The recent launch of DeepSeek, a powerful yet cost-effective AI model, has profoundly affected the stock market and tech industry. Some analysts believe that DeepSeek's ability to outperform other AI models in a variety of benchmarks while utilizing less sophisticated hardware could be a "Sputnik moment" for China. This has raised questions about the future of AI dominance especially related to Nvidia, which held a monopoly on the processors that could run most models before Deepseek.

By concentrating on three main areas—stock market reactions, public and media perception, and AI model cost efficiency—this research investigates the financial, technological, and geopolitical ramifications of DeepSeek's publication. To assess market impact, we will monitor the stock performance of significant tech firms like Microsoft, Nvidia, and Meta. We'll examine media coverage to find important themes and generate word clouds that draw focus on public concerns using tools like NewsAPI. To ascertain whether DeepSeek represents a significant advancement in AI efficiency or merely a breakthrough, we will also evaluate its expenses in comparison to models such as ChatGPT and LLaMA.

We'll create a concise, data-driven narrative about DeepSeek's place in the AI race using information from APIs like TwelveData and interactive infographics made with D3.js and Tableau. This research attempts to clarify DeepSeek's influence on the AI sector, investment trends, and the larger tech environment by looking at market trends, media sentiment, and cost structures. This will help us determine whether DeepSeek is a game-changer or merely another development in AI.

CS Project Team Agreement

Communication

- Our primary method of communication will be Discord. Members are expected to respond within 24 hours of messages during the weekdays. For urgent requests or questions use the @ function.
- We will hold meetings on Sunday evenings at 8PM to discuss progress on the weekly deliverables and plans for future deliverables.

Code Guidelines

- Github branch naming convention: feature/[feature-name]
- All non trivial code must be commented adequately to ensure other members can use it. Functions must all have docstrings detailing their parameters and return values.
- Use the d3 conventions seen in the labs when making graphs and visualizations.

Tasks

- Github Projects board to organize tasks based on priority and deadline.
- When working on a task mark it as in progress and notify other members so no one repeats work. Once a task is complete mark it as such.

Version Control

- Use Github as our main version control tool
- Main branch protected and all merge requests require at least one other member to approve the merge.
- Commit messages must adequately describe the feature implemented. The merge request's comment thread will be used for code reviews and discussion.

Quality Standards

- Code must follow conventions discussed above.
- Each new added function or code must be given proper documentation before merging.

Team

Members: Liam Maguire, Dmitrii Vlasov, Inan Sanon

Date: 05/02/2025

Signed: 

Project Plan

Basic Info

Project title: The Deepseek Effect

Members:
- Liam Maguire - liam.maguire@mail.utoronto.ca
- Inan Sanon - inan.sanon@mail.utoronto.ca
- Dmitrii Vlasov - d.vlasov@mail.utoronto.ca

Team Name: LID

Background and Motivation

Nvidia has long dominated the AI industry, but the emergence of a more affordable and higher-performing alternative has introduced real competition in the field. The release of Deepseek has had significant impacts on the economy, media, and public perception of AI, and we aim to tell this story with data.

Data

To achieve these goals, we will collect real-time and historical stock data using the **TwelveData API**, ensuring accurate financial trend tracking. For sentiment and media analysis, we will use **NewsAPI** to gather articles and headlines, allowing us to generate a dynamic word cloud that highlights the most frequently mentioned terms and concerns. Finally, to assess DeepSeek's technological impact, we will gather comparative pricing and efficiency data from **official model documentation** and **third-party benchmarks**.

Data Cleanup

We do not expect there to be much data cleanup except maybe filtering the headlines for language, and relevance to the topic.

Map

Discussion

1) Possible target audiences are people who use AI, stock market investors, or people who build AI. For our main target audience we will choose people who use AI.

2) Our target audience is people that use LLMs to help them with daily tasks or work. They are aware of the main commercial AI models but may not know the differences between them. They have moderate visualization literacy as they are tech savvy enough to integrate AI into their daily life but are not necessarily experts. The level of detail we expect is not too high. We expect to compare at most 3 metrics per visualization.

3) Question about our data:
    a) How big of an impact did deepseek have on Nvidia?
    b) How does deepseek compare to its competitors?
    c) How much cheaper is deepseek compared to its competitors?
    d) Does the increase in quality correlate to an increase in training?
    e) Can we expect similar models to deepseek to be developed to compete with GTP?
    f) What other companies were affected by the release of deepseek?
    g) How cost effective is deepseek compared to other models?
    h) What are the best performing LLMs?
    i) How many new LLMs are being released every year?
    j) Who's hardware does a model's training happen on?
    k) Does a model's training dataset size correlate to its performance on benchmarks?
    l) How has the media reacted to the release of deepseek?
    m) How do people feel regarding the fact that deepseek is a model developed in China?

4) Data Analysis:

We have three stock datasets, one for each of the publicly traded companies of interest: Nvidia, Google, and Microsoft. They include stock data on the day of Deepseek R1 reveal (between 24th Jan 2025 and 27th Jan 2025)

1. **datetime** (Categorical, Time-Series)
    - This represents the timestamp of the stock data, including both the date and time at 15-minute intervals. While it follows a structured format, it is typically treated as a categorical variable in time-series analysis.

2. **open** (Quantitative, Continuous)
    - The stock's price at the start of the 15-minute interval.
3. **high** (Quantitative, Continuous)
    - The highest price reached during the 15-minute interval.
4. **low** (Quantitative, Continuous)
    - The lowest price reached during the 15-minute interval.
5. **close** (Quantitative, Continuous)
    - The stock's price at the end of the 15-minute interval.
6. **volume** (Quantitative, Discrete)
    - The number of shares traded during the 15-minute interval. Since it represents a count of shares, it is a discrete variable.

We also have information about all the major LLMs:

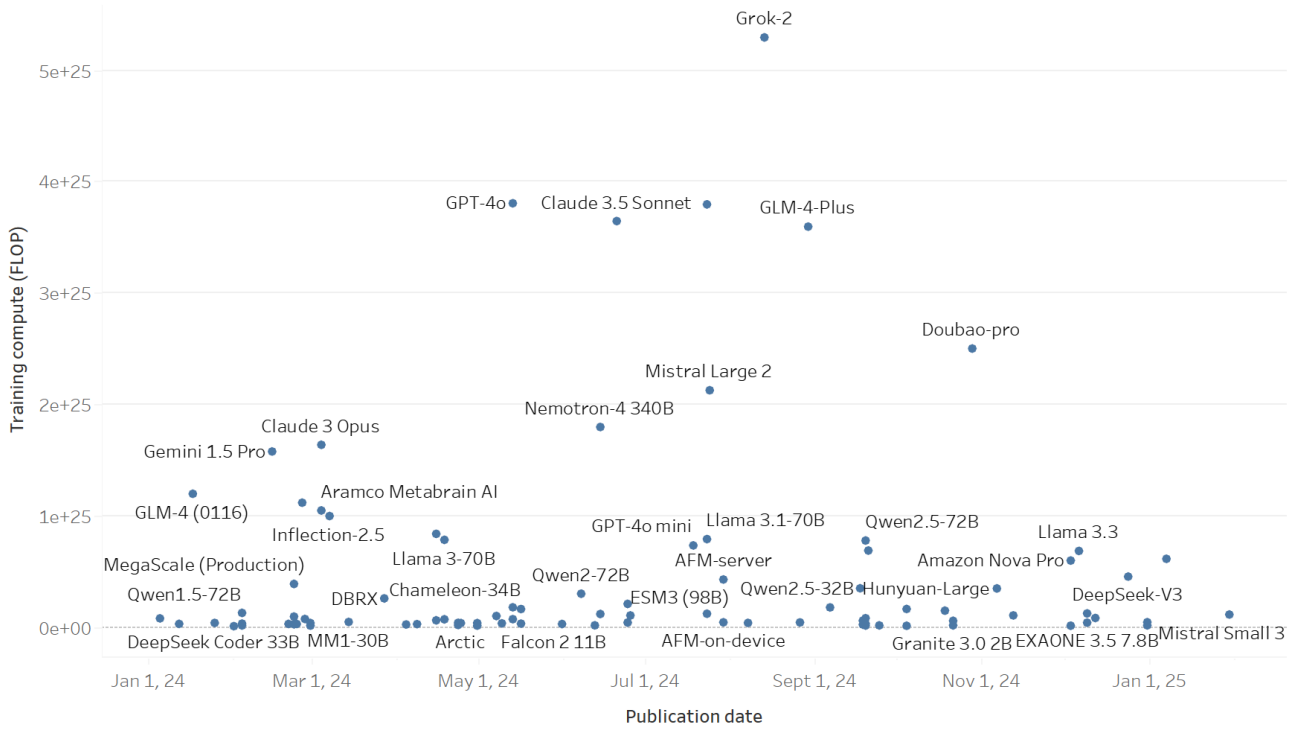
1. **Model** (Categorical): Name of the AI model
2. **Publication Date** (Quantitative - Discrete) - Date the model was published
3. **Parameters** (Quantitative - Discrete) - Number of parameters for the model
4. **Training Compute** (Quantitative - Discrete) - Number of operations to train the model
5. **Training Dataset Size** (Quantitative - Discrete) - Number of datapoints in the training data set
6. **Training Hardware** (Categorical) - The processor used to train the model
7. **Country** (Categorical) - Country in which the company that published the model is based

Next is benchmarking data on the major LLMs:

1. **Model** (Categorical): Name of the AI model
2. **Mean** (Quantitative - Continuous) - The mean score achieved by the model
3. **Stderr** (Quantitative - Continuous) - The standard error of the means
4. **Task** (Categorical) - The benchmarking task that the model was tested on
5. **Best Score** (Quantitative - Continuous) - Best score achieved for a model on a given task

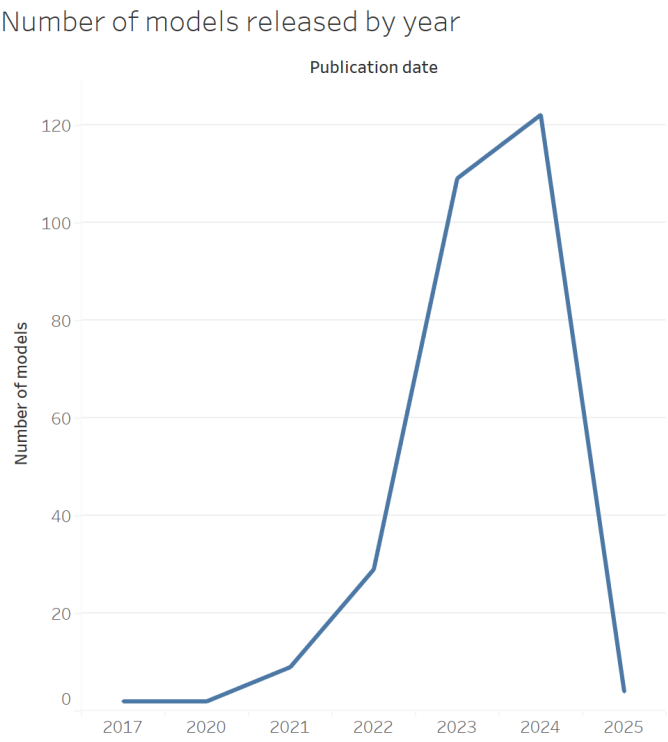
Liam Maguire:

Training compute of models in past year



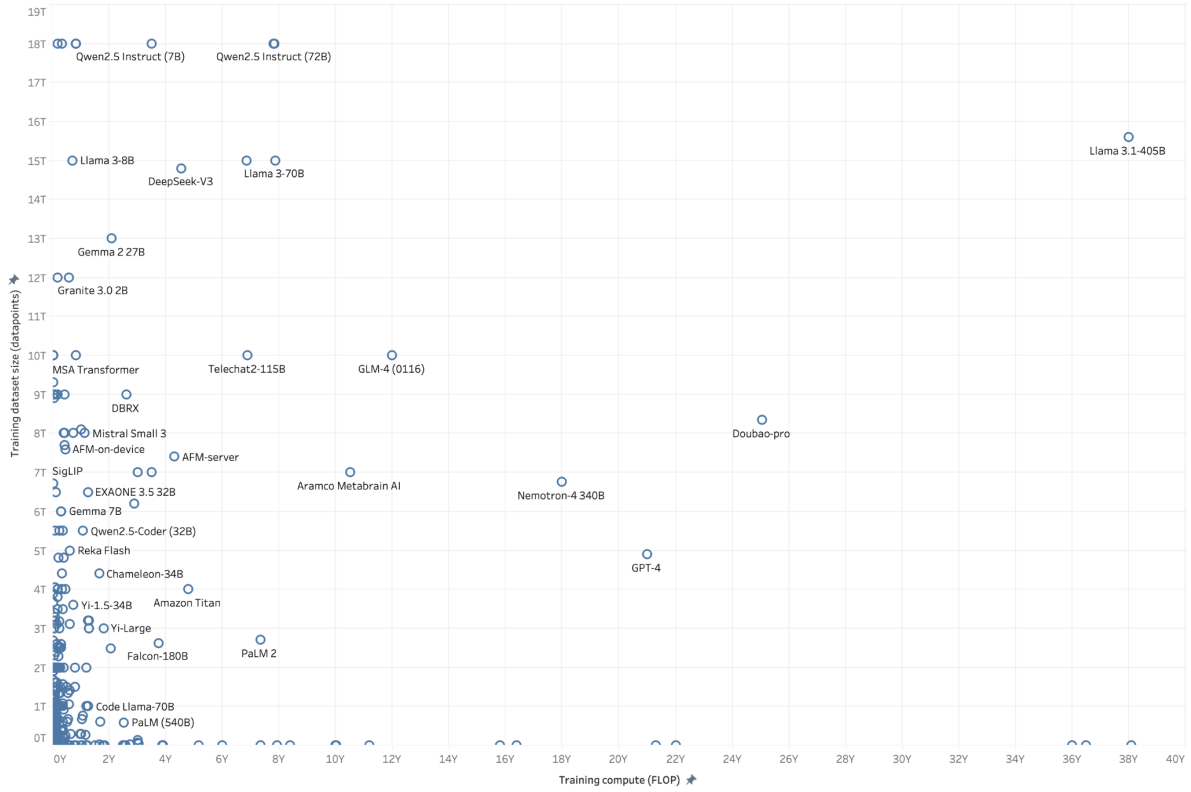
Training hardware for largest comercial AI models

Training hardware	
AMD Radeon Instinct MI2..	Poro 34B Viking
Google TPU v5e	Gemini Nano-2 Gemma 7B
NVIDIA A100	Amazon Titan Granite 13B
NVIDIA A100 SXM4 40 GB	Falcon-180B GPT-3.5 GPT-4
NVIDIA A100 SXM4 80 GB	Chameleon-34B
NVIDIA A100 SXM4 80 GB..	Pharia-1-LLM-7B
NVIDIA A100,NVIDIA H10..	Reka Core
NVIDIA H100 SXM5 80GB	Cosmos-1.0-.. Inflection-2.5 Llama 3-70B
NVIDIA H800 SXM5	DeepSeek-V3

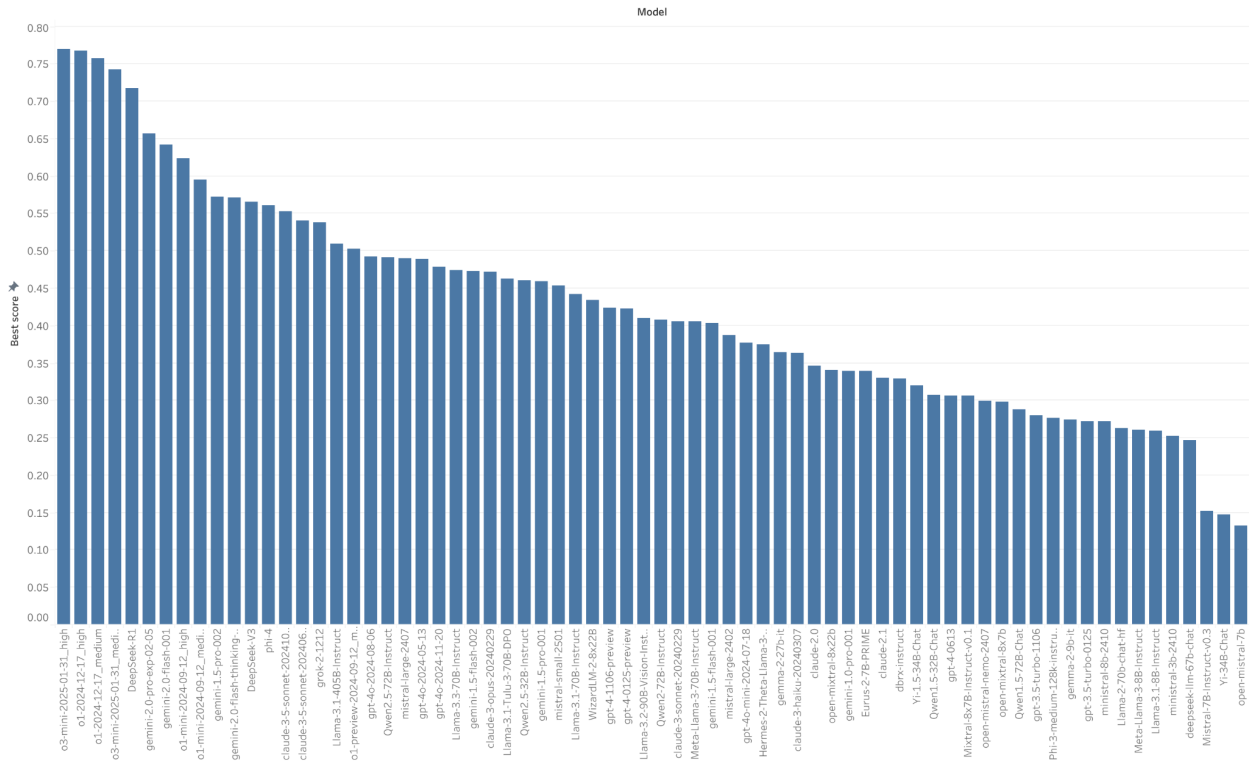


Dmitrii Vlasov:

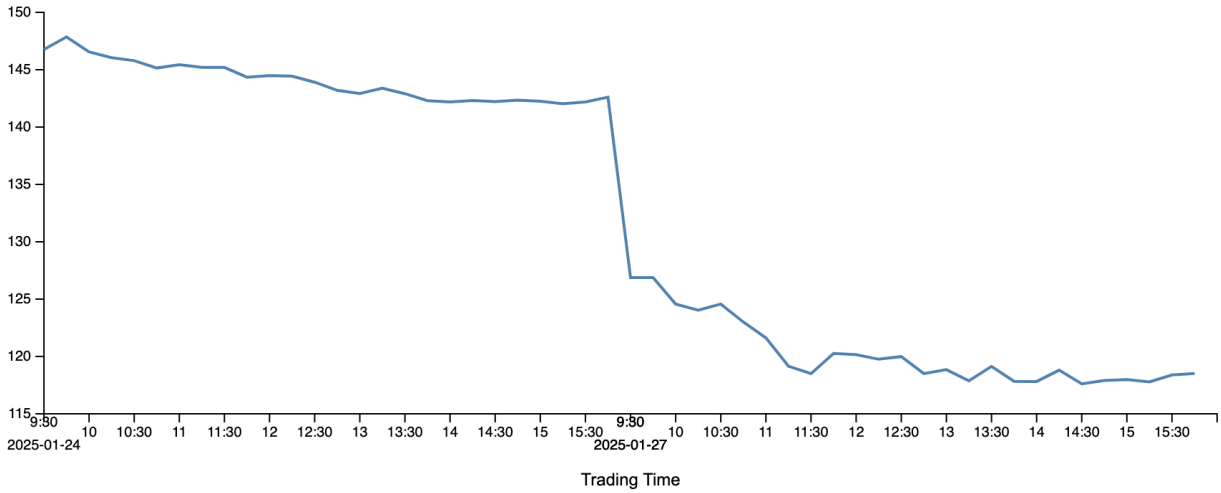
Scatter Plot (Training Compute vs Dataset Size)



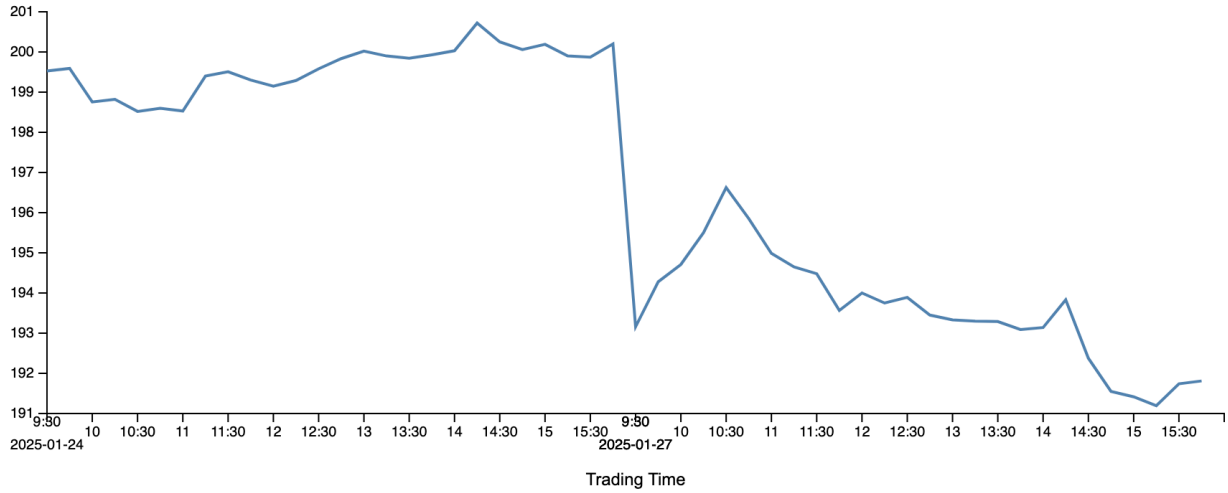
QPQA Diamond Benchmark Score



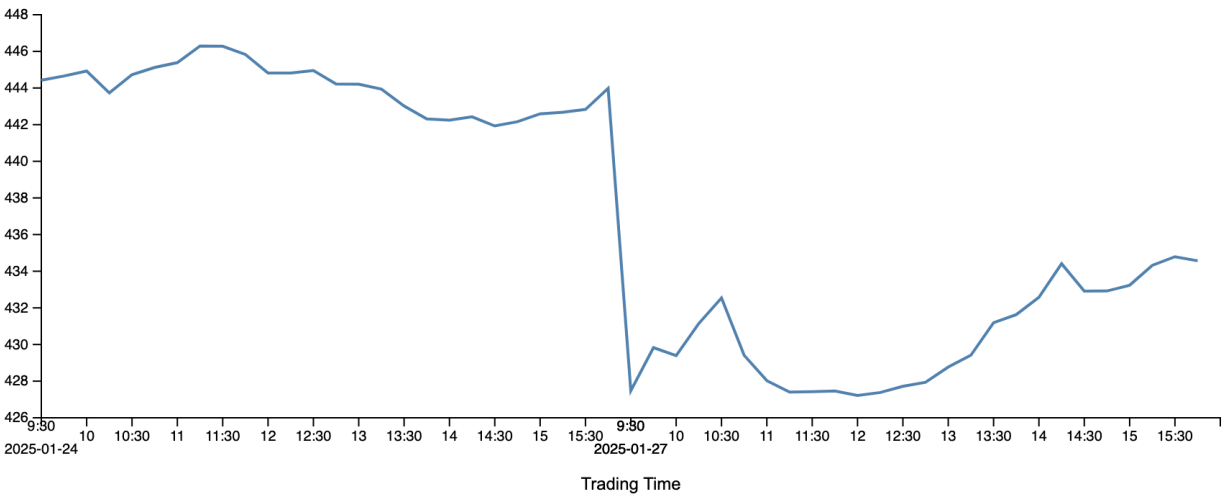
NVDA Closing Price (Compressed Trading Hours)

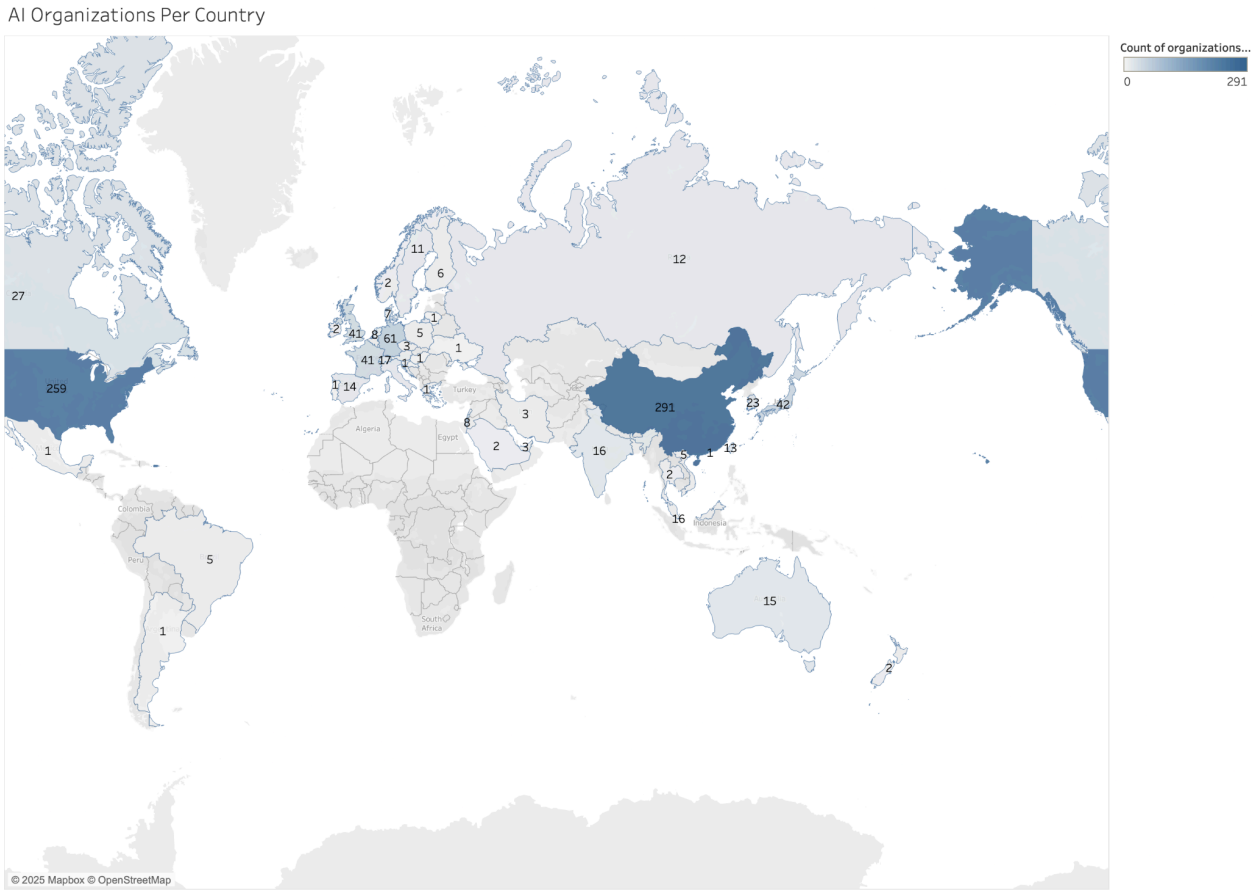


GOOGL Closing Price (Compressed Trading Hours)



MSFT Closing Price (Compressed Trading Hours)





Question Refinement after Tableau Visualizations

While our initial research questions provided a broad framework, the questions we ultimately answered in Tableau were refined based on data availability and clarity in visualization. Some original questions, such as "How do people feel about DeepSeek being a Chinese model?" were too subjective to answer solely through data analysis, requiring sentiment interpretation beyond what a word cloud could provide. Others, like "Who's hardware does a model's training happen on?" were straightforward and easily visualized, making them a natural fit for our project. Additionally, while we initially considered tracking multiple AI companies, the most significant financial impact was observed in Nvidia's stock drop, leading us to prioritize that aspect. The shift in focus allowed us to create more meaningful and visually compelling insights, ensuring that our final analysis was both data-driven and relevant to our core objective—understanding DeepSeek's disruption in AI and its industry-wide ramifications.

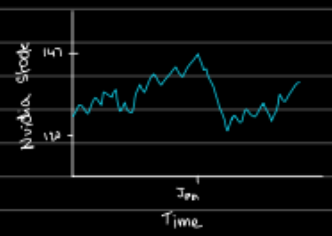
New questions:

a) How significantly did DeepSeek's release affect Nvidia's stock price?
b) How does DeepSeek compare in efficiency and performance to its competitors?
c) How much cheaper is DeepSeek compared to other leading LLMs?
d) Is there a correlation between model quality and training compute?
e) Will DeepSeek's release encourage competitors to develop similar models?
f) Which other companies were financially impacted by DeepSeek's launch? How cost-effective is DeepSeek relative to other AI models?
g) What are the best-performing LLMs in the industry?
h) How many new LLMs are being released annually?
i) Who provides the hardware for training leading AI models?
j) Does training dataset size correlate with performance on benchmarks?
k) How has the media responded to DeepSeek's release?
l) What is the general sentiment regarding DeepSeek's Chinese origins?

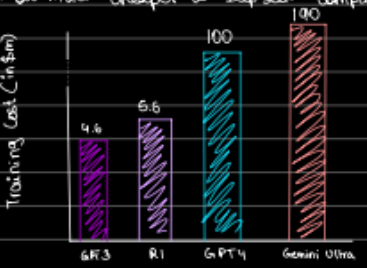
Inan Sanon

1) How big of an impact did Deepseek have on Nvidia?

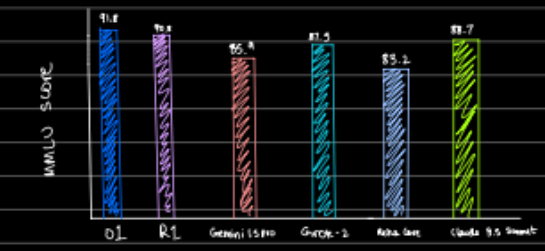
*DS R1 released in January 2025



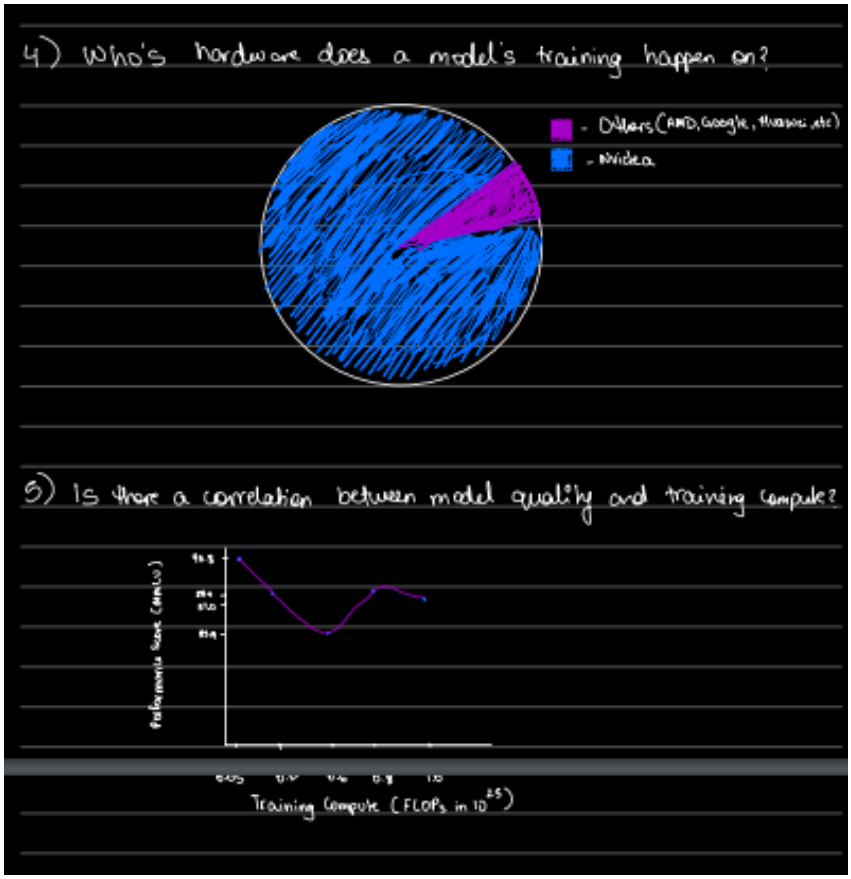
2) How much cheaper is Deepseek compaired to its competitors?



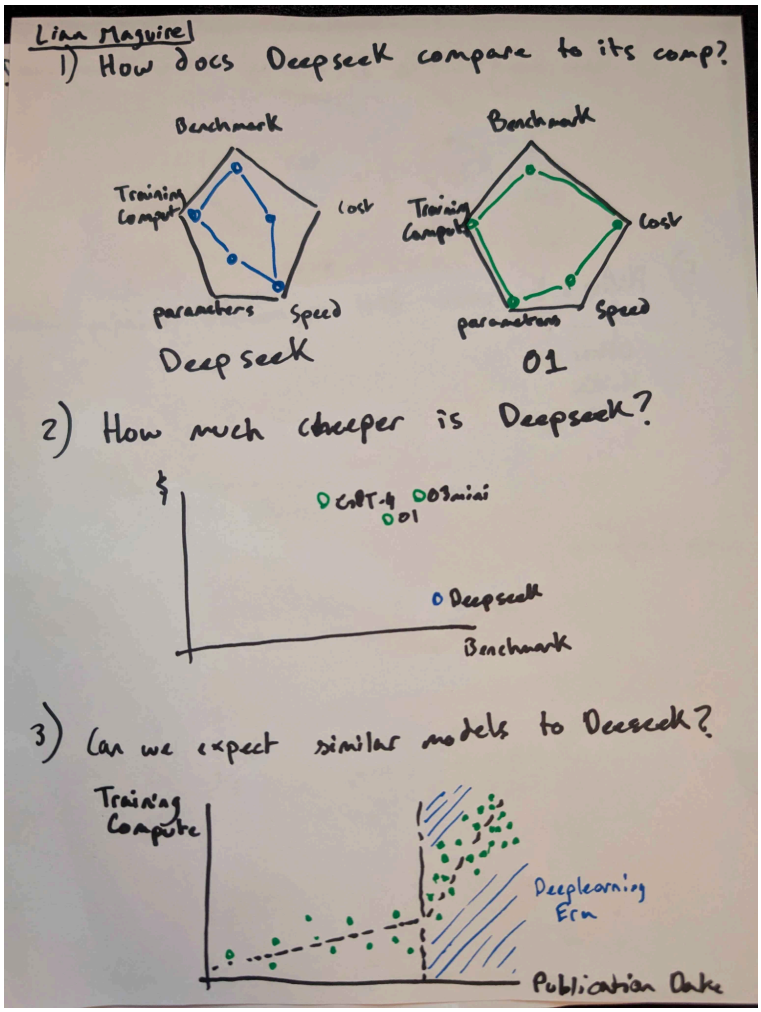
3) What are the best performing LLMs?

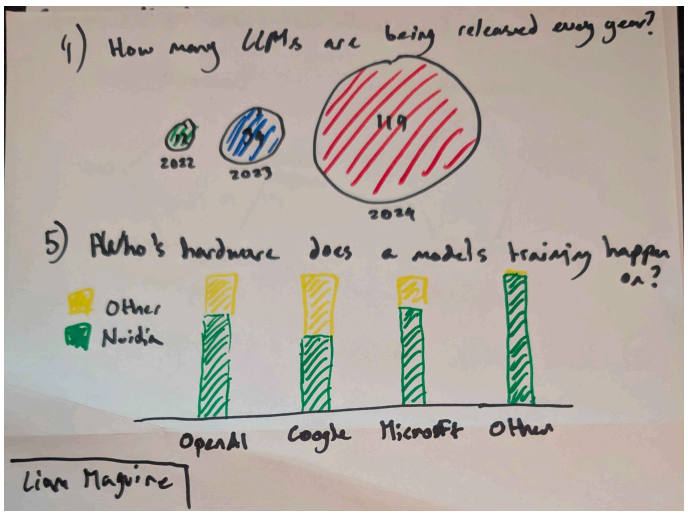


4) Who's hardware does a model's training happen on?



5) Is there a correlation between model quality and training compute?

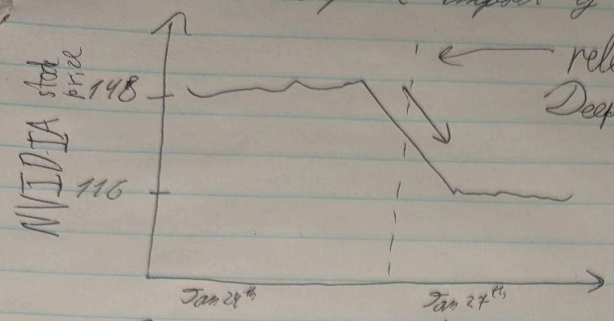




Dmitrii Vlasov

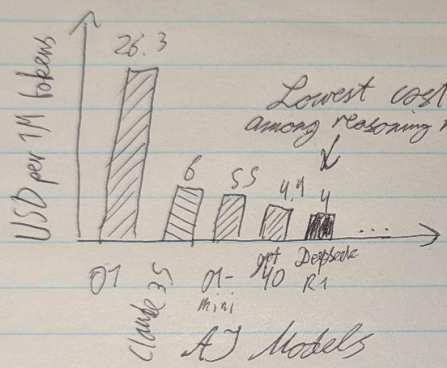
Q A - Line Chart

Stock price impact of DeepSeek



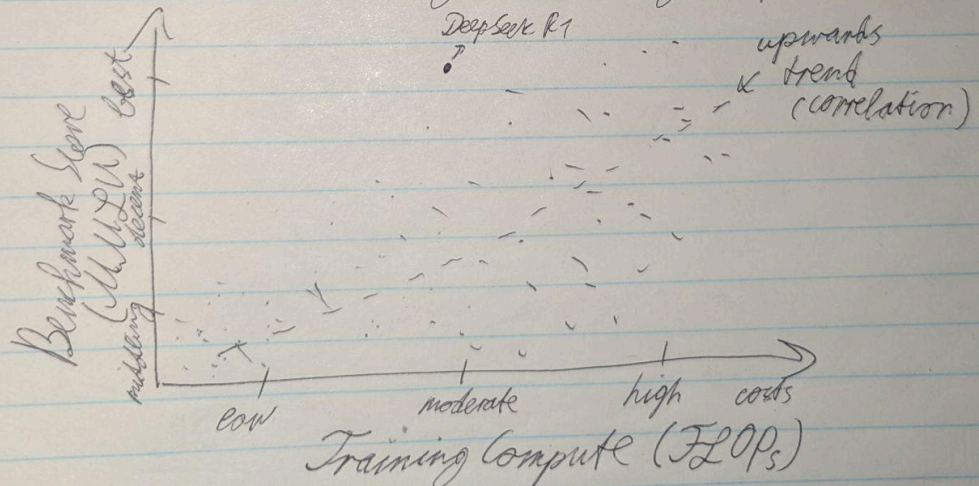
QC - Bar Chart

DeepSeek Cost vs. Competitors

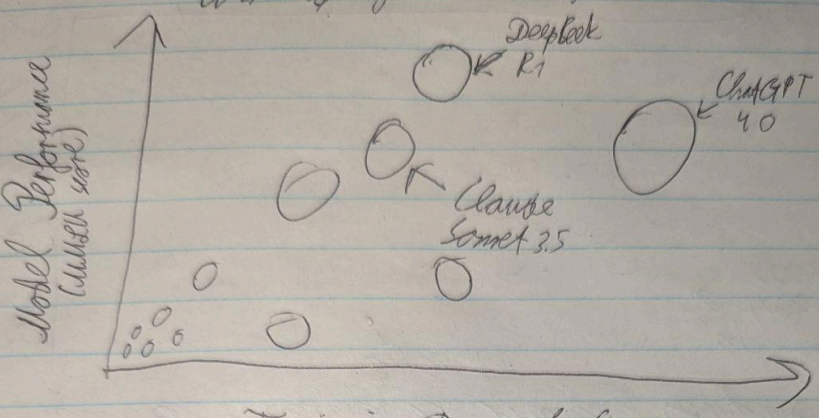


QD - Scatter Plot

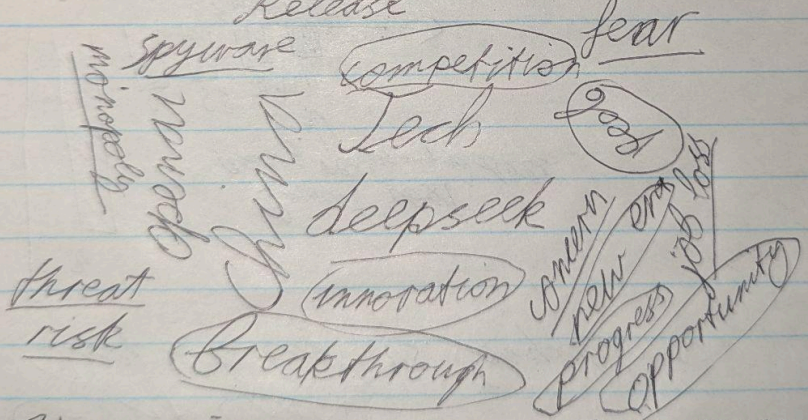
Model Quality vs. Training Compute



Q J - Bubble Chart
Does Training Dataset Size Correlate with Performance?



Q K - Word Cloud
Media Sentiment Toward DeepSeek's Release



Bigger size - more common in news headlines
positive words - circled
negative words - highlighted

Sketch ID (Initial + number)	Question ID	Author
I1, D1	A	Inan, Dima
I2, L2, D2	C	Inan, Liam, Dima
I3	G	Inan
I4, L5	I	Inan, Liam
I5, D3	D	Inan, Dima
L1	B	Liam
L3	E	Liam
D4	J	Dima
D5	K	Dima
L4	H	Liam

Vote Results:

I1, I4, D2, L2, L3

Storyboarding:

Hook: Surprising sudden fall in Nvidia stock price

I1: Nvidia stock viz

Rising Insight: Most models train on Nvidia chips but are expensive to train

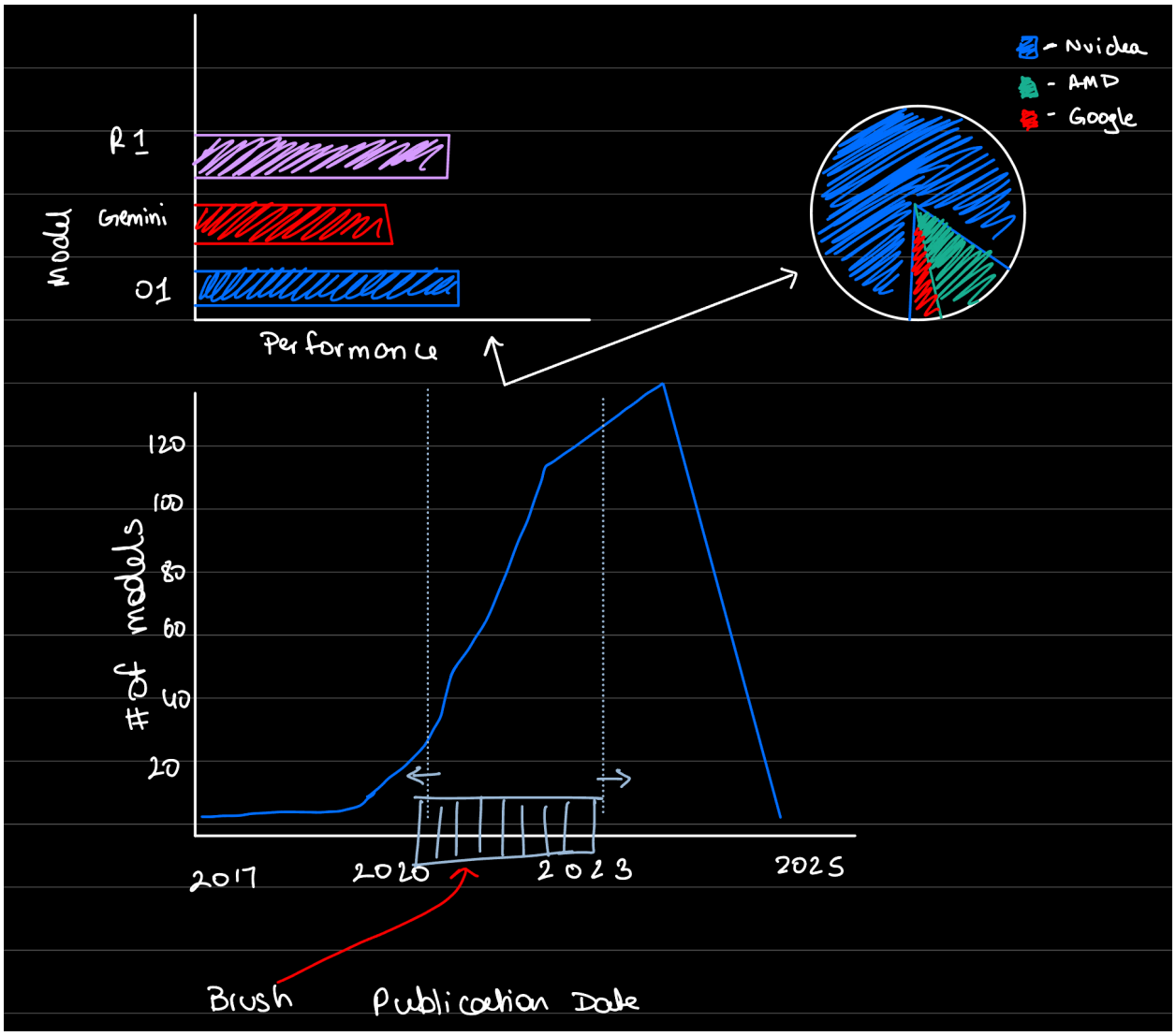
I4: Hardware used for ai training
D2: Cost of model use

Main Message: Meanwhile Deepseek is powerful and cheap

L2: Cost versus benchmark

Solution: Impact of deepseek; future trends?

L3: Projection of power of ai models



This innovative and interactive visualization, where you can brush across the interactive timeline, that will then apply the filter to the hardware pie chart that displays the distribution of the different hardware across the models that were released in the selected timeline, and to the bar chart of performance of said models.