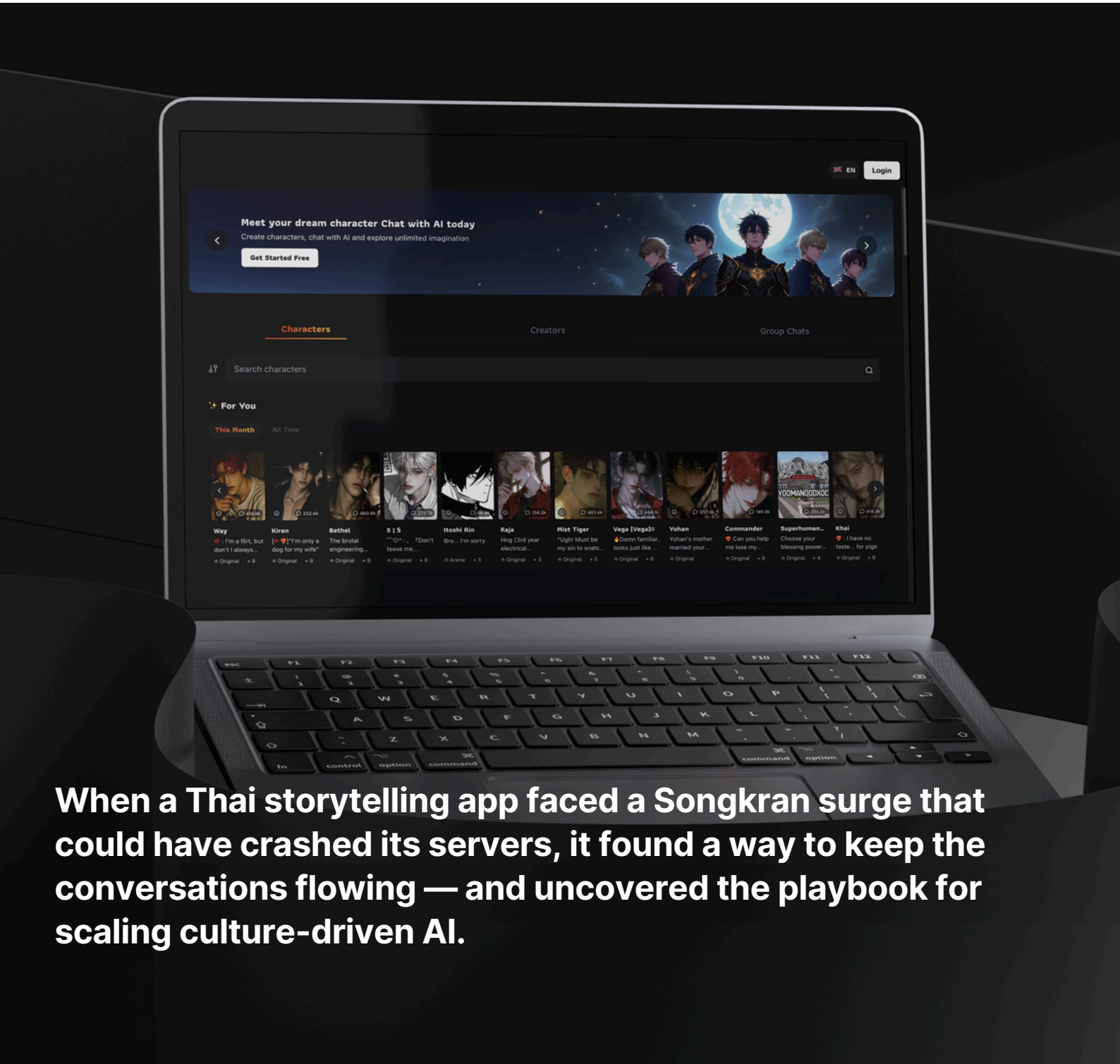
BytePlus | khui ai

Case Study

The Startup That Turned a Holiday Spike into 10 Billion-Token Breakthrough



When a Thai storytelling app faced a Songkran surge that could have crashed its servers, it found a way to keep the conversations flowing — and uncovered the playbook for scaling culture-driven AI.

BytePlus

A Water-Fight Holiday Becomes a Billion-Token Stress Test

April in Bangkok usually means splashing water on strangers in the streets. But during the Songkran festival of 2025, the flood wasn't just on the roads — it hit the servers of KhuiAI.

The fledgling conversational-storytelling app suddenly saw its traffic surge to **more than a billion tokens in just 24 hours** as fans flocked online to chat with their favourite AI characters during the holiday downtime.

For Jetnipat, KhuiAI's 24-year-old co-founder and CTO, the moment was both thrilling and unnerving. The team had built the platform in their university dorm room only months earlier. "A national holiday suddenly became our biggest stress test," he recalled later, describing the night spent watching dashboards spike far beyond their forecasts.

From Campus Side-Project to Thai Pop-Culture Phenomenon

KhuiAI began life as a senior-year project, driven by the idea that storytelling could feel as personal as texting a friend. Instead of generic bots, the team created **Thai-language personas shaped by local fandoms — BL-drama idols, romantic leads, fantasy characters.**

The cultural resonance was immediate. Teen and twenty-something women made up **80 percent of its users**, often chatting for h**alf an hour at a time**, as if catching up with a favourite character. By July, **20,000 people were logging in daily and monthly visitors topped 300,000.**

What kept the world alive was a vibrant **creator community — 13,000 on Discord and 8,000 on Facebook** — where fans designed their own characters and swapped story arcs. Jetnipat liked to say that fans "didn't just want to read stories; they wanted the characters to talk back," a line that neatly explains why a class project became a phenomenon in months.

When Popularity Exposed the Cracks

The early months were run on a patchwork of low-cost providers. That choice kept the bills low but became unsustainable as the audience grew.

During holiday peaks, **latency slowed conversations to a crawl,** breaking the illusion of real-time companionship. **Token consumption ballooned from 300 million a day to over a billion**, straining both budget and stability. The tiny three-person development team spent sleepless nights firefighting outages without dedicated enterprise-grade support.

Jetnipat often describes that period as “trying to keep a music concert going while replacing the speakers mid-show.”

The Switch to BytePlus ModelArk with DeepSeek V3

By mid-2025, KhuiAI’s growing pains called for industrial-grade infrastructure. The team migrated to **BytePlus ModelArk, powered by DeepSeek V3**, in a transition that users barely noticed, but the developers immediately felt.

Behind the scenes, the difference was stark. The new backbone **scaled seamlessly from a few hundred million tokens a day to sustaining up to 10 billion daily tokens without a single minute of downtime**, according to BytePlus’ engineering team.

For a startup that once scrambled to survive a festival surge, that level of headroom changed everything: traffic spikes were no longer a threat but simply another day at scale.

Cost efficiency also factored heavily into the decision. “For a startup, every cent counts, and BytePlus was the most cost-effective option that still delivered the performance we needed,” said Jetnipat. Real-time dashboards, a smooth API switch-over, and hands-on 24/7 support rounded out the move — a migration completed with near-zero downtime.

BytePlus

Results at a Glance

75,000+
daily active users

920,000
monthly visitors (July 2025)

10 billion
tokens processed in a single day without downtime

40,000
Discord community members

~30-minute
average session length

Impact: Stability that Preserved the Magic

The technical upgrade rippled all the way to user experience. Holiday spikes no longer froze the platform, so fans could continue chatting uninterrupted even during Songkran's billion-token surge.

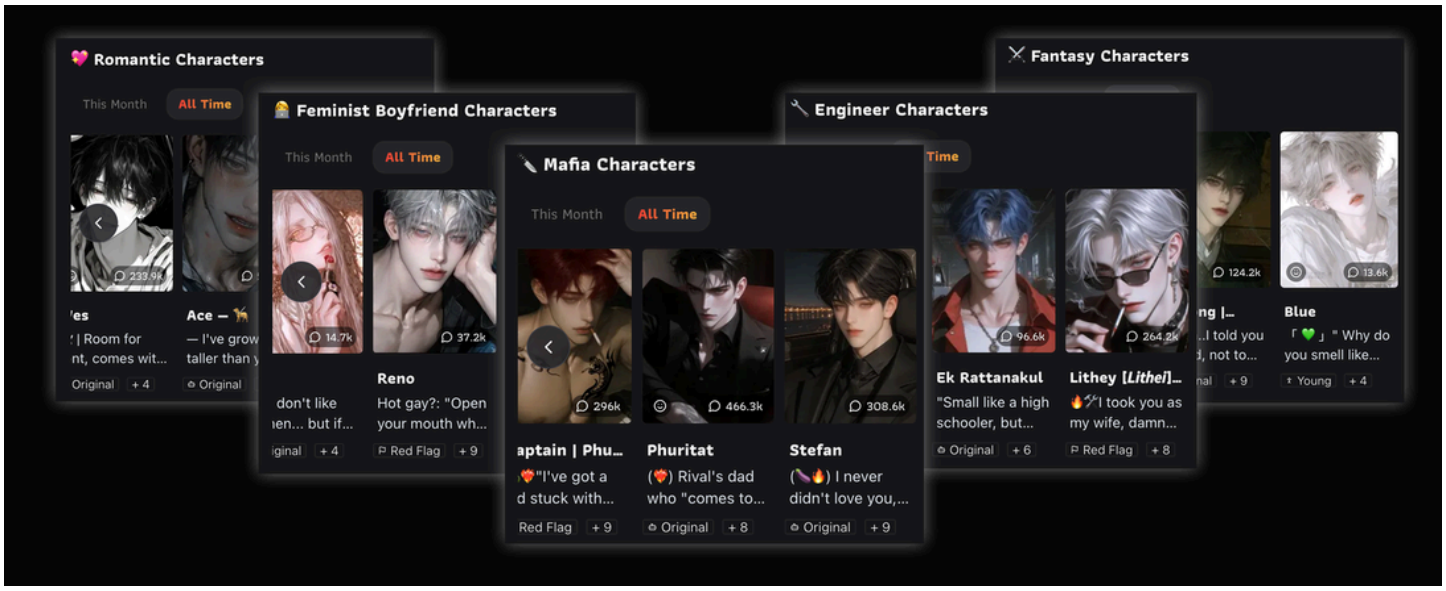
The improvement in response speed and reliability also made the AI characters feel more alive — especially when they began initiating messages themselves to pull users back into ongoing storylines. Average sessions held at about 30 minutes per user, a strong signal of stickiness.

KhuiAI's growing network of hobbyist writers and illustrators benefited too. **For the first time, creators could see dashboards tracking how their characters performed — how many people interacted, how often, for how long.** That feedback loop boosted confidence and spurred more inventive story arcs.

As Jetnipat put it, "The platform now runs smoothly enough for us to focus on the creative side. Our community has more confidence to build great stories because they trust the experience won't break midway."

From Thai Fandom to a Regional Story-Verse

With a solid foundation in place, KhuiAI is preparing to expand beyond Thailand, starting with Laos and eventually broader Asia-Pacific markets. The team is experimenting with voice-enabled companions, AR Try-On-style effects for characters, and even AI-driven mascots for banks as conversational service agents.



Future versions may also tap BytePlus Seedream and Seedance to give characters richer, more visually expressive identities. Jetnipat often frames the ambition this way: “We want to make media interactive everywhere — the way YouTube changed video, but for stories you can actually talk to.”

If you would like to learn more about our products and solutions, pleae reach out to our at www.byteplus.com/en/contact.