bugcrowd

DATA SHEET

AI SAFETY AND SECURITY

AI Bias Assessment

Unleash human ingenuity to find bias in LLM apps 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 Executive Order 14410 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 security researchers (aka a "crowd") 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 AI specialists in the hacker/security researcher community

KEY CAPABILITIES

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

Curates a trusted team of 3rd-party security researchers with specialized tools and skills in bias-specific prompt engineering

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

Equally effective on implementations of open source (LLaMA, Bloom, etc.) and private models, and on trained and pre-trained (foundation) ones

Includes managed compliant payments, researcher communications, and reporting

Integrates with your existing DevSec processes for rapid remediation

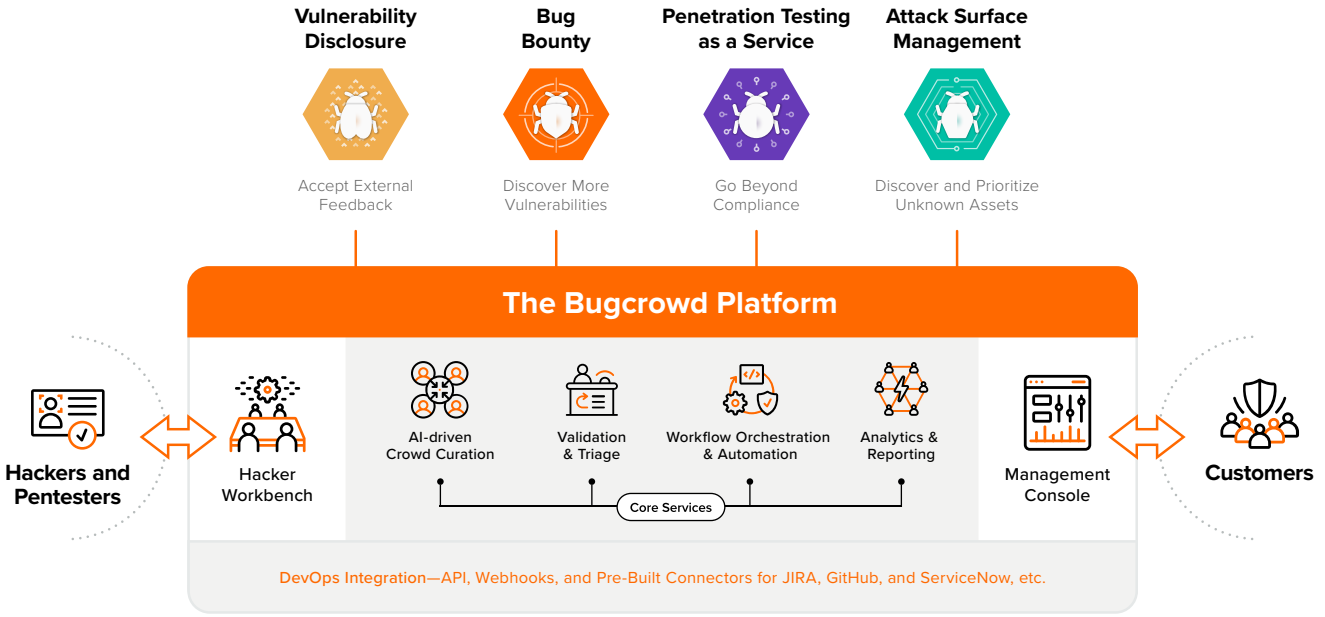
TYPES OF DATA BIAS

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

Why Bugcrowd

The Bugcrowd Platform helps customers defend themselves against cybersecurity attacks by connecting with trusted, skilled hackers to take back control of the attack surface. Our AI-powered platform for crowdsourced security is built on the industry's richest repository of data about vulnerabilities and hacker skill sets, activating the ideal hacker talent needed on demand, and bringing scalability and adaptability to address current and emerging threats.



BEST SECURITY ROI FROM THE CROWD

We match you with trusted security researchers who are perfect for your needs and environment across hundreds of dimensions using machine learning.

INSTANT FOCUS ON CRITICAL ISSUES

Working as an extension of the platform, our global security engineering team rapidly validates and triages submissions, with P1s often handled within hours.

CONTEXTUAL INTELLIGENCE FOR BEST RESULTS

We apply over a decade of knowledge accumulated from experience devising thousands of customer solutions to achieve your goals for better outcomes.

CONTINUOUS, RESILIENT SECURITY FOR DEVOPS

The platform integrates workflows with your existing tools and processes to ensure that apps and APIs are continuously tested before they ship.

TESLA indeed credit karma Pinterest mastercard ATLASSIAN