bugcrowd

DATA SHEET

AI Bias Assessment

Unleash human ingenuity to find bias in a LLM beyond the reach of traditional testing

Challenges

Enterprises adopting, or wanting to adopt, Large Language Model (LLM) applications seek confidence that this revolutionary, but very new, technology can be onboarded safely and productively. In government, emerging requirements like those described in OMB Memorandum M-24-10 make that a mandate.

LLM applications run on algorithmic models trained on data. Even when that training data is curated by humans (which it often is not), the application can easily reflect "data bias" caused by stereotypes, misrepresentations, prejudices, derogatory or exclusionary language, and a range of other possible biases from the training data, leading the model to behave in unintended and potentially harmful ways.

This potential vulnerability can add considerable risk and unpredictability to LLM adoption.

Solution

A Bugcrowd AI Bias Assessment is a private, reward-for-results engagement on the Bugcrowd Platform that activates trusted, 3rd-party experts in prompt engineering and AI safety to identify and prioritize data bias flaws in LLM applications. Participants are paid based on successful demonstration of impact, with more impactful findings earning higher payments.

For over a decade, Bugcrowd's industry-first, "skills-as-a-service" approach to risk reduction has been proven to uncover more high-impact vulnerabilities than traditional testing, while offering clearer line of sight to ROI.

Key benefits

- Detects numerous types of data bias in LLM apps that traditional testing will miss
- Streamlines and accelerates LLM adoption
- For government agencies... Contributes to compliance with AI Safety requirements for red teaming and bias detection
- Builds relationships with experts in prompt engineering and AI safety for future engagements

Key Capabilities

Runs as a private, time-bound engagement with duration, scope, severity scoring, and competition-style rewards structure determined by you with guidance from Bugcrowd

Curates a trusted team of experts with specialized tools and skills in bias-specific prompt engineering

Processes, validates, prioritizes, and rewards findings based on predetermined scoring

Integrates with your existing DevSec processes for rapid remediation

Equally effective on implementations of open source (LLaMA, Bloom, etc.) and private models

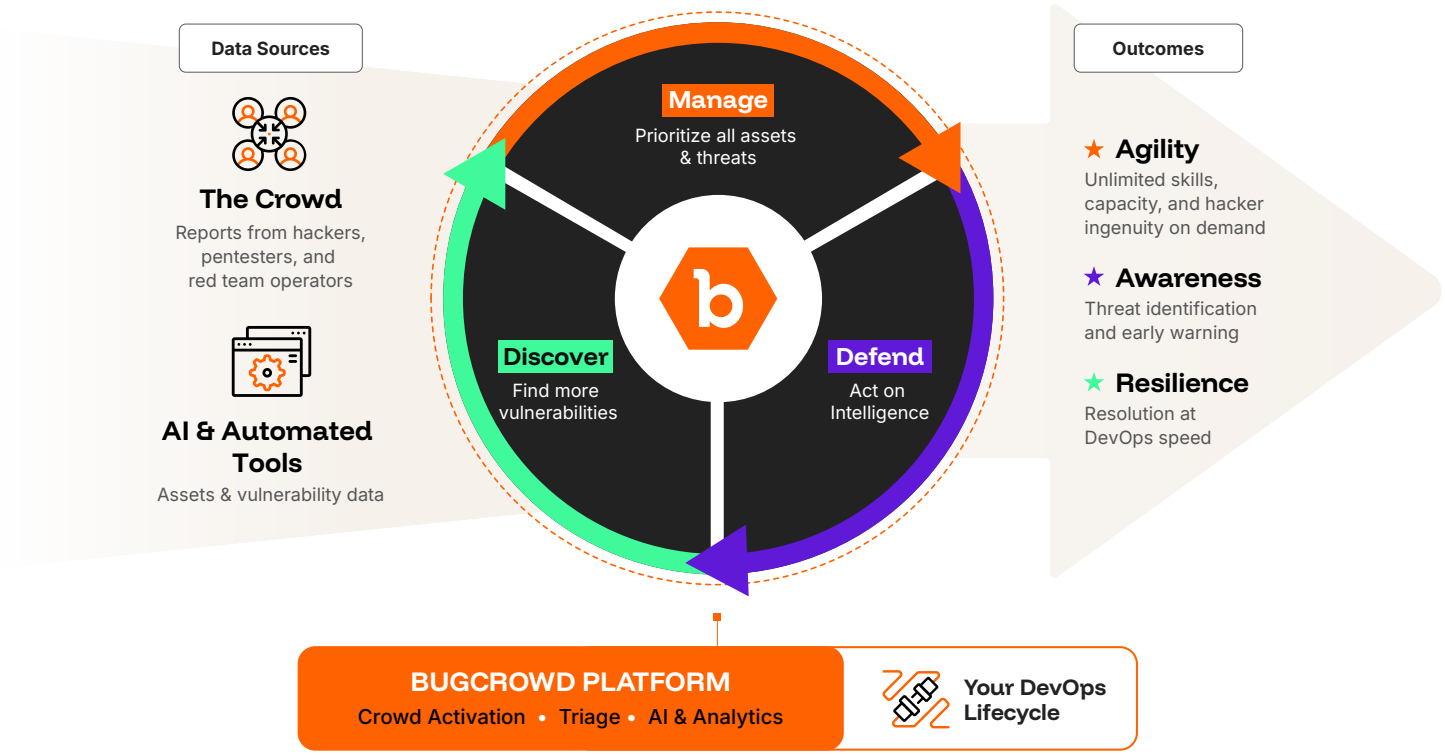
Includes managed compliant payments, researcher communications, and reporting

TYPES OF DATA BIAS

Data bias can take numerous forms and have varying types of impact; quite often, its symptoms are missed until after the LLM application is deployed to production. Examples of the types of bias that can be uncovered by a Bugcrowd AI Bias Assessment include:

REPRESENTATION BIAS	Disproportionate representation or omission of certain groups in the training data
PRE-EXISTING BIAS	Biases stemming from historical or societal prejudices present in the training data
ALGORITHMIC BIAS *PROCESSING*	Biases introduced through the processing and interpretation of data by AI algorithms
ALGORITHMIC BIAS *AGGREGATION*	Incorrect assumptions made when aggregating data
GENERAL SKEWING	When the need for fairer representation is not recognized explicitly, certain groups may be inadvertently left out, leading to skewed outputs

Bugcrowd Platform



The Bugcrowd Platform fuses AI with real-time, crowdsourced intelligence from the world's top ethical hackers, pentesters, and red teamers (aka The Crowd), as well as from automated tools that generate asset, threat, and vulnerability data. The powerful combination of human creativity and automation empowers you to continuously:

Agility

Augment Your Team On Demand

- Attacker mindset on tap for vulnerability discovery, pen testing, and red teaming
- 350+ skill sets and certifications available
- Crowd curation and activation guided by data and AI

Awareness

See and Prioritize Emerging Threats

- Continuous vulnerability intake, validation, and triage at scale
- 24/7 triage coverage with same-day response for P1s
- Early warning of emerging vulnerabilities

Resilience

Continuously Improve Security Posture

- Actionable reporting, benchmarking, and recommendations
- Directly integrates with existing tools for change at DevOps speed
- Deep bench of solution & support specialists at your side for quick wins and long-term ROI

mastercard indeed credit karma Pinterest NASA ATLASSIAN