AI Inference Platform-as-a-Service (PaaS) Market by Deployment (Private Cloud, Public Cloud, Hybrid Cloud), Application (Gen AI, Machine Learning, NLP, Computer Vision), Vertical (BFSI, IT & Telecom, Retail & E-commerce), Region - Global Forecast to 2030

Market Report | 2025-10-03 | 303 pages | MarketsandMarkets

AVAILABLE LICENSES:

- Single User $4950.00
- Multi User $6650.00
- Corporate License $8150.00
- Enterprise Site License $10000.00

Report description:

The AI inference PaaS market is projected to reach USD 18.84 billion in 2025 and USD 105.22 billion by 2030, recording a CAGR of 41.1% during the forecast period. The market is witnessing strong growth fueled by the rising need for real-time decision-making and the increasing integration of AI inference with industry-specific SaaS platforms. Sectors such as finance, retail, and healthcare leverage real-time insights to improve fraud detection, customer engagement, and clinical decision support, driving adoption of scalable inference services. At the same time, embedding inference capabilities into SaaS offerings allows enterprises to unlock tailored AI solutions without heavy infrastructure investments. These trends are expanding the addressable market and positioning AI inference PaaS as a core enabler of digital transformation.

https://mnmimg.marketsandmarkets.com/Images/ai-inference-platform-as-a-service-paas-market-img-overview.webp

"Private cloud segment is projected to record the second-highest CAGR between 2025 and 2030"
The private cloud segment is expected to grow at the second-highest CAGR in the AI inference PaaS market during the forecast period, driven by the increasing demand for data security, compliance, and customized infrastructure among enterprises. Sectors such as BFSI, healthcare, and government prioritize private cloud deployments due to strict regulatory frameworks and the data sensitivity involved. AI inference on private clouds allows organizations to retain full control over data, reduce latency, and achieve high performance with dedicated resources. Vendors are responding with hybrid and private cloud offerings that combine scalability with governance, enabling enterprises to deploy large language models (LLMs) and machine learning workloads

securely. Moreover, the rising adoption of sovereign AI initiatives in Europe and Asia-Pacific further strengthens demand for private cloud-based inference platforms.

"Machine learning segment is expected to hold a major share of the AI inference PaaS market in 2025"
The machine learning segment is likely to account for a significant share of the AI inference PaaS market in 2025, driven by its widespread adoption across end-use industries, such as finance, healthcare, retail, and manufacturing. Enterprises increasingly leverage machine learning algorithms for predictive analytics, fraud detection, customer personalization, and operational optimization, creating steady demand for scalable inference solutions. The ability of PaaS offerings to support real-time inference, automated model deployment, and cost-efficient scalability makes them a preferred choice for machine learning applications. Furthermore, the availability of pre-trained models, APIs, and managed infrastructure on cloud platforms is lowering entry barriers for SMEs and startups.

"Europe is anticipated to hold a significant market share in 2025"
Europe is projected to hold a strong position in the AI inference PaaS market in 2025, supported by advanced digital infrastructure, rising adoption of AI technologies, and increasing investments in sovereign AI initiatives. Countries such as the UK, Germany, and France are leading in AI adoption across industries, particularly in BFSI, automotive, and healthcare. The emphasis on data privacy and compliance, especially under GDPR, shapes the demand for secure and localized inference platforms, with global players and regional cloud providers expanding offerings tailored to these requirements. Growth in Europe is also driven by significant investments in cloud infrastructure and partnerships between hyperscalers and European institutions. In May 2024, Amazon announced major investments to expand cloud operations and a European sovereign cloud project, directly enhancing local compute capacity and enabling enterprises to access compliant inference services within the region. This move reflects a broader trend of hyperscalers localizing infrastructure to address Europe's sovereignty concerns. Alongside Amazon, Microsoft Azure, and Google Cloud are strengthening their European presence, while local providers, such as OVHcloud and Deutsche Telekom, are capturing enterprises prioritizing domestic hosting and trusted AI deployment.

Extensive primary interviews were conducted with key industry experts in the AI inference PaaS market space to determine and verify the market size for various segments and subsegments gathered through secondary research. The breakdown of primary participants for the report is shown below.
The study contains insights from various industry experts, from component suppliers to Tier 1 companies and OEMs. The break-up of the primaries is as follows:
- By Company Type: Tier 1 - 50%, Tier 2 - 30%, and Tier 3 - 20%
- By Designation: C-level Executives - 20%, Directors - 30%, and Others - 50%
- By Region: North America - 40%, Europe - 20%, Asia Pacific- 30%, and RoW - 10%

The AI inference PaaS market is dominated by a few globally established players, such as Microsoft (US), Amazon Web Services, Inc. (US), Google Cloud (US), Oracle (US), IBM (US), Alibaba Cloud (China), Salesforce, Inc. (US), Tencent Cloud (China), Baidu, Inc. (China), Together AI (US), CoreWeave (US), Predibase (US), Vectara (US), Prem AI (US), and Baseten (China), among others. The study includes an in-depth competitive analysis of these key players in the AI inference PaaS market and their company profiles, recent developments, and key market strategies.

Research Coverage:
The report segments the AI inference PaaS market based on deployment (public cloud, private cloud, and hybrid cloud), application (generative AI, machine learning, natural language processing, and computer vision), and vertical (healthcare, BFSI, automotive, retail & e-commerce, media & entertainment, government & defense, IT & telecom, and other verticals). It also discusses the market's drivers, restraints, opportunities, and challenges. It gives a detailed view of the market across four main regions (North America, Europe, Asia Pacific, and RoW). The report includes an ecosystem analysis of key players.

Key Benefits of Buying the Report:

- Analysis of key drivers (surging adoption of generative AI and large language models, increasing preference for cloud-native AI architectures, rising need for real-time decision making), restraints (high cost of AI accelerators and service pricing volatility, vendor lock-in concerns, data privacy and regulatory restrictions), opportunities (availability of on-demand inference for SMEs and startups, rise in sovereign AI and regional cloud partnerships, integration of AI inference platforms with industry-specific SaaS solutions), challenges (latency and bandwidth issues in cloud-only setups, complexities in managing AI models in dynamic production environments)
- Service Development/Innovation: Detailed insights on upcoming technologies, research and development activities, and new launches in the AI inference PaaS market
- Market Development: Comprehensive information about lucrative markets through the analysis of the AI inference PaaS market across varied regions
- Market Diversification: Exhaustive information about new products and services, untapped geographies, recent developments, and investments in the AI inference PaaS market
- Competitive Assessment: In-depth assessment of market shares, growth strategies, and product offerings of leading players, such as Microsoft (US), Amazon Web Services, Inc. (US), Google Cloud (US), Oracle (US), IBM (US), Alibaba Cloud (China), Salesforce, Inc. (US), Tencent Cloud (China), Baidu, Inc. (China), and Together AI (US)

Table of Contents:

1 INTRODUCTION 26
1.1 STUDY OBJECTIVES 26
1.2 MARKET DEFINITION 26
1.3 STUDY SCOPE 27
1.3.1 MARKETS COVERED AND REGIONAL SCOPE 27
1.3.2 INCLUSIONS AND EXCLUSIONS 28
1.3.3 YEARS CONSIDERED 29
1.4 CURRENCY CONSIDERED 29
1.5 LIMITATIONS 29
1.6 STAKEHOLDERS 30
2 RESEARCH METHODOLOGY 31
2.1 RESEARCH DATA 31
2.1.1 SECONDARY DATA 32
2.1.1.1 List of key secondary sources 32
2.1.1.2 Key data from secondary sources 33
2.1.2 PRIMARY DATA 33
2.1.2.1 List of primary interview participants 34
2.1.2.2 Breakdown of primaries 34
2.1.2.3 Key data from primary sources 35
2.1.2.4 Key industry insights 36
2.1.3 SECONDARY AND PRIMARY RESEARCH 36
2.2 MARKET SIZE ESTIMATION 37
2.2.1 BOTTOM-UP APPROACH 38
2.2.1.1 Approach to arrive at market size using bottom-up analysis
(demand side) 39
2.2.2 TOP-DOWN APPROACH 39
2.2.2.1 Approach to arrive at market size using top-down analysis
(supply side) 40
2.3 MARKET BREAKDOWN AND DATA TRIANGULATION 41

2.4 RESEARCH ASSUMPTIONS 42
2.5 RESEARCH LIMITATIONS 42
2.6 RISK ANALYSIS 43
3 EXECUTIVE SUMMARY 44
?
4 PREMIUM INSIGHTS 49
4.1 ATTRACTIVE OPPORTUNITIES FOR PLAYERS IN AI INFERENCE PAAS MARKET 49
4.2 AI INFERENCE PAAS MARKET, BY DEPLOYMENT AND APPLICATION 50
4.3 AI INFERENCE PAAS MARKET, BY VERTICAL 50
4.4 AI INFERENCE PAAS MARKET, BY GEOGRAPHY 51
5 MARKET OVERVIEW 52
5.1 INTRODUCTION 52
5.2 MARKET DYNAMICS 52
5.2.1 DRIVERS 53
5.2.1.1 Surging adoption of generative AI and large language models 53
5.2.1.2 Increasing preference for cloud-native AI architectures 53
5.2.1.3 Rising need for real-time decision making 54
5.2.2 RESTRAINTS 54
5.2.2.1 High cost of AI accelerators and service pricing volatility 54
5.2.2.2 Vendor lock-in concerns 55
5.2.2.3 Data privacy and regulatory restrictions 56
5.2.3 OPPORTUNITIES 56
5.2.3.1 Availability of on-demand inference for SMEs and startups 56
5.2.3.2 Rise in sovereign AI and regional cloud partnerships 57
5.2.3.3 Integration of AI inference platforms with industry-specific SaaS solutions 58
5.2.4 CHALLENGES 59
5.2.4.1 Latency and bandwidth issues in cloud-only setups 59
5.2.4.2 Complexities in managing AI models in dynamic production environments 59
5.3 TRENDS/DISRUPTIONS IMPACTING CUSTOMER BUSINESS 60
5.4 VALUE CHAIN ANALYSIS 61
5.5 ECOSYSTEM ANALYSIS 63
5.6 INVESTMENT AND FUNDING SCENARIO 65
5.7 PORTER'S FIVE FORCES ANALYSIS 65
5.7.1 INTENSITY OF COMPETITIVE RIVALRY 66
5.7.2 BARGAINING POWER OF SUPPLIERS 66
5.7.3 BARGAINING POWER OF BUYERS 67
5.7.4 THREAT OF SUBSTITUTES 67
5.7.5 THREAT OF NEW ENTRANTS 67
5.8 KEY STAKEHOLDERS AND BUYING CRITERIA 68
5.8.1 KEY STAKEHOLDERS IN BUYING PROCESS 68
5.8.2 BUYING CRITERIA 69
5.9 PATENT ANALYSIS 70
?
5.10 REGULATORY LANDSCAPE 73
5.10.1 REGULATORY BODIES, GOVERNMENT AGENCIES, AND OTHER ORGANIZATIONS 73
5.10.2 REGULATIONS 77
5.10.3 STANDARDS 79

5.11 PRICING ANALYSIS 80
5.11.1 PRICING RANGE OF AI INFERENCE PAAS OFFERED BY KEY PLAYERS, BY DEPLOYMENT, 2024 81
5.11.2 AVERAGE SELLING PRICE OF AI INFERENCE PAAS, BY APPLICATION, 2024 82
5.12 TECHNOLOGY ANALYSIS 83
5.12.1 KEY TECHNOLOGIES 83
5.12.1.1 Machine learning 83
5.12.1.2 Cloud computing 83
5.12.2 COMPLEMENTARY TECHNOLOGIES 84
5.12.2.1 Big data analytics 84
5.12.3 ADJACENT TECHNOLOGIES 84
5.12.3.1 High-performance computing (HPC) 84
5.13 CASE STUDY ANALYSIS 85
5.13.1 FORETHOUGHT OPTIMIZES AI INFERENCE AND SCALABILITY USING AWS SAGEMAKER 85
5.13.2 DOCUSIGN BOOSTS PRODUCTIVITY USING NVIDIA TRITON INFERENCE SERVER ON AZURE 85
5.13.3 SMEG UK LTD DELIVERS SMARTER CUSTOMER SERVICE THROUGH ORACLE GENERATIVE AI SOLUTIONS 86
5.13.4 CERN BUILDS LARGE-SCALE AI MODELS FOR SCIENTIFIC DISCOVERY USING OCI DATA SCIENCE 86
5.13.5 FIREWORKS AI BOOSTS AI MODEL PERFORMANCE WITH OCI AI INFRASTRUCTURE 87
5.14 KEY CONFERENCES AND EVENTS, 2025-2026 87
5.15 IMPACT OF 2025 US TARIFF ON AI INFERENCE PAAS MARKET 88
5.15.1 INTRODUCTION 88
5.15.2 PRICE IMPACT ANALYSIS 90
5.15.3 KEY TARIFF RATES 90
5.15.4 IMPACT ON COUNTRIES/REGIONS 91
5.15.4.1 US 91
5.15.4.2 Europe 92
5.15.4.3 Asia Pacific 92
5.15.5 IMPACT ON VERTICALS 93
6 AI INFERENCE PAAS MARKET, BY DEPLOYMENT 95
6.1 INTRODUCTION 96
6.2 PUBLIC CLOUD 97
6.2.1 RISING ADOPTION OF GEN AI AND LARGE LANGUAGE MODELS ACROSS INDUSTRIES TO ACCELERATE SEGMENTAL GROWTH 97
6.3 PRIVATE CLOUD 99
6.3.1 GROWING FOCUS ON ENTERPRISE CONTROL AND INTELLECTUAL PROPERTY PROTECTION TO FUEL SEGMENTAL GROWTH 99
6.4 HYBRID CLOUD 100
6.4.1 INCREASING COMPLEXITY OF AI WORKLOADS TO CONTRIBUTE TO SEGMENTAL GROWTH 100
7 AI INFERENCE PAAS MARKET, BY APPLICATION 103
7.1 INTRODUCTION 104
7.2 GENERATIVE AI 105
7.2.1 RULE-BASED MODELS 107
7.2.1.1 Strong focus on operational efficiency, governance, and regulatory compliance to bolster segmental growth 107
7.2.2 STATISTICAL MODELS 108
7.2.2.1 Increasing need for low-latency, real-time predictions to foster segmental growth 108
7.2.3 DEEP LEARNING 108
7.2.3.1 Rapid advances in hardware, optimized serving frameworks, and access to pre-trained models and repositories to drive market 108
7.2.4 GENERATIVE ADVERSARIAL NETWORKS (GANS) 109
7.2.4.1 Mounting demand for synthetic content, creative AI applications, and high-fidelity simulation environments to fuel

segmental growth 109
7.2.5 AUTOENCODERS 109
7.2.5.1 Rise in fraud detection and cybersecurity use cases to boost segmental growth 109
7.2.6 CONVOLUTIONAL NEURAL NETWORKS (CNNS) 109
7.2.6.1 Increasing adoption of visual AI across industries to contribute to segmental growth 109
7.2.7 TRANSFORMER MODELS 110
7.2.7.1 Growing demand for contextual AI and LLM-based productivity tools to augment segmental growth 110
7.3 MACHINE LEARNING 110
7.3.1 INCREASING AVAILABILITY OF STRUCTURED AND UNSTRUCTURED DATA TO ACCELERATE SEGMENTAL GROWTH 110
7.4 NATURAL LANGUAGE PROCESSING 111
7.4.1 RISING ADOPTION OF CHATBOTS, VOICE ASSISTANTS, AND TEXT ANALYTICS TO BOLSTER SEGMENTAL GROWTH 111
7.5 COMPUTER VISION 112
7.5.1 MOUNTING DEMAND FOR AUTOMATION OF SURVEILLANCE, MANUFACTURING, AND HEALTHCARE TO FUEL SEGMENTAL GROWTH 112
8 AI INFERENCE PAAS MARKET, BY VERTICAL 114
8.1 INTRODUCTION 115
8.2 HEALTHCARE 116
8.2.1 GROWING FOCUS ON COST-EFFICIENT CLINICAL WORKFLOWS AND DIGITAL HEALTH TO BOOST SEGMENTAL GROWTH 116
8.3 BFSI 123
8.3.1 RISING EMPHASIS ON REAL-TIME FRAUD DETECTION AND RISK ASSESSMENT TO FUEL SEGMENTAL GROWTH 123
8.4 AUTOMOTIVE 129
8.4.1 MOUNTING DEMAND FOR CONNECTED, AUTONOMOUS, AND ELECTRIC VEHICLE TECHNOLOGIES TO DRIVE MARKET 129
8.5 RETAIL & E-COMMERCE 135
8.5.1 INCREASING ADOPTION OF AI-DRIVEN MARKETING AND SALES ANALYTICS TO CONTRIBUTE TO SEGMENTAL GROWTH 135
8.6 MEDIA & ENTERTAINMENT 142
8.6.1 ESCALATING DIGITAL CONTENT CONSUMPTION TO ACCELERATE SEGMENTAL GROWTH 142
8.7 GOVERNMENT & DEFENSE 148
8.7.1 INCREASING INVESTMENT IN MODERNIZATION AND SOVEREIGN AI INITIATIVES TO FOSTER SEGMENTAL GROWTH 148
8.8 IT & TELECOM 154
8.8.1 RISING NEED TO OPTIMIZE INTERNAL OPERATIONS AND DELIVER INTELLIGENT SERVICES TO BOLSTER SEGMENTAL GROWTH 154
8.9 OTHER VERTICALS 160
9 AI INFERENCE PAAS MARKET, BY REGION 166
9.1 INTRODUCTION 167
9.2 NORTH AMERICA 168
9.2.1 MACROECONOMIC OUTLOOK FOR NORTH AMERICA 169
9.2.2 US 172
9.2.2.1 Rising deployment of AI-powered applications in industries to boost market growth 172
9.2.3 CANADA 173
9.2.3.1 Increasing government investment in compute infrastructure to accelerate market growth 173
9.2.4 MEXICO 174
9.2.4.1 Strong focus on building data center cluster to meet surging demand for enterprise cloud services to drive market 174
9.3 EUROPE 174
9.3.1 MACROECONOMIC OUTLOOK FOR EUROPE 175
9.3.2 GERMANY 179
9.3.2.1 Mounting demand for intelligent, scalable, and locally compliant platforms in manufacturing sectors to augment market growth 179

9.3.3 UK 179
9.3.3.1 Rising implementation of policies to strengthen compute capacity and enhance digital sovereignty to fuel market growth 179
9.3.4 FRANCE 180
9.3.4.1 Growing emphasis on scaling compute resources and strengthening data sovereignty to bolster market growth 180
?
9.3.5 ITALY 181
9.3.5.1 Increasing investment in sovereign compute infrastructure to contribute to market growth 181
9.3.6 SPAIN 181
9.3.6.1 Government-backed digital transformation initiatives to accelerate market growth 181
9.3.7 POLAND 182
9.3.7.1 Strategic investments in compute infrastructure and government-led digitalization to support market growth 182
9.3.8 NORDICS 183
9.3.8.1 High commitment to sustainable digital infrastructure and advanced connectivity to boost market growth 183
9.3.9 REST OF EUROPE 183
9.4 ASIA PACIFIC 184
9.4.1 MACROECONOMIC OUTLOOK FOR ASIA PACIFIC 184
9.4.2 CHINA 188
9.4.2.1 Large-scale AI adoption and strategic investments in computing infrastructure to contribute to market growth 188
9.4.3 SOUTH KOREA 189
9.4.3.1 Rapid digital transformation and rise of industry-leading technology firms to foster market growth 189
9.4.4 JAPAN 190
9.4.4.1 Large-scale investments in AI platforms to accelerate market growth 190
9.4.5 INDIA 190
9.4.5.1 Booming startup ecosystem and government-led digital initiatives to fuel market growth 190
9.4.6 AUSTRALIA 191
9.4.6.1 Robust cloud infrastructure and growing ecosystem of AI adopters to bolster market growth 191
9.4.7 INDONESIA 192
9.4.7.1 Increasing digital-first population and integration of AI into core industries to augment market growth 192
9.4.8 MALAYSIA 192
9.4.8.1 Expanding data center footprint to expedite market growth 192
9.4.9 THAILAND 193
9.4.9.1 Growing focus on digital sovereignty and data localization to bolster market growth 193
9.4.10 VIETNAM 193
9.4.10.1 Proliferating infrastructure investment and strong ecosystem of digital infrastructure to foster market growth 193
9.4.11 REST OF ASIA PACIFIC 194
9.5 ROW 194
9.5.1 MACROECONOMIC OUTLOOK FOR ROW 195
9.5.2 SOUTH AMERICA 198
9.5.2.1 Growing demand for scalable and cost-effective solutions to support digital transformation to drive market 198
9.5.3 AFRICA 198
9.5.3.1 South Africa 199
9.5.3.1.1 AI-driven transformation in healthcare and e-commerce sectors to accelerate market growth 199
9.5.3.2 Other African countries 199
9.5.4 MIDDLE EAST 200
9.5.4.1 Bahrain 202
9.5.4.1.1 Robust digital infrastructure and progressive regulatory environment to augment market growth 202

9.5.4.2 Kuwait 202
9.5.4.2.1 Increasing investment in digital infrastructure and government support for technological innovation to drive market 202
9.5.4.3 Oman 203
9.5.4.3.1 Strong commitment to diversifying the economy and enhancing technological capabilities to foster market growth 203
9.5.4.4 Qatar 203
9.5.4.4.1 Innovative smart city initiatives and commitment to digital transformation to accelerate market growth 203
9.5.4.5 Saudi Arabia 204
9.5.4.5.1 Increasing investment in AI infrastructure and focus on digital transformation to expedite market growth 204
9.5.4.6 UAE 204
9.5.4.6.1 Rising deployment of AI to enhance operational efficiency to contribute to market growth 204
9.5.4.7 Rest of Middle East 205
10 COMPETITIVE LANDSCAPE 206
10.1 OVERVIEW 206
10.2 KEY PLAYER STRATEGIES/RIGHT TO WIN, 2021-2025 206
10.3 REVENUE ANALYSIS, 2021-2024 209
10.4 MARKET SHARE ANALYSIS, 2024 210
10.5 COMPANY VALUATION AND FINANCIAL METRICS 213
10.6 BRAND COMPARISON 215
10.7 COMPANY EVALUATION MATRIX: KEY PLAYERS, 2024 216
10.7.1 STARS 216
10.7.2 EMERGING LEADERS 216
10.7.3 PERVASIVE PLAYERS 216
10.7.4 PARTICIPANTS 216
10.7.5 COMPANY FOOTPRINT: KEY PLAYERS, 2024 218
10.7.5.1 Company footprint 218
10.7.5.2 Region footprint 219
10.7.5.3 Deployment footprint 220
10.7.5.4 Application footprint 221
10.7.5.5 Vertical footprint 222
10.8 COMPANY EVALUATION MATRIX: STARTUPS/SMES, 2024 224
10.8.1 PROGRESSIVE COMPANIES 224
10.8.2 RESPONSIVE COMPANIES 224
10.8.3 DYNAMIC COMPANIES 224
10.8.4 STARTING BLOCKS 224
10.8.5 COMPETITIVE BENCHMARKING: STARTUPS/SMES, 2024 226
10.8.5.1 Detailed list of key startups/SMEs 226
10.8.5.2 Competitive benchmarking of key startups/SMEs 227
10.9 COMPETITIVE SCENARIO 229
10.9.1 PRODUCT LAUNCHES/ENHANCEMENTS 229
10.9.2 DEALS 230
11 COMPANY PROFILES 233
11.1 KEY PLAYERS 233
11.1.1 MICROSOFT 233
11.1.1.1 Business overview 233
11.1.1.2 Products/Solutions/Services offered 234
11.1.1.3 Recent developments 235
11.1.1.3.1 Product launches/enhancements 235

11.1.1.3.2 Deals 236
11.1.1.4 MnM view 237
11.1.1.4.1 Key strengths/Right to win 237
11.1.1.4.2 Strategic choices 237
11.1.1.4.3 Weaknesses/Competitive threats 238
11.1.2 AMAZON WEB SERVICES, INC. 239
11.1.2.1 Business overview 239
11.1.2.2 Products/Solutions/Services offered 240
11.1.2.3 Recent developments 241
11.1.2.3.1 Product launches/enhancements 241
11.1.2.3.2 Deals 241
11.1.2.4 MnM view 242
11.1.2.4.1 Key strengths/Right to win 242
11.1.2.4.2 Strategic choices 243
11.1.2.4.3 Weaknesses/Competitive threats 243
11.1.3 GOOGLE CLOUD 244
11.1.3.1 Business overview 244
11.1.3.2 Products/Solutions/Services offered 245
11.1.3.3 Recent developments 246
11.1.3.3.1 Product launches/enhancements 246
11.1.3.3.2 Deals 246
?
11.1.3.4 MnM view 247
11.1.3.4.1 Key strengths/Right to win 247
11.1.3.4.2 Strategic choices 248
11.1.3.4.3 Weaknesses/Competitive threats 248
11.1.4 ORACLE 249
11.1.4.1 Business overview 249
11.1.4.2 Products/Solutions/Services offered 251
11.1.4.3 Recent developments 252
11.1.4.3.1 Product launches/enhancements 252
11.1.4.3.2 Deals 252
11.1.4.4 MnM view 253
11.1.4.4.1 Key strengths/Right to win 253
11.1.4.4.2 Strategic choices 253
11.1.4.4.3 Weaknesses/Competitive threats 254
11.1.5 IBM 255
11.1.5.1 Business overview 255
11.1.5.2 Products/Solutions/Services offered 256
11.1.5.3 Recent developments 258
11.1.5.3.1 Product launches/enhancements 258
11.1.5.3.2 Deals 258
11.1.5.4 MnM view 259
11.1.5.4.1 Key strengths/Right to win 259
11.1.5.4.2 Strategic choices 259
11.1.5.4.3 Weaknesses/Competitive threats 259
11.1.6 ALIBABA CLOUD 260

11.1.6.1 Business overview 260
11.1.6.2 Products/Solutions/Services offered 261
11.1.6.3 Recent developments 261
11.1.6.3.1 Product launches/enhancements 261
11.1.6.3.2 Deals 262
11.1.6.3.3 Other developments 262
11.1.7 SALESFORCE, INC. 263
11.1.7.1 Business overview 263
11.1.7.2 Products/Solutions/Services offered 265
11.1.7.3 Recent developments 265
11.1.7.3.1 Product launches/enhancements 265
11.1.7.3.2 Deals 266
11.1.8 TENCENT CLOUD 267
11.1.8.1 Business overview 267
11.1.8.2 Products/Solutions/Services offered 267
11.1.9 BAIDU, INC. 269
11.1.9.1 Business overview 269
11.1.9.2 Products/Solutions/Services offered 270
11.1.9.3 Recent developments 271
11.1.9.3.1 Deals 271
11.1.10 TOGETHER AI 272
11.1.10.1 Business overview 272
11.1.10.2 Products/Solutions/Services offered 272
11.1.10.3 Recent developments 273
11.1.10.3.1 Product launches/enhancements 273
11.1.10.3.2 Deals 273
11.2 OTHER PLAYERS 274
11.2.1 COREWEAVE 274
11.2.2 PREDIBASE 275
11.2.3 VECTARA 276
11.2.4 PREM AI 277
11.2.5 BASETEN 278
11.2.6 C3.AI, INC. 279
11.2.7 CLOUDFLARE, INC. 280
11.2.8 XFERENCE SRL 281
11.2.9 H2O.AI 282
11.2.10 DATAROBOT, INC 283
11.2.11 CEREBRAS 284
11.2.12 CLOUDERA, INC. 285
11.2.13 GROQ, INC. 286
11.2.14 SAMBANOVA, INC. 287
11.2.15 LATENT AI 288
11.2.16 MODULAR INC 289
11.2.17 FIREWORKS AI, INC. 290
11.2.18 DEEP INFRA 291
11.2.19 REPLICATE 292
11.2.20 ANYSCALE, INC 293

11.2.21 FEATHERLESS.AI 294
11.2.22 RAFAY SYSTEMS, INC. 295
12 APPENDIX 296
12.1 INSIGHTS FROM INDUSTRY EXPERTS 296
12.2 DISCUSSION GUIDE 296
12.3 KNOWLEDGESTORE: MARKETSANDMARKETS? SUBSCRIPTION PORTAL 299
12.4 CUSTOMIZATION OPTIONS 301
12.5 RELATED REPORTS 301
12.6 AUTHOR DETAILS 302

SCOTTS INTERNATIONAL

AI Inference Platform-as-a-Service (PaaS) Market by Deployment (Private Cloud, Public Cloud, Hybrid Cloud), Application (Gen AI, Machine Learning, NLP, Computer Vision), Vertical (BFSI, IT & Telecom, Retail & E-commerce), Region - Global Forecast to 2030

Market Report | 2025-10-03 | 303 pages | MarketsandMarkets

To place an Order with Scotts International:

□ - Print this form

□ - Complete the relevant blank fields and sign

□ - Send as a scanned email to support@scotts-international.com

ORDER FORM:

Select license	License	Price
	Single User	$4950.00
	Multi User	$6650.00
	Corporate License	$8150.00
	Enterprise Site License	$10000.00
	VAT	
	Total	

*Please circle the relevant license option. For any questions please contact support@scotts-international.com or 0048 603 394 346.

□** VAT will be added at 23% for Polish based companies, individuals and EU based companies who are unable to provide a valid EU Vat Numbers.

Email* Phone*

First Name* Last Name*

Job title*

Company Name* EU Vat / Tax ID / NIP number*

Address* City*

Zip Code*

Country*

Date 2026-02-09

Signature