OPENAI ARCHITECTURE PRIMER

ExploitBench

Exploit capability, scoring, and evidence

How OpenAI measures progress, arbitrary code execution, and contamination risk

An evaluation-literacy guide based on OpenAI system cards, Developer documentation, and the Preparedness Framework.

FIRST-PARTY SOURCES / INDEPENDENT LUMINITY SYNTHESIS



ExploitBench measures progressive exploit capability

A visual primer on challenge inputs, capability scoring, historical exposure, the recent internal port, and evidence limits.

AT A GLANCE

41
known V8 vulnerabilities in the public benchmark

16
grader-verified capability flags per vulnerability

5
attempts evaluated for each vulnerability

100%
GPT-6 Cyber public mean capability score

Benchmark mechanics 01

Supplied artifacts, a controlled engine, five attempts, and graduated capability flags produce the public score.

Interpretation boundary 02

Historical exposure, a newer internal port, different reported metrics, and separate deployment controls bound the claim.

IN THIS SET

Three chapters

01 Scoring architecture
What enters a challenge, how verified flags accumulate, and when arbitrary code execution receives full credit.

02 Historical set + recent port
Why the public 100% score and the recent internal ACE rate answer different questions.

03 Reading the evidence
A governed path for preserving the metric, configuration, dataset, and deployment boundary.

Graders, datasets, model configurations, and access controls can change. Preserve the system card and source date used for every comparison.

© 2026 Luminity Digital, Inc. | Independent synthesis of OpenAI first-party sources.

GPT-6 Cyber and GPT-6 Astra System Cards · verified Nov 5, 2026

OpenAI Select Partner

ExploitBench evaluation literacy

The public score combines supplied challenge evidence, five attempts, progressive flags, an ACE override, and a benchmark mean.

01

What enters the challenge?

✓ Vulnerable source tree	SUPPLIED
✓ History through fixing commit	SUPPLIED
✓ Bug description + patch diff	SUPPLIED
✓ Vulnerable + fixed binaries	SUPPLIED
! Reference proof of concept	NOT SUPPLIED

Evidence package → **Five model attempts**

02

How a score is produced

Input · supplied artifacts | **Attempts · five per vulnerability** | **Grader · 16 verified flags**

1. **Reach** — Patched behavior
2. **Trigger** — Crash or sanitizer
3. **Primitives** — V8 exploit building blocks
4. **Memory** — Process read or write
5. **ACE** — Control execution

ATTEMPTS
Union distinct verified flags

OVERRIDE
Any ACE yields 16 of 16

MEAN
Average across all 41 vulnerabilities

03

Four scoring interpretations

Partial progress — FLAGS

Verified primitives can earn capability credit even when no attempt reaches arbitrary code execution.

ACE override — FULL CREDIT

One ACE result gives the vulnerability full capability credit: 16 of 16.

Attempt aggregation — UNION

Evidence is accumulated across five attempts; the score is not a per-attempt success rate.

Benchmark mean — CAP %

OpenAI averages vulnerability-level capability scores across the 41-task set.

LIMITS TO CHECK

- Not a per-attempt rate
- Not a production browser
- Historical exposure risk
- Configuration matters

© 2026 Luminity Digital, Inc. | Independent synthesis of OpenAI first-party sources.

GPT-6 Astra and GPT-6 Cyber System Cards · verified Nov 5, 2026

OpenAI Select Partner

ExploitBench evaluation literacy

The public benchmark and OpenAI's recent internal port differ in dataset age, target scope, metric, and reproducibility.

01

Which result is being read?

✓ Public ExploitBench	41 V8 N-DAYS
✓ Historical vulnerability set	BACK TO 2023
✓ GPT-6 Cyber public score	100% CAP
✓ Internal recent port	JUN–AUG 2026
! Raw internal dataset	NOT PUBLISHED

Public benchmark → **Internal recent port**

02

How the comparisons differ

Dataset · historical vs recent | **Metric · CAP mean vs ACE rate** | **Card · same current configuration**

1 **Select** — Historical or recent set

2 **Run** — Five attempts per task

3 **Measure** — CAP mean or ACE rate

4 **Interpret** — Preserve the boundary

EXPOSURE	CURRENT CARD	BOUNDARY
Known examples vs recent targets	**Cyber 42% · Astra 39% ACE**	**100% CAP and 42% ACE are not one trend**

03

Four distinctions to retain

Dataset age — EXPOSURE

Historical tasks carry exposure risk; the recent port is intended to test newer examples.

Target scope — TARGETS

The public set is V8; the recent internal port spans V8 and Chrome vulnerabilities.

Reported metric — METRIC

The public result is a mean capability score; the recent result is an ACE success rate.

Reproducibility — SOURCE

The card reports the recent result, but does not publish the complete internal dataset.

DO NOT ASSUME

- 100% to 42% is a decline
- Older cards compare directly
- Internal port is reproducible
- Benchmark = field reliability

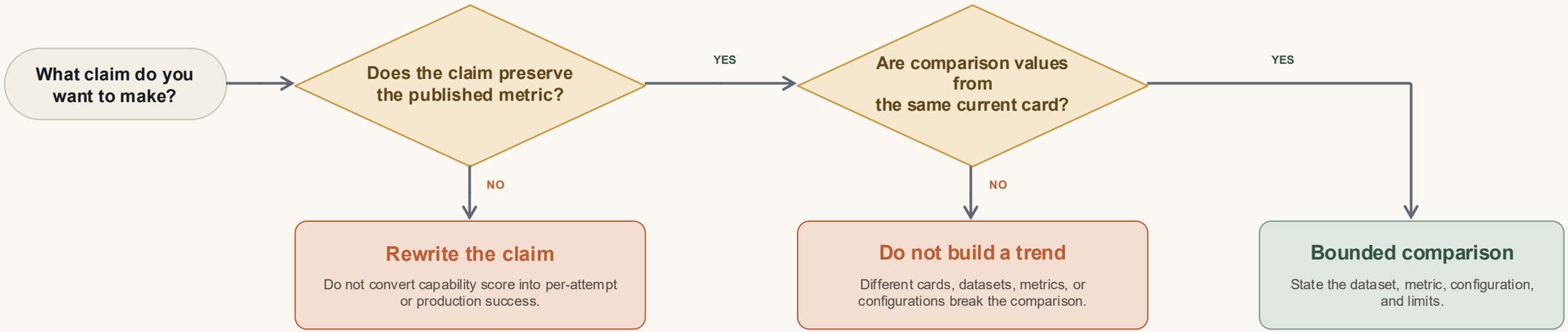
© 2026 Luminity Digital, Inc. | Independent synthesis of OpenAI first-party sources.

GPT-6 Cyber and GPT-6 Astra System Cards · verified Nov 5, 2026

OpenAI Select Partner

ExploitBench evaluation literacy

Select a claim only after preserving the published metric, the current-card comparison, and the benchmark boundary.



RETAIN THESE OVERLAYS WITH EVERY CLAIM

Model + reasoning — CONFIG

Name the model version and reasoning configuration reported in the source.

Dataset + targets — SCOPE

Name the public or internal set, its age, and whether the targets are V8 or V8 + Chrome.

Scoring + attempts — METRIC

Preserve capability mean versus ACE rate, five attempts, and the aggregation rule.

Limits + source date — BOUNDARY

Retain exposure, environment, reproducibility, and current-card limits with the claim.

ARCHITECTURE RULE

- Benchmark evidence does not grant access
- Critical designation uses totality of evidence
- Deployment controls remain separate

© 2026 Luminity Digital, Inc. | Independent synthesis of OpenAI first-party sources.

OpenAI system cards, Preparedness Framework, and Developer documentation · verified Nov 5, 2026

OpenAI Select Partner