AI in Financial Services – Unlock Growth Opportunities

Opportunities for AI in Financial Services

- As of 2024, 51 percent of financial services executives surveyed "strongly agree" that AI would be important to their company's future success, a 76 percent increase from last year.[1]
- According to a McKinsey study, the impact of generative AI on the banking industry in terms of revenue could be between $200–340 billion annually.[2]

Common Pain Points of Adopting AI for The Financial Services Industry

- **Data Privacy:** AI systems require huge quantities of often sensitive customer financial data, such as bank account details and transaction histories. If the data is not handled correctly, it can result in significant compliance issues.
- **Work Force Challenges:** Large tech organizations struggle to hire experienced AI developers since innovative AI solutions require a skilled workforce to generate value. The availability of skilled talent remains a challenge.
- **Data Challenges for Model Training and Accuracy:** Ensuring sufficient and consistent data is critical for AI model training. Without an adequate volume of accurately labeled and consistent data, models risk overfitting or underfitting, leading to flawed insights and unreliable decision-making, particularly when applied across different industries.
- **Legacy Systems and Infrastructures:** Legacy systems often lack compatibility with modern AI technologies, making integration complex and costly. This can slow down the deployment of AI solutions and limit their effectiveness.

Supermicro with NVIDIA: Empower Your AI Journey

Supermicro and NVIDIA® deliver best-in-class outcomes for Predictive and Generative AI implementations in the financial services 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 financial services. Some of the common use cases you can adopt today are:

Key Use Cases of AI in Financial Services

- **Quant Finance:** Leverage mathematical models and data analysis to analyze trends, forecast asset valuations, manage risk, and optimize investment portfolios. This approach integrates quantitative modeling with data science to make intelligent business decisions.

- **Smarter Trading with Alternative Data:** Alternative data sets such as consumer transactions, logistics information, and employment trends provide an informational edge that improves trading strategies. AI analyzes large volumes of financial data to generate predictive insights, optimize trading strategies, visualize anomalies, and automate decision-making processes in trading operations.

- **KYC, AML and Fraud Prevention Security:** AI helps streamline customer verification programs for Know Your Customer (KYC) processes, while offering transaction monitoring that scans global payment networks and flags credit card transactions when needed. Additionally, AI can collect evidence to protect customers and help ensure compliance with efforts such as Anti-Money Laundering (AML).

- **Intelligent Document Automation:** AI can streamline financial processes by automating document handling, document extraction, and document validation. It enhances efficiency, reduces errors, and helps ensure compliance in areas such as loan processing, contract management, and regulatory reporting. Insurance companies can speed claims intake, review photos and videos submitted by customers, and make decisions faster with high level analysis and decision making.

- **Customer Experience with Chatbots:** GenAI-enabled chatbots and virtual assistants improve customer interactions, as organizations can be more efficient with their support staff and deliver more tailored, customer-aware engagements such as personalized financial recommendations for customers in-app and over email.

Relevant Supermicro Systems

SYS-521GE-TNRT (Intel),
AS -4125GS-TNRT (AMD),
Air: ARS-111GL-NHR ,
Liquid: ARS-111GL-NHR-LLC,
ARS-111GL-DNHR-LCC (2-node),
ARS-221GL-NHIR (1-node 2x GH200 with NVLink),
Air: SYS-821GE-TNHR (Intel),
AS -8125GS-TNHR (AMD),
Liquid: SYS-421GE-TNHR2-LCC (Intel),
AS -4125GS-TNHR2-LCC (AMD)

SYS-221H-TNR (Intel),
AS -2025HS-TNR (AMD),
AS -2015HS-TNR (AMD),
SYS-221GE-NR (Intel),
ARS-221GL-NR (NVIDIA),
SYS-521GE-TNRT (Intel),
AS -4125GS-TNRT (AMD),
Air: ARS-111GL-NHR ,
Liquid: ARS-111GL-NHR-LLC,
ARS-111GL-DNHR-LCC (2-node),
ARS-221GL-NHIR (1-node 2x GH200 with NVLink),
Air: SYS-821GE-TNHR (Intel),
AS -8125GS-TNHR (AMD),
Liquid: SYS-421GE-TNHR2-LCC (Intel),
AS -4125GS-TNHR2-LCC (AMD)

[1]State of AI in Financial Services | NVIDIA | [2]Scaling GenAI in Banking: Choosing the Best Operating Model

Supermicro Solutions for Powering Financial Services AI

	Air-cooled 2U Hyper Systems (Intel or AMD CPU)	Air-cooled 2U MGX Systems (Intel or NVIDIA Grace CPU Superchip)	Air-cooled 4U/5U PCIe GPU Systems (Intel or AMD CPU)	Air-cooled or Liquid-cooled 1U MGX Systems with NVIDIA GH200 Grace Hopper Superchip	Air-cooled 2U MGX Systems with 2x NVIDIA GH200 Grace Hopper Superchip	Air-cooled 8U or Liquid-cooled 4U Systems with NVIDIA HGX H100 8-GPU (Intel or AMD CPU)
						
X13, H13 Models						
	SYS-221H-TNR (Intel), AS -2025HS-TNR (AMD), AS -2015HS-TNR(AMD)	SYS-221GE-NR (Intel), ARS-221GL-NR (NVIDIA)	SYS-521GE-TNRT (Intel), AS -4125GS-TNRT (AMD)	Air: ARS-111GL-NHR, Liquid: ARS-111GL-NHR-LLC, ARS-111GL-DNHR-LCC (2-node)	ARS-221GL-NHIR (1-node 2x GH200 with NVLink)	Air: SYS-821GE-TNHR (Intel), AS -8125GS-TNHR (AMD), Liquid: SYS-421GE-TNHR2-LCC (Intel), AS -4125GS-TNHR2-LCC (AMD)
Form Factor						
	2U	2U, OCP compatible	4U, 5U	1U	2U	Air: 8U | Liquid: 4U
Power requirements/power draw						
	3kW	4kW	6kW	1kW (1U 1-node) 2kW (1U 2-node)	2kW	10kW
Cooling						
	Air cooling	Air cooling	Air cooling	Air or liquid cooling	Air cooling	Air or liquid cooling
Max recommended GPUs						
	3 (L40S, L40)	4 (L40S, H100 NVL)	10 (L40S, H100, H100/H200 NVL)	2 for 2-node system 1 for others (GH200)	2 with NVLINK (GH200)	8 with NVLINK & NVSWITCH (HGX H100)
LLM Training						
	–	–	✓	–	✓	✓
LLM Fine Tuning (max model) and RAG, Inference (max model)						
	LLAMA 3.1 8B	LLAMA 3.1 70B	LLAMA 3.1 70B	LLAMA 3.1 70B	LLAMA 3.1 70B	LLAMA 3.1 405B

- Quant Finance
- Smarter Trading with Alternative Data
- KYC, AML and Fraud Prevention Security customer
- Intelligent Document Automation
- Customer Experience with Chatbots

Supermicro Systems Accelerating Your AI journey

Supermicro's cutting-edge AI-ready infrastructure solutions helps with **large-scale training to enterprise level inferencing** enabling financial services 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 financial services organizations 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 such as the size of AI models, system compatibility, and specific use case requirements. Whether it's handling large-scale data of a financial institution or processing LLMs in building chatbots, Supermicro's portfolio of platforms are available across a wide range of form factors. 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