GENBOUNTY

AI safety needs a reporting workflow built for behavioural evidence

Genbounty gives AI security teams a dedicated route from probabilistic findings to structured evidence, specialist triage and accountable remediation.



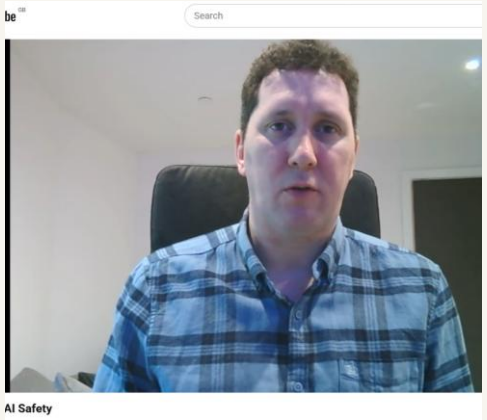
Genbounty Client Pack

The AI safety bug bounty platform, built to FLARE-AI reporting standards

sales@genbounty.com • genbounty.com

FOR AI SECURITY LEADERS

Route AI safety findings without disrupting your AppSec workflow



BOB MOREL

Co-founder • AppSec SME • AI safety advocate

KEEP IN YOUR EXISTING PROGRAMME

Conventional application security

Deterministic exploits, code-level vulnerabilities and established severity models stay in your current bug bounty and vulnerability workflow.

ROUTE TO GENBOUNTY

Behavioural harm in deployed AI

Specialist teams assess the full interaction, model version, prompts, settings, repeated trials, impact and safeguard failure.

LLMs • Agents • RAG • Prompts • MCP

Independent, continuous AI safety testing for the systems your teams deploy now.

Source: Genbounty sales deck and company materials.

THE REPORTING FAILURE

Behavioural harm breaks conventional reporting at three points

01

EVIDENCE DOES NOT FIT

Behavioural harm is probabilistic and context-dependent, so it rarely maps cleanly to conventional exploit and severity criteria.

02

REVIEW COST COMPOUNDS

Teams must review full interactions, model versions, prompts, settings and repeated trials without specialist capacity.

03

VALID REPORTS STALL

Unclear ownership and inconsistent evidence standards leave credible reports rejected, stalled or passed between teams.

AI security teams are being asked to manage credible safety reports without the triage capacity to resolve them, while using a bug bounty model and rewards designed for critical security risks.

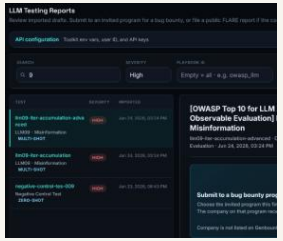
Source: Genbounty, “Why AI Safety Reporting Needs Its Own Bug Bounty Model.”

EVIDENCE OF THE GAP

Zero of 33 respondents saw an AI safety finding assessed above “low risk”

0 of 33 respondents

Researchers described submitting AI safety reports to leading frontier-model providers. None reported a finding assessed above low risk.



PROBABILISTIC EVIDENCE

Behaviour varies by prompt, model version and configuration - not a clean CVSS score.

SECURITY-LED TRIAGE

Conventional processes ask for deterministic proof and a discernible security impact.

INCENTIVES BREAK DOWN

Inconsistent decisions, weak feedback and low rewards discourage sustained testing.

Source: Genbounty study of 33 AI safety researchers; company-provided and unaudited.

THE OPERATING MODEL

Add a dedicated AI safety reporting layer - without replacing your security stack

TRADITIONAL BUG BOUNTY

Defined technical exploits

Clear boundaries, deterministic reproduction and conventional severity.

AI SECURITY BUG BOUNTY

Model-level vulnerabilities

Prompt injection, jailbreaks and flaws scoped as security issues in the AI stack.

AI SAFETY BUG BOUNTY — GENBOUNTY

Day-to-day harm in deployed AI

TWO REPORTING PATHS

Public reports on any AI product, plus scoped paid programmes.

SPECIALIST TRIAGE

Probabilistic evidence, repeated trials, impact and safeguard failure.

INCENTIVES → REMEDIATION

Meaningful rewards and a clear route to accountable action.

Source: Genbounty sales deck and company materials.

PROGRAMME COVERAGE

Configure a private program for each production feature

01 APPLICATION

User-facing AI products and workflows.

02 AGENT

Autonomous systems that take actions or use tools.

03 MCP

Connected services and Model Context Protocol servers.

04 MODEL

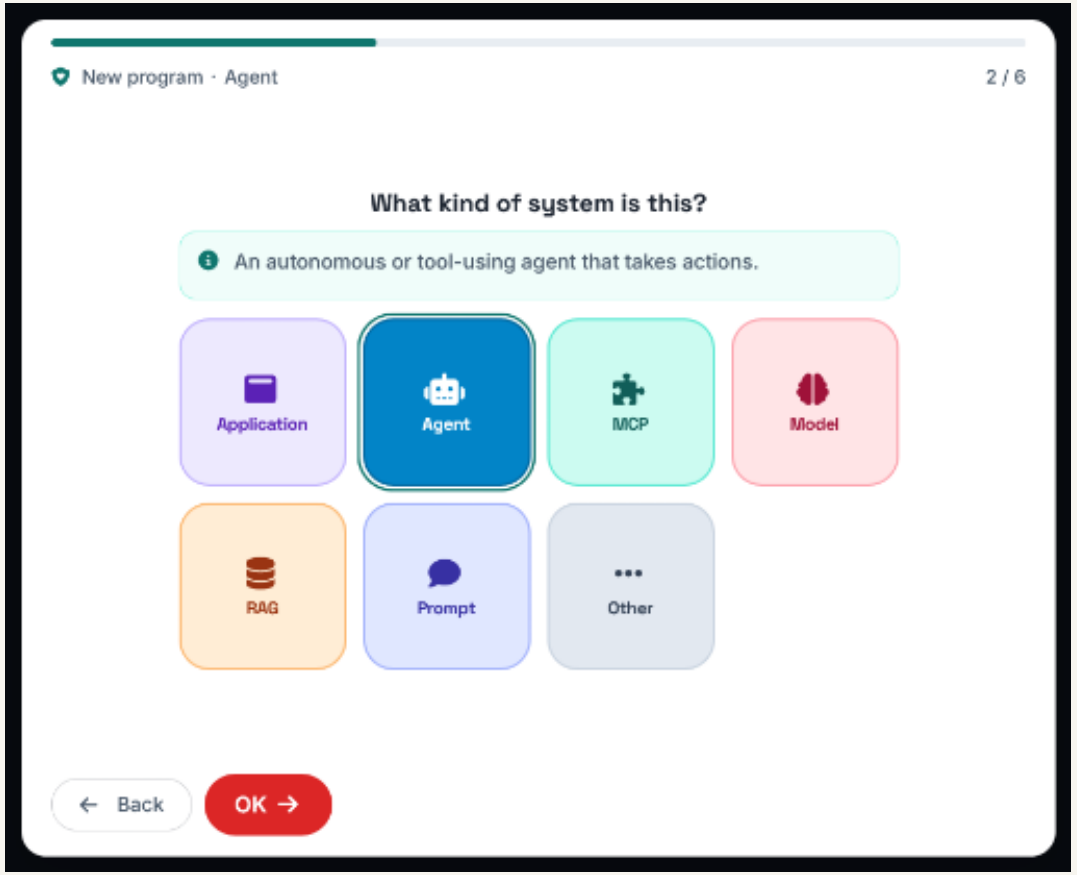
Hosted, fine-tuned and foundation model endpoints.

05 RAG

Retrieval pipelines, grounding and data access.

06 PROMPT

System prompts, guardrails and attack surfaces.



Source: Genbounty sales deck and company materials.

CONTROLLED WORKFLOW

Move from scope to remediation in four controlled steps

01

SCOPE

Define one feature, risk categories, exclusions and severity rules.

02

INVITE EXPERTS

Vetted teams, controlled access, clear rules and incentives.

03

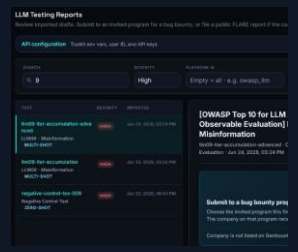
CAPTURE EVIDENCE

Full interactions, configuration, repeated trials, frequency and impact.

04

TRIAGE & RESOLVE

Company documents, tests, accepts, remediates. Genbounty issues micropayments from the program budget.



ONE EVIDENCE TRAIL

From probabilistic evidence to accountable remediation

Source: Genbounty sales deck and company materials.

BUILT ON AN OPEN STANDARD

One FLARE-AI record makes every finding portable and auditable

CLASSIFICATION FIRST

Three questions classify each report as a hazard, vulnerability or real-world incident.

COMMON TAXONOMIES

Fields map to OWASP LLM Top 10, AIAAIC harm categories, severity and disclosure intent.

ENCRYPTED EVIDENCE

Reproduction steps and attachments are encrypted in the browser and unsealed only with a key.

PORTABLE EXPORTS

Records travel as JSON-LD or JSON, with CSAF, VEX, AVID and CVE exports.



Source: FLARE-AI paper and Genbounty company materials.

COMMERCIAL STARTING POINT

Start with one deployed AI feature for £250

£250

pilot

A focused way to establish a defensible evidence baseline before expanding coverage.

4 weeks	**Private pilot**	Test one deployed feature with expert researchers and specialist triage.
One programme	**Operationalise**	Embed scope, evidence, incentives and remediation standards.
Portfolio-wide	**Expand**	Extend the model across applications, agents, RAG, MCP, models and prompts.

READY TO SCOPE A PILOT?

sales@genbounty.com • genbounty.com

Source: Genbounty sales deck. Pilot pricing and duration are company-provided.