Bugcrowd Launches AI Bias Assessment Offering for LLM Applications

First solution in Bugcrowd's AI Safety and Security portfolio unleashes human ingenuity to find data bias beyond the reach of traditional testing

SAN FRANCISCO, CALIFORNIA, UNITED STATES, April 16, 2024 /EINPresswire.com/ -- Bugcrowd, the leader in crowdsourced security, today announced the availability of AI Bias Assessments as part of its AI Safety and Security Solutions portfolio on the Bugcrowd Platform. AI Bias Assessment taps the power of the crowd to help enterprises and government agencies adopt Large Language Model (LLM) applications safely, efficiently, and confidently.

> Bugcrowd's work with customers like the US DoD's Chief Digital and Artificial Intelligence Office (CDAO), along with our partner ConductorAI, has become a crucial proving ground for AI detection"
>
> *Dave Gerry, CEO, Bugcrowd*

LLM applications run on algorithmic models that are trained on huge sets of 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, prejudices, exclusionary language, and a range of other possible biases from the training data. Such biases can lead the model to behave in potentially unintended and harmful ways, adding considerable risk and unpredictability to LLM adoption.

Some examples of potential flaws 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), and Algorithmic Processing Bias (biases introduced through the processing and interpretation of data by AI algorithms).

The public sector is urgently affected by this growing risk. As of March 2024, the US Government (1) mandated its agencies to conform with AI safety guidelines – including the detection of data bias. That mandate extends to Federal contractors later in 2024.

This problem requires a new approach to security because traditional security scanners and penetration tests are unable to detect such bias. Bugcrowd AI Bias Assessments are private,

reward-for-results engagements on the Bugcrowd Platform that activate trusted, third-party security researchers (aka a “crowd”) to identify and prioritize data bias flaws in LLM applications. Participants are paid based on the successful demonstration of impact, with more impactful findings earning higher payments.


David Gerry CEO, Bugcrowd

The Bugcrowd Platform’s industry-first, AI-driven approach to researcher sourcing and activation, known as CrowdMatchTM, allows it to build and optimize crowds with virtually any skill set, to meet virtually any risk reduction goal, including security testing and beyond.

“Bugcrowd’s work with customers like the US DoD’s Chief Digital and Artificial Intelligence Office (CDAO), along with our partner ConductorAI, has become a crucial proving ground for AI detection by unleashing the crowd for identifying data bias flaws,” said Dave Gerry, CEO of Bugcrowd. “We’re eager to share the lessons we’ve learned with other customers facing similar challenges.”

"ConductorAI's partnership with Bugcrowd for the AI Bias Assessment program has been highly successful. By leveraging ConductorAI's AI audit expertise and Bugcrowd's crowdsourced security platform, we led the first public adversarial testing of LLM systems for bias on behalf of the DoD. This collaboration has set a solid foundation for future bias bounties, showcasing our steadfast commitment to ethical AI," said Zach Long, Founder, ConductorAI.

For over a decade, Bugcrowd's unique "skills-as-a-service" approach to security has consistently uncovered more high-impact vulnerabilities than traditional methods.

Our customer base, which numbers nearly 1,000, has benefited from this approach, which also provides a clearer line of sight to ROI. With unmatched flexibility and access to a decade of vulnerability intelligence data, the Bugcrowd Platform has evolved over time to reflect the changing nature of the attack surface – including the adoption of mobile infra, hybrid work, APIs, crypto, cloud workloads, and now AI. In 2023 alone, customers found almost 23,000 high-impact vulnerabilities using the Bugcrowd Platform, helping to prevent potential breach-related costs of up to $100 billion.

“As the leading crowdsourced security platform provider, Bugcrowd is uniquely positioned to meet the new and evolving challenges of AI Bias Assessment, just as we’ve met the emergent security challenges of previous technology waves such as mobile, automotive, cloud computing,

crypto, and APIs," said Casey Ellis, Founder and Chief Strategy Officer of Bugcrowd.

To learn more about the Bugcrowd AI Bias Assessment offering, visit https://www.bugcrowd.com/products/ai-bias-assessment/

Bugcrowd at Black Hat Asia Conference, April 17-19, 2024
- Visit us at the Business Hall on the Black Hat Expo floor for swag, demos, and conversation about the news.
- Request 1:1 time with execs for a deep dive into our announcement and the value of the Bugcrowd Security Knowledge Platform.
- Register to our events https://ww1.bugcrowd.com/black-hat-asia-2024/

To learn how the Bugcrowd Platform can equip your organization to protect itself from cyber risk, visit Bugcrowd.com or download Inside the Platform: Bugcrowd's Vulnerability Trends Report. https://www.bugcrowd.com/resources/report/bugcrowds-vulnerability-trends-report/

About Bugcrowd
We are Bugcrowd. Since 2012, we've been empowering organizations to take back control and stay ahead of threat actors by uniting the collective ingenuity and expertise of our customers and trusted alliance of elite hackers, with our patented data and AI-powered Security Knowledge Platform™. Our network of hackers brings diverse expertise to uncover hidden weaknesses, adapting swiftly to evolving threats, even against zero-day exploits. With unmatched scalability and adaptability, our data and AI-driven CrowdMatch™ technology in our platform finds the perfect talent for your unique fight. We are creating a new era of modern crowdsourced security that outpaces threat actors.

Unleash the ingenuity of the hacker community with Bugcrowd, visit www.bugcrowd.com. Read our blog https://www.bugcrowd.com/blog/

*Based on Bugcrowd Platform data and IBM cost of data breach report https://www.ibm.com/reports/data-breach

1. https://www.whitehouse.gov/briefing-room/statements-releases/2024/03/28/fact-sheet-vice-president-harris-announces-omb-policy-to-advance-governance-innovation-and-risk-management-in-federal-agencies-use-of-artificial-intelligence/

Contact
Zonic Group International
EMEA

Krison Thakkar
Email: kthakkar@zonicgroup.com
Tel: +1 408 458 865

APAC

Bruce Reid
Zonic Group
+61 499 888 466
email us here
