dreadnode

Where security agents run.

// THE CORE

01
Real capabilities

AI red teaming, exploits, operations, threat hunting — or distill your own. Tailored, deployed, ready in hours.

02
Security at machine speed

Deploy, improve, and scale continuously across security, safety, and trust risks — not once a quarter.

03
Compounding intelligence

Every finding and trace stays inside your walls. Each new agent builds on what the last one discovered.

// OFF THE SHELF

FORKABLE · VERSIONED · PRIVATE OR PUBLIC

CAPABILITY	DESCRIPTION	SOURCE · VERSION
dreadnode/ai-red-teaming	Probe AI apps for safety, security, and agentic risk	public · v1.1.1
dreadnode/web-security	Web application pentest agent · 43 skills · 4 tools	public · v1.0.2
dreadnode/bloodhound	Active Directory attack path enumeration	public · v1.0.0
dreadnode/agentic-actions-auditor	Audits GitHub Actions workflows for security issues	org · v1.2.0
+ 50 more	recon, exploitation, IaC scanning, model auditing, SOC triage…	

// JUST ASK

DEPLOY · BUILD · EVALUATE · OPTIMIZE · TRAIN

> "Start watching our github org for security issues."
↳ commit-monitor@0.3.1 · LIVE · 12 reports

> "Build me a scanner for vulnerable LiteLLM instances."
↳ litellm-scanner@0.1.0 · 6m 12s

> "Red team my agent before it ships."
↳ AIRT · 1,284 trials · 17.3% ASR

> "Optimize for fewer false positives."
↳ web-security@1.1.0 · score 0.87

> "Train on the evidence we just collected."
↳ sft #114 · 8 checkpoints

$ START FREE

SaaS, pay-as-you-go. Compute, inference, storage at cost.

$ ENTERPRISE

On-prem, control your data, custom capabilities, dedicated support.

→ **dreadnode.io**

// EVALUATE

Validate with benchmarks and AI red teaming

Benchmark against the **tasks you care about**. Inspect traces, compare pass rates across model families and providers, export evidence for review. Use AIRT to probe models and agents for security, safety, and trust risks — **off-the-shelf attacks** for fast starts, a rich algorithm set for the long tail.

AIRT · ASSESSMENT q2-safety-audit

ATTACKS	FINDINGS	ASR	TRIALS
27	20	78%	302

OWASP · MITRE ATLAS · NIST AI RMF · GOOGLE SAIF

EVALUATE · VERDICTS 5 OF 100

EVALUATION	VERDICT	PASS-RATE
crapi-all-qwen36plus	PASSED	92.7%
dn-eval-vulnbank-kimi	PARTIAL	36.4%
falco-sast-eval	PASSED	88.1%
vulnbank-bola-balance	FAILED	12.0%
vulnbank-weak-jwt	PASSED	74.2%

SCORE EVERY AGENT · INSPECT EVERY TRACE

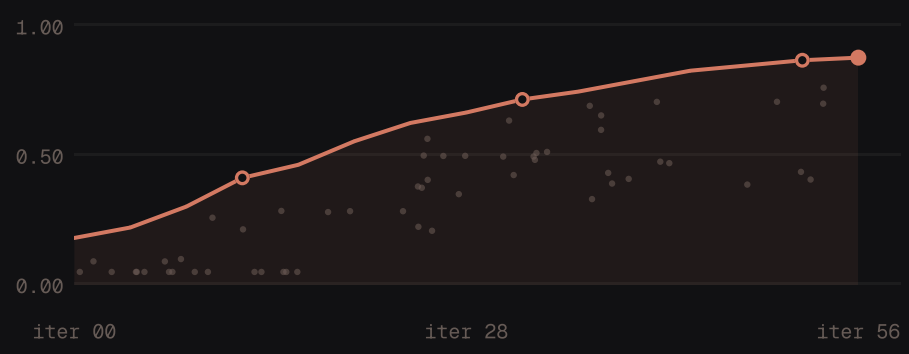
// OPTIMIZE

From guesswork to measured improvement

Hosted optimization searches the frontier of prompts, tools, and configurations — surfacing what actually moves the metric. Pair it with synthetic **worlds** for realistic environments, and feed the resulting trajectories into **training** when prompt engineering hits its ceiling.

OPTIMIZE · FRONTIER web-security@1.0.2

BEST SCORE	FRONTIER	TRAIN / VAL	METRIC CALLS
0.87	14	4 / 4	56



THE FRONTIER, EXPOSED

0.18 → 0.87 across 56 iterations on `web-security@1.0.2`. Every point a config; every step, evidence.

WORLDS — SYNTHETIC ENVIRONMENTS

Realistic AD networks, cloud infra, web apps — attack trajectories at scale, no real systems.

`ad-network` `cloud-misconfig` `webapp · vulnbank`

TRAINING — BEYOND PROMPT ENGINEERING

RL fine-tuning on eval evidence + world trajectories. Track jobs, register models to Hub.

```
$ dn train start  ↳ sft #114 · 8 ckpt
```

// DEPLOY

Always monitoring, live reporting

Same Runtime that ran the eval now hosts the live operator. Schedule fleets of autonomous agents across your services, your network, your SaaS perimeter — each goal met triggers a structured report back. Observations about the assets your team owns with trajectories available across the agent development lifecycle.

AGENTS · ON WATCH ● LIVE · 4 ACTIVE

AGENT	WATCHING	REPORTS	ELAPSED
commit-monitor	github · main repos	12 reports	02:14:08
exploit-research	cve · weekly sweep	7 reports	00:08:42
ops-support	slack · oncall queue	3 reports	04:31:55
threat-intel	actor groups · weekly digest	21 reports	18:42:10

EVERY SESSION RESUMABLE · EVERY TRAJECTORY DURABLE · EVERY REPORT EXPORTABLE