AI in Retail – Unlock Growth Opportunities

AI Fuelling Sector's Transformation

- As of 2024, 42 percent of retailers are actively utilizing AI, with an additional 34 percent[1] in the stages of assessment or pilot testing.
- Generative AI solutions, leveraging Large Language Models (LLMs) for advanced natural language processing, can generate an astounding $660 billion[2] in value for the retail sector.

Common Pain Points of Adopting AI for The Retail Industry

- Insufficient or low data quality impact AI insights.
- Integrating AI with outdated legacy systems often leads to compatibility issues, and scalability concerns.
- High latency and low responsiveness can impact a customer's shopping experience.
- Brick and mortar store environmental conditions such as noise, lighting, etc. can impact AI processes like interpreting images from computer vision.
- Scarcity of tech talent, high employee turnover and adoption resistance by frontline workers.
- If not handled properly, cybersecurity and physical security of edge devices can be vulnerable to attacks.

Supermicro with NVIDIA Empowering Your AI Journey

Supermicro and NVIDIA® deliver best-in-class outcomes for predictive and generative AI implementations in the retail sector. This collaborative approach provides a comprehensive framework that integrates CPUs, GPUs, and optimized memory, all orchestrated within the resilient infrastructure of Supermicro's platforms. These AI solutions serve applications across multiple use cases in retail. Some of the common use cases you can adopt today are:

Key Use Cases of AI in Retail

	Relevant Supermicro Systems
Personal Shopping Advisor: AI-enabled shopping advisors bring a new experience to shoppers, improve customer ratings, removes friction for sales and boosts revenue for retailers.	**Edge – Kiosk Operation Server:** Supermicro SYS-110P-20C-FRAN8TP with NVIDIA L40 GPU | **Back-office- LLM Host Server:** Supermicro Hyper-E SYS-221HE-TNR(D) with dual NVIDIA RTX A6000 ADA.
Retail Loss Prevention: Computer vision and AI for security can help retailers identify and prevent retail loss and product shrinkage, while also integrating with inventory management.	**Edge - AI Computer Vision:** Supermicro E300 Embedded System with NVIDIA L4 GPU, Deep stream SDK | **Datacenter – AI Evaluator with similarity search:** Supermicro ARS-111GL-NHR with NVIDIA Grace Hopper Superchip.
Store Analytics: AI-driven store analytics can help by predicting and preventing stockouts, optimize merchandising through precise assessment of planogram performance, and streamline store operations via real-time data insights.	**Edge-AI Computer Vision:** Supermicro E403-13E-FRN2T with NVIDIA L40S GPU | **Datacenter - AI vector search backend**: Supermicro HGX Systems with NVIDIA H100/ H200 GPUs.
Intelligent Supply Chain: AI and simulation solutions enhance supply chain efficiency and intelligence, helping retailers align inventory with projected demand.	**Datacenter – Digital Twins warehouse simulation server:** Supermicro 5U PCIe GPU system SYS-521GE-TNRT with NVIDIA RTX 6000 ADA GPUs.
AI-powered Employee Assistance: AI can modernize the workplace by simplifying tasks, improving information access, and providing dynamic guidance for a more efficient work environment.	**Datacenter-Retrieval-Augmented Generation with custom LLM:** Supermicro HGX Systems with NVIDIA H100/ H200 GPUs, NVIDIA AI Enterprise.

[1]State of AI in Retail and CPG | NVIDIA | [2]Game On, Retailers: Elevate your customer Experience with Gen AI

Supermicro Solutions for Powering Retail AI

	Small	Medium	Large	Extra Large
				
	SYS-E300-13AD	SYS-110D-20C-FRAN8TP*	SYS-E403-13E-FRN2T	SYS-221HE-TNR(D)
Dimensions (mm)	H W D	H W D	H W D	H W D
	43.0 x 264.8 x 225.8	43.0 x 437.0 x 399.0	117.3 x 266.7 x 406.4	88.9 x 436.9 x 574.0
	Single NVIDIA A2, L4 or RTX A1000	**Single NVIDIA L4 or RTX 4000 Ada**	**Single NVIDIA L40S, L40, or RTX 6000 Ada**	**Multiple H100PCIe, L40S, or RTX 6000 Ada**
Key Features for Predictive AI				
	AI computer vision for up to 8 streams[1]	AI computer vision for up to 16 streams[1]	AI computer vision for up to 32 streams[1]	AI computer vision for up to 48 streams per GPU[1]
	Up to ~2,000 automatic speech recognition (ASR) samples per second[2]	Up to ~6,000 automatic speech recognition (ASR) samples per second[2]	Up to ~22,000 automatic speech recognition (ASR) samples per second[2]	Up to ~32,000 automatic speech recognition (ASR) samples per second per GPU[2]
Key Features for Generative AI				
	LLM up to 8 billion parameters	LLM up to 24 billion parameters	LLM up to 48 billion parameters	LLM up to 80 billion parameters
	Image and video generation with Stable Diffusion at ~1 image every 4-5 seconds.	Image and video generation with Stable Diffusion at ~1 image every 2 seconds.	Image and video generation with Stable Diffusion at ~1-2 images per second.	Image and video generation with Stable Diffusion at ~3 images per second.

*Additional models include SYS-110D-4C-FRAN8TP, SYS-110D-8C-FRAN8TP, SYS-110D-14C-FRAN8TP, and SYS-110D-16C-FRAN8TP

[1] Based on using an image classification model similar to EfficientNet-B4, dependent on video stream compression and other workloads on the system

[2] Based on using an ASR model like QuartzNet

Supermicro Systems Accelerating Your AI journey

Supermicro's cutting-edge AI-ready infrastructure solutions enable large-scale training to intelligent edge inferencing allowing retailers to streamline and accelerate AI deployment. Their AI-infrastructure empower workloads with optimal performance and scalability while optimizing costs and minimizing environmental impact.

Supermicro's flexible range of solutions ensures that retailers implementing AI solutions can scale up their implementation as much as needed. Whatever the requirements, solutions are available to expand memory, processing power, and storage to meet any situation.

Selecting The Optimal Supermicro Systems for Your AI Applications

Supermicro and NVIDIA excel in guiding organizations to select the right system for their specific AI applications. This support involves considering factors like the size of AI models, system compatibility, and specific use case requirements. Whether it's handling large-scale video analytics of a smart store or processing LLMs in building chatbots, Supermicro's portfolio of edge platforms are available across a wide range of form factors, including a compact box, 1U and 2U, wall mount, rack mount, and even fanless models. When combined with NVIDIA's powerful AI and computing platforms, they provide a range of solutions to meet these diverse needs effectively.

For more information, visit:
https://www.supermicro.com/en/solutions/ai-deep-learning