Reproducibility and Validity of Cyber Reasoning System Benchmarks

Stavros Kousidis

Master thesis · Published: 2026-10-15

Research question

How reliably do existing benchmarks measure the vulnerability discovery and remediation capabilities of Cyber Reasoning Systems under reproducible and comparable evaluation conditions?

Description

Cyber Reasoning Systems (CRSs) can autonomously inspect software and support vulnerability discovery or remediation. Public benchmarks and competition results make such systems measurable, but do not automatically establish that reported performance is reproducible, comparable or representative of previously unseen software.

You will investigate one existing benchmark and at least one accessible CRS or agent-based security-analysis system. Develop an evaluation methodology that distinguishes actual task success from unsupported reports, environmental failures and artifacts of the benchmark setup. The focus is on what the measurements justify, not on producing a new scoreboard.

Scope and evaluation

Select a feasible subset of publicly documented vulnerability tasks and record the software, tool access, compute budget and evaluation settings. Repeat trials and analyze variability. Investigate at least one possible source of bias, such as benchmark contamination, differing budgets, ground-truth quality or unreliable patch validation. Compare the measured conclusions with a fixed baseline or reference procedure. Use only authorized, isolated test environments.

A preliminary feasibility study should determine whether the chosen CRS can be deployed within the available resources.

Artifact licensing note: Before fixing the benchmark, check the licenses and redistribution rights for benchmark code, agent/model components, vulnerable software snapshots, and datasets. Publish independently developed evaluation scripts under an agreed license; where third-party artifacts cannot be redistributed, document how authorized users can obtain them instead.

Expected outcome

A reproducible evaluation protocol, experimental results, and a critical assessment of the validity and limitations of benchmark-based CRS comparisons.

Starting points

- CyberGym benchmark and dataset.
- AI Cyber Challenge open-source archive.
- Zhang et al., *SoK: DARPA's AI Cyber Challenge* (2026).
- Wang et al., *CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale.*

Contact: stavros.kousidis@th-koeln.de
Code publication: open source by default; exceptions require prior agreement (policy).