Instagram Uses AI to Warn Users Who Are About to Post Potentially Offensive Captions

|
3 hours ago
Instagram rolled out a feature in July in which it used artificial intelligence to detect comments that might be considered offensive, warning the people who are posting those comments and giving them a chance to reconsider. That feature was extended to captions Monday.

The Facebook-owned photo- and video-sharing network said in a blog post that when its AI detects a caption on a feed post to be potentially offensive, the user about to post that caption will see a prompt warning them that it is similar to others that were reported for bullying.

The user will then be given the opportunity to edit his or her caption before posting it.

The feature began rolling out in select countries Monday, with plans to expand it globally in the coming months.

Instagram wrote in its blog post, “In addition to limiting the reach of bullying, this warning helps educate people on what we don’t allow on Instagram, and when an account may be at risk of breaking our rules.

https://www.adweek.com/digital/instagram-uses-ai-to-warn-users-who-are-about-to-post-potentially-offensive-captions/