Microsoft Corp

Microsoft to rank 'safety' of AI models sold to cloud customers

Tech group's corporate clients use its 'leaderboard' to assess rival offerings by likes of OpenAI and Elon Musk's xAI

Microsoft ranks AI models by three metrics — quality, cost and throughput, which is how quickly a model can generate an output © Reuters

Rafe Uddin in Seattle and **Cristina Criddle** in San Francisco

Published JUNE 7 2025

Microsoft will start ranking artificial intelligence models based on their safety performance, as the software group seeks to build trust with cloud customers as it sells them AI offerings from the likes of OpenAI and Elon Musk's xAI.

Sarah Bird, Microsoft's head of Responsible AI, said the company would soon add a "safety" category to its "model leaderboard", a feature it launched for developers this month to rank iterations from a range of providers including China's DeepSeek and France's Mistral.

The leaderboard, which is accessible by tens of thousands of clients using the Azure Foundry developer platform, is expected to influence which AI models and applications are purchased through Microsoft.

Microsoft currently ranks three metrics: quality, cost and throughput, which is how quickly a model can generate an output. Bird told the Financial Times that the new safety ranking would ensure "people can just directly shop and understand" AI models' capabilities as they decide which to purchase.

The decision to include safety benchmarks comes as Microsoft's customers grapple with the potential risks posed by new AI models to data and privacy protections, particularly when deployed as autonomous "agents" that can work without human supervision.

Microsoft's new safety metric will be based on its own ToxiGen benchmark, which measures implicit hate speech, and the Center for AI Safety's Weapons of Mass Destruction Proxy benchmark. The latter assesses whether a model can be used for malicious purposes such as building a biochemical weapon.

Rankings enable users to have access to objective metrics when selecting from a catalogue of more than 1,900 AI models, so that they can make an informed choice of which to use.

"Safety leader boards can help businesses cut through the noise and narrow down options," said Cassie Kozyrkov, a consultant and former chief decision scientist at Google. "The real challenge is understanding the trade-offs: higher performance at what cost? Lower cost at what risk?"

Alongside Amazon and Google, the Seattle-based group is considered one of the largest "hyperscalers" that together dominate the cloud market.

Microsoft is also positioning itself as an agnostic platform for generative AI, signing deals to sell models by xAI and Anthropic, rivals to start-up OpenAI which it has backed with roughly $14bn in investment.

Last month, Microsoft said it would begin offering xAI's Grok family of models under the same commercial terms as OpenAI.

The move came despite a version of Grok raising alarm when an "unauthorised modification" of its code led to it repeatedly referencing "white genocide" in South Africa when responding to queries on social media site X. xAI said it introduced a new monitoring policy to avoid future incidents.

"The models come in a platform, there is a degree of internal review, and then it's up to the customer to use benchmarks to figure it out," Bird said.

There is no global standard for AI safety testing, but the EU's AI Act will enter force later this year and compel companies to conduct safety tests.

Some model builders including OpenAI are dedicating less time and money to identify and mitigate risks, the FT previously reported citing several people familiar with the start-up's safety processes. The start-up said it had identified efficiencies without compromising safety.

Bird declined to comment on OpenAI's safety testing, but said it was impossible to ship a high quality model without investing a "huge amount" in evaluation and that processes were being automated.

Microsoft in April also launched an "AI red teaming agent" that automates the process of stress testing computer programmes by launching attacks to identify vulnerabilities. "You just specify the risk, you specify the attack difficulty . . . And then it's off attacking your system," Bird said.

There are concerns that without adequate supervision AI agents could take unauthorised actions opening the owners up to liabilities.

"The risk is that leader boards can lull decision makers into a false sense of security," said Kozyrkov. "Safety metrics are a starting point, not a green light."

Follow the topics in this article

Technology sector

Artificial intelligence

Microsoft Corp

Rafe Uddin

Cristina Criddle