CASE STUDY

Major social media platform turns

moderation chaos into strategic defense

A leading social media platform turned overwhelmed mod teams into strategic defenders with AI user profile moderation—removing harmful profiles 500x faster while humans focus on what they do best.

Content moderation was supposed to get easier as this major social media platform grew. Instead, it kept getting harder. It felt like an endless game of cat and mouse with bad actors who always found new ways around their defenses. The bigger their user base got, the more overwhelmed their moderation teams became.

Their existing approach wasn't working. Third-party tools provided temporary relief, but spammers and bad actors kept finding workarounds.

"We were hiring more moderators, but it felt like we were just putting a band-aid on a broken system," explains the platform's engineering team.

"Third-party tools would slow down bad actors, but they always found workarounds. Our team was drowning in reports."

The scale was overwhelming with millions of user reports flooded in daily and moderators stuck reviewing obvious spam. Meanwhile, the subtle, nuanced cases that needed human judgment got buried in the noise.

"Every delayed removal was potential brand damage," says a platform team member.

"Users didn't feel safe. They expect harmful content to be removed immediately, not hours or days later."

Breaking the cycle

When the social platform discovered an AI-powered User Profile Moderation solution, it felt like a lifeline. It was exactly what they'd been hoping for – the AI could automatically catch the obvious spam, so their moderators could focus on the complex cases that required real judgement.

"We'd tried using only AI solutions, but they kept missing context, and depending solely on people to moderate meant we were constantly overwhelmed," explains the engineering team. *"This combined both."*

Within two weeks, User Profile Moderation was integrated into their existing workflow. Training the AI model took another 2-4 weeks, meaning they were seeing automated decisions within six weeks of project start.

"Six weeks from start to automation—that's unheard of in our industry," says the team. *"Most moderation overhauls take months or years."*

alorica

Learning while working

The magic of User Profile Moderation is that it improves automatically. While human experts handle nuanced cases, the AI is quietly learning from their expertise and getting better at its own job.

Because AI handles repetitive, low-value tasks automatically, there's no maintenance required from their engineering team. Moderators can focus on strategic decisions that improve the overall system, and the system learns 24/7 just by seeing our analysts work, allowing both sides get better at their jobs.

"Our moderators went from feeling overwhelmed to feeling strategic," explains the platform's leadership. *"Instead of burning out on obvious spam, they're solving complex cases that actually impact user safety."*

The transformation

User Profile Moderation now processes over **8.5 million reports monthly**, with an average resolution time of just **12.2 seconds**. The platform sees **9x fewer false positives** compared to their previous system.

"We couldn't have imagined this working so well," says the engineering leader. "It's [AI moderation] lightning fast, saves us money, and does a better job than anything we've used before."

The AI equivalent of 850 specialists now handles the bulk of routine moderation, while human experts tackle the cases that require contextual understanding, cultural nuance, or complex judgment calls.

Strategic defense

Today, the social platform operates with confidence that harmful profiles get removed before they can cause damage. The self-learning system continuously seals the loopholes that bad actors try to exploit, staying ahead of evolving threats.

"We've transformed from reactive firefighting to proactive defense," reflects the engineering team. "Bad actors used to stay one step ahead of us. Now we're ahead of them."

The User Profile Moderation solution continues to evolve as new threats emerge. The social platform appreciates having a solution that adapts automatically without requiring constant engineering resources.

> “In social media, threats evolve daily. Having a system that learns and improves without our intervention means we can focus on building features users love, not just fighting the bad guys. That's the kind of strategic advantage that matters."
>
> — Social Media Platform Team