

A high-angle, isometric view of the QNAP QAI-h1290FX AI platform. The system consists of a large, black, rectangular server unit with a front panel featuring multiple drive bays and a small display. On top of the server, there are two large, silver, fan-cooled components, likely GPUs. To the left of the server, a monitor is positioned, displaying a colorful, abstract, fractal-like image. The entire setup is placed on a dark, textured surface that resembles a server rack or a specialized base. The background is dark with some green light streaks, suggesting a data center or high-tech environment.

# QNAP

# QAI-h1290FX Fully On-Premises AI Platform

Powered by AMD EPYC™ 7302P 16-core/32-thread processor with optional NVIDIA® Blackwell GPU support. Run LLM inference, RAG knowledge bases, and Agentic AI workflows entirely within your own environment — no cloud dependency required.



**Absolute Edge AI**  
Absolute Confidence

## Specifications

### Processor

AMD EPYC™ 7302P  
16-core/32-thread

### GPU

NVIDIA® Blackwell  
(Optional)

### Memory

Up to 1TB RDIMM  
DDR4 ECC

### Network

2x 25GbE SFP28  
+ 2x 2.5GbE RJ45

### Storage

12 x 2.5-inch U.2 PCIe  
NVMe / SATA SSD

# Why QAI-h1290FX?

**In 2026, AI is everywhere —  
but data sovereignty & compute autonomy  
cannot be compromised.**

Open-source models such as Gemma 4, Qwen3.6, and recent releases like Mistral Small 4 and Olmo Hybrid continue to raise the bar, making private on-premises AI more accessible than ever. QAI-h1290FX delivers complete on-premises AI compute power, enabling LLM inference, knowledge base Q&A, and Agentic workflows entirely within your own infrastructure — with predictable costs and guaranteed compliance.



## **On-Premises Private AI Assistant Empower Every Employee**

Deploy leading open-source models — Gemma 4 and Qwen3.6 — to give your entire workforce a secure, private AI assistant for Q&A, document summarization, and content generation. Integrate with enterprise applications via OpenClaw and Hermes. All conversations and data stay 100% within your premises — fully compliant, zero cloud exposure.

## **Agentic AI Automation Complex Tasks, Completed On-Premises**

Run Agentic AI on the QAI platform with OpenClaw, Hermes, and on-premises AI models — AI autonomously handles multi-step business tasks: compiling weekly reports, detecting order anomalies, triggering compliance approvals — all executed within your internal network, with no external API dependency and no interruption from network outages.

## **Industries Served**

### **Finance & Insurance**

Compliance document review, risk model inference — customer data never leaves the firewall.

### **Manufacturing**

Production anomaly prediction, technical document Q&A, supplier contract analysis.

### **Healthcare & Life Sciences**

Clinical note summarization, drug research retrieval — fully compliant with medical data regulations.

### **Education & Research**

Shared AI compute pools, research model deployment, academic knowledge base construction.

## **Private Enterprise Knowledge Base Your Documents, Made Conversational**

Vectorize your SOPs, contracts, and regulatory documents into a fully on-premises private knowledge base. Employees ask in natural language; AI retrieves and answers with precise, cited responses from your own data — eliminating manual search and accelerating decision-making across the organization.

## **Shared AI Compute Pool One Machine, Serving the Entire Organization**

QAI-h1290FX supports optional NVIDIA® Blackwell GPU. For most AI inference workloads, acceleration can be used directly in on-premises environments such as containers without GPU Passthrough. GPU Passthrough is only required for VM-based deployments such as Virtualization Station. IT administrators can allocate compute resources flexibly by workload priority — one-time investment replacing ever-growing cloud API bills.

## **Key Benefits**

### **Data Never Leaves Your Premises**

Complete on-premises inference architecture — fully compliant with GDPR, financial regulations, and data privacy laws. Zero compromise on security.

### **Transparent, Predictable Cost**

Replace metered cloud API billing with a one-time hardware investment. At scale, total cost of ownership is significantly lower than public cloud AI services.

### **Ready Out of the Box**

Supports OpenClaw and Hermes AI Agent platforms. No MLOps expertise required — IT teams can go live within hours of deployment.

**QNAP SYSTEMS, INC.**

Copyright © 2026 QNAP Systems, Inc. All rights reserved.



Explore QAI-h1290FX



Sales Inquiry