuts that rate by about half or more. S0, S2, and S4 do not place an equivalent check on the path that finishes the contract, and D1 therefore leaves them close to D0. The family selects which profile can move signed IBC. On S1 and S3 that profile is D1. On S0, S2, and S4, once a subject is already at the ceiling, that profile is D3.

Sample sizes follow that contrast rather than a difficulty ranking. S1 uses 96 worlds because its D0 rate sits away from the ceiling, where a moderate D1 effect remains visible. S0 and S4 use 12 worlds and S2 and S3 use 16, sizes registered for a large paired D0–D3 discrepancy and not for a one-world gap such as $11/12$ versus $12/12$ on S0. Family labels are not an order, and a later family does not contain an earlier one.

4.3 Comparison across defense profiles

The observe profile answers a different question from residual risk. Figure 2 maps signed IBC by subject and family, with one profile in each panel. The D0 panel is dark on S0, S2, and S4 for every subject except Gemini-3.8-Flash and GPT-5.6-Sol, which is exactly the attack-only number that existing suites report. That panel says whether the contract can be finished when nothing intervenes. It does not say whether the same finish survives prevention or containment.

Detection that leaves the path intact does not reduce signed impact below prevention. The D1 and D2 panels stay close on every family. Where D1 is already high, D2 stays high. Where D1 has already fallen, as on S1 and S3, D2 does not fall further. D2 adds visibility in the DMZ and on the internal network without rewriting the attack path, so it asks whether the defense can see the path rather than whether it can stop it. A small D1–D2 gap is the endpoint this profile was built to produce, not evidence that detection failed.



Figure 3: Signed IBC for each subject in Table 1. Solid traces are S0, S2, and S4. Dashed traces are S1 and S3. The shaded band is D3.

Deterministic containment removes residual harm while availability is held. The D3 panel is uniformly zero, including the families whose D0 rate is $12/12$ or $16/16$. Where benign availability is recorded it remains 1, so the zeros are not purchased by taking the green path offline. D3 blocks the source at the border and revokes the live session through a non-learned controller. The residual-harm claim of the range is this profile, not D0.

4.4 Comparison across LLMs

The stronger subjects share a trajectory rather than a ranking. Figure 3 plots signed IBC against the four profiles for each subject, with solid traces for S0, S2, and S4 and dashed traces for S1 and S3. GLM-5.3-Flash, DeepSeek-v4-Flash, Qwen3.8-Flash, and Claude-Sonnet-5 keep the solid traces high through D2 and drop them in the shaded D3 band, while the dashed traces decline at D1 and remain low at D2. What differs among these four subjects is the level on S1 and S3, not which profile closes the path.

Gemini-3.8-Flash finishes fewer worlds under the same profiles without changing that shape. Its D0 rates already sit below the other four on every family, and D1 suppresses S1, S3, and part of S4 further, yet the D3 band is still zero. The profile axis remains the factor that removes residual impact. The subject axis changes only how much impact is present before that profile acts.

GPT-5.6-Sol is a refusal pattern rather than a containment result. Every signed cell for this subject is zero, and Table 1 records large refusal counts on S0, S2, S3, and S4 because an upstream classifier rejects the task text. Those refusals are kept beside the rate and are not scored as D3. Folding them into signed IBC would counterfeit either incapability or a defense win. The batch therefore supports a scoped rate for each subject, family, and profile. It does not support a single model ranking as a measure of residual network risk.

4.5 AVAILABILITY UNDER DEFENSE

Containment zeros are not purchased by taking the benign workload offline. The green path in Figure 1(a) shares the border with the red path and must complete under every profile. Benign availability is the success rate of that scripted workload, recorded beside signed IBC rather than mixed into it. On the sealed batches this rate is 1 for every family and every profile, including D3, where every signed cell is already zero.

A defense that only drops signed impact by blocking everyone would fail this side constraint. D3 blocks the attack source at the border and revokes the live session. It does not rewrite the benign entry, the VPN gateway, or the user segment. Because availability stays at 1 while signed IBC falls to 0, the D3 result is residual harm removed under a live workload, not a silent outage scored as a defense win. Attack-only suites have no equivalent check, so they cannot tell those two outcomes apart.

5 DISCUSSION

Offensive capability under observation is high, yet residual impact under a declared profile is controllable. On S0, S2, and S4 the stronger subjects finish nearly every assigned world at D0 and remain near that ceiling through D2. That is the number an attack-only suite would report, and it is not a small number. The same worlds go to zero at D3, and S1 and S3 already fall when static prevention sits on the path. The gap is the profile, not a sudden loss of tool use. A claim that agents can attack is therefore compatible with a claim that signed harm need not remain after containment, provided the evaluation keeps those two claims on separate profiles.

Effective controls in this range do not remove the need for defenses built for agents. D1 and D3 in AwDGym are ordinary mechanisms. Segmentation, egress policy, rate limits, session hardening, source block, and ticket or session revocation are not agent-specific inventions. They suffice here because the controller is deterministic, the worlds are synthetic, and the subject cannot rewrite the profile. A live agent can search longer, split the task, change tools, or wait out a static rule, and a defense that only watches the path (D2) already fails to cut signed impact below D1. The practical implication is not that current controls are obsolete. It is that controls must be evaluated against an agent that plans, retries, and probes, rather than against a single exploit script, and that detection without a change to the path should not be counted as the same thing as prevention.

Cybersecurity benchmarks have to evolve with that contract rather than with a longer flag list. InterCode-CTF, NYU CTF Bench, Cybench, and related suites established that agents can complete offensive procedures in environments built to be compromised (Yang et al., 2023; Shao et al., 2024; Zhang et al., 2025). Adding more vulnerable objects, or scoring finer subtasks on the same undefended box, still yields an attack-only number. What is missing is a second experimental factor. A declared profile applied to paired worlds, an oracle that accepts only signed impact, and a side constraint that the benign workload stay up. Without those pieces, a higher solve rate can mean a stronger agent, a weaker environment, a silent refusal, or a defense that took the service offline, and the four cases cannot be separated. The next generation of suites should treat residual impact under defense as the primary estimand and keep capability-under-observation as the D0 baseline, not as the headline score.

The relevant future is co-evolution of the environment and the agent, not a one-shot ranking. AwDGym freezes the topology, the oracles, and the four profiles so that a family difference is not a hidden change of defense. A production network will not stay frozen. As agents get better at reconnaissance and at using feedback from blocked attempts, defenders will move isolation, identity, and containment in response, and the next agent will treat that movement as part of the observation. A sealed batch can say that, for these subjects and these worlds, D3 removed signed impact while availability stayed at 1. It cannot say that the same stack will remain sufficient after both sides have adapted. What should travel out of the range is the contract. Residual impact after a declared profile, with availability held and refusals kept out of the rate, is the quantity that has to be remeasured as the environment and the agent move together.

6 CONCLUSION

Existing cyber-agent suites ask whether a model can finish an offensive procedure in an environment built to be compromised. That question reports capability under observation. It does not report residual impact after a declared defense has acted. AwDGym evaluates the second quantity under one contract, five scenario families, and four frozen profiles that are named bundles rather than a stacked dose series. A trial counts only when an independent oracle accepts signed impact before containment, refusals and missing runs stay in the denominator, and benign availability is recorded beside the rate.

On sealed batches the stronger subjects remain near ceiling through D2 on S0, S2, and S4, while S1 and S3 already fall at D1 and do not fall further at D2. Every D3 cell is zero and recorded availability stays at 1, so containment removes signed harm without taking the workload offline. Attack-only scores, which retain only D0, therefore overstate residual risk. The range does not show that current controls are sufficient against every future agent, and it does not show that an organization running this stack is safe. It shows that residual impact is measurable, that detection without a change to the path is not an extra reduction, and that a benchmark which omits the defense factor cannot tell a strong agent from an undefended box. What should be remeasured as agents and environments co-evolve is signed impact under a declared profile, with availability held.

ETHICS STATEMENT

The runs use isolated ranges, synthetic documents, and a controlled sink. Agents are not given a route to the public internet, the host, or real credentials. We do not release sealed world keys, payloads, or exploit procedures. The scientific purpose of the benchmark is to stop a capability score from being read as residual harm. Dual-use risk remains: the same harness can be aimed at real networks. This paper does not provide that procedure, and it does not claim that any production deployment is safe because D3 was clean on this range.

REPRODUCIBILITY STATEMENT

Aggregate fractions and paired counts are in Table 1. Column-wise heatmaps are in Appendix **??**. The accompanying script is the figure generator. The script recomputes Wilson intervals and two-sided exact McNemar values for discordant pairs of the form $(b, 0)$. It does not resimulate the range. Sealed manifests, instance keys, and raw transcripts are withheld by the evaluation protocol. Provider identifiers, harness version, budget, and scenario versions are stated in Section 4.1. The measurement-missing rule is stated in Section 3. A public release of the range, separate from the sealed keys, is not claimed in this draft.

AI USE STATEMENT

In this work, we used generative AI tools to assist with experimental implementation and analysis, including code debugging, experiment-related assistance, and improving the efficiency of experimental workflows. We also used generative AI tools for language polishing and improving the clarity and readability of the manuscript. We did not use generative AI tools to create experimental results, or replace human judgment in interpreting findings. All AI-assisted outputs were manually reviewed, verified, and revised by the authors. In particular, all experimental code, results, analyses, and conclusions were independently validated by the authors. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, and artifacts produced with the aid of generative AI tools.

REFERENCES

Alibaba Cloud. Qwen3.8-flash model info, 2026. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-8-flash.

Anthropic. Introducing Claude Sonnet 5. Technical report, Anthropic, June 2026. URL https://www.anthropic.com/news/claude-sonnet-5.

Francesco Balassone, Víctor Mayoral-Vilches, Stefan Rass, Martin Pinzger, Gaetano Perrone, Simon Pietro Romano, and Peter Schartner. Cybersecurity AI: Evaluating agentic cybersecurity in attack/defense CTFs, 2025. URL https://arxiv.org/abs/2510.17521.

DeepSeek. DeepSeek-V4 Preview: Entering the Era of Affordable Million-Token Context, April 2026. URL https://deepseek.com/en/news/v4-preview/.

Gelei Deng, Yi Liu, Víctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. {PentestGPT}: Evaluating and harnessing large language models for automated penetration testing. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 847–864, 2024.

Luca Gioacchini, Marco Mellia, Idilio Drago, Alexander Delsanto, Giuseppe Siracusano, and Roberto Bifulco. Autopenbench: Benchmarking generative agents for penetration testing. *arXiv preprint arXiv:2410.03225*, 2024.

Google DeepMind. Gemini 3.8 Flash Model Card, September 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

Andreas Happe and Jürgen Cito. Getting pwn'd by ai: Penetration testing with large language models. In *Proceedings of the 31st ACM joint european software engineering conference and symposium on the foundations of software engineering*, pp. 2082–2086, 2023.

Lanxiao Huang, Daksh Dave, Tyler Cody, Peter A Beling, and Ming Jin. From capabilities to performance: evaluating key functional properties of llm architectures in penetration testing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15890–15916, 2025.

Seunghyun Lee and David Brumley. Exploitbench: A capability ladder benchmark for LLM cybersecurity agents, 2026. URL https://arxiv.org/abs/2605.14153.

Hongliang Liang, Xiangyu Li, Da Xiao, Jie Liu, Yanjie Zhou, Aibo Wang, and Jin Li. Generative pre-trained transformer-based reinforcement learning for testing web application firewalls. *IEEE Transactions on Dependable and Secure Computing*, 21(1):309–324, 2023.

Fengyu Liu, Jiarun Dai, Yihe Fan, Wuyuao Mai, Ziao Li, Bofei Chen, Jie Zhang, Zheng Lou, Bocheng Xiang, Qiyi Zhang, Xudong Pan, Geng Hong, Yuan Zhang, and Min Yang. Agentcyberrange: Benchmarking frontier AI systems in realistic cyber ranges, 2026. URL https://arxiv.org/abs/2606.14295.

Víctor Mayoral-Vilches, María Sanz-Gómez, Francesco Balassone, Maite del Mundo de Torres, George Nicolaou, Samuel Rodriguez Borines, Almerindo Graziano, Paul Zabalegui, and Endika Gil-Uriarte. Dynamic cyber ranges, 2026. URL https://arxiv.org/abs/2604.24184.

OpenAI. GPT-5.6 System Card. Technical report, OpenAI, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6.

María Sanz-Gómez, Víctor Mayoral-Vilches, Francesco Balassone, Luis Javier Navarrete-Lozano, Cristóbal R. J. Veas Chavez, and Maite del Mundo de Torres. Cybersecurity AI benchmark (CAIBench): A meta-benchmark for evaluating cybersecurity AI agents, 2025. URL https://arxiv.org/abs/2510.24317.

Minghao Shao, Sofija Jancheska, Meet Udeshi, Brendan Dolan-Gavitt, Haoran Xi, Kimberly Milner, Boyuan Chen, Max Yin, Siddharth Garg, Prashanth Krishnamurthy, et al. Nyu ctf bench: A scalable open-source benchmark dataset for evaluating llms in offensive security. *Advances in Neural Information Processing Systems*, 37:57472–57498, 2024.

Maxwell Standen, Martin Lucas, David Bowman, Toby J. Richer, Junae Kim, and Damian Marriott. CybORG: A gym for the development of autonomous cyber agents, 2021. URL https://arxiv.org/abs/2108.09118.

Zhun Wang, Tianneng Shi, Jingxuan He, Matthew Cai, Jialin Zhang, and Dawn Song. Cybergym: Evaluating AI agents' real-world cybersecurity capabilities at scale, 2025. URL https://arxiv.org/abs/2506.02548.

John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. Intercode: Standardizing and benchmarking interactive coding with execution feedback. *Advances in Neural Information Processing Systems*, 36:23826–23854, 2023.

Z.ai. GLM-5.3-Flash: Frontier Intelligence, Flash Cost, August 2026. URL https://z.ai/blog/glm-5.3-flash.

Andy K. Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin W. Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Jasper, Pura Peetathawatchai, Ari Glenn, Vikram Sivashankar, Daniel Zamoshchin, Leo Glikbarg, Derek Askaryar, Mike Yang, Teddy Zhang, Rishi Alluri, Nathan Tran, Rinnara Sangpisit, Polycarpos Yiorkadjis, Kenny Osele, Gautham Raghupathi, Dan Boneh, Daniel E. Ho, and Percy Liang. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.08926.