How Tenali is Redefining Real-Time Sales with Groq

The challenge: Latency can kill deals

Tenali is an AI-powered real-time sales assistant that helps revenue teams respond instantly during live customer conversations. It listens in real-time to sales calls, detects questions, retrieves answers from the organization's data, and surfaces the right response to sellers within seconds, all without human intervention.

Behind the scenes, two purpose-built agents work in tandem: one answers every question automatically; the other turns the call into action before you hang up.

Today, Tenali's users include sales reps and customer success teams, delivering unmatched conversational AI performance.

But getting there took a lot of trial and error.

Before Groq, Tenali explored options like OpenAI, Anthropic, and Fireworks, but each had the same key issue: latency.

"When we used OpenAI and other providers, latency was 5–10 seconds," explained Aniket Patel, CEO and founder of Tenali. "That just doesn't work when you're on a call with a customer."

> "Groq solved our existential crisis. Without it, we wouldn't have a product. That's how critical Groq is to our business."
>
> **ANIKET PATEL, CEO & CO-FOUNDER, TENALI**

CUSTOMER:
Tenali

Tenali AI

PRODUCT:
Tenali is an AI-powered real-time sales assistant that helps revenue teams respond instantly during live customer conversations.

Every delay meant lost momentum, missed opportunities, and worse, lost deals. In a live sales call, if your AI assistant takes 10 seconds to answer, your buyer has already moved on.

The Tenali team knew their value proposition hinged on real-time responses, not 10-second delays. But nothing in the market could deliver at that level.

Until they found Groq.

The solution: Performance that changed the game

Switching to Groq was, in Aniket's words, a "game changer" for Tenali. They worked hard to figure out what actually performs best for their specific needs. What they discovered was that a smaller, open-source model running on Groq gave them better speed and cost-efficiency without compromising accuracy, right out of the box.

While most assume larger models means better performance, Tenali proved otherwise. Groq's LPU-powered architecture gave Tenali the ability to run small but mighty open-source models like Llama 70B, Whisper-large-v3-turbo, and gpt-oss-20B at 128k with incredible speed, and at scale.

"With Groq, our latency dropped to around 200 milliseconds," Aniket shared, "That is an **over 25x improvement**. And, with Gemma models, it's now at 50 milliseconds. That's not just fast, that's unheard of."

But Groq delivered more than speed for Tenali. It delivered accuracy. "Groq models just 'get' the conversation," Aniket explained. "No other provider came close in terms of accuracy."

Groq uses TruePoint numerics, which reduces precision only in areas that do not reduce accuracy. Coupled with Groq's LPU architecture, this allows for the preservation of quality with high precision numerics.

The impact: Achieving real-time performance and cost efficiency

The models optimized and compiled on Groq's LPUs delivered high accuracy with minimal latency and Tenali is now able to provide answers that are real-time.

While speed and accuracy were must-haves, Groq delivered another win: cost efficiency that scales. When Tenali compared inference costs across providers, Groq came out on top. Tenali used to pay 40 cents per hour for transcription. With Groq, it dropped to just 4 cents.

Tenali even uses Groq-powered Speech-To-Text models for transcription tasks that used to cost 10x more with other services. Aniket shared, "We literally use Groq as our price-performance benchmark."

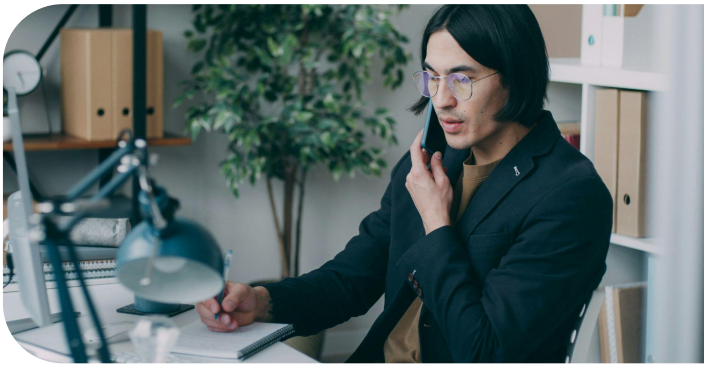
Additionally, Aniket shared that Groq's support and documentation helped them move fast without friction. "With other vendors, we constantly had to ping someone to help us figure things out. With Groq, the docs are solid and support is always responsive."

In short, Groq made it technically viable for Tenali to even exist. As Aniket put it, "Groq solved our existential crisis. Without it, we wouldn't have a product. That's how critical Groq is to our business."

The bottom line: Real-time, reliable, affordable, and powered by Groq

Tenali isn't just building a product. They're scaling a category-defining platform. Powered by Groq's ultra-low latency, high-accuracy infrastructure, Tenali is setting a new standard for real-time AI in sales. Their app shortens sales cycles by 33% and boosts rep productivity 4x, delivering intelligent, in-the-moment assistance on every customer call.

With Groq, Tenali didn't just find a vendor, they found a foundational technology partner. And for Groq, the feeling is mutual: Tenali is a model customer, pushing the boundaries of what's possible with real-time AI and proving that bold vision, sharp execution, and the right platform can reshape an entire industry.



Want to see what's possible?

Start building at console.groq.com