Harnessing Generative AI for Investment Excellence

Tao Jin, Founder and CTO
Pyligent AI

New York, 2024.10

Financial Machine Learning vs Gen AI

Financial ML vs. Gen AI

Purpose

- Primarily focused on predictive modeling, optimization, and risk management
- Primarily focused on generating new content, such as text, images, or code.

Data

- Typically requires labeled data for supervised learning or unlabeled data for unsupervised learning
- Often relies on large amounts of text data to learn patterns and generate new content.

Outputs

- Produces predictions, classifications, or optimized solutions
- Generates formatted text, code, original content that is similar to the training data

Key Techniques

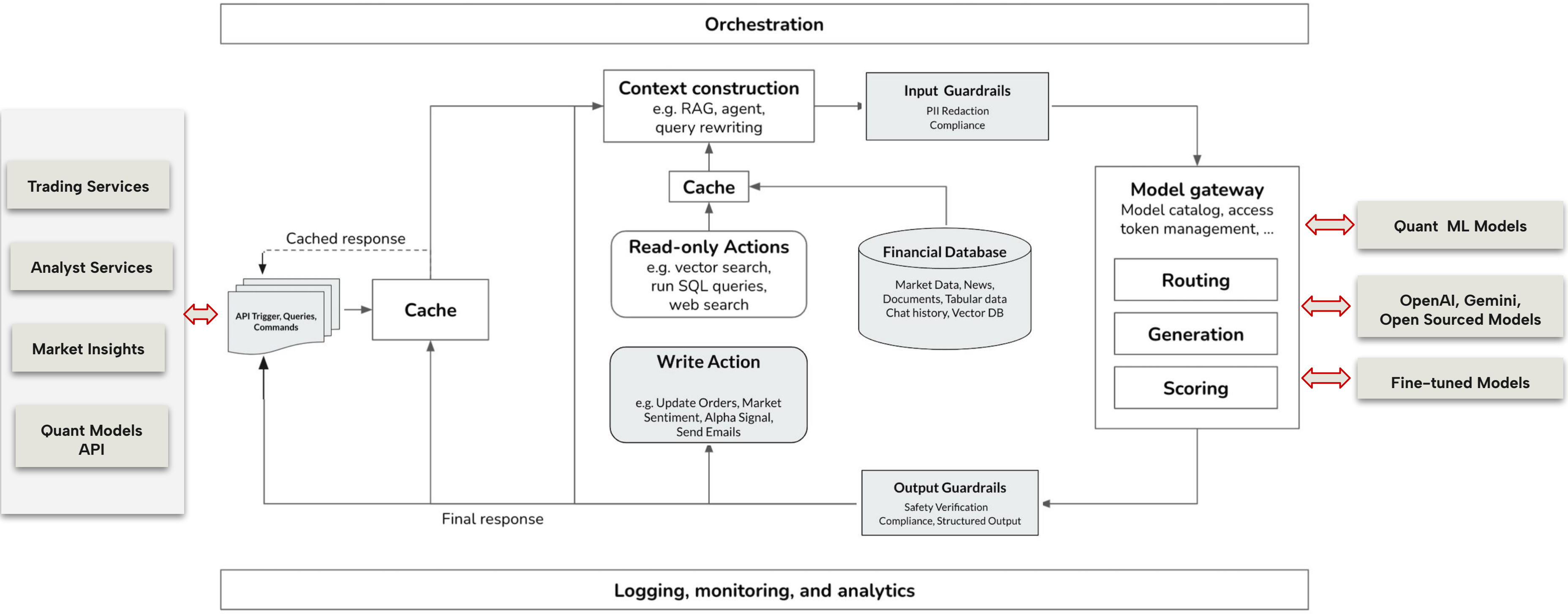
- Regression, classification, clustering, dimensionality reduction, bootstrap, bagging, ensemble
- Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), transformer-based Foundation models

Feature	Quant Machine Learning Models	Foundation Models
Purpose	Primarily designed for financial tasks like price prediction, risk assessment, and portfolio optimization.	General-purpose models capable of a wide range of tasks, including text generation, translation, and summarization.
Training Data	Typically focused on financial data, such as stock prices, market indices, and company financials.	Trained on massive amounts of text data from a variety of sources, such as books, articles, and code.
Architecture	Often based on traditional machine learning algorithms like regression, classification, and time series analysis.	Typically based on transformer architectures, which are designed to process sequential data.
Customization	Highly customizable to specific financial tasks and datasets.	Can be fine-tuned for specific tasks, but may require more data and computational resources.
Interpretability	Often more interpretable due to the use of traditional algorithms and feature engineering techniques.	Can be less interpretable due to their complex architectures and large number of parameters.
Computational Resources	Generally require less computational power compared to foundation models.	Often require significant computational resources, especially for large-scale models.

How Gen AI play a role in Quant Finance?

"Generative AI is poised to revolutionize quantitative finance by providing novel tools for **data augmentation** , **market insights** , and **risk assessment** . Its ability to generate realistic financial scenarios and **extract meaningful features from unstructured data (and structured data)** will empower quants to make more informed and innovative decisions."

Gen AI Platform Architecture

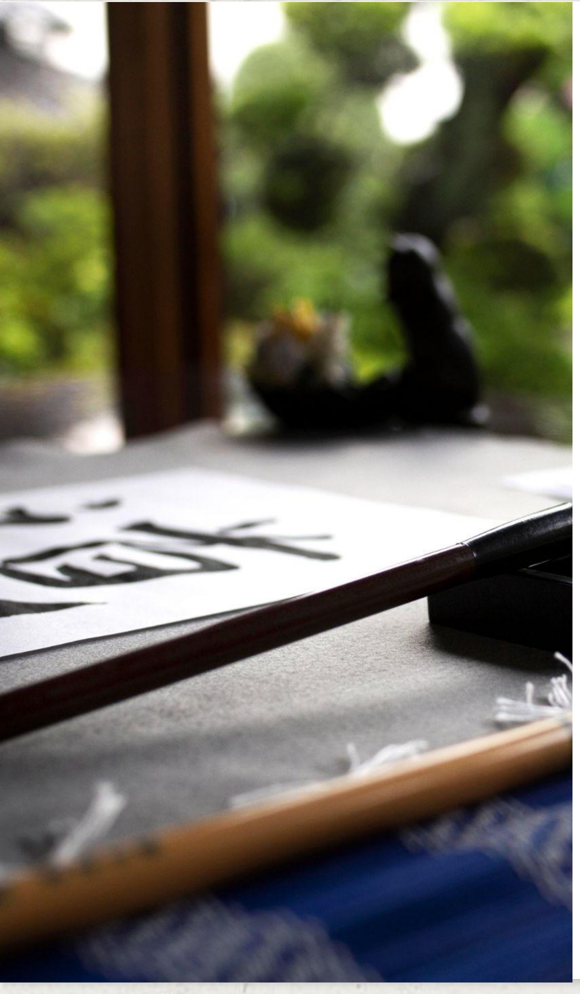


Content Services
RAG Services, Agentic RAGs
Trading, Analyst, Market Insights

Guardrails
Migrate AI risks, Regulator Compliance

Model Gateway
Response the different queries and interface with the customized models in a unified and secure manner.

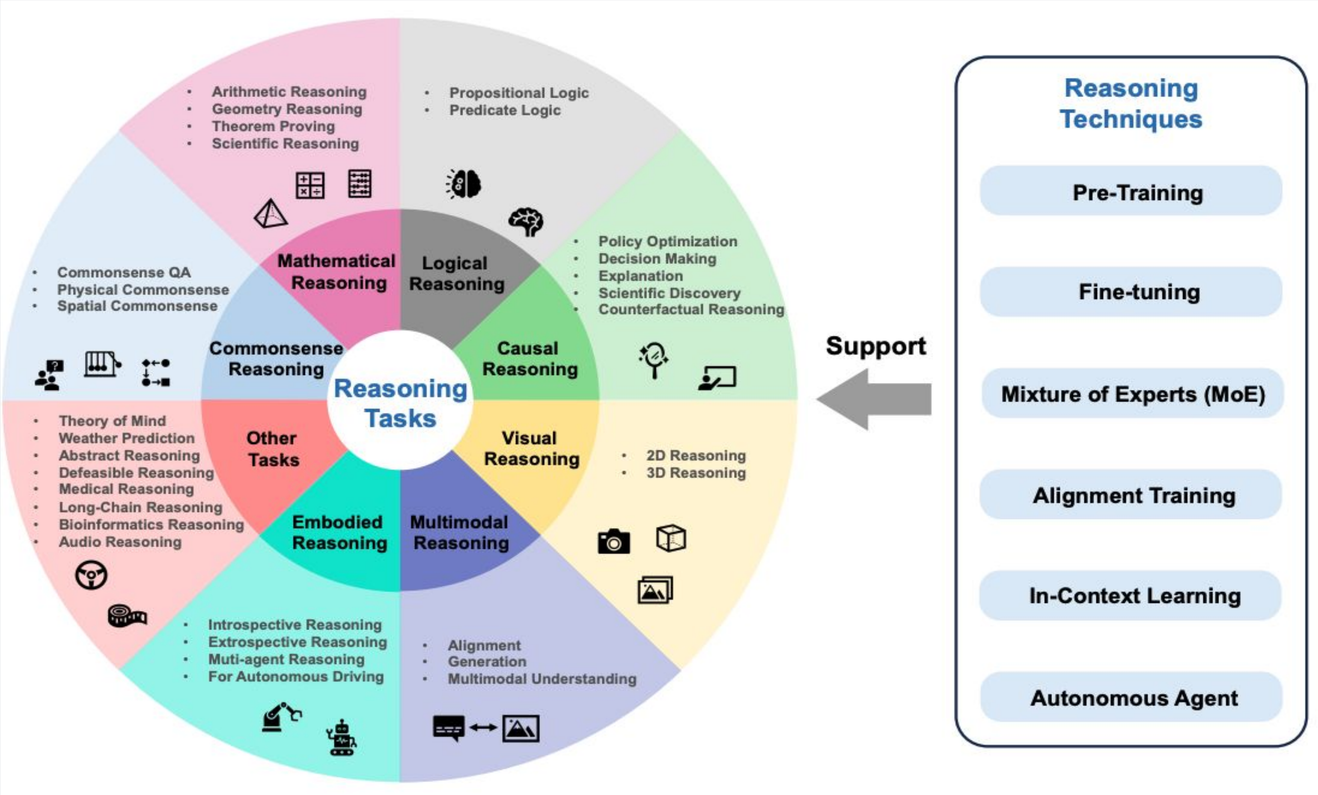
Cache Services
Reduce latency in production



Gen AI Agents

- Generative AI **Agents** combines the power of large language models (LLMs) and retrieval-augmented generation (RAG) with financial data, letting users query/execute action from diverse domain knowledge bases.
- **Conversational and QA Agents** : Query Market Information
- **Data Analytics Agents** : Produce the Market Insights, Tabular data Process
- **Trading Services, Analyst Services and CS support**
- Agent Types: **multi-agent collaboration, Self-Improving Agent, Task-oriented Agent**

Reasoning in Gen AI



Don't **Overestimated** Reasoning in LLM/Gen AI due to memorization rather than true understanding.

Risk Assessment:

Evaluating the impact of different factors on financial risk.

Portfolio Optimization:

Finding optimal asset allocations based on various constraints.

Algorithmic Trading:

Developing automated trading strategies that consider complex market dynamics.

Financial Forecasting:

Predicting future market trends and events.

Curate the Financial Dataset for Gen AI application

Category	Traditional Financial Data	Data for Training Financial LLMs
Financial Statements	Income statement, balance sheet, cash flow statement	Same
Market Data	Stock prices, Economic indicators	Same
Company Fundamentals	Assets, Liabilities, Sales, Costs/earnings, Macro variables	Same
Alternative Data	N/A	News articles, Analyst reports, Social media sentiment, Legal and regulatory documents, Structured data

Key Differences:

- **Unstructured Data:** Financial LLMs require a significant amount of unstructured data, such as text and social media content, in addition to traditional structured data.
- **Data Variety:** LLMs benefit from a wider variety of data, including news articles, analyst reports, and legal documents, to learn complex relationships and patterns.
- **Data Volume:** Training large language models requires massive amounts of data to achieve high performance.

Fine-tuning Domain Specific LLM Models

Key Considerations:

- **Model Selection:** Consider factors such as model size, capabilities, and computational requirements.
- **Dataset Quality:** Ensure the training data is relevant, diverse, and high-quality.
- **Fine-Tuning Techniques:**
  - **Prompt Engineering:** Carefully craft prompts to guide the LLM's responses.
  - **Few-Shot Learning:** Provide a few examples of desired outputs to help the LLM learn the task.
  - **In-Context Learning:** Provide examples and instructions directly within the input prompt.
  - **Transfer Learning:** Leverage pre-trained models on related tasks to accelerate training.
  - **Parameter-Efficient Fine-Tuning (PEFT):**
    - **LoRA:** Add low-rank matrices to the model's attention layers.
    - **QLoRA:** Quantize the model's weights to reduce memory usage and inference time.
    - **Adapter Tuning:** Add small, trainable modules to the model's layers.

Fine-tuning Domain Specific LLM Models

Evaluation:

- **Metrics:**
  - **Text Generation:** BLEU, ROUGE, METEOR, CIDEr, CIDEr-D
  - **Question Answering:** Accuracy, F1-score, Exact Match
  - **Summarization:** ROUGE, METEOR, CIDEr
  - **Classification:** Accuracy, Precision, Recall, F1-score
  - **Sentiment Analysis:** Accuracy, F1-score, Cohen's Kappa
- **Benchmarks:**
  - **GLUE Benchmark:** General Language Understanding Evaluation benchmark
  - **SuperGLUE Benchmark:** SuperGLUE Benchmark
  - **SQuAD:** Stanford Question Answering Dataset
  - **WikiText-103:** A large-scale text corpus for language modeling
- **Human Evaluation:** Assess the quality of the model's outputs through human evaluation.
- **LLM as a Judge:** Use another LLM to evaluate the output of the target LLM, providing a more automated and objective assessment.

Cost of Generative AI Models in Production

Key Components:

- **Container Service:** Kubernetes
- **Compute Resources:** GPUs, CPUs
- **Networking and DNS:** Routing, Ingress & Egress
- **Load Balancing:** Distributing traffic
- **Autoscaling:** Adjusting resources based on demand
- **Logging & Telemetry:** Collecting and analyzing data
- **CI:** Continuous integration for updates and rollbacks

Summary:

Deploying a Generative AI model involves more than just a VM with a GPU. It requires a comprehensive infrastructure to ensure efficient and scalable operation. By considering these components, organizations can better estimate and manage the TCO of their Generative AI deployments.

Cost of Generative AI Models in Production

Key Takeaways:

- Managed services offer a simplified solution with a higher fixed cost.
- Self-built solutions can be more cost-effective in the long run, but require significant engineering expertise and ongoing maintenance.
- The cost of engineering labor can significantly impact the total cost of ownership for self-built solutions.
- Consider factors such as scalability, reliability, and security when choosing between managed services and self-built solutions.

Hypothetical Scenario:

Assume a hypothetical scenario of deploying an LLM on a single A100 80GB GPU.

- **Managed Service:** Using a managed service like Hugging Face Inference, the total cost per hour would be $4.
- **Self-Built Solution:** Deploying the LLM on a cheaper compute service like AWS EC2, the total cost per hour would be approximately $2.40.

Cost	Managed Service	Self-Built Solution
Compute (1x A100 80GB)	-	$1.89/hr
Storage (500GB)	-	$0.09/hr
Domain / IP (1x)	-	$0.01/hr
Networking (100GB)	-	$0.05/hr
Load Balancer (1x)	-	$0.03/h
Container Services (k8s)	-	$0.28/hr
Logging and Monitoring (25GB)	-	$0.05/hr
Total	$4/h	$2.40/h
Total + 1 Engineer*	$4/h	$70.9/h

Use Case : Building Embedding dataset for Nvidia

Purpose: For fine-tune the Embedding Model to improve accuracy of Retrieval-Augmented Generation/ Gen AI performance

- **Problem to solve:** Represent relationships or similarities between sentences.
- **Fine-tuning**
  - **Matryoshka Representation Learning (MRL)** is a technique designed to create embeddings that can be truncated to various dimensions without significant loss of performance.
  - **Base Model:** BAAI/bge-base-en-v1.5
  - **Evaluation:** InformationRetrievalEvaluator
  - **Results** : Optimizing the model dimension of 128 allows us to keep 99% of the performance of the 768 dimension model but reducing the storing cost by 6x and improving the cosine similarity search.
- **Data Source:** Gather relevant data from various sources, here is 2024_08_28_Nvidia_10Q_SEC_Filling
- **Data Cleaning:** Remove noise, inconsistencies, and irrelevant information from the text data.
- **Pair or Triplet Formation:**
  - **Positive Pairs:** Identify pairs of sentences that are semantically similar or related.
  - **Triplets:** Create triplets where the anchor sentence is similar to the positive sentence but dissimilar to the negative sentence.
  - **Similarity Scores:** If applicable, assign similarity scores to pairs based on their semantic relationship.
  - **Dataset Formatting:** Structure the dataset in a consistent format, such as CSV or JSON, to facilitate processing.

{"**question** ": "<What kind of applications now support RTX technology?>", "**context** ": "<Many leading 3D design and content creation applications developed by our ecosystem partners now support RTX, allowing professionals to accelerate and transform their workflows with NVIDIA RTX GPUs and software.>"}

Use Case : market Insights Examples

NVIDIA Corporation and Subsidiaries
Condensed Consolidated Statements of Cash Flows
(In millions)
(Unaudited)

	Six Months Ended	
	Jul 28, 2024	Jul 30, 2023
Cash flows from operating activities:		
Net income	$ 31,480	$ 8,232
Adjustments to reconcile net income to net cash provided by operating activities:		
Stock-based compensation expense	2,164	1,576
Depreciation and amortization	843	749
Gains on investments in non-affiliated entities and publicly-held equity securities, net	(264)	(45)
Deferred income taxes	(3,276)	(1,881)
Other	(288)	(102)
Changes in operating assets and liabilities, net of acquisitions:		
Accounts receivable	(4,133)	(3,239)
Inventories	(1,380)	861
Prepaid expenses and other assets	(12)	(592)
Accounts payable	801	789
Accrued and other current liabilities	3,314	2,675
Other long-term liabilities	584	236
Net cash provided by operating activities	29,833	9,259
Cash flows from investing activities:		
Proceeds from maturities of marketable securities	8,098	5,111
Proceeds from sales of marketable securities	164	—
Purchases of marketable securities	(15,047)	(5,343)
Purchases related to property and equipment and intangible assets	(1,346)	(537)
Acquisitions, net of cash acquired	(317)	(83)
Purchases of investments in non-affiliated entities	(534)	(456)
Proceeds from sales of investments in non-affiliated entities	105	—
Other	—	21
Net cash used in investing activities	(8,877)	(1,287)
Cash flows from financing activities:		
Proceeds related to employee stock plans	285	247
Payments related to repurchases of common stock	(14,898)	(3,067)
Repayment of debt	(1,250)	(1,250)
Payments related to tax on restricted stock units	(3,389)	(1,179)
Dividends paid	(344)	(199)
Principal payments on property and equipment and intangible assets	(69)	(31)
Net cash used in financing activities	(19,665)	(5,479)
Change in cash, cash equivalents, and restricted cash	1,291	2,493
Cash, cash equivalents, and restricted cash at beginning of period	7,280	3,389
Cash, cash equivalents, and restricted cash at end of period	$ 8,571	$ 5,882
Reconciliation of cash, cash equivalents, and restricted cash to the Condensed Consolidated Balance Sheet:		
Cash and cash equivalents	$ 8,563	$ 5,783
Restricted cash, included in prepaid expenses and other current assets	8	99
Total cash, cash equivalents, and restricted cash	$ 8,571	$ 5,882
Supplemental disclosure of cash flow information:		
Cash paid for income taxes, net	$ 7,449	$ 328

Page 8, Nvidia 10 Q SEC Filing, 2024/08/24

Key Observations:

- Strong Cash Flow from Operating Activities:
- Significant Capital Expenditures
- Share Repurchases and Dividends
- Increased Cash and Cash Equivalents

Metric	Jul 28, 2024	Jul 30, 2023
Net Income	$31,480 million	$8,232 million
Cash Flow from Operating Activities	$29,833 million	$9,259 million
Capital Expenditures	$15,047 million	$5,343 million
Share Repurchases	$14,898 million	$3,067 million
Dividends Paid	$344 million	$199 million
Cash and Cash Equivalents	$8,571 million	$5,882 million

ML Model

Generate SQL for "Calculate financial ratios:"

```
SELECT
    f.year,
    f.quarter,
    (f.total_revenue / f.total_assets) AS return_on_assets
FROM financial_statements f;
```

Database

Use Case : Fine-Tuning Models

NVIDIA Corporation and Subsidiaries
Condensed Consolidated Statements of Cash Flows
(In millions)
(Unaudited)

	Six Months Ended	
	Jul 28, 2024	Jul 30, 2023
Cash flows from operating activities:		
Net income	$ 31,480	$ 8,232
Adjustments to reconcile net income to net cash provided by operating activities:		
Stock-based compensation expense	2,164	1,576
Depreciation and amortization	843	749
Gains on investments in non-affiliated entities and publicly-held equity securities, net	(264)	(45)
Deferred income taxes	(3,276)	(1,881)
Other	(288)	(102)
Changes in operating assets and liabilities, net of acquisitions:		
Accounts receivable	(4,133)	(3,239)
Inventories	(1,380)	861
Prepaid expenses and other assets	(12)	(592)
Accounts payable	801	789
Accrued and other current liabilities	3,314	2,675
Other long-term liabilities	584	236
Net cash provided by operating activities	29,833	9,259
Cash flows from investing activities:		
Proceeds from maturities of marketable securities	8,098	5,111
Proceeds from sales of marketable securities	164	—
Purchases of marketable securities	(15,047)	(5,343)
Purchases related to property and equipment and intangible assets	(1,346)	(537)
Acquisitions, net of cash acquired	(317)	(83)
Purchases of investments in non-affiliated entities	(534)	(456)
Proceeds from sales of investments in non-affiliated entities	105	—
Other	—	21
Net cash used in investing activities	(8,877)	(1,287)
Cash flows from financing activities:		
Proceeds related to employee stock plans	285	247
Payments related to repurchases of common stock	(14,898)	(3,067)
Repayment of debt	(1,250)	(1,250)
Payments related to tax on restricted stock units	(3,389)	(1,179)
Dividends paid	(344)	(199)
Principal payments on property and equipment and intangible assets	(69)	(31)
Net cash used in financing activities	(19,665)	(5,479)
Change in cash, cash equivalents, and restricted cash	1,291	2,493
Cash, cash equivalents, and restricted cash at beginning of period	7,280	3,389
Cash, cash equivalents, and restricted cash at end of period	$ 8,571	$ 5,882
Reconciliation of cash, cash equivalents, and restricted cash to the Condensed Consolidated Balance Sheet:		
Cash and cash equivalents	$ 8,563	$ 5,783
Restricted cash, included in prepaid expenses and other current assets	8	99
Total cash, cash equivalents, and restricted cash	$ 8,571	$ 5,882
Supplemental disclosure of cash flow information:		
Cash paid for income taxes, net	$ 7,449	$ 328

Page 8, Nvidia 10 Q SEC Filing, 2024/08/24

Fine-tuned Models

ChatGPT

What is the primary source of NVIDIA's cash inflows?

- **Answer** : Net income from operations, which increased from $8,232 million in the six months ended July 30, 2023, to $31,480 million in the six months ended July 28, 2024.
- **Context** : The company's strong sales growth and effective cost management led to a significant increase in net income, which is the primary driver of cash inflows from operating activities.

What is the primary source of NVIDIA's cash inflows?

The primary source of NVIDIA's cash inflows comes from its operating activities, particularly driven by robust revenues in its **Data Center** and **AI** businesses. NVIDIA's AI-powered chips and systems, widely used in data centers for cloud computing and generative AI, have contributed heavily to its revenue growth. In the second quarter of fiscal 2025, the company reported record revenue of $30 billion, more than doubling year-over-year, with **Data Center** revenue alone at $10.32 billion(NVIDIA Investor Relations)(NVIDIA Newsroom).

Use Case: Trading Strategy with Gen AI

Reinforcement Learning Strategies:
States Space Setting:

S = [P_t, N_t,C_t, I_t], $S \in \mathbf{R}^{30}$. For any time t, the model has the following elements:

- $P_t \in \mathbf{R}^{30}$, the Dow 30 Stocks price vector at time t;
- N_t, for each Dow 30 stock, the holding share numbers at time t;
- C_t, at the time t, the Dow 30 portfolio balance (Cash Value);
- $I_t = (MACD_t, RSI_t, ADI_t, ADX_t)$, for each Dow 30 stock, its related trends indicator vector at time t;

So the total Dow 30 portfolio trading states spaces will be 30(P_t) + 30(N_t) + 1(C_t) + 4*30(I_t) = 181.

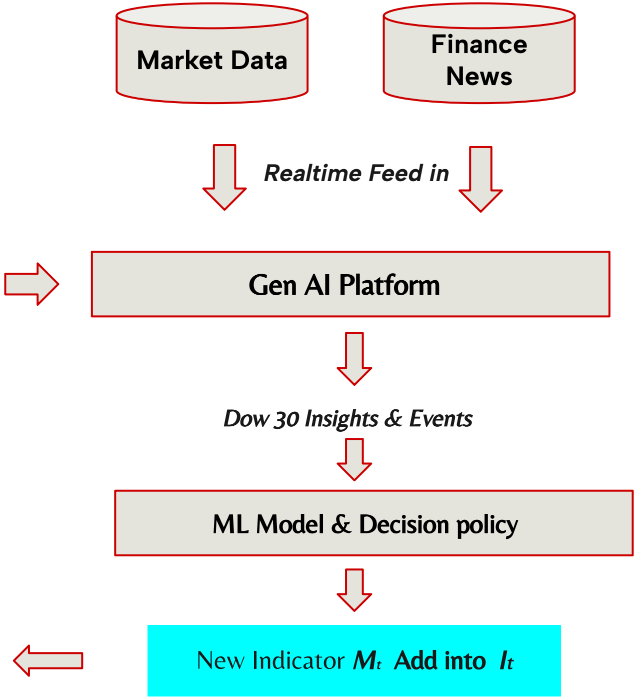
Algorithm 1: Reinforcement Trading Algorithm

Result: Best Trading Strategy to maximize the Dow 30 Portfolio value

1. Processing Stock data;
2. Building the train/valid data set by adding price,trends indicators and etc. information ;
3. Environment Setting: Parameters setting, Train/Trade/Valid Environment(build states,action) ;
4. Initialization Environment;
5. Training $model = A2C('MlpPolicy', env\_train)$

while *timestamps* **do**
    Train the A2C model ;
    model.learn()
end

6. Validation Model ;
7. Compute the Sharpe Ratio ;



The goal for trading algorithm is to maximize the Dow 30 portfolio's positive cumulative cash value. At the time t, the portfolio value is C_t, based on the policy at the the new state , the trading algorithm should maximize the portfolio value:

$$r(s_{t+1} \mid s_t, a_t)) = (C_{t+1} + P_{t+1}^T N_{t+1}) - (C_t + P_t^T N_t) - Cost_t$$

Reward Function:

$$R_{\bar{a}} = (P_{t+1}^{\bar{a}} - P_t^{\bar{a}})^T N_t^{\bar{a}}$$

For maximizing portfolio value, the algorithm should find a strategies to maximize ***hold*** Reword and ***buy*** Reward and minimize ***Sell*** Reward

https://github.com/Pyligent/RL_Algo_Trading/blob/main/RL_Algo_Trading.pdf

Challenges

- Still in the early stages
- Integrate with Quant Models and legacy system
- Reasoning capabilities
- Real-time alpha generation
- High production cost
- Validation and Hallucination
- Compliance and Regulations