# AI Inference Platform-as-a-Service (PaaS) Market worth $105.22 billion by 2030 - Exclusive Report by MarketsandMarkets™

NEWS PROVIDED BY
**MarketsandMarkets →**
Oct 03, 2025, 10:01 ET

DELRAY BEACH, Fla., Oct. 3, 2025 /PRNewswire/ -- **The global AI inference PaaS market is anticipated to be valued at USD 18.84 billion in 2025 and USD 105.22 billion by 2030, registering a CAGR of 41.1% during the forecast period according to a new report by MarketsandMarkets™.** The growth of the AI inference PaaS market is attributed to the surging adoption of generative AI and large language models (LLMs), which demand scalable, low-latency infrastructure for real-time deployment. As enterprises shift toward cloud-native AI architectures, PaaS providers emerge as critical enablers by offering flexible, cost-efficient, high-performance inference environments. Furthermore, the increasing integration of inference capabilities with industry-specific SaaS platforms is expanding use cases across sectors such as finance, retail, and healthcare, accelerating overall market adoption and growth.

**Download PDF Brochure: https://www.marketsandmarkets.com/pdfdownloadNew.asp?id=102780827**

**Browse in-depth TOC on " AI Inference Platform-as-a-Service (PaaS) Market "**

**270 – Tables**
**62 – Figures**
**302 – Pages**

**AI Inference Platform-as-a-Service (PaaS) Market Scope:**

| Report Coverage | Details |
|---|---|
| Market Revenue in 2025 | $ 18.84 billion |
| Estimated Value by 2030 | $ 105.22 billion |
| Growth Rate | Poised to grow at a CAGR of 41.1% |
| Market Size Available for | 2021–2030 |
| Forecast Period | 2025–2030 |
| Forecast Units | Value (USD Million/Billion) |
| Report Coverage | Revenue Forecast, Competitive Landscape, Growth Factors, and Trends |
| Segments Covered | By Deployment, Application, Vertical and Region |
| Geographies Covered | North America, Europe, Asia Pacific, and Rest of World |
| Key Market Challenge | Latency and bandwidth issues in cloud-only setups |
| Key Market Opportunities | Availability of on-demand inference for SMEs and startups |
| Key Market Drivers | Surging adoption of generative AI and large language models |

**By deployment, the public cloud segment is projected to account for the largest market share in 2025.**

The public cloud segment is anticipated to capture the largest market share in 2025, driven by its scalability, cost efficiency, and wide industry accessibility. Hyperscale providers, such as AWS, Microsoft Azure, and Google Cloud, have built robust infrastructures with advanced GPU and TPU resources, making them the preferred choice for deploying large-scale AI inference workloads. Public cloud models enable enterprises to rapidly operate generative AI, NLP, and computer vision applications without heavy upfront investment in infrastructure. The pay-as-you-go pricing model attracts SMEs and startups, who benefit from flexible cost structures and seamless

integration with AI toolchains. With the rise of generative AI and LLM-driven applications requiring massive inference capabilities, public cloud providers continue to dominate, offering specialized AI accelerators, pre-trained APIs, and managed inference services that effectively address enterprise and developer needs.

**IT & telecom segment is likely to grow at a high CAGR in the AI inference PaaS market from 2025 to 2030.**

The IT & telecom sector is expected to register the highest CAGR in the AI inference PaaS market during the forecast period, fueled by rapid digitization, 5G deployment, and the rising demand for AI-powered customer experience management. Telecom operators leverage inference PaaS to optimize network performance, predict traffic loads, and deliver real-time analytics for seamless connectivity. In parallel, IT service providers adopt inference platforms to scale AI-enabled cloud services, enhance cybersecurity, and support enterprise clients in deploying AI workloads at speed. The integration of AI inference with edge computing unlocks new opportunities in low-latency applications, such as autonomous networks, IoT analytics, and immersive digital services. With increasing partnerships between telecom operators and hyperscalers and growing demand for sovereign AI in regional cloud ecosystems, the IT & telecom sector is emerging as a critical growth engine for AI inference PaaS adoption worldwide.

**Inquiry Before Buying: https://www.marketsandmarkets.com/Enquiry_Before_BuyingNew.asp?id=102780827**

**North America is expected to account for the largest market share in 2030.**

North America is likely to hold the largest share of the **AI inference PaaS industry** in 2030, supported by its advanced cloud infrastructure, strong presence of hyperscale providers, and early adoption of AI technologies across industries. The US leads the region, with tech giants, such as AWS, Microsoft Azure, and Google Cloud, offering robust inference services tailored to generative AI, machine learning, and computer vision applications. The BFSI, healthcare, and media & entertainment sectors are among the heaviest users of inference PaaS, deploying it for fraud detection, medical imaging, personalized recommendations, and real-time analytics. A mature ecosystem of AI startups, venture capital investments, and research institutions further strengthens the innovation pipeline, ensuring continuous demand for inference capabilities.

Regulatory frameworks, such as the US NIST AI Risk Management Framework and Canada's AI governance initiatives, drive trust and responsible adoption, particularly in sensitive sectors, including finance and healthcare. Moreover, enterprises in North America are shifting toward hybrid and multi-cloud inference strategies to balance performance, compliance, and cost. The region is also witnessing significant adoption of sovereign AI frameworks, with enterprises emphasizing data localization and AI security. With strong enterprise AI budgets, high penetration of generative AI applications, and growing collaborations between hyperscalers and industry verticals, the region is expected to maintain its leadership position, serving as the hub for innovation and commercialization in the global AI inference PaaS market.

**Key Players**

Key companies operating in the **AI inference PaaS companies** include Microsoft (US), Amazon Web Services, Inc. (US), Google (US), Oracle (US), and IBM (US).

**Get 10% Free Customization on this Report: https://www.marketsandmarkets.com/requestCustomizationNew.asp?id=102780827**

**Browse Adjacent Market: Semiconductor and Electronics Market** Research Reports &Consulting

**See More Latest Semiconductor Reports:**

**IoT Technology Market** by Node Component (Sensor, Memory Device, Connectivity IC, Processor, Logic Devices), Software Solution (Remote Monitoring, Data Management), Platform, Service, End-use Application, Geography - Global Forecast to 2030

**Proximity Sensor Market** by Technology (Inductive, Capacitive, Magnetic, Photoelectric/Optical, Ultrasonic), Product Type (Fixed & Adjustable distance), Range (<10 MM, 10-20 MM, 21-40 MM, >40 MM), Output and Region - Global Forecast to 2030

**About MarketsandMarkets™**

MarketsandMarkets™ has been recognized as one of America's Best Management Consulting Firms by Forbes, as per their recent report.

MarketsandMarkets™ is a blue ocean alternative in growth consulting and program management, leveraging a man-machine offering to drive supernormal growth for progressive organizations in the B2B space. With the widest lens on emerging technologies, we are proficient in co-creating supernormal growth for clients across the globe.

Today, **80% of Fortune 2000 companies rely on MarketsandMarkets**, and **90 of the top 100 companies in each sector trust us to accelerate their revenue growth**. With a **global clientele of over 13,000 organizations**, we help businesses thrive in a disruptive ecosystem.

The B2B economy is witnessing the emergence of $25 trillion in new revenue streams that are replacing existing ones within this decade. We work with clients on growth programs, helping them monetize this $25 trillion opportunity through our service lines – TAM Expansion, Go-to-Market (GTM) Strategy to Execution, Market Share Gain, Account Enablement, and Thought Leadership Marketing.

Built on the 'GIVE Growth' principle, we collaborate with several Forbes Global 2000 B2B companies to keep them future-ready. Our insights and strategies are powered by industry experts, cutting-edge AI, and our **Market Intelligence Cloud, KnowledgeStore**™, which integrates research and provides ecosystem-wide visibility into revenue shifts.

To find out more, visit **www.MarketsandMarkets™.com** or follow us on **Twitter** , **LinkedIn** and **Facebook** .

**Contact:**
**Mr. Rohan Salgarkar**
MarketsandMarkets™ INC.
1615 South Congress Ave.
Suite 103, Delray Beach, FL 33445
USA: +1-888-600-6441
Email: **sales@marketsandmarkets.com**
Visit Our Web Site: **https://www.marketsandmarkets.com/**
Research Insight: **https://www.marketsandmarkets.com/ResearchInsight/ai-inference-platform-as-a-Service-paas-companies.asp**
Content Source: **https://www.marketsandmarkets.com/PressReleases/ai-inference-platform-as-a-service-paas.asp**

Logo: **https://mma.prnewswire.com/media/1868219/MarketsandMarkets_Logo.jpg**

SOURCE MarketsandMarkets