

# Applying AI to the Investment Management Process: A Generational Opportunity



---

Brendan Erhard  
[Erhardbr@gmail.com](mailto:Erhardbr@gmail.com)  
January 2025

Hi, my name is **Brendan Erhard**. I am a global technology investor and I am an **AI-obsessed developer** looking to apply AI to investment management.

I have 14 years of experience as a **public equity investor** across several successful portfolio management franchises. I am currently an analyst on a global technology fund with AUM exceeding \$4 billion, with top-quartile performance during my tenure on the strategy, starting November 2015 (initial assets of \$500m). The strategy has won several industry awards for outstanding performance and has compounded capital at ~19% per year during my tenure.

I'm also a skilled Python and JavaScript developer, and have built and deployed many full-stack apps that integrate LLM technology from the leading AI providers. I am a **Tier 4 OpenAI developer** (based on model usage and tenure), and I have experience with AWS/Azure and AI frameworks, among many other technologies. I've played a leading role in developing and implementing AI applications at my current organization.

My GitHub profile can be accessed here: <https://github.com/bme3412>, and my AI project portfolio can be accessed here: <https://bme-ai-investing.vercel.app/> .

In the following slides I will provide comprehensive details from **my experience using AI technologies and LLM frameworks to create compelling workflows** that have enhanced my investment process.

I also provide ideas for future projects that I believe can deliver tremendous value.

Throughout the slide deck, anything in **blue font** is a prompt fed into a Large Language Model, and where appropriate I provide responses to these prompts.

I have also included an appendix at the end of the slide deck that provides a high level overview of LLM technology.

My emphasis on this presentation is to provide **demonstrated evidence and discussion of how LLM technology can be used in investment management**, along with thoughts and lessons learned along the way.

But first I wanted to set the scene with some **introductory thoughts and best practices for using LLMs**.

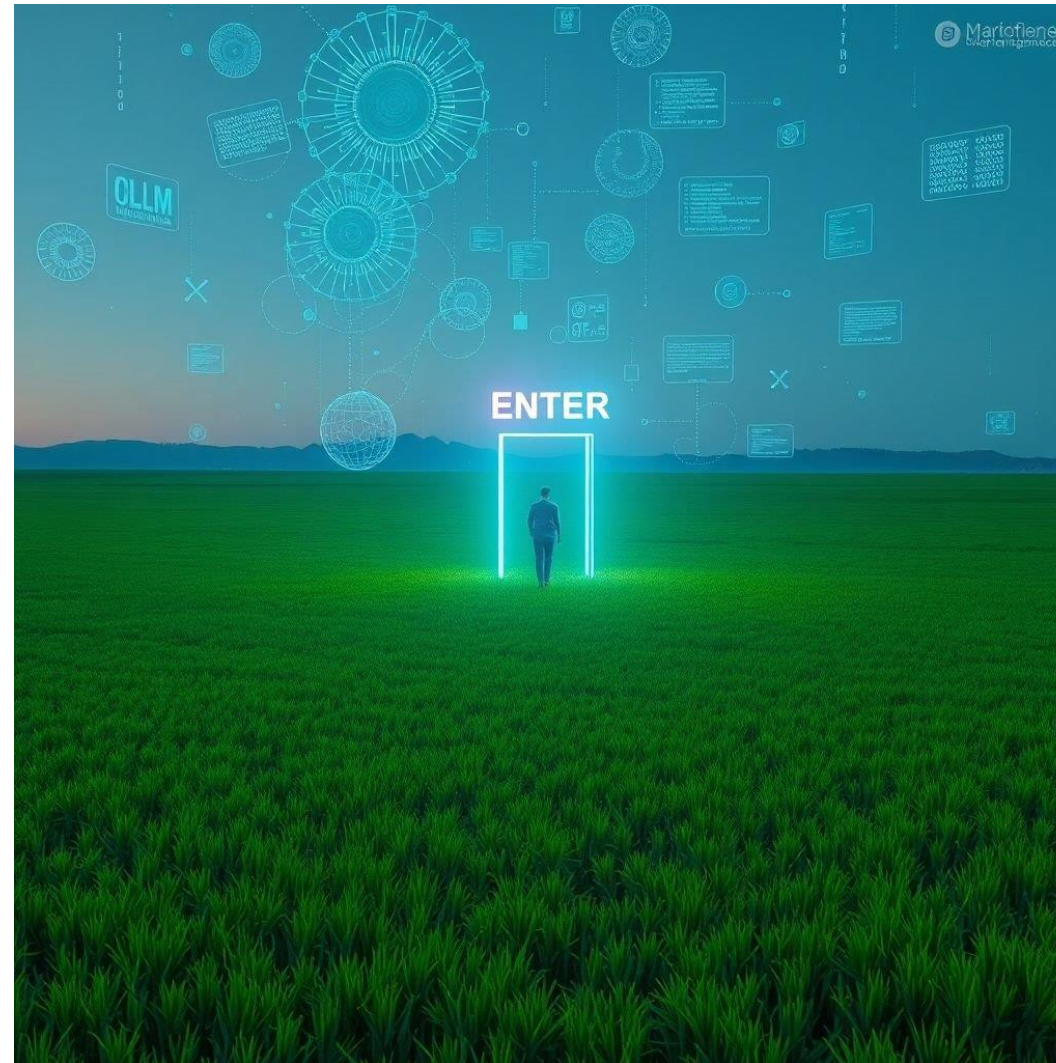
If there is 1 quote I can think of that **best sums up the opportunity for using AI, and not just for investing**, it's this one, from Andrej Karpathy, who was Director of AI and Autopilot Vision at Tesla, and also co-founded OpenAI.

"There has never been a green pasture of this size with this low barrier to entry"



In fact, I thought the quote was so important, I used a stable diffusion text-to-image model to illustrate it:

I made the image abstract on purpose; this is not a stock image. It was created by AI



Text-to-image model used:

Flux-Schnell provided by Black Forest Labs (German AI startup); cost to run \$0.003 per image, ~ 2 seconds

<https://blackforestlabs.ai/>

# Best Practices & High Level Philosophy From Using LLMs

1. ChatGPT and Anthropic Claude web interfaces are great, but **knowing how to use LLM APIs enables users to create workflows and applications** that are so much more valuable than summarizing content.
2. Text analysis and generation (with NLP techniques), code generation, and audio transcription are **the most important modalities for using LLMs as part of the investment process**. Computer vision models can be used with LLM models to extract and parse information
3. Responses to LLM prompts are **context and input dependent**; feeding more data and providing more detailed prompts results in better outcomes. The models aren't always right, so **iterate and experiment**.
4. Hallucinations can be mitigated by using **evaluation techniques and concepts such as type validation (Pydantic) and structured outputs**. Some workflows can be constructed such that the LLMs can cite excerpts from content used (ex: RAG techniques and Claude Anthropic: <https://docs.anthropic.com/en/prompt-library/cite-your-sources>)

## ...Best Practices [Continued]

5. LLM models **keep getting better over time** (and cheaper), and the ecosystem continues to evolve. Stay up to date, and follow the developer community
6. At the start of a project, it's beneficial to provide a description of the intended outcomes, and then ask the LLM how to improve the process. **LLMs are great at brainstorming new approaches, identifying blind spots, and systems engineering.**

# End to End Examples

1. Clarity 2.0 Investment Co-Pilot using RAG
2. Structured Report Generation
3. Industry Primer Generation with Google Gemini Deep Research
4. Contextualizing Quant Scores
5. Document Parser
6. Audio Transcription using OpenAI Whisper
7. Auto-DCF
8. Semantic Screening
9. Bloomberg Workflows



# 1. Clarity 2.0 - Investment CoPilot

Clarity 2.0 is an investment co-pilot built using Python and JavaScript, along with OpenAI, Langchain (LLM framework) and a Pinecone vector database

Users of Clarity 2.0 can write queries about any of the Big Tech companies, and will receive a comprehensive response that incorporates commentary from earnings calls, investor presentations, and reported quarterly financials.

Data Scope: Incorporates 5+years of data (quarterly earnings reports, SEC filings, financials) from 10 large-cap technology companies\*

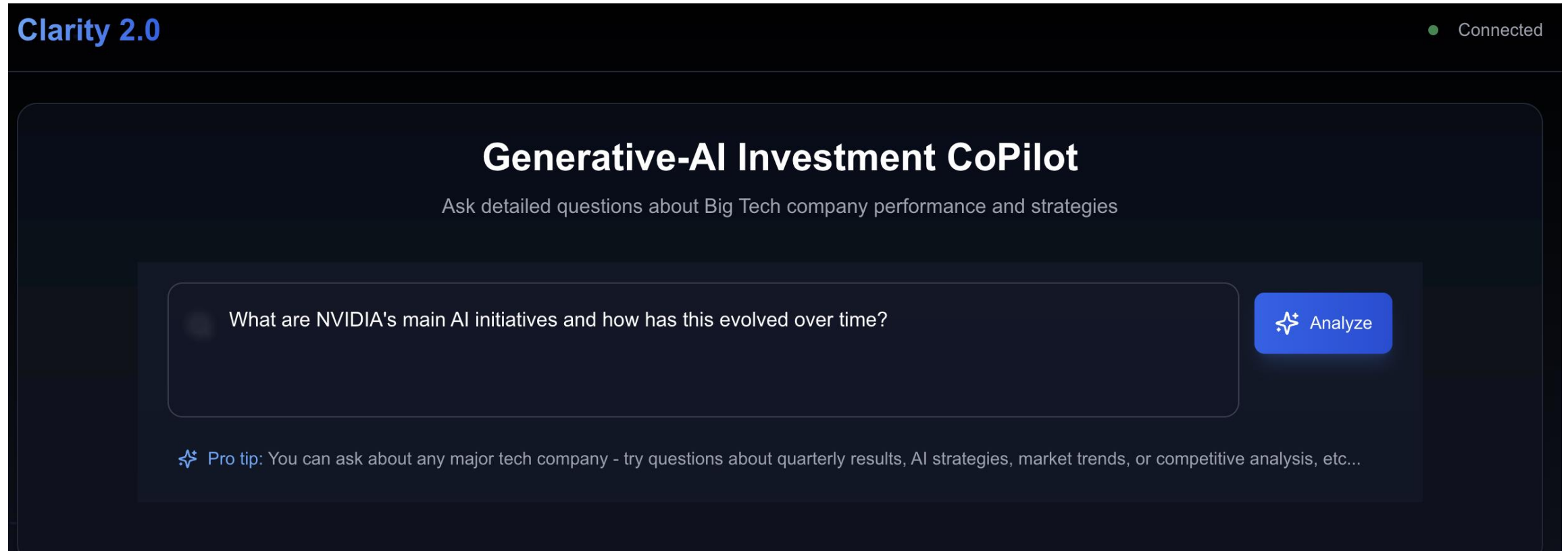
The project is deployed online and users can write their own prompts.

Feel free to try it out! <https://investment-copilot-eight.vercel.app/>

\*AAPL, AMD, AMZN, AVGO, CRM, GOOGL, META, MSFT, NVDA, ORCL

# Clarity 2.0 - Investment CoPilot

Example prompt:



The screenshot shows the Clarity 2.0 interface. At the top left, it says "Clarity 2.0" in blue. At the top right, there is a green dot and the text "Connected". The main content area has a dark background with the title "Generative-AI Investment CoPilot" in white. Below the title is the subtitle "Ask detailed questions about Big Tech company performance and strategies". In the center, there is a text input field containing the prompt: "What are NVIDIA's main AI initiatives and how has this evolved over time?". To the right of the input field is a blue button with a star icon and the text "Analyze". Below the input field, there is a "Pro tip" section with a star icon and the text: "Pro tip: You can ask about any major tech company - try questions about quarterly results, AI strategies, market trends, or competitive analysis, etc..."

# Clarity 2.0 - Investment CoPilot

Example response (abbreviated):

Analysis:

NVIDIA's historical initiatives and evolution in AI have been significant, with the company consistently focusing on expanding its technological capabilities and market reach in the AI sector. Let's break down the details:

1. Market Size and TAM: NVIDIA has been positioning itself to capture a significant share of the AI market, which is projected to be worth \$500 billion by 2025. This includes applications across various verticals such as autonomous driving, drug discovery, and robotics.
2. Revenue and Growth Rates: NVIDIA's data center revenue, which includes contributions from the automotive vertical, exceeded \$1 billion last year. The company's revenue from AI-related initiatives has been steadily increasing, driven by the adoption of its AI frameworks and platforms across multiple sectors.
3. Market Share and Positioning: Almost every automotive company working on AI collaborates with NVIDIA, signifying its strong market presence. NVIDIA's AI infrastructure is used by nearly 80 vehicle manufacturers, including OEMs, new energy vehicles, and Tier 1 suppliers.

# Clarity 2.0 - Investment CoPilot

- **How does it work?**

- The user sends a query, such as “What were the revenue numbers for Apple last quarter?”
- The application converts this query into a vector representation (“embedding”), then uses a vector database to find the most relevant documents.
- Finally, it uses an LLM (Large Language Model) to generate a thorough answer referencing the retrieved documents

## **Creation of embeddings**

- Each piece of text (e.g., transcripts, financial data) and user query is transformed into a numerical vector by an embedding model.
- These vectors capture semantic meaning, so text with similar content ends up close together in vector space.

## **Upsert to vector database**

- Newly embedded documents (transcripts or financial data) are “upserted” into the vector database—meaning they are inserted or updated if they already exist.
- This ensures the database is always current, storing all content in a format optimized for semantic similarity searches.

# Clarity 2.0 - Investment CoPilot

## **Semantic search**

- When a user queries the system, the query is also embedded and compared (via cosine similarity or a similar metric) to the stored vectors in the database.
- The highest-scoring matches are deemed most relevant, allowing the system to retrieve only the excerpts or data points most likely to answer the user's question.

## **LLM API as orchestration layer**

- Finally, the application uses an LLM (like OpenAI's GPT) to orchestrate the workflow, from analyzing the user's intent, to retrieving information, to synthesizing a coherent, human-like answer.
- This single LLM "orchestration layer" reduces complexity, as the model can dynamically decide which parts of the data to use and how to structure the final response.

## 2. Structured Report Generation

- LM APIs can be used to extract specific financial metrics and KPIs from unstructured documents, while Pydantic models can **enforce strict data validation and typing** - ensuring extracted numbers are properly formatted as floats/integers and fall within expected ranges for metrics like revenue, margins, or growth rates.
- The combination can parse earnings call transcripts by having the LLM identify key sections (prepared remarks vs Q&A), speaker roles (CEO, CFO, analysts), and notable quotes, with **Pydantic models structuring this into a consistent format** that captures speaker metadata, timestamp information, and conversation flow.
- For financial reports like 10-Ks, the LLM can extract relevant sections and tables while **Pydantic models can define strict schemas for financial statements - ensuring balance sheets balance, income statement items sum correctly, and year-over-year comparisons maintain consistency in their structure** and calculations.
- When analyzing multiple documents over time, Pydantic models can **enforce consistent formatting of time-series data extracted by the LLM**, making it easier to track metrics across quarters/years and ensuring proper handling of things like fiscal vs calendar year reporting differences.
- <https://openai.com/index/introducing-structured-outputs-in-the-api/>

### 3. Industry Primer Generation with Google Gemini Deep Research

- In December 2024, Google announced the release of **Deep Research, an agentic research tool** as part of the Gemini Advanced subscription offering
- Deep Research uses AI to **conduct automated online research** based on a topic provided by the user as part of a prompt. Deep Research then searches the internet, finds compelling information related to the topic, and then **writes a research report with proper citations of sources used**. The user can then ask follow-up questions about the report.
- In the following slides, I provide an example prompt, with screenshots of the output, along with a link to the generated research report.
- Deep Research searched and parsed through 34 websites about my prompt (investment implications for co-packaged optics), and wrote an 8-page report, **all in < 5 minutes**. This likely saved me at least 2 hours of work.

More details here:

<https://blog.google/products/gemini/google-gemini-deep-research/>

### 3. Industry Primer Generation with Google Gemini Deep Research

Initial Prompt:

**Please provide a primer on co-packaged optics for an audience of investment professionals. Describe the technology, the potential market opportunities, potential winners and losers across the technology industry and supply chain, and identify companies to invest in or which could be at risk.**

Gemini Deep Research provides an initial research plan based on the prompt. The user can make changes to the research plan. Once approved, the research process begins, and the report is generated with sources used.

After reading through the initial research report, I then wanted more company-specific details, so I wrote a new prompt.

Follow Up Prompt:

**Can you also integrate from company SEC filings and investor presentations? I want more specifics on companies**



# 3. Industry Primer Generation with Google Gemini Deep Research

Gemini **Advanced** ▾  
1.5 Pro with Deep Research



Here's the plan I've put together. Let me know if you need any changes before I start researching.

## Co-Packaged Optics Primer for Investors



### Research Websites

Please provide a primer on co-packaged optics (CPO) for an audience of investment professionals. Describe the technology, the potential market opportunities, potential winners and losers across the technology industry and supply chain, identify companies to invest in or which could be at risk.

(1) Find research reports and articles explaining co-packaged optics (CPO) technology.

(2) Find information on the potential market size and growth opportunities for CPO.

(3) Find information on potential winners and losers in the technology industry (e.g., chip manufacturers, optics companies, system integrators) due to the adoption of CPO.

(4) Find information on potential winners and losers in the supply chain (e.g., material suppliers, equipment manufacturers) due to the adoption of CPO.

(5) Find publicly traded companies involved in the development or manufacturing of CPO technology.

(6) Find companies that may be negatively impacted by the adoption of CPO.

(7) Find information on current CPO market financials and expected growth.



### Analyze Results



### Create Report

### 3. Industry Primer Generation with Google Gemini Deep Research

The screenshot displays the Google Gemini Deep Research interface. On the left, a sidebar contains a 'Start research' button, a message stating 'I've completed your research. Feel free to ask me follow-up questions or request changes.', and a card for the report titled 'Co-Packaged Optics Primer for Investors' dated 'Jan 19, 9:40 AM'. At the bottom of the sidebar is an 'Ask Gemini' input field with a microphone icon. The main content area on the right shows the report's title 'Co-Packaged Optics: An Emerging Opportunity in the Data Center Market' and an 'Introduction' section. The introduction text discusses the growing demand for high-speed data transmission and the role of co-packaged optics (CPO) in data center infrastructure. A blue 'Open in Docs' button is visible in the top right corner of the report area.

Start research

I've completed your research. Feel free to ask me follow-up questions or request changes.

Co-Packaged Optics Primer for Investors  
Jan 19, 9:40 AM

Ask Gemini

Co-Packaged Optics Primer for Investors Open in Docs

#### Co-Packaged Optics: An Emerging Opportunity in the Data Center Market

##### Introduction

The demand for high-speed data transmission is accelerating, driven by the exponential growth of data centers and cloud services, artificial intelligence (AI), high-performance computing (HPC), and the increasing penetration of connected devices. This surge in data traffic necessitates innovative solutions to address bandwidth limitations, power consumption, and latency challenges. Co-packaged optics (CPO) has emerged as a promising technology to revolutionize data center interconnect architecture and potentially disrupt the existing market landscape. CPO is essential for meeting the growing bandwidth demands of data-intensive applications and enabling the next generation of data center infrastructure. This analysis provides an in-depth look at CPO technology, its market opportunities, potential winners and losers across the technology industry and supply chain, and identifies key companies for investment consideration.

To gather the information presented in this analysis, a comprehensive research process was

### 3. Industry Primer Generation with Google Gemini Deep Research

Research Report, Prompt 1 :

- 34 websites searched, generated 8-page report with 1,900 words; identified 9 companies to potentially invest in

Research Report, Prompt 2:

- 128 websites searched, generated 9-page report with 2,400 words; identified 8 companies to potentially invest in

## 4. Using AI to Unite Quant and Fundamental Investing

- Goal of project: Use LLMs to rapidly parse changes in quant scores, to **better identify trading signals**, and to effectively contextualize and **communicate implications to non-quant investment professionals**
- Background: Quant factor scores are compiled weekly, and changes in quant scores do provide valuable trading signals. But the drivers of the underlying factors that results in changes to composite quant scores **are not fully understood**.
  - With 500 stocks in the S&P 500 Index, and 2,000 stocks in the Russell 2000 – as well as different quant models across strategies – bandwidth and resource constraints makes it **difficult for these trading signals to be understood or acted upon in a timely manner**.
  - This presents an opportunity for **AI to be used to come up with a repeatable and scalable process in which the language capabilities of the LLM APIs can be used to provide a workflow that creates a best-efforts qualitative commentary** for weekly quant changes
- The following few slides have example screenshots

# Weekly Quant Scores

<b>AAPL</b> Rating: 2 (Buy) Status: No change Alpha: +0.04 <a href="#">View Details</a>	<b>MSFT</b> Rating: 3 (Hold) Status: No change Alpha: +0.13 <a href="#">View Details</a>	<b>NVDA</b> Rating: 3 (Hold) Status: Downgrade Alpha: +0.43 <a href="#">View Details</a>	<b>AMD</b> Rating: 2 (Buy) Status: Upgrade Alpha: -0.33 <a href="#">View Details</a>	<b>ADBE</b> Rating: 2 (Buy) Status: No change Alpha: -0.01 <a href="#">View Details</a>
<b>SNPS</b> Rating: 4 (Sell) Status: Downgrade Alpha: +0.24 <a href="#">View Details</a>	<b>CDNS</b> Rating: 2 (Buy) Status: No change Alpha: +0.03 <a href="#">View Details</a>	<b>INTC</b> Rating: 3 (Hold) Status: No change Alpha: -0.27 <a href="#">View Details</a>	<b>CSCO</b> Rating: 3 (Hold) Status: No change Alpha: -0.07 <a href="#">View Details</a>	<b>IBM</b> Rating: 4 (Sell) Status: No change Alpha: +0.16 <a href="#">View Details</a>

**Rating Change Summary: NVDA**

2024-01-08 Current: 3 (Hold)	2024-01-15 Previous: 2 (Buy)
---------------------------------	---------------------------------

Rating Change: +1

Alpha Change: +0.43

[Generate Detailed Analysis](#) [View Factor Details](#)

- When "Generate Detailed Analysis" is clicked, it initiates a workflow where the LLM AI (uses OpenAI GPT-4o model) will read in the synthetic data, compare the data from the most recent week to the week prior, and then generate content that attempts to explain changes in the quant score week to week. See next page for generated commentary

## Prompt Used:

Analyze changes in a stock's performance by:

1. Review the overall rating changes from [previous rating] to [current rating] and explain what caused this shift.
2. Break down the most significant changes in: - Most improved areas - Areas showing decline - Key performance metrics
3. For each major change identified: - What business decisions or market conditions drove this change? - How do current numbers support or challenge these changes? - Which key financial metrics warrant investor attention?

Consider: - Top 3 highest performing components and their scores - Bottom 3 lowest performing components and their scores - Notable trends or patterns in the data

Output should include: - Clear explanation of what the rating changes mean for investors - Specific data points that support the analysis - Recommendations for metrics to monitor going forward

Write your analysis in clear, non-technical language that explains what these changes mean for the company's overall health and future prospects.

# Example Response:

## Investment Implications

### 1. **\*\*Concrete Business Changes Driving the Rating Move:\*\***

The downgrade from 2 to 3, driven by an alpha decline of 0.43, indicates a perceived reduction in the company's ability to generate excess returns relative to a benchmark. The primary factor changes suggest that NVDA has seen improvements in several value-related metrics, yet these improvements were not enough to prevent the downgrade. Specifically, the increase in Value from 1.10 to 1.56, with significant improvements in Cash Flow Conversion (from 0.93 to 1.34), EBITDA Growth (from 1.00 to 1.39), FCF Valuation (from 1.03 to 1.51), ROIC (from 1.03 to 1.44), and Buybacks (from 0.99 to 1.39), suggests that NVDA has enhanced its operational efficiency and capital allocation strategies. However, these improvements might not have been sufficient to offset other negative factors impacting the overall rating.

### 2. **\*\*Metrics Supporting or Challenging the Current Trend:\*\***

The highest scoring components—Accruals, Inventory Turnover, and DSO—all scoring 2.00, indicate strong operational efficiency and effective working capital management. These metrics support the positive changes in value-related factors. However, the lowest scoring components—insider transactions (1.18), SI % (1.27), and Cash Flow Conversion (1.34)—highlight areas of concern. Insider transactions scoring low could suggest a lack of confidence from insiders, while a low SI % indicates potential market skepticism. Despite improvements, Cash Flow Conversion remains one of the lowest scoring metrics, challenging the sustainability of the positive trend in value metrics.

### 3. **\*\*Key Financial Metrics for Investors to Monitor:\*\***

Investors should closely monitor the following key financial metrics:



## 5. Document Parser

- A simple web app (Flask, Streamlit, etc) can be created that enables the user to upload documents, and then the web app will parse the document on the back end using an LLM API, and enable the user to ask questions about the document
- Documents can include anything from annual reports, earnings call transcripts, investor presentations, sell-side research, industry primers, articles, etc. The LLM API will be used to generate responses to user queries
- The system can rapidly process large volumes of financial documents like 10-K reports, earnings transcripts, and analyst reports, extracting key metrics and trends that would take hours to review manually.
- The technology can identify patterns across multiple documents and time periods, helping spot trends in company performance, management communication style changes, or shifts in business strategy that might not be immediately obvious to human readers.

## 6. Audio Transcription and Report Generation

- Whisper is an AI speech recognition tool created by OpenAI that can convert audio to text with high accuracy. It's particularly good at handling business language and financial terms commonly used in earnings calls.
- The system can process a variety of audio file formats and works with recordings of any length, making it ideal for quarterly earnings calls that typically run 45-60 minutes.
- Unlike traditional transcription services that might take hours or require human transcriptionists, Whisper can process an earnings call in minutes, allowing for quick analysis and distribution of the information.
- The tool handles multiple speakers well, which is essential for earnings calls that typically include the CEO, CFO, and various analysts asking questions.
- Once transcribed, the text can be easily searched, analyzed, or imported into other tools for sentiment analysis or to extract key financial metrics and commitments.
- <https://openai.com/index/whisper/>

## 7. Automated DCF

- A web application can allow users to upload Excel files containing financial statements, with the **LLM API intelligently identifying and extracting key financial metrics** like revenue, EBIT, capex, and working capital - even when files have varying formats or naming conventions across different companies.
- The app can **integrate the extracted historical financials into a structured DCF framework**, where the LLM helps generate initial growth and margin projections based on historical trends, industry comparables, and management guidance found in the uploaded data.
- Users can interact with an intuitive interface to **flex key DCF assumptions (growth rates, margins, discount rates)**, with the LLM providing real-time commentary on how these adjustments compare to industry benchmarks and historic performance patterns found in the uploaded data.
- For sensitivity analysis, **the web app can allow users to create multiple scenarios (base, upside, downside)**, with the LLM suggesting reasonable ranges for key variables based on the company's historical volatility and industry patterns found in the data.

## 8. Semantic Screening

- Semantic screening leverages LLMs to understand natural language queries like "**Find companies with new CEOs who came from different industries**" - the system can parse news, filings, and press releases to identify not just CEO changes, but understand the context of each transition (internal vs external hire, industry background, circumstances of previous CEO's departure).
- Beyond binary "yes/no" classifications of CEO transitions, the LLM can extract rich semantic details - **identifying cases where CEOs signaled strategic shifts in their first earnings calls, came from key competitors, or had specific experience** (e.g., turnaround specialists, digital transformation experts).
- Natural language queries can be progressively refined based on results - **if a user asks for "companies undergoing significant change," the system can suggest specific patterns it recognizes (new CEO, strategic reviews, activist campaigns)** and allow users to focus on the most relevant signals.

## 9. LLM-enhanced Bloomberg workflows

- LLMs can analyze company fundamentals and market data from Bloomberg to identify typical activist target characteristics, then **help generate Bloomberg formulas ({FLDS XLTP}) that screen for these patterns** - like declining ROIC, high cash balances, or conglomerate discounts.
- For function building, the LLM can suggest relevant Bloomberg fields and proper syntax for complex queries - for example, **constructing BQL formulas that combine ownership structure analysis (institutional holders, insider stakes) with valuation metrics and performance data that typically attract activist attention.**
- Rather than just rigid screens, **the LLM can help craft adaptive Bloomberg queries** that look for historical patterns preceding activist campaigns - such as correlating specific corporate events (spin-off announcements, CEO changes) with subsequent activist involvement to identify similar situations.

## Case Studies: Evidence of LLMs Improving Outcomes

The University of Chicago's Booth School of Business published a working paper in January 2025 that explores the use of AI in investment management. The paper, titled 'AI and Information Advantage in Investment Management', discusses the potential of AI to provide an information advantage to investors.

When GPT-4o summarizes earnings calls to match investor expertise level, **aligned AI increases one-year portfolio returns by 9.6% for more sophisticated investors and by 1.7% for less sophisticated ones**, thereby widening the gap by 7.9%

[https://bfi.uchicago.edu/wp-content/uploads/2025/01/BFI\\_WP\\_2025-02.pdf](https://bfi.uchicago.edu/wp-content/uploads/2025/01/BFI_WP_2025-02.pdf)

## Case Studies: Evidence of LLMs Improving Outcomes

In a wide-ranging December 2024 interview with Bloomberg, the CEO of systematic investor AQR spoke of integrating Artificial Intelligence into his firm's investment process.

For Asness, **AI is 'annoyingly better' than humans at many tasks**, and he sees it as a way to improve the firm's investment process. He envisions AI as being a far more dynamic approach than quant factor investing, and views the integration of AI and massive alternative data sets as an arms race that will intensify over time

[https://www.youtube.com/watch?v=qK4t4boN\\_VE](https://www.youtube.com/watch?v=qK4t4boN_VE)

## Case Studies: Evidence of LLMs Improving Outcomes

Goldman Sachs CEO David Solomon made remarks at the Cisco AI Summit in January 2025 about the immense capabilities and potential threats posed by AI.

Solomon mentioned that **AI can write '95% of an IPO prospectus accurately and people are required for the last 5% of the report** – the only part that “matters”. Normally, the task of drafting an initial registration prospectus for an initial public offering known as an S1, took a six-person team two weeks to prepare. AI can complete this task “in minutes”.

<https://www.businessinsider.com/goldman-sachs-ceo-david-solomon-ai-ipo-investment-bank-analyst-2025-1>



# Appendix and Review

1. Review of the Technology
2. Further Resources
3. Technologies I use and am familiar with
4. Example Prompts



# 1. Review of LLMs

# So what is a Large Language Model?

- Large Language Models are AI systems trained on vast amounts of text data that can understand and generate human-like text, which can then be applied to many use cases.
- At their core, LLMs work by breaking down text into smaller pieces (tokens), and then learning patterns from billions of examples
- For instance, after seeing millions of times that "I feel under the \_\_\_\_" is often followed by "weather", they learn this pattern
- These models can perform an impressive range of tasks: writing code, answering questions, summarizing documents, translating languages, and even basic reasoning - all through their ability to understand and generate text patterns
- The "Large" in LLM refers to both the massive amount of training data (often hundreds of billions of words) and the model size itself (often hundreds of billions of parameters)
- Modern LLMs (like GPT-4 or Claude) can maintain context over long conversations and understand nuanced instructions, making them useful for everything from customer service to creative writing to programming assistance

# The leading AI companies offer APIs which can be used to build applications

- An **API (Application Programming Interface)** enables different services/programs and applications to work together seamlessly
- APIs serve as building blocks for new applications, letting developers create powerful software without reinventing the wheel
- They democratize access to advanced technology - instead of needing to build complex AI models, developers can tap into powerful tools like GPT-4 through OpenAI's API, or use cloud services through AWS APIs to handle massive computing tasks
- Other AI companies that offer APIs include Anthropic, Google, Microsoft, Cohere, AI21 Labs. AI models can also come in different form factors, with image generation APIs offered by StabilityAI, Midjourney, Runway, Adobe Firefly. Voice and speech APIs are offered by AssemblyAI, ElevenLabs, Deepgram, etc. Obviously, just scratching the surface

# Documentation: A guide to using AI models to build applications

- AI Models and APIs come with documentation, which can be thought of as a menu for developers to access various endpoints of the API. It's a guide for how to use the API
- AI companies also often publish "cookbooks" which are example applications built by employees to demonstrate potential use cases; below image shows how to call the API in Python

Create a human-like response to a prompt 

```
1 from openai import OpenAI
2 client = OpenAI()
3
4 completion = client.chat.completions.create(
5     model="gpt-4o-mini",
6     messages=[
7         {"role": "system", "content": "You are a helpful assistant."},
8         {
9             "role": "user",
10            "content": "Write a haiku about recursion in programming."
11        }
12    ]
13 )
14
15 print(completion.choices[0].message)
```

OpenAI documentation: <https://platform.openai.com/docs/overview>

OpenAI cookbook: <https://cookbook.openai.com/>

Anthropic documentation: <https://docs.anthropic.com/en/home>

1/20/2025 Anthropic cookbook: <https://docs.anthropic.com/en/docs/resources/cookbook>

## 2. Further Resources

Investing with LLM Blog: <https://investment-llm.vercel.app/>

AI Projects: <https://bme-ai-investing.vercel.app/>

ResumeLLM: <https://llm-resume-bme.vercel.app/>

Github: <https://github.com/bme3412>

# 3. Technologies I Have Experience With

- Python - Jupyter Notebooks, Multi-Thread Processing, OOP
- JavaScript - React, Node.js, Next.js
- LLM APIs - OpenAI (Tier 4 Developer), Anthropic, Google Gemini
- AI frameworks - LangChain, LlamaIndex, Microsoft CoPilot
- Vector Databases - Pinecone, Chroma, FAISS
- Type Validation - Pydantic, Pydantic AI
- Cloud Computing - AWS, Azure
- Databases - SQL, MongoDB
- Version Control - Git, GitHub
- Visualization - Matplotlib, Seaborn, bokeh, d3.js
- Machine Learning
- Web App Development - Flask, Django, Streamlit, FastAPI
- Web Scraping - BeautifulSoup, Selenium, Firecrawl.dev
- NLP - NLTK, SpaCY

## 4. Example Prompts

- An existing code base can be fed into ChatGPT or Claude (literally just copying and pasting code into the application, provided there are proper guardrails for IP) with the prompt:
  - "Please refactor this code to make it more efficient, to add better logic and flow, and provide explanations to changes that are being made" and the model will provide a response
  - Or, you could also add a prompt such as "Please provide better documentation in the form of docstrings to make the code easier to understand" and the response will be an improved code base with supporting docstrings.
- You could also include a prompt "Please convert the following code from Matlab to Python and explain changes being made".
- The output itself can be featured in prompts: Please write a program that reads in this excel file, convert it to a dataframe and save it as a CSV file, using f-strings to come up with an appropriate file name



# 4. Example Prompts

- Summarizing content is just one example of what LLM APIs can do, but at the very least does demonstrate that the LLMs do understand the contextual meaning of the text.
- Through a series of prompts, you can create a custom workflow that saves a lot of time, expands bandwidth, and relies on the LLMs to do heavy lifting.
- An example would be creating a Python Flask app which would allow the user to upload an academic paper, and in addition to summarizing content, the user could write prompts such as:
  - 1) Please provide a critique of the paper and proposed trading signal.
  - 2) What are the paper's strengths and weaknesses?
  - 3) Are there any blind spots in the author's viewpoints?
  - 4) How could I take the hypothesis and convert to code?
  - 5) How feasible would it be to replicate in code, and provide on a scale of 1 to 10 how impactful a potential trading signal could be?
  - 6) How should I structure my file directory to ensure code scalability?
- Rather than asking static questions, if asked in a sequence where a prompt builds upon the response to another prompt, the LLM can be used to solve problems through dynamic workflows, especially with code generation.

# 4. Example Prompts

- In addition to image generation, LLMs can ingest images, and prompts can then be written about the image. Anthropic Claude is great at this.
  - An example workflow could be:
    - 1) Take a screenshot of a visualization
    - 2) copy/paste the image into Claude, and
    - 3) write a prompt such as: "I really like this visualization. Please suggest Python code that could be used to replicate the image. It's ok to use synthetic or randomly generated data, I just want ideas for code to replicate the plot"



# Thank You!

Email: erhardbr@gmail.com

LinkedIn: <https://www.linkedin.com/in/brendan-erhard/>