CIC灼识

Global AI Inference Chip Industry Report

Executive Summary

Training and inference are two primary computing tasks for AI chips. Training involves processing massive amounts of data and optimizing parameters to build models before they are used in real-life applications. As a result, training was the main focus during the early stages of AI SoC industry development. However, as AI models, especially LLMs, have advanced in both performance and practicality, demands have extended – the industry is now placing greater emphasis on real-world applications, with AI inference chips gaining increasing traction.

Table of Contents

1. Market Overview

1.1 Market Definition

As the AI chip market is undergoing a fundamental shift from a training-centric to an inference-centric paradigm, the demand for AI inference chips is surging.

AI inference chips are designed for high energy efficiency and low latency, enabling near-instantaneous outcomes.

1.2 Market Structure

AI inference chips can be deployed in the cloud, edge, and on-device scenarios, with each scenario demanding tailored chip designs for different requirements:

Cloud AI inference chips are typically used in data centers, responsible for handling large-scale, high-density, and high-concurrency centralized inference tasks. As such, these chips prioritize high computing power, broad applicability, flexibility, and scalability.

Edge AI inference chips are deployed in edge servers, gateways, or base stations closer to data sources. They perform real-time local inference, which on the other hand demands a careful balance

between strong performance and power efficiency to ensure low latency, data security, and operational stability.

On-device AI inference chips are used directly in end-user devices, such as consumer electronics like smartphones, smart vehicles, and smart home applications.

1.3 Market Size and Growth

In 2024, the global AI inference chip market reached RMB 606.7 billion. The segment breakdown is as follows:

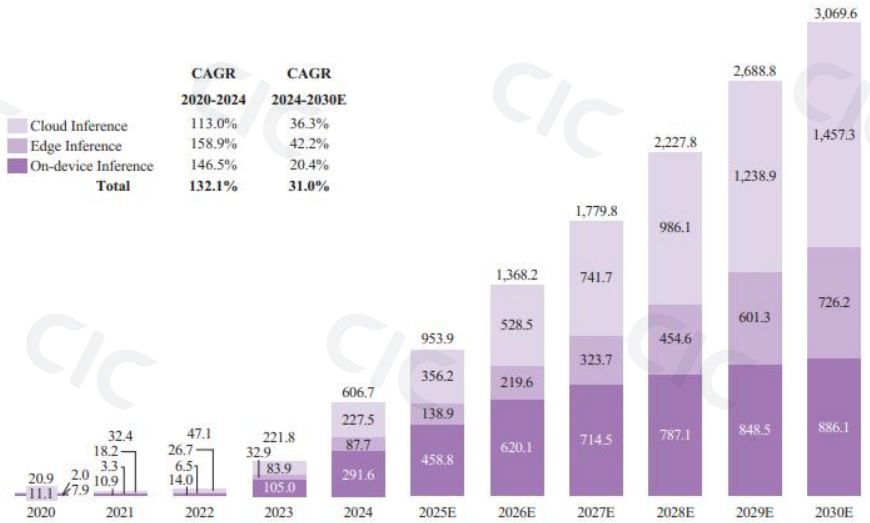
On-device Inference: RMB 291.6 billion

Cloud Inference: RMB 227.5 billion

Edge Inference: RMB 87.7 billion

Global AI Inference Chip Market Size by Cloud, Edge, and On-device, 2020-2030E

(RMB Billion)



In 2024, the China AI inference chip market reached RMB 160.8 billion.

The segment breakdown is as follows:

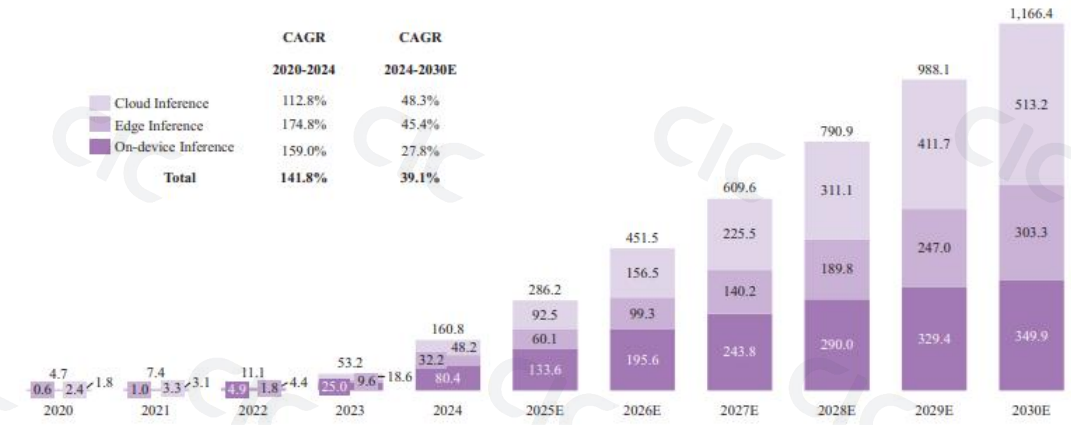
On-device Inference: RMB 80.4 billion

Cloud Inference: RMB 48.2 billion

Edge Inference: RMB 32.2 billion

China's AI Inference Chip Market Size by Cloud, Edge, and On-device, 2020-2030E

(RMB Billion)



Source: CIC Reports, interviews with market participants, industry publications, government statistics, listed companies' public filings, news, etc.

2. Key Growth Drivers and Trends

2.1 Key Drivers

AI-enabled smart devices increase: The global penetration rate of AI-enabled smart devices grew from under 1% in 2020 to 9.4% in 2024, and is expected to exceed 44% by 2030, fueling greater demand for edge-based and on-device AI inference computing, positioning edge and on-device AI inference chips as critical enabler of this intelligent transformation.

Surging data volumes and low latency demands: Real-time applications like smart driving, robotics, and industrial control demand millisecond-level processing that traditional cloud architectures often fail to deliver due to latency and bandwidth constraints. By deploying AI inference chips directly at the edge, systems can process data locally, enabling instant, reliable, and coordinated responses. Consequently, edge inference has evolved into essential infrastructure, fueling the rapid rise of increasingly intelligent devices.

Data compliance driving localized processing: Stricter global data regulations have transformed data into a high-stakes strategic asset. To balance operational efficiency with rigorous compliance,

enterprises are increasingly adopting localized processing over cloud-based transfers. Consequently, edge and on-device AI inference chips have emerged as critical infrastructure, providing the security and closed-loop data handling necessary to meet modern regulatory requirements.

2.2 Future Outlook

Global AI inference chip market is projected to reach RMB 3,069.6 billion by 2030, at a CAGR of 31.0% (2024-2030E).

China's AI inference chip market is projected to reach RMB 1,166.4 billion by 2030, at a CAGR of 39.1% (2024-2030E).

About CIC

CIC is a professional consulting firm offering tailored end-to-end support across the full investment and financing lifecycle. The firm boasts a world-leading track record in guiding landmark first-in-sector IPOs across global markets, alongside unrivaled reach and in-depth coverage capabilities across specialized niche market segments.

CIC helps enterprises refine scalable business models and craft compelling capital narratives to enable seamless access to global capital markets, while serving as a trusted due diligence partner to investment institutions. It delivers granular industry insights and direct access to subject matter experts, empowering clients to identify high-value opportunities and mitigate critical risks effectively.

CIC team maintains deep, real-time market intelligence across a diverse set of sectors—including financial services, artificial intelligence, big data, internet, high technology, healthcare, education, entertainment, consumer goods, transportation and logistics, energy and power, environmental and building technology,

chemicals, industrial manufacturing, and agriculture—delivering unparalleled access to sector-specific, actionable insights.

CIC Reports & Industry Overview

At CIC, we employ a rigorous, multi-method research framework, combining primary and secondary sources to underpin our analysis. Primary research involves in-depth engagements with industry thought leaders and practitioners, particularly in supply chain finance. Secondary research synthesizes publicly available datasets from authoritative bodies, including the National Bureau of Statistics of the People's Republic of China, the State Administration of Financial Regulation (SAFR, formerly the China Banking and Insurance Regulatory Commission), the China Securities Regulatory Commission (CSRC), and public company filings. We apply proprietary data analytics frameworks to process collected information, validating findings through cross-referencing data from multiple research streams to ensure analytical rigor and reliability.

All statistical data presented is verifiable and grounded in information available as of the date of this report.

Extracts are refined summaries of in-depth CIC industry research reports, highlighting supply and demand trends, key growth drivers, R&D trends and future outlook, etc. of various segmented fields, integrating multi-dimensional insights such as expert interviews, market surveys and industry data analysis.

Disclaimer

This report (the "Report") is prepared by CIC based on information available as of the date hereof. The Report is furnished strictly for informational and reference purposes only and is not intended to, nor shall it be construed as, being definitive or conclusive. Nothing contained herein shall constitute or be deemed to constitute investment advice, a recommendation, or an offer, solicitation or inducement to engage in any investment activity. CIC hereby expressly disclaims any and all liabilities for any loss, damage or claims of any nature howsoever arising, whether directly or indirectly, from the use of or reliance upon any information contained in the Report.

Contact CIC

For more information about this report or to learn more about CIC services, please visit CIC official website, or email us at marketing@cninsights.com.