Jul - Sep 2025

Earlier this year, we shared that TikTok has been making ongoing improvements to our safety technologies and enhancements to our content moderation processes. This quarter's report sees those investments continue to bear out, as we share our:

- Highest-ever rate of violative content removed by automated technologies: 91%
- Highest-ever rate of proactive removal rate (content removed before it's reported to us): 99.3%
- Highest-ever volume of content removed in under 24 hours: 94.8%

These advances are helping us get violative content down faster, reducing the likelihood of our community seeing it. In addition, as more of these tasks get handled by technology, our safety teams can focus more of their time on the work that most benefits from human expertise—like handling appeals, consulting external experts, and responding to fast-moving events.

As we continually strengthen our trust and safety strategies, there may be variations in the data. This quarter, for example, we refined our automated models for reviewing comments in line with our updated Community Guidelines. This change makes it easier to take a broader range of actions against violative or inappropriate comments—from restricting visibility to removal—so that we can protect the community while supporting self-expression. As a result, we have removed fewer comments this quarter.

Finally, this report features an updated "Total removals and restorations" chart that includes the two new LIVE demonetization metrics that we introduced last quarter.

Apr - Jun 2025

Through ongoing advancements in our moderation technologies, we continue to see improvements in how we proactively detect violative content and restrict it at scale. In the second quarter of 2025, more than 99% of the violative content we removed was taken down before someone reported it to us, and more than 90% was removed before it got a single view. Thanks to these improvements, our teams are exposed to less content. For example, over the last year we've seen a 76% decrease in the volume of shocking and graphic content seen by moderators.

We are also reporting a significant increase in the number of LIVE streams we suspended since the first quarter of the year, thanks to improved moderation technologies. As we have strengthened our enforcement strategies, we have seen only a small increase in the number of

LIVEs restored following appeal, suggesting that these improvements are accurately targeting violative LIVEs.

We're expanding the LIVE safety data we report to reflect our actions to enforce our monetization guidelines. These guidelines help reward creators who stream safe, authentic, and high-quality content, while preventing streams that go against these guidelines. In the second quarter of 2025, we took action, including warnings and demonetization, on 2,321,813 LIVE sessions and 1,040,356 LIVE creators for violating our LIVE monetization guidelines. Warnings help educate creators if their content may violate our guidelines so they can make a change. If no changes are made, we may then suspend a creator's access to earn through monetization features on LIVE. As part of our ongoing efforts to bring greater transparency to how we protect the TikTok LIVE experience, we will start to include this data in future reports.

Jan - Mar 2025

We protect TikTok by investing in a combination of advanced moderation technologies and thousands of trust and safety professionals who work together to enforce our policies. Through ongoing enhancements to our safety technologies, we take a proactive approach to detecting violative content and restricting it at scale.

In the first three months of this year, over 99% of violating content was removed before someone reported it to us, and over 90% was removed before gaining any views. The vast majority of violations (94%) were removed within 24 hours. This was also a quarter where automated moderation technologies removed more violative content than ever—over 87% of all video removals. In addition, TikTok's moderation technologies helped identify violative livestreamed content faster and more consistently. Over 19M LIVEs were stopped this quarter, a 50% increase from the previous quarter. At the same time, we continued to prioritize moderation accuracy, with the number of appeals remaining steady amid the increase in automated moderation.

We also began testing large language models (LLMs) to further support proactive moderation at scale. LLMs can comprehend human language and perform highly specific, complex tasks. This can make it possible to moderate content with high degrees of precision, consistency, and speed. We've started to pilot LLMs to help enforce our rules for comments on TikTok. In preliminary testing, LLM-driven comment moderation has been shown to help improve both the detection of violative comments and the accuracy of removals.

Ongoing advancements in AI and other moderation technologies can also benefit the overall well-being of content moderators by requiring them to review less content. It also provides moderators with better tools to do this critical moderation work. For instance, over the course of 2024, the

number of videos that moderators removed for Shocking and Graphic policy violations decreased by 60% as our moderation technologies took on more of these potentially distressing videos. Meanwhile, technologies like AI helped make it easier to moderate nuanced areas like misinformation by extracting specific misinformation "claims" from videos for moderators to assess directly or route to our fact-checking partners.

We continue to invest in a range of safety technologies to strengthen moderation accuracy, support our safety team and their well-being, while improving the scale and speed of our trust and safety efforts. Learn more about our safety teams, strategies, and technologies here.

Oct - Dec 2024

The last quarter of 2024 capped off what was one of the biggest electoral years in history, with a record number of elections taking place in countries around the world. We drew on our experience protecting our platform through more than 150 elections over the previous four years and invested significantly in trust & safety and security measures, so that TikTok remained a place for discovery, creativity, and entertainment.

From October-December 2024, we introduced new and updated initiatives to connect people to reliable election information. For example, we worked with local electoral commissions, election experts, and fact-checkers to bring authoritative information to our in-app Election Centers in 16 countries. We also expanded our fact-checking program to six new markets and added two new IFCN-accredited fact-checking partners. We published blog posts outlining our election integrity work across Germany, Croatia, Romania, and Japan, and regularly updated our U.S. Elections Integrity Hub to bring ongoing transparency to our actions.

We continued providing more insight into the ways we counter attempts to manipulate our platform's integrity. For instance, we published extensive information about how we counter deceptive behaviors like fake engagement and covert influence operations. We also reported 19 covert influence operations we disrupted from October-December as well as over 30,000 accounts we removed that were associated with previously disrupted networks which had attempted to return to our platform. As a continuation of our transparency efforts, this report now includes details about the comments we remove from fake accounts. In the fourth quarter of 2024, we removed 157,158,547 comments from accounts that violated our policies on fake content and engagement.

We continue to invest in enhancing our detection and moderation technologies, so that we can catch violative content quickly and accurately. Automated removals have increased by 10% compared to the same period last year, which helps us detect and remove more content

proactively. This quarter, over 98% of content was removed before a user reported it to us — a 2% increase from the year before.

Jul - Sep 2024

This report reflects our updated Community Guidelines that took effect in May and provide our community with more transparency about our rules and how we enforce them. After consulting with dozens of organizations globally, we overhauled how we organize our policies, simplified the language we use, and added granularity to help everyone, from creators to researchers, easily access the information they need. We have refreshed the data visualizations in this report to reflect these updates.

We update our Community Guidelines on an ongoing basis. There are different types of updates that our community might see. For example, we might make changes that are more minor, as in our May update when we revised our policy titles and replaced any ampersands (&) with the word “and.” Sometimes our changes are more substantive, reflecting emerging issues or trends in online behavior. This most recent update included specific language about our AI-generated content policies, consistent with growing interest in AI and AI tools across the internet.

In addition to revising our Community Guidelines, TikTok evaluates and updates our enforcement tools on a regular basis. One of TikTok's eight Community Principles is to apply our content polices consistently, equitably, and transparently. In order to do that, we regularly assess the performance of our automated moderation technology with the goal of strengthening accuracy and speed, and enabling a wide-range of creativity and expression on our platform. As we detailed in our previous report, these investments are paying off, with clear improvements in metrics like our proactive removal rate and number of videos restored. We remain committed to strengthening our policies and their enforcement, and doing so transparently through publications like the Community Guidelines Enforcement Report.

Apr - Jun 2024

TikTok has a long-standing commitment to transparency, and we’re proud of our continuing work to ensure that this report provides updated and helpful insights into our Community Guidelines enforcement. For instance, in our last report, we introduced details about video comment removals, and we’re now expanding this report with new charts that illustrate comment removal data.

With over a billion people and millions of pieces of content posted to our platform every day, we continue to prioritize and enhance TikTok's automated moderation technology as such technology enables faster and consistent removal of content that violates our rules. We invest in technologies that improve content understanding and predict potential risks so that we can take action on violative content before it's viewed. These technical investments also reduce the volume of content that moderators review, helping minimize human exposure to violative content. As a result, automated technology now removes 80% of violative videos, up from 62% a year ago. At the same time, our proactive detection rate has continued to improve; this quarter saw our highest result so far at 98.2% of videos removed before a user report. And, the proportion of total videos restored after removal has decreased by more than half from a year ago.

Our priority is keeping our global community safe and secure, and we continue to invest in robust measures that actively enforce and defend the security of our platform. To provide greater insight into the work our security professionals around the globe are doing to uphold the safety, integrity, and reliability of TikTok, we have launched a new, stand-alone quarterly report detailing our efforts.

Jan - Mar 2024

As we continue to bring transparency to how we operate and protect the platform, this report now includes details about the video comments we remove for violating our Community Guidelines. In the first quarter of 2024, we removed **976,479,946** video comments for violating our Community Guidelines. This represents **1.6%** of total video comments published during that period. Creators also used our extensive comment safety tools to filter or remove **3,381,646,722** comments from their videos during that period. In future reports, this data will be incorporated into charts below.

We're also launching a new, stand-alone report on our efforts to combat online child sexual abuse and exploitation (CSEA) and the sharing of child sexual abuse material (CSAM), as we make our platform inhospitable for those that seek to cause harm to children. The report provides further insights into our work to identify, remove and report instances of CSAM to the National Center for Missing and Exploited Children.

Protecting platform integrity during a historical year of elections

With more than 2 billion people around the world heading to the polls in this historical year of elections, we continue our efforts to protect the integrity of our platform in order to maintain a creative, safe and authentic environment for our community.

We have increased transparency about our approach to election-related content with new and updated resources on how we protect the integrity of elections, combat harmful misinformation,

and counter influence operations. We have also published information about the ongoing steps we're taking to protect our community through the US elections in November of this year. We will continue to take voluntary and unprecedented steps to build a safe and secure environment for US users that's free from outside manipulation.

We have also launched a new dedicated Transparency Report for more frequent and granular reporting about the covert influence operations we have disrupted. This will replace the quarterly disclosures made here. The first dedicated report, published last month, covers operations disrupted between January and May 2024.

Oct - Dec 2023

As our platform continues to grow, we remain focused on implementing robust safety strategies. We continue to respond vigorously to the evolving ways that people attempt to circumvent our rules. In addition, we are investing in proactively and expediently removing violative content and accounts. As part of our commitment to transparency, we're now providing more detail on how often videos have been viewed before being removed.

Israel-Hamas war

Since the start of the Israel-Hamas war on October 7, 2023, we provided regular updates on our extensive actions to safeguard our community, including mobilizing significant resources and personnel, introducing product features, and working with experts.

To stay accountable to the communities we serve, we have also been transparent about the efficacy of our efforts through regular disclosures on content removals. From the start of the war through to the end of 2023, we removed more than 1.5 million videos and suspended more than 46,000 livestreams in Israel and Palestine for violating our Community Guidelines, including content promoting Hamas, hate speech, terrorism and misinformation.

We also remain vigilant against deceptive behaviors. From October 7 through to the end of 2023, we removed more than 169 million fake accounts globally, and we have removed about 1.2 million bot comments on content tagged with hashtags related to the war.

Jul - Sep 2023

Building on our commitment from the last report, we are publishing our second quarterly update on TikTok LIVE enforcement. TikTok LIVE continues to be an important part of the TikTok

experience, fostering real-time entertainment between creators and their communities. Despite the increase in suspended livestream sessions this quarter, the proportion of total livestreams suspended remained consistent at 1.5%, indicating that our enforcement measures are scaling with the platform's growth.

Apr - Jun 2023

The information shared in this report reflects to our updated Community Guidelines, with Youth Safety and Well-Being remaining a top priority for us. We are deeply committed to ensuring that TikTok is a safe and positive experience for people under the age of 18.

As part of this commitment, our policies strictly prohibit content that may jeopardize the well-being of young people, including the risk of exploitation or potential psychological, physical, or developmental harm. We prioritize youth safety and well-being throughout our policy development process, which includes the prohibition of child sexual abuse material (CSAM), youth abuse, bullying, dangerous activities and challenges, exposure to overly mature themes, and the consumption of alcohol, tobacco, drugs, or regulated substances. In cases where we identify instances of youth exploitation on our platform, we will take immediate action, including the banning of the offending account and any associated accounts.

We're also adding enforcement information for TikTok LIVE, which allows users and creators to interact in real-time. In the second quarter of 2023, we suspended 8,074,632 LIVE sessions. This represents 1.5% of total LIVE sessions during that period. Of this number, 252,045 suspensions were successfully appealed. As an important part of the TikTok experience, we are excited to share more about our LIVE moderation efforts and we will continue to expand on this in future reports.

Jan - Mar 2023

In this report, we're sharing new insights as we continue to bring transparency to how we operate and protect the platform. First, we're providing insight into the language capabilities of TikTok's moderators whose work not only contributes to a safer platform but also plays a vital role in enhancing our automated moderation systems. Second, we're providing data on our response time to remove violative content reported by our community using our in-app and online reporting tools.

During this report period, we updated TikTok's Community Guidelines. These guidelines went into effect in April 2023 and will be reflected in our next report.

Oct - Dec 2022

We are focused on working to create a safe, positive and joyful experience for our community. We continue to innovate in our approach to keeping people safe, which includes removing content, making some content ineligible for recommendation into For You feeds, introducing our Content Levels system, and updating our current account enforcement system to help ensure that each person's For You feed is full of content that is appropriate and entertaining for them.

To help us achieve this at scale, we invest in technology to automatically detect and remove violative content, which helps reduce the volume content moderators review and enables more people to focus on nuanced and contextual areas, such as misinformation, hate speech, and harassment. When it comes to enforcement, our approach aims to:

- Minimise views of content which violates our Community Guidelines
- Prioritise expeditious removal of especially egregious content, such as child sexual abuse material (CSAM) and violent extremism
- Ensure consistency and fairness in our enforcement to reduce incorrect removals

To better achieve this vision, over the last year we've continued to invest in machine learning as well as an upgrade of the architecture that looks at video, text, and audio information along with the possible severity of the violation and the potential reach of the video (e.g. based on following) in order to determine the best course of action. For instance, content may be removed automatically, made ineligible for recommendation into For You feeds, or directed to a moderator for further review. We've also introduced a new enforcement system that is more transparent for creators. We believe these improvements are helping us build a greater experience for creators while continuing to advance our commitment to safety for the overall platform. Taken together, these efforts have helped improve accuracy and efficiency in our enforcement, which is reflected in the data in this report.

Jul - Sep 2022

This report has been expanded to bring additional transparency to our actions, progress, and challenges, and to stay accountable to our community. In an effort to provide useful and relevant information, we now provide additional insight on video removals across 50 markets, which represents approximately 90% of overall removal volume.

In addition, this report now includes information about the covert influence operations we identified and removed from our platform from July-September. When we investigate and remove these operations, we focus on behaviour and assessing linkages between accounts and

techniques to determine if actors are engaging in a coordinated effort to mislead TikTok's systems or our community. In each case, we believe that the people behind these activities coordinate with one another to misrepresent who they are and what they are doing. We know that covert influence operations will continue to evolve in response to our detection and networks may attempt to reestablish a presence on our platform. For example, in addition to the networks disclosed in this report, 365 additional accounts were identified as belonging to previously removed networks and were subsequently removed. As new deceptive behaviours emerge, we'll continue to evolve our response, strengthen enforcement capabilities, and publish our findings.

Apr - Jun 2022

TikTok uses a combination of people and technology to enforce our Community Guidelines and keep our community safe and welcoming. To do this effectively at scale, we continue to invest in technology-based flagging and moderation. We rely on automated moderation when our systems have a high degree of confidence that content is violative so that we can expeditiously remove violations of our policies. As a result, our overall protective detection efforts have improved.

Leveraging machine learning has been especially impactful when it comes to our countering harmful misinformation. We expanded our capacity to iterate rapidly on our systems given the fast changing nature of misinformation, especially during a crisis or event (e.g. the war in Ukraine or an election). We also improved our ability to detect known misleading audio and imagery to reduce manipulated content on the platform. From these investments, this quarter we're seeing gains in the right direction when it comes to enforcing our integrity and authenticity policies: proactive removal of videos improved from 83.6% in Q1 to 89.1% in Q2; removal of videos at zero views improved from 60.8% in Q1 to 74.7% in Q2; and video removals in under 24 hours improved from 71.9% to 83.9%.

Jan - Mar 2022

Responding to the war in Ukraine

As a platform, we've responded to the war in Ukraine with increased safety and security measures to help ensure people can express themselves and share their experiences while we work aggressively to counter harmful misinformation and other violations of our policies. The war has challenged us to confront a complex and rapidly changing environment. We've increased investments with our fact-checking partners who help assess the accuracy of content. We've continued to evolve our detection methods to help take swifter action against accounts that attempt to scam or mislead our community through unoriginal livestreams and video content. And

we began a pilot of our state-controlled media policy by labeling content from some accounts in order to bring important context to viewers. In response to Russia's 'fake news' law, from March 6 we suspended livestreaming and new content to our video service in Russia given the safety implications of this law. Content posted by accounts based outside of Russia is not currently available for distribution in Russia.

From February 24 through the end of the first quarter, March 31, 2022, we took the following steps to safeguard our community:

- Our safety team focused on the Ukraine war removed 41,191 videos, 87% of which violated our policies against harmful misinformation. The vast majority (78%) were identified proactively.
- Our fact-checking partners helped assess 13,738 videos globally.
- We added prompts on 5,600 videos informing viewers that content could not be verified by fact checkers.
- We labeled content from 49 Russian state-controlled media accounts.
- We identified and removed 6 networks and 204 accounts globally for coordinated efforts to influence public opinion and mislead users about their identities.

Oct - Dec 2021

We use a combination of technology and people to identify and remove violations of our Community Guidelines, and we continually invest in our policies, products, and partnerships to support the overall safety of our community and platform. In the fourth quarter of 2021, we continued to enhance our machine learning moderation techniques, which resulted in improvements to enforcement consistency and the accuracy of our automated removal systems. As a result of ongoing improvements to audio and speech recognition and detection, we meaningfully improved proactive detection and removals for violations of our suicide, self-harm, and dangerous acts policy, hateful behavior policy, and harassment and bullying policy. Additionally, ongoing optimisations of our minor safety detection models helped improve consistency in enforcement. The number of videos reinstated after being appealed decreased from the previous quarter, which we attribute to a reduction in false positives and an improvement in the accuracy of our detection and automated removals.