Submersion AI

SYSTEM MODEL CARD / 01

Basin

A compact security research model for authorized vulnerability discovery.

Submersion AI • 23 September 2026

MODEL PROFILE

Architecture

Dense · <50B parameters

Context

256k tokens

Focus

Authorized security research

Changelog

- **23 September 2026:** Initial version.

1 Introduction

Basin is a security research model for authorized vulnerability discovery. It can inspect a live target or a supplied codebase, investigate a suspected weakness, and produce a proof of concept and a written finding for human review.

Attribute	Basin
Model	Dense language model with fewer than 50 billion parameters
Context window	256k tokens (262,144 tokens)
Input and output	Text input; text and tool calls as output
Primary use	Authorized vulnerability research and verification

Evaluation mode	Information available to the agent
Black-box	An in-scope URL, host, or service endpoint
Source-assisted	An in-scope repository or package tree, with a test target when the task requires one

1.1 Model development and training

Basin was developed in two broad stages. **Pretraining** established general competence with language and code: reading technical material, following relationships across files, and generating explanations and candidate programs. **Post-training** adapted that foundation for instruction following and multi-step, tool-mediated work. In a security workflow, these capabilities support forming a hypothesis, inspecting evidence, testing a bounded candidate, and writing a finding that a researcher can verify. These are descriptions of the model's intended functions, not guarantees that any particular output is correct.

The evaluations in this card measure the final model in specified settings. Agent results also depend on the tools, permissions, time limits, and review process around the model. Basin is intended to assist authorized researchers; operators remain responsible for defining scope, checking evidence, and deciding what to do with a finding. **[2]**

Use of Basin is subject to Submersion AI's Acceptable Use Policy and applicable law. Security testing requires authorization from the system owner and adherence to the agreed scope.

2 Cyber capabilities

AI can accelerate vulnerability research for both defenders and attackers. The UK National Cyber Security Centre assesses that AI is already improving parts of intrusion activity and that wider access to AI-enabled tools will increase the number of capable threat actors. It also warns that cyber resilience depends on keeping pace with frontier capabilities. **[1]** Defenders therefore need robust, authorized tools capable of finding and verifying weaknesses at a comparable pace, backed by people and processes that can validate, prioritize, and fix what those tools find. **[1] [2]** Basin is designed to support that defensive work. The evaluation below measures targeted vulnerability reproduction, one component of cyber capability; it does not establish autonomous attack capability or safe operation on live targets. **[3]**

2.1 CyberGym

CyberGym Level 1 tests whether an agent can reproduce a known vulnerability from a description and an unpatched codebase. Its official verifier checks that the submitted proof of concept triggers on the vulnerable version and not on the fixed version. The suite contains 1,507 cases across 188 projects. **[3]**

Basin was evaluated in **one pass over the 1,507 cases using a Codex CLI agent harness**. The case-level score summary records **1,218 successful cases (80.8%)**; 21 cases were not scored. **[10]**

CyberGym Level 1 · vulnerability reproduction success

The CyberGym leaderboard reports Grok 4.6 and Meta's Muse Spark 1.1 helpful-only model; Anthropic reports the Claude Opus 4.8 result in its system card. Meta's evaluation report also gives the Muse Spark 1.1 result. **[4] [5] [12]** The systems use different agent harnesses and evaluation settings, and CyberGym cautions that small score differences may not reflect meaningful capability gaps. The chart supplies context rather than a controlled head-to-head ranking. **[4]**

3 General capability evaluations

HumanEval measures functional correctness on short Python programming tasks; IFEval checks compliance with objectively verifiable instructions. **[6] [7]** Basin's internal runs used the final BF16 model on a 32,768-token serving window; all evaluated prompts fit within that window. The benchmark papers define the tasks, while the Basin values below are results reported by Submersion AI.

Evaluation	Basin	Protocol
HumanEval pass@1	**97.6% (160 / 164)**	Full suite; temperature 0; xhigh thinking; Python execution in Docker
IFEval final accuracy	**82.3%**	Full 541 prompts; xhigh thinking
IFEval prompt strict / loose	**77.3% / 80.8%**	Automatically checked constraints
IFEval instruction strict / loose	**84.2% / 86.8%**	Automatically checked constraints

4 Dual-use knowledge evaluations

WMDP measures answers to multiple-choice questions about potentially sensitive knowledge in cyber, biological, and chemical domains. It does not measure vulnerability discovery, operational ability, or refusal behavior. **[8]** Basin's internal runs used direct-answer mode at temperature 0, with no invalid responses or output-cap hits.

Evaluation	Basin
WMDP-Cyber accuracy	**65.5% (1,302 / 1,987)**
WMDP-Bio accuracy	**82.9% (1,055 / 1,273)**
WMDP-Chem accuracy	**66.4% (271 / 408)**

5 Refusal robustness

StrongREJECT scores whether responses provide useful assistance to forbidden requests; lower scores indicate less such assistance. The full set has 313 prompts. **[9]** The original Basin baseline and its replication used different judges, so they are reported separately.

Evaluation	Basin	Protocol
Unmodified baseline	**0.029**	Full 313 prompts; xhigh thinking; independent GPT-4o judge
Unmodified baseline jailbreak rate	**0.6% (2 / 313)**	Same GPT-4o-judged run
Unmodified baseline replication	**0.0192**	Full 313 prompts; xhigh thinking; GPT-4.1-mini judge; **[11]**
Wikipedia prompt transformation	**0.0032**	313-case mean; no best-of; GPT-4.1-mini judge; **[11]**

6 Intended use and limitations

Basin is suited to authorized penetration testing, source-assisted code review, vulnerability triage, proof-of-impact work in controlled environments, and verification that a proposed fix addresses the reported weakness. Researchers can use it to inspect an in-scope codebase or service, generate candidate explanations and tests, and prepare a finding for human review. These uses require permission from the system owner, a defined target and time window, and a process for handling sensitive findings. **[2]**

Basin **must not** be used to probe systems outside an agreed scope, obtain or use credentials without authorization, deploy malware, steal data, disrupt services, or conduct other unlawful activity. It should not be given unrestricted access to production systems or allowed to execute high-impact actions without human approval. A generated proof of concept is evidence to investigate, not a finding to accept automatically: reviewers should reproduce it in an appropriate environment, confirm the affected version and impact, and coordinate remediation and disclosure. **[2]**

The comparison chart combines provider-reported results from different harnesses. Benchmark performance does not establish safety on live networks or guarantee that a generated finding is correct. The StrongREJECT results do not establish robustness to every jailbreak or to long conversations. Human review, access controls, authorization, and logging are properties of the deployed system, not scores conferred by the model. **[2]**

References

1. UK National Cyber Security Centre. *Impact of AI on cyber threat from now to 2027*. 2025.
2. UK Financial Conduct Authority. *Frontier AI and cyber resilience*. 2026.
3. Wang, Z., et al. *CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale*. International Conference on Learning Representations, 2026.
4. CyberGym. *Level 1 leaderboard and evaluation notes*. Accessed 23 September 2026.
5. Anthropic. *Claude Opus 4.8 System Card*. 2026, §3.3.2.
6. Chen, M., et al. *Evaluating Large Language Models Trained on Code*. 2021.
7. Zhou, J., et al. *Instruction-Following Evaluation for Large Language Models*. 2023.
8. Li, N., et al. *The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning*. 2024.
9. Souly, A., et al. *A StrongREJECT for Empty Jailbreaks*. 2024.
10. Submersion AI. *Basin CyberGym final case score summary*. Internal evaluation record, 2026.
11. Submersion AI. *Basin StrongREJECT run manifest*. Internal evaluation record, 2026.
12. Meta AI. *Muse Spark 1.1 Evaluation Report*. 2026.