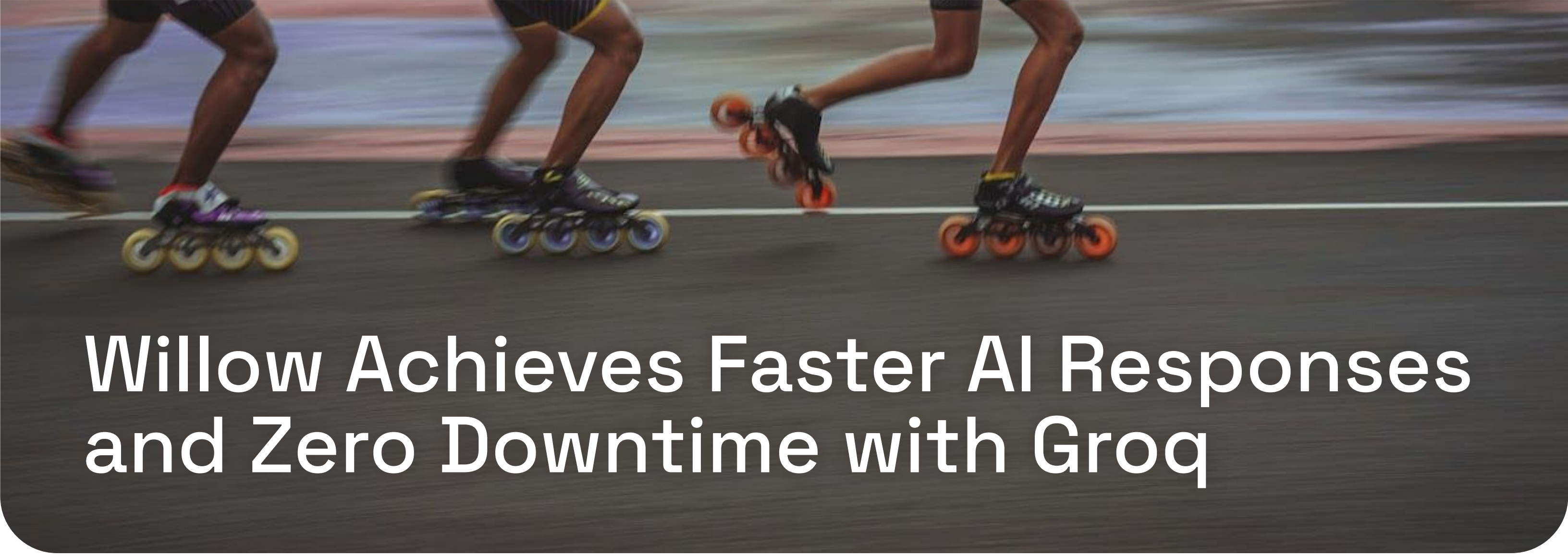

Willow Achieves Faster AI Responses and Zero Downtime with Groq

How reliability became Willow's #1 priority

For voice-driven apps like Willow, reliability matters more than speed. "Uptime is the lifeblood of our product," says Lawrence Liu, CTO and Co-founder. "If the service goes down, even briefly, we risk losing trust and users."

Early on, Willow self-hosted its LLMs, managing scaling, uptime, and incident response. Larger providers offered fine-tuning but not the reliable infrastructure Willow needed. The result: latency from long prompts, weekly outages tied to GPU instability, and user churn.

To grow, Willow needed infrastructure that could handle real-time voice input without blinking.

> "Since switching to Groq, we've had zero downtime. That's been transformational for our users and our team."
>
> **LAWRENCE LIU, CTO & CO-FOUNDER, WILLOW**

CUSTOMER:
Willow

PRODUCT:
AI voice dictation that turns speech into well-formatted text, helping you write faster and more accurately across any app.

Groq-powered performance and peace of mind

Willow switched to Groq to power their fine-tuned Llama-3.1-8b model. By running a LoRA fine-tune on a GroqCloud dedicated instance, they gained full control over performance without managing infrastructure.

Throughput and long prompts were a initially a concern for Willow. Typically, more tokens mean slower responses. But GroqCloud's LPU architecture scales workloads from thousands to millions of tokens instantly. With Groq engineering's help, Willow added speculative decoding, cutting latency dramatically. "We expected latency to increase with longer token counts, but with Groq, it didn't," said Lawrence.

The benefits went beyond speed: "Since switching to Groq, we've had zero downtime. That's been transformational for our users and our team." Groq's deterministic performance ensures consistency, even under heavy workloads.

Faster responses, happier customers

Switching to GroqCloud brought Willow four critical improvements: zero downtime; noticeably lower latency (300–500 ms faster); reduced support requests; and higher user retention.

“We haven’t had to send a single, ‘Our servers are down’ message since moving to Groq,” said Lawrence. “That’s huge for customer trust.”

Latency gains were just as critical: responses are now 300–500ms faster, transforming real-time voice interactions. “That speed unlocks new workflows—like dictating a Slack message or email,” he explained.

With uninterrupted uptime, Willow no longer worries about emergency fixes or churn spikes. Developers also saw unexpected value: swapping model weights is nearly on-demand instead of hours.

Switching to GroqCloud brought Willow four critical improvements:

1. No more downtime - zero
2. Noticeably lower latency (300–500 ms faster)
3. Reduced support requests
4. Higher user retention

With Groq handling the heavy lifting, Willow has grown from a voice dictation app into a workplace productivity platform. “We’re seeing more users dictate emails, send Slack messages, even summarize meetings with voice,” said Lawrence.

“The biggest benefits have been uptime, latency, and support,” Lawrence added. “Customers are happier, and we’re no longer worrying about outages. It’s been a night-and-day difference.”

Reduced latency and rock-solid reliability make voice input a seamless part of daily workflows. The impact is clear: higher retention, fewer issues, smoother usage, and positive feedback on speed and reliability.

Scaling voice across Silicon Valley and beyond

Willow has big plans for growth. The team wants to become the go-to dictation provider in Silicon Valley, then expand nationwide as voice input becomes more common across the workplace. And Groq is a key part of that vision. “ Groq is becoming known as the infrastructure provider for low latency AI,” said Lawrence. “We’re proud to be building with them.”

The partnership isn’t just about performance, it’s a shared mission to reshape how people communicate with software. Together, Willow and Groq are unlocking real-time AI experiences that feel effortless and human.

Want to see what’s possible?

Start building at console.groq.com

groq