AI Financial Bubble and Hype Cycles

By: Jurgen Shimani | MSDS*640 Ethics, Privacy, and Social Justice

---

Introduction

Since the public release of ChatGPT in 2022, artificial intelligence has become integrated in our everyday lives. The platform surpassed over 100 million users within its first months, marking one of the fastest-growing platfroms, overshadowing social media services such as TikTok and Instagram (Alexandra Andhov, 2025). With such unprecedented growth, it is only natural that artificial intelligence has drawn a lot of attention from investors. This momentum has contributed to what many would describe as the "AI boom", during which major technological conglomerates, including Google, Microsoft, IBM, and NVIDIA, to increase in market validation in an astronomical rate. This rapid expansion has fueled speculation that an "AI Bubble" has emerged, characterized by intense competition among firms seeking to position themselves as the market leaders and by substantially investing in startups claiming to be the next ChatGPT or Claude. If an AI bubble is indeed forming, its scale could potentially compete with the housing market crash of 2008, given the amount of capital being directed toward companies identifying themselves as AI startups. The current state of the investment climate reflects a willingness among venture capitalists and institutional investment companies to fund virtually any enterprise associated with Large Language Models, raising concerns about inflated valuations and long-term sustainability. However, this level of investment also reflects a strong belief among investors in the potential of AI technologies, especially those of LLMs. The willingness to allocate billions of dollars to fund AI-focused ventures suggests that many stakeholders view these innovations as having substantial long-term value, despite the short-term risks they could pose. The central question remains whether the current landscape mirrors previous speculative periods, such as the dot-com bubble or the cryptocurrency surge, and whether the rapid growth of the industry may ultimately lead to a significant market correction.

The Ethical Issue

With such significant capital at stake and considering the outcomes of previous market bubbles, it is often ordinary people who bear the greatest impact when a crash occurs. The potential for substantial financial losses within a short period raises important ethical questions regarding responsibility and accountability. Furthermore, organizations like Open-AI-operating as a non-profit organization until 2025, were exempt from tax obligations, thereby reducing the direct fiscal contribution while relying on "donations" by billionaires, functioning as effective investing strategies instead (Alexandra Andhov, 2025). This dynamic places a substantial burden on taxpayers, who indirectly support these sectors while taking on the risk associated with potential market instability. In addition to financial risks, the environmental impact that these AI models have, is becoming a growing concern. The energy demands of training and operating advanced models are substantial, with estimates suggesting that more than 32 billion dollars' worth of GPU power consumption and an annual energy use that exceeds 8.4 TWh. This level of consumption translates to roughly 3.25 gigatons of CO_2 emitted into the ozone layer, the equivalent of five billion cross-country flights within the United States. Such figures highlight the significant, not only financial, but also ecological costs associated with the current pace of AI innovations (Oviedo Felipe,2025).

What I Investigated

OpenAI's Infrastructure and Operational Costs

Training, maintaining, and operating large-scale LLMs such as ChatGPT requires an extraordinary level of computational resources. Estimates suggest that the GPU requirements of GPT-4 alone amount to approximately $800 million in graphical processing costs, in addition to roughly $40 million in amortized hardware and energy expenditures. Research and development expenses, including staff and equity compensations, amount to 29-49% of total amortized costs. Notably, the amortized costs of training these models have increased by 2.4per year since 2016. Across most frontier AI systems, the largest cost drivers are the AI accelerator chips, which are crucial in training these models, and also personnel costs. Server components account for 15-22% of the total expenditure, while cluster-level interconnect typically add another 9-13%. Energy consumption, though often the focus of public discussion, is in fact only a relatively small amount, represent-

ing only 2-6% of the total costs(Cottier Ben, 2024). Despite being the smallest component, this nevertheless reflects on the immense financial burden that LLMs require to scale. These costs extend well beyond the models themselves. The broader AI infrastructure ecosystem is undergoing a rapid expansion. For example, NVIDIA sold approximately $15 billion in GPUs between May 2022 and April 2023 alone, with a quarterly revenue projection doubling from $4 to $8 billion over the same period. Major hyperscalers, such as Amazon, Google, and Microsoft, also increased data-center capital expenditure from $78 billion in 2022 to $120 billion in 2023, a 54% increase year-over-year(Chien A. Andrew, 2023). Power consumption associated with AI companies also represents the scale of this growth. Training and operating a complex LLM is estimated to require more than $32 billion worth of GPU power annually, which translates approximately to 8.4 TWh of electricity. This corresponds to an estimated 3.25 gigatons of CO_2, using average U.S grid emissions (Chien A. Andrew, 2023). Collectively, these figures highlight the immense financial and environmental burden associated with sustaining a state-of-the-art LLM. The operational footprint of AI companies is expanding rapidly, underscoring the significant infrastructural, economic, and ecological challenges that come with advancing these AI systems.

API Usage and Adoption

In a study discussed by Irugalbandara Chandra in Scaling down to scale up: A cost-benefit analysis of replacing OpenAI's GPT-4 with self-hosted open source SLMs in production.", the authors note that more than two million developers have adopted large language models, with OpenAI's APIs characterized as the preferred cloud-based solution. However, these figures primarily reflect developer-level usage, rather than organizational adoption, and therefore do not fully capture enterprise-scale integration(Irugalbandara Chandra, 2023). The study also highlights the scalability of LLM-powered applications, reporting that systems averaging 1,000 requests per day can scale to over 360,000 requests per day in real-world environments(Irugalbandara Chandra, 2023). A platform like ChatGPT itself processes approximately 2.5 billion queries per day as of July 2025, surpassing search engine platforms such as Bing, Yahoo, etc., which collectively handle only 10 billion queries per day(Oviedo Felipe, 2025). Despite this immense user volume, the majority of ChatGPT's consumer-facing queries contribute little to no revenue to the company, largely because: 1. Many users access the service for free or through a relatively low-cost subscription, like $20 per month. 2. A substantial portion of user traffic is used for data collection, model fine-tuning, and reinforcement learning, all of

which improve the system over time by providing data. Overall, while user engagement demonstrates significant demand, with LLMs replacing search engines and widespread adoption, the economic value derived from such traffic remains modest compared to the operational costs of running a frontier LLM system. However, as AI models continue to integrate more deeply into our daily lives, this economic landscape is about to shift. The long-term value of large language models does not primarily stem from casual, day-to-day consumer usage, but rather from enterprise adoption, where organizations rely on these systems to support their development and deployment. These enterprise clients represent the core revenue stream for AI companies, spending millions of dollars daily on large-scale automation, model deployment, and customized inference.

Customer Costs for Using OpenAI Services

Large language models such as OpenAI's GPT-4 API use a token-based pricing model, with different rates applied to input and output tokens. For GPT-4, the API costs $0.03 per 1,000 input tokens, and $0.06 per 1,000 output tokens(Irugalbandara Chandra, 2023). A request that includes both 1,000 input and output tokens will cost about $0.09. For organizations with moderate usage patterns, these expenses can scale rapidly. A concrete illustrative example comes from Duolingo, one of the few organizations that publicly post their monthly and daily users on social media, and also acknowledges the use of GPT-4 API. According to a recent official Reddit post by Duolingo, the platform houses roughly 46.6 million active daily users, but only 10 million of them are active subscribers. Due to Duolingo Max being a subscription-based service, and the reputation that it has on the internet, let's assume that out of 10.3 million subscribers, only half of them use the Max version. Assuming each user generates 25 input and 25 output tokens, and applying the GPT-4 pricing per token, Duolingo would incur an estimated $11,250 per day in API costs. This estimate is conservative, with actual usage varying widely, with some users generating significantly lower numbers of tokens and others more, but this highlights the substantial revenue generated by the API pricing model. Duolingo is just one of the publicly known examples, with many other enterprises integrating GPT-X models without disclosing specific usage metrics, suggesting that the total industry-wide spending is considerably higher. The cost structure consists of three primary components: prompt cost, proportional to prompt length, generational cost, proportional to output length, and, lastly, sometimes a fixed per-query cost (Chen Lingjiao,2023). For processing large volumes of text, GPT-4 costs $30 for 10 million tokens, representing a

substantially higher expense than using one of the older alternatives, such as GPT-J, which is $0.2 for the same volume. However, due to the competitive pressures, marketing, and demand of state-of-the-art performance, most large enterprises opt to use the newer versions of these LLM models, rather than the older, but cheaper alternative. At the operational level, OpenAI's daily ChatGPT operation costs are estimated to be over $700,000, though this number references overall performance costs rather than per-customer charges(Chen Lingjiao,2023). The tokenization model also introduces language-dependent cost variations, with some languages experiencing a 5higher cost than English due to less tokenization training(Ahia Orevaoghene, 2023).

Data Sources:
•Duolingo Users Data Source: https://www.reddit.com/r/duolingo/comments/1
text = *Monthly*%20*active*%20*users*, 10.3*M*%20(9.5; %207.4)
https : *//www.youtube.com/watch?v* = *−cdJQ*8*UyVLAlist* = *PLHu*8*Vs* − *bmiQopbjOUgoAdmyjnUz*2*t*1*c*1*W*

Key Findings

Historical Stock Price Graph Analysis

AI companies and technology giants have experienced remarkably high valuations in recent years, with firms such as Microsoft, Amazon, Google, and NVIDIA reaching unprecedented stock prices. This rapid growth has led to speculations and fueled the public concern that an economic bubble may be forming, raising questions about whether these companies are becoming significantly overvalued. In this section, I examined the stock performance of these four organizations over the past five years and analyzed their annual return rates to assess whether their valuations appear to be increasing at an unsustainable pace.

Interestingly, the return rates before and after the rise of generative AI show relatively little change to the stock price for most major technological companies. The notable exception is NVIDIA, which experienced a drastic surge in value. This is unsurprising, though, due to NVIDIA being effectively the primary supplier of high-performance AI accelerator chips, which significantly outperform traditional GPUs for large-scale model training. With the rapid expansion of LLMs and the broader AI ecosystem, NVIDIA's unique position as the dominant leader of hardware providers, with limited to no competition, has naturally led to a massive appreciation by investors. While Google has begun developing its own in-house AI chips (TPUs), it continues to heavily rely on NVIDIA's hardware, as they currently offer superior performance and support. This dynamic places NVIDIA in an advantageous position when it comes to the AI ecosystem, resulting in

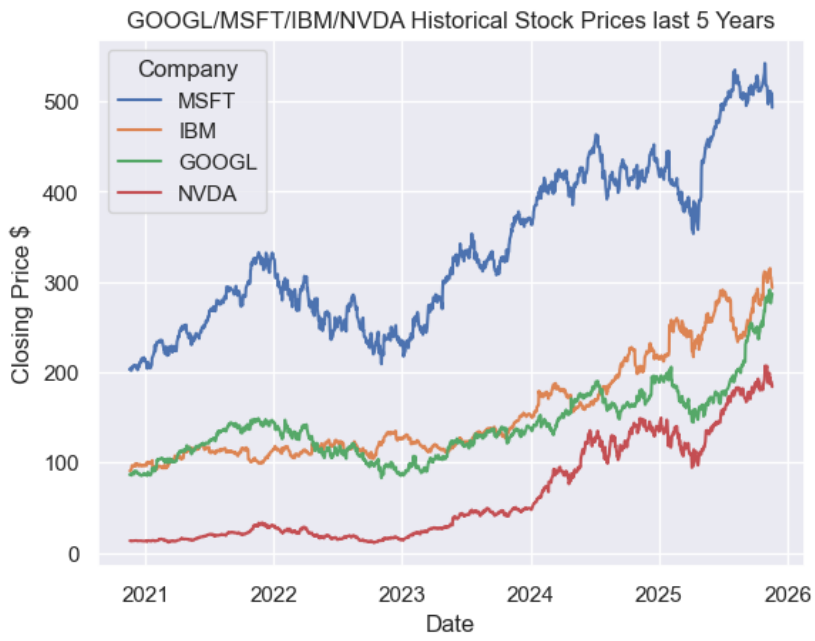

Figure 1: Historical stock prices of Microsoft, Google, Nvidia, and IBM.

growth that exceeds that of many AI software companies.

Relationship between US economy and AI companies

In a recent video, finance commentator Andrei Jikh provides an insightful discussion on the flow of capital within the AI ecosystem and the resulting in economic dynamics. He highlights how leading AI companies are generating billions of dollars in revenue while simultaneously operating at substantial losses, raising the question about the sustainability of their business models and the scale of ongoing investment(Jikh, 2025). Jikh explains that two key concepts can help us understand this phenomenon. The first relates to the sheer volume of capital circulating within the AI ecosystem. As illustrated in Figure 3, the financial loop operates as follows: NVIDIA invests $100 million in OpenAI, and in turn, uses that capital to expand its data center capacity by purchasing cloud infrastructure from providers such as Amazon, Oracle, Microsoft, etc. Observing the surge of demand for their services by AI companies, these institutions, in return, scale their infrastructure horizontally by acquiring additional high-performance computing resources- largely buy purchasing NVIDIA hardware. Consequentially, the capital returns to NVIDIA, creating a self-reinforcing cycle, where no new money is flowing inside the bubble. This is really concerning due to the unattainability of the model. At first glance, this

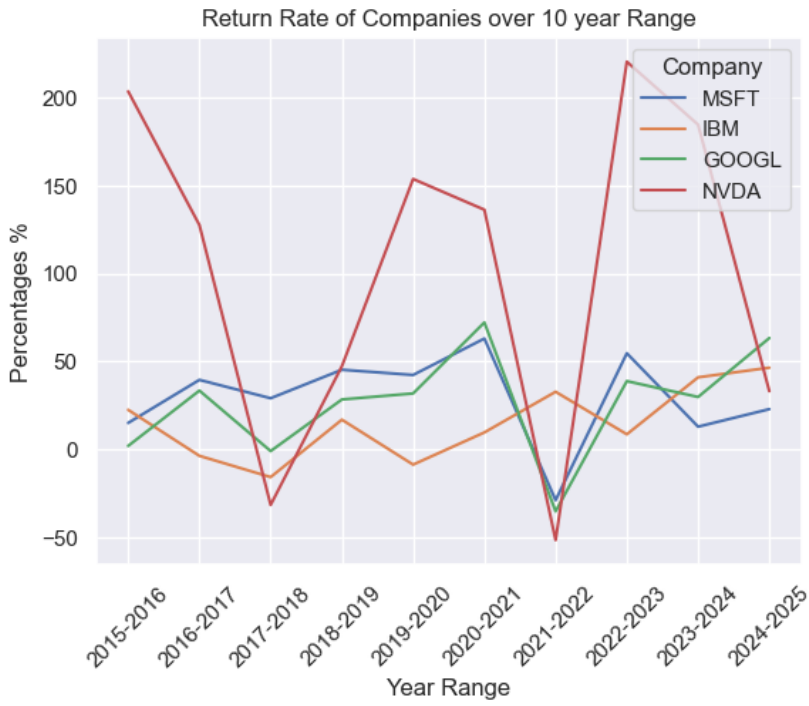

Figure 2: Return rates of companies from 2015-2025 to check if there is a big difference between returns over the last five years.

cycle flow of capital may suggest that AI sector is exhibiting characteristics of a bubble. However, Jikh highlights an additional structural factor that significantly shapes the AI ecosystem: the implicit governmental support for leading the AI firms. In an interview with Sam Altman, CEO of OpenAI, Altman is asked whether insufficient capital could prompt a government bailout. Although he does not answer explicitly, his remarks strongly suggest that such support would be likely. This position is not unreasonable. Given the geopolitical war between the United States of America and China in artificial intelligence, no party want to fall behind, so the governments have strategic incentives to prevent domestic AI leaders from falling. Another dimension of the implicit safety net lies in the broader U.S economy. A substantial portion of the S&P500 is approximately composed of 40% of only AI-driven companies, whose rapid growth has significantly increased their weight in the index. Because index funds allocate proportionally more to firms as their market capitalisation grows, the performance of AI companies now has a massive influence on the overall market stability. A severe downturn in the AI sector would therefore trigger a parallel decline on the S&P500, potentially participating in a crash bigger than the 2008 housing market collapse. For this reason, major AI companies continue to deploy their capital aggressively, often expanding their investments despite operating at a net loss. Their willingness

to do so reflects an underlying expectation that, should their financial resources become insufficient, external support, particularly from the government, would intervene to prevent a collapse. This implicit guarantee reduces the perceived risk of overspending and reinforces the rapid, investment-driven growth characteristics of the AI ecosystem.

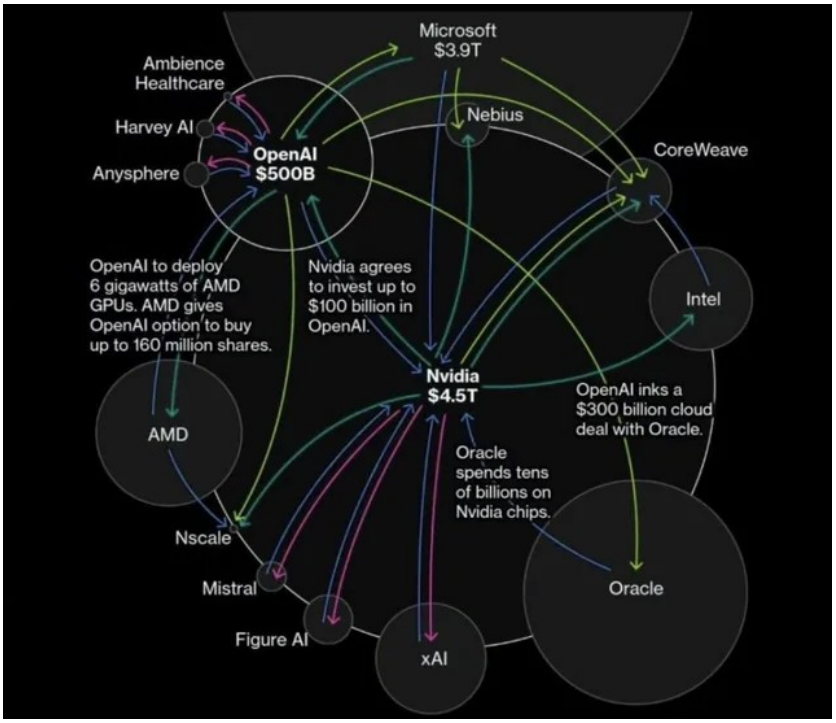

Figure 3: The circle of investment circulating in the AI space.

Key Insight:
[Use margin notes for key takeaways or interesting observations]

What This Means

Based on the data and the sources examined, it appears that artificial intelligence currently represents the technological frontier and is the currency of tomorrow, shaping the future across multiple industries. Although significant challenges may lie ahead, particularly the accumulation of public and private debt and the potential for market corrections, the strategic imperative to lead AI development outweighs these short-term economic risks. For the U.S. government, maintaining global leadership in AI is viewed as more valuable than the immediate financial costs, as technology dominance of today is likely to determine and shape the geopolitical influence in the world for decades to come. Because most countries lack the resources and infrastructure necessary

to create and maintain these LLM models, reliance will naturally shift toward the nations and their companies capable of delivering the most advanced, accessible and cost-efficient AI systems. As a result, geopolitical and economic influence will align with those who can deliver superior AI capabilities, further reinforcing the strategic importance of leadership in this domain.

Recommendations

We have examined how government support, funded through taxpayer money, plays a significant role in sustaining large AI companies that are currently operating at a substantial loss in pursuit of LLM dominance. Ethically, this raises critical questions about the impact on ordinary citizens who do not have the resources to invest tons of capital in these technological AI organizations. As highlighted by the U.S. Government Accountability Office(GAO), rising government debt can directly affect the public through higher living costs, increased borrowing expenses, and reduced fiscal flexibility (U.S. Government Accountability Office, n.d.). Given the scale of federal investment in AI firms, many of which have no clear short-term path to profitability, these financial pressures may ultimately be shifted onto citizens through higher taxes and escalating prices across essential goods and services. In this context, it becomes ethically important for individuals and policymakers to recognize the current technological transition and proactively prepare for its economic consequences by leveraging and capitalizing the current window of rapid AI growth in the near future, which can mitigate some of the foreseeable societal burdens that could arise from the government's long-term commitments to the AI sector.

Questions:
Are AI companies still worth investing in after all the hype so far?

Conclusion

Based on the research conducted thus far, an imminent "AI bubble" appears to be unlikely. The rapid technological growth of large language models(LLMs) and the substantial infrastructure required to develop and maintain them have created an environment in which only a small group of major technology firms can feasibly support these systems. As AI becomes increasingly more integrated into everyday life, most organizations seeking to adopt such technologies are expected to rely on the existing api models rather than developing their own, give the considerable financial and technical barriers involved, putting the AI

industry leaders at a big advantage. Current evidence suggests that the AI ecosystem is still growing and remains an attractive sector for investment, contrary to some media outlets' narratives that discourage the average investor by framing the field as overinflated. Some companies may appear overvalued from the perspective of traditional financial analysis, largely because their recent growth lacks historical precedent. However, such assessments often overlook the transformative technological shifts that advanced LLMs are expected to generate. Evaluating these firms solely through the lens of present-day financial metrics fails to account for the long-term structural impact that they are going to have on future markets and industries. Although it is difficult to predict how long this period of stability and expansion will persist, particularly given the lack of reliable metrics for evaluating the pace of AI innovation and advancement, the field has progressed from limited conversational agents to highly capable LLMs within just five years. This trajectory indicates that AI is likely to continue growing and reshaping a wide range of domains as its capabilities evolve.

Sources

- Andhov, Alexandra, et al. "OpenAI's Transformation: From a Non-profit to a 157 Billion Valuation." Business Law Review 46.1 (2025).
- Andrei Jikh. (2025, November). The Al Bubble Is A Lot Worse Than You Think. YouTube. https://www.youtube.com/w cdJQ8UyVLA
- Cottier, Ben, et al. "The rising costs of training frontier AI models (2024)." URL https://arxiv. org/abs/2405.21015 2405.
- Chen, Lingjiao, Matei Zaharia, and James Zou. "Frugalgpt: How to use large language models while reducing cost and improving performance." arXiv preprint arXiv:2305.05176 (2023).
- Ahia, Orevaoghene, et al. "Do all languages cost the same? Tokenization in the era of commercial language models." arXiv preprint arXiv:2305.13707 (2023).
- Oviedo, Felipe, et al. "Energy Use of AI Inference: Efficiency Pathways and Test-Time Compute." arXiv preprint arXiv:2509.20241 (2025).
- TWh, Terawatt-Hours. "GenAI: Giga $$$, TeraWatt-Hours, and GigaTons of CO2." A Computational Inflection for Scientific Discovery 66.8 (2023): 4.
- Irugalbandara, Chandra, et al. "Scaling down to scale up: A cost-benefit analysis of replacing openai's gpt-4 with self-hosted open source slms in production." arXiv preprint arXiv:2312.14972 (2023).
- U.S. Government Accountability Office. (n.d.). How could federal debt affect you? Retrieved December 9, 2025, from https://www.gao.gov/americas-fiscal-future/how-could-federal-debt-affect-you