5 Under the Radar AI Infrastructure Startups Powering the Consumer AI Boom

From real time voice AI to generative media, these five startups are building the inference layer powering the next wave of consumer AI applications.

SAN FRANCISCO, DE, UNITED STATES, February 3, 2026 /EINPresswire.com/ -- As enterprise AI spending hits $37 billion and consumer apps process billions of daily requests, a new generation of infrastructure startups is quietly powering the boom.

> We built infrastructure with opinions. We made hundreds of decisions so our customers don't have to."
>
> *Eftal Yurtseven, CEO of each::labs*

The AI infrastructure market is experiencing unprecedented growth. Enterprise AI spending reached $37 billion in 2025, a 3.2x increase from the previous year, while consumer AI applications now process over 4 billion daily prompts according to Andreessen Horowitz's latest analysis of the top 100 consumer AI apps.

Behind every viral AI app lies a critical layer that rarely makes headlines: the inference infrastructure that runs these models in production. Here are five companies quietly powering the fastest-growing AI products in the world.

Cerebrium: The Real-Time AI Backbone

Cerebrium has carved out a specialized niche in real-time voice and video AI infrastructure. While most platforms optimize for throughput, Cerebrium obsesses over latency, achieving cold start times as low as 2 seconds and network latency under 50 milliseconds. The platform powers production workloads for Tavus (AI video avatars), Deepgram (speech-to-text), and Vapi (voice assistants). Founded by Michael Louis and Jonathan Irwin, their custom container runtime delivers 40% compute savings compared to previous solutions.

Beam: Serverless GPUs Without the Wait

Beam built its own container runtime called beta9, designed for launching GPU-backed containers in under one second. Add a decorator to your Python function, specify your GPU, and deploy with a single command. Hundreds of teams run production workloads on Beam, including Coca-Cola and Geospy. The open-source approach has built a thriving developer

community contributing improvements and sharing deployment patterns.

each::labs: The Contrarian Bet on Consumer AI

While most AI infrastructure companies chase enterprise contracts, each::labs made a deliberate bet in the opposite direction: consumer AI builders.
"We made a deeply contrarian choice," wrote CEO Eftal Yurtseven. "Everyone told us to follow the money. The safe path. Chase enterprise."
The San Francisco-based company provides end-to-end generative media infrastructure, including visual models, audio models, text models, and a drag-and-drop workflow system.

The results: 13x revenue growth in 12 months and 260% net revenue retention. Perhaps most remarkably, 5% of Andreessen Horowitz's Top 100 Consumer AI Apps now run on each::labs infrastructure.

"We built infrastructure with opinions," Yurtseven explains. "We made hundreds of decisions so our customers don't have to."
The company achieved this growth with zero salespeople. Every customer came through word-of-mouth. Founded by Eftal Yurtseven, Ferhat Budak, and Canberk Sinangil, their obsession with developer experience created what one customer described as "infrastructure that gets out of your way."

Modal: Making GPUs Feel Like Functions

Modal built the most developer-friendly way to run AI workloads. AI infrastructure should feel as simple as writing a Python function, with no Dockerfiles, no YAML, and no infrastructure management. Thousands of customers run workloads on the platform. Founded by Erik Bernhardsson (previously at Spotify) and Akshat Bubna, Modal's custom runtime enables sub-second cold starts that competitors struggle to match.

BentoML: Open-Source Goes Enterprise

BentoML took a different path: open source first. The framework for packaging and deploying ML models has been adopted by thousands of developers, now translating into enterprise adoption. Founded in 2018 by Chaoyu Yang, the company built a loyal developer community contributing to the open-source project while enterprise customers pay for managed services.

The Infrastructure Layer That Matters

These five companies represent a broader shift. As model capabilities have commoditized, the battlefield has moved to infrastructure. The winners won't be determined by who has the biggest models, but by who can run them fastest, cheapest, and most reliably. Cerebrium owns real-time voice and video. Beam offers the fastest serverless GPU experience. each::labs

provides the integrated stack for generative media. Modal makes infrastructure feel like Python functions. BentoML gives enterprises the control they need.

The AI infrastructure market is projected to reach $96.6 billion by 2027. The companies building the picks and shovels for the AI gold rush may ultimately capture more value than the prospectors themselves.

Eftal Yurtseven
eachlabs
email us here
Visit us on social media:
LinkedIn
X
