DeepInfra Series B Announcement - Media FAQ

This document is intended as a media FAQ for DeepInfra's Series B announcement, providing background, key metrics, and clear answers to the most common questions from reporters and analysts. It is designed to support press coverage, briefings, and interviews by ensuring consistency across messaging on DeepInfra's growth, technology, and market position.

Company + Growth

Q: What is DeepInfra and what problem are you solving?
DeepInfra is a purpose-built inference cloud platform for high-throughput AI inference, enabling enterprises and developers to run models at scale with high performance, low latency, and strong cost efficiency. As AI moves from experimentation to production, DeepInfra provides the infrastructure needed to make real-world AI deployments economically viable and reliable.

Q: What led you to found DeepInfra?
DeepInfra was founded by Nikola Borisov and team, drawing on their experience scaling large, real-time systems at imo, which serves over 200 million users globally. That experience highlighted the challenges of running compute-intensive workloads efficiently at scale. As AI workloads began to follow a similar trajectory, it became clear that existing cloud infrastructure was not optimized for high-performance, cost-efficient inference—leading to the creation of DeepInfra.

Q: How big is the platform today?
DeepInfra powers production-scale AI inference, supporting 190+ models across text, image, speech, embeddings, and multimodal use cases. The platform operates across eight global data centers, serves thousands of developers, and processes over four trillion tokens per week. DeepInfra is built to handle both generation and embedding workloads at scale.

Q: How fast is the company growing?
DeepInfra has tripled revenue since the beginning of 2026, reflecting strong and accelerating demand for production-grade AI inference infrastructure.

Q: What's driving that growth?
Growth is driven by the shift toward agentic and always-on AI systems, which dramatically increase token consumption and require continuous, high-volume inference. Enterprises are also moving toward open-source models and OpenAI-compatible infrastructure to improve economics, increase flexibility, and reduce vendor lock-in. In addition, platforms like OpenClaw and other agent-driven workloads represent a growing share of inference demand.

Product + Technology

Q: Where is your infrastructure located?
DeepInfra operates across eight global data centers in the United States, with plans to expand into Europe, APAC, and the Middle East. This distributed footprint enables low-latency inference and geographic flexibility for customers worldwide.

Q: How does DeepInfra approach open-source AI models?
DeepInfra prioritizes open-source models and provides infrastructure that enables customers to deploy, fine-tune, and scale them efficiently. The platform offers OpenAI-compatible APIs, making it easy to integrate and migrate workloads while maintaining full control over performance and cost.

Q: What differentiates DeepInfra from hyperscalers or other AI infra startups?
DeepInfra is purpose-built for inference and owns and operates its GPU infrastructure across its data centers, rather than relying on rented or shared capacity. This enables deeper optimization, more predictable performance, and lower cost per token.

Unlike general-purpose cloud providers, DeepInfra is vertically integrated for high-throughput inference workloads and designed specifically for agentic, continuous-use AI systems. The platform emphasizes open-source model support and OpenAI-compatible APIs, reducing friction for developers and avoiding proprietary lock-in.

Q: How does DeepInfra ensure performance and reliability at scale?
DeepInfra leverages distributed GPU clusters, bare-metal optimization, and co-located infrastructure to deliver high throughput and low latency. The platform is engineered for high-volume generation and embedding workloads, with built-in monitoring, redundancy, and scaling capabilities to ensure consistent performance under peak demand.

Q: What role do NVIDIA GPUs play in your platform?
DeepInfra works closely with NVIDIA to optimize performance on its latest GPU architectures, including Blackwell and Vera Rubin. This collaboration enables significant improvements in inference efficiency and cost per token while maintaining compatibility with cutting-edge open-source models.

Q: How does DeepInfra ensure the security of AI workloads?
DeepInfra is designed with enterprise-grade security, including isolated environments, zero data retention, and strict access controls. The platform is SOC 2 and ISO 27001 compliant, ensuring that customer data remains protected in transit and at rest while meeting enterprise and government security requirements.

Funding

Q: Who are your investors?
The Series B was co-led by 500 Global and Georges Harik, with participation from A.Capital Ventures, Crescent Cove, Felicis, NVIDIA, Peak6, Samsung, Supermicro, and Upper90.

Q: How much are you raising in the Series B?
DeepInfra raised $107 million in total, consisting of $67 million in equity and $40 million in debt financing.

Q: What is DeepInfra's valuation following the Series B?
DeepInfra is not disclosing its valuation. The round reflects strong investor conviction in the company's growth, technology, and role in powering production-scale AI infrastructure.

Q: What will you use the Series B funding for?
The funding will expand global compute capacity, accelerate GPU infrastructure growth, and further optimize the platform for high-throughput inference. It will also support developer tooling, next-generation model support, and enterprise expansion, including growth in EU, APAC and the Middle East, and sovereign AI deployments.

Q: Why raise now?
The AI market is shifting from training to inference at scale. As more AI systems move into production, inference is becoming the primary driver of cost and performance constraints. This creates a critical opportunity for infrastructure built specifically to optimize inference, making it the right time to accelerate investment.

Market + Positioning

Q: Are you seeing more enterprise or developer adoption?
DeepInfra is seeing strong adoption across both enterprises and developers, with the majority of usage coming from fast-growing AI startups. These companies are building copilots, agentic workflows, and large-scale AI systems that require high-volume, reliable inference. Developers are also adopting DeepInfra for its performance, cost efficiency, and ease of integration.

Q: What is the long-term vision?
DeepInfra aims to be the leading global platform for efficient, secure, and scalable AI inference—enabling enterprises, developers, and governments to run open-source and agent-driven AI systems at production scale with predictable performance and cost.