Market Overvaluation and Infrastructure Constraints in Generative Artificial Intelligence Systems

Received: 03.03.2026
Revised: 05.03.2026
Accepted: 10.05.2026
Published: 27.05.2026

Aykut Remziev[1], **Riza Rizaev**[2]

[1,2]*Konstantin Preslavski University of Shumen, Shumen, Bulgaria*

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (`https://creativecommons.org/licenses/by/4.0/`).

Abstract: *This paper examines the phenomenon of rapid market expansion in the Artificial Intelligence (AI) sector in recent years, questioning the existence of an economic "bubble." Current investment flows into leading tech companies and startups are analyzed, drawing parallels with the "Dot-com" crisis of the early 2000s. The study focuses on the discrepancy between the market capitalization of AI companies and the actual Return on Investment (ROI) they currently generate. The risks associated with the Hype Cycle, high hardware infrastructure costs, and regulatory challenges are discussed. The article concludes that while AI is a fundamental technology with long-term potential, current market valuations show signs of overheating, suggesting a potential market correction in the short term.*

Keywords: *Artificial Intelligence, AI bubble, Generative AI, Return on Investment (ROI), Dot-com bubble, tech sector, economic cycle*

1. INTRODUCTION

The technology sector is undergoing a period of radical transformation driven by the emergence of large language models (LLMs). Since the end of 2022, investor interest in companies such as NVIDIA, Microsoft, and Alphabet has reached levels that many analysts compare to the dot-com era of the late 1990s. The question facing economists and engineers alike is whether the technology's added value justifies trillions of dollars in market capitalization. In this context, artificial intelligence proves to be not merely a new technological tool, but a systemic factor that simultaneously transforms production processes and distorts market valuation mechanisms.

The combination of limited model transparency, the difficulty of measuring real value creation, and aggressive media positioning creates an environment in which investment decisions are increasingly based on expectations rather than verifiable economic outcomes. This necessitates an interdisciplinary analysis that considers generative AI not only as a software innovation, but as a complex system subject to physical, energy, and financial constraints. The observed imbalance between the speed of capital inflows and the pace of real economic deployment of AI systems raises a fundamental question about the sustainability of the current growth model. While financial markets price in future profits decades ahead, industrial and corporate adoption of generative AI remains fragmented, experimental, and often limited to pilot projects.

This asynchrony between financial expectations and operational reality creates conditions for the accumulation of systemic risk, whereby even a moderate slowdown in technological or regulatory development could lead to a sharp revaluation of assets. Figure 1 illustrates the advantages and disadvantages of investments in AI.

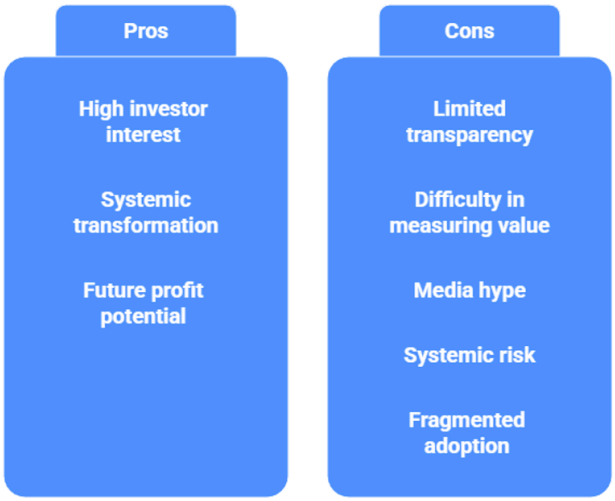


Fig. 1. Investments in AI

2. ANATOMY OF THE AI BUBBLE

The phenomenon of technological bubbles is not new; it follows a predictable pattern of social and economic behavior. At the core of the current situation surrounding Generative AI (GenAI) lies the so-called Gartner Hype Cycle, which defines five phases in the life cycle of an innovation.[1]

– **Positioning at the "Peak of Inflated Expectations"**

At present, GenAI has moved beyond the "Technology Trigger" phase (initiated by the release of ChatGPT) and is firmly situated at the Peak of Inflated Expectations. At this stage, public discourse is dominated by excessive optimism, and the technology is portrayed as a panacea for virtually every business problem - from code generation to strategic planning. The main characteristics of this phase in the AI context include: "AI-washing": The widespread addition of the "AI" label to existing software products in order to artificially boost their market valuation, often without any real change in their underlying architecture.

– **Fear of Missing Out**

Institutional investors and corporate boards make decisions under pressure from the fear of falling behind competitors, leading to irrational capital allocation.

*Corresponding author.
E-mail address:

– **Psychology of market mania**

The bubble is fueled by so-called narrative economics. When a technology is so complex that it remains a "black box" for the average investor, its value is determined not by financial statements but by stories about the future. Here we observe a classic deviation from fundamental analysis, captured by Amara's Law: "We tend to overestimate the effect of a technology in the short run and underestimate it in the long run." At present, the market is pricing in the short-term effects of AI as if they have already transformed the global economy, while ignoring the time required for real structural deployment.

– **Is the "Trough of Disillusionment" approaching**

The next phase in the cycle is the Trough of Disillusionment. It occurs when: Technical limitations become visible: Model hallucinations, lack of true logical reasoning, and copyright issues begin to outweigh the initial "wow effect." Maintenance costs exceed benefits: Companies realize that maintaining specialized LLM infrastructures is more expensive than hiring people to perform the same tasks. Mathematically, the risk of a "burst" can be expressed through the gap between expected profit growth (g) and the actual rate of technology adoption. If the market has priced in g = 30%, while the real business sector adopts the technology at 5%, a correction becomes inevitable.[2]

– **Comparison with historical precedents**

If we examine Railway Mania (1840s) or the dot-com bubble (1990s), we see the same pattern: Infrastructure oversaturation: Far more railways / fiber-optic cables / GPUs are built than the current economy can effectively utilize. Capital rotation: Money shifts from companies that "make things" to those that "sell dreams."

Figure 2 presents a schematic representation of the Gartner Hype Cycle.

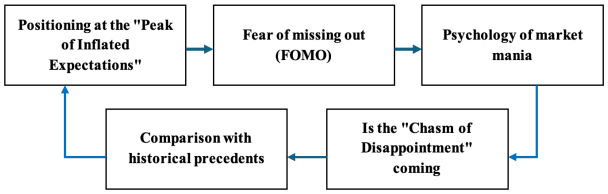


Fig. 2. Media noise cycle

3. FINANCIAL INDICATORS OF SPECULATIVE OVERVALUATION

When share prices rise much faster than the actual net profit generated by the product, speculative overvaluation is present. The formula for the P/E ratio is:

$$P/E = \frac{Market\ Value\ Per\ Share}{Earnings\ Per\ Share} \quad (1)$$

For many AI startups and hardware suppliers, this ratio exceeds historical norms, relying on future (as yet unrealized) revenues from automation and software services. The cost problem: the "hardware wall". While the software capabilities of generative AI appear limitless, its existence is anchored in the harsh physical reality of semiconductors and energy infrastructure. The "hardware wall" is a term describing the point at which the marginal cost of computational power begins to exceed the economic benefit of the resulting model. Table 1 illustrates the relationship between financial valuations and infrastructural constraints.[4]

Table 1. Connection between financial assessments and infrastructure constraints.

Market Indicator	Market Assumption	Physical Reality	Potential Risk
Profit Growth	Exponential	Linear/Limited Valuation	Valuation Adjustment
Service price	Decreasing	Increasing (energy, HBM)	Margin compression
Scalability	Infinite	Hardware-Limited	Hardware Wall
Return	Short-Term	Long-Term/Uncertain	Investment Risk

4. INFRASTRUCTURAL CONSTRAINTS AND THE HARDWARE WALL OF AI

Although public discourse around generative artificial intelligence is dominated by software breakthroughs and abstract algorithmic capabilities, its practical realization is inextricably tied to the material infrastructure of modern computing systems. Every advance in models is a function of real physical resources - semiconductors, energy, cooling, and network connectivity. This dependence places a ceiling on the economic scalability of AI and gives rise to the so-called "hardware wall" - the point at which additional investment in computational power yields diminishing returns and increasing systemic risk.[3]

– ***Capital Expenditures***

We are currently witnessing a historic asymmetry between the costs of building AI capacity and the revenues generated from it. Companies such as Microsoft, Alphabet, and Meta are investing tens of billions of dollars each quarter in the acquisition of specialized chips (primarily NVIDIA H100 and Blackwell). According to analyses by Goldman Sachs (2024), the technology sector would need to generate more than $600 billion in annual revenues merely to cover current hardware and infrastructure costs. By comparison, as of the end of 2024, direct revenues from GenAI services (such as ChatGPT Plus, Copilot, etc.) are estimated to represent only a small fraction of this amount.

– ***The Energy Paradox and Logistical Constraints***

The hardware wall also has a purely physical dimension electricity. The average energy consumption of a single LLM query (e.g., GPT-4) is approximately ten times higher than that of a standard Google search. This leads to: Energy shortages: Many data centers in the United States and Europe can no longer obtain approval for expansion due to insufficient capacity in national power grids. Cooling requirements: The need for massive water-based cooling systems for servers confronts companies with environmental and regulatory barriers.

– ***Obsolescence and Depreciation***

Unlike the infrastructure of the dot-com bubble (optical fiber cables with a lifespan of 20–30 years), graphics processing units (GPUs) have a technological obsolescence cycle of only 2 to 3 years. If the AI services market undergoes a correction, companies may be

left with hundreds of thousands of chips whose market value could collapse almost instantaneously. This transforms "assets" on corporate balance sheets into potential dead capital that cannot be resold at purchase price. [6]

– ***The Law of Diminishing Returns***

Within engineering circles, there is growing debate over whether so-called scaling laws the hypothesis that more data and more compute always led to better models - remain valid. There are indications that training new models now requires ten times more resources to achieve only 1-2% improvements in accuracy. This erosion of efficiency is a classic sign of an approaching technological plateau, which often precedes the bursting of speculative bubbles.[5]

– ***Architectural Breakthrough: High Bandwidth Memory (HBM)***

While the computational power of GPUs continues to grow exponentially, the speed at which data can be transferred from memory to compute cores has failed to keep pace. This creates the so-called "memory wall" - a condition in which the most powerful chips in the world remain idle while waiting for data. Architectural Breakthrough: High Bandwidth Memory (HBM). To overcome this limitation, manufacturers have shifted toward the use of HBM (High Bandwidth Memory). Unlike standard DDR5 memory, HBM employs vertically stacked layers (3D stacking) connected directly to the processor via microscopic interconnects (TSVs). Economic dependence: HBM manufacturing is extremely complex and expensive, with low yield rates. This makes memory the costliest component of AI servers after the GPU itself. Supplier oligopoly: The HBM market is controlled by just three companies - SK Hynix, Samsung, and Micron. Any disruption in their production directly threatens the entire AI ecosystem.

– ***Scaling Parameters vs VRAM***

Large language models (LLMs) require enormous amounts of VRAM simply to be loaded into memory. For example, a model with 175 billion parameters (such as GPT-3) requires hundreds of gigabytes of memory just to store its weights in 16-bit format. The approximate formula for the memory required to store a model (excluding activations and gradients during training) is:

$$M \approx P \times B \tag{2}$$

where:

M is the required memory in bytes;
P is the number of parameters (e.g., 70 billion);
B is the number of bytes per parameter (typically 2 for FP16 or 1 for INT8).

– ***Risk to Investors***

Dependence on HBM introduces an additional layer of fragility into the "AI bubble." If memory prices continue to rise due to shortages, the profit margins of companies offering AI services will erode even faster. This renders AI applications economically inaccessible for small and medium-sized enterprises, limiting the market potential primarily to the largest corporations.[7]

– ***The Interconnect Bottleneck***

Even with fast memory, data transfer between GPU clusters (via NVLink or InfiniBand) presents a significant logistical challenge. The costs of building networking infrastructure capable of supporting such speeds are often underestimated in market analyses, yet they represent a substantial portion of capital expenditures (CAPEX). Figure 3 presents a schematic overview of the hardware constraints facing generative AI. [8]

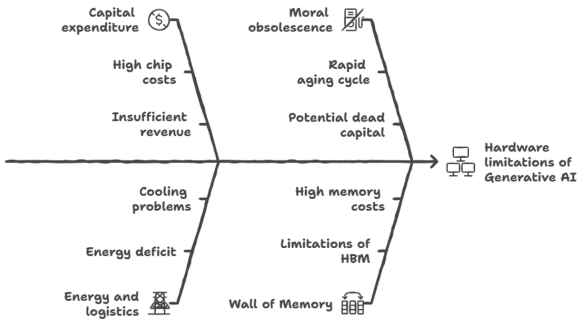


Fig. 3. Hardware limitations of generative AI

5. STRUCTURAL SIMILARITIES AND DIFFERENCES IN CAPITAL DYNAMICS

To assess whether the current euphoria surrounding artificial intelligence represents a speculative bubble or a sustainable technological transition, it must be placed within a historical context. A comparison with the dot-com bubble of the late 1990s reveals similar patterns of capital concentration, inflated expectations, and a weak link between market valuations and real profitability. At the same time, however, there are significant differences stemming from higher technological maturity, the greater role of corporate giants, and the centralization of critical infrastructure. This section analyzes these structural parallels and divergences that shape the nature and potential outcome of the current AI market cycle. Table 2 presents the parallels between the situation in 1999 and that of 2024–2025. Figure 4 illustrates the life cycle of technological bubbles.[9]

Table 2. Comparison of models and situations.

Indicator	Dot-com Bubble	AI Bubble
Key Asset	Fiber Optics / Domains	GPUs (NVIDIA) / Chips
Slogan	"The Internet is changing everything"	"AI is the new electricity"
Venture Capital	Unlimited for .com companies	Massive rounds for LLM developers
Monetization	Low (traffic reliance)	Low (high operating costs)

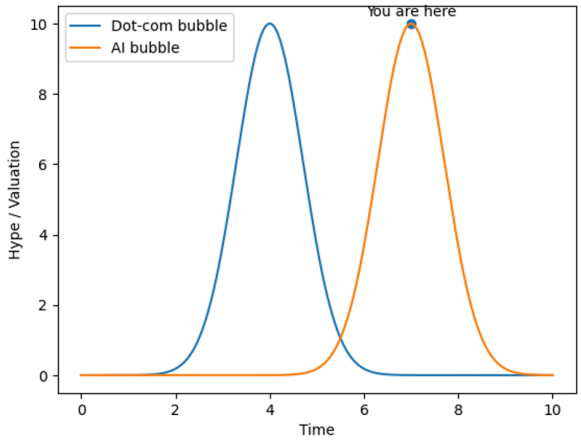


Fig. 4. Life cycle of technology bubbles

6. CONCLUSIONS

The current euphoria surrounding generative AI exhibits all the hallmarks of a classic financial bubble: a disconnect between prices and fundamentals (excessive P/E ratios), investment-driven FOMO (fear of missing out), and massive expenditures without a clear path to profitability for most market participants. An inevitable market

consolidation is expected. Companies that fail to deliver real, functional products that reduce costs for end users will collapse as soon as cheap investor capital dries up. The future of the sector belongs not to media hype, but to energy-efficient and narrowly specialized AI applications. The historical precedent of the dot-com bubble shows that such technological manias do not destroy innovation, they refine it. After the bubble bursts, only a small number of resilient players remain, combining technological advantage with disciplined business models. In the same way, the current correction in the AI sector will not mark the end of artificial intelligence, but rather the end of the illusion that every AI solution is economically justified. True value will be created where AI is an indispensable tool, not a universal excuse for inflated valuations.

REFERENCES

[1] **Docheva, A., Denev, D.** *Using AI to Predict Radiation Levels in Real Time*, Conference Proceedings of the International Scientific Conference "Radiation Safety in the Modern World", vol. 1, National Military University "Vasil Levski", Veliko Tarnovo, 2025, ISSN 2738-7607, pp. 213–223.

[2] **Denev, D., Nikolova, S.** *A Novel Algorithm for Industrial Network Reliability Estimation and AI-Based System Optimization*, Proceedings of XXXIII National Conference with International Participation TELECOM, 2025, ISBN 978-1-6654-6564-9, ISSN 2837-5246, DOI: 10.1109/TELECOM66943.2025.11304081, pp. 1–4.

[3] **Dimitrova, E., Denev, D.** *The Development of Artificial Intelligence and its Applications in the Modern World*, Conference with International Participation MATTEH 2024, Conference Proceedings, vol. 2 Communication and Computer Technologies, ISSN 1314-3921, pp. 153–160.

[4] **NVIDIA Corporation, Microsoft Corporation.** *Financial Reports of NVIDIA and Microsoft*, 2023–2024.

[5] **Hatib, I., Denev, D.** *The Modern Role of Artificial Intelligence in Transforming Education through Personalized Learning and Adaptive Platforms*, XVIII International Scientific Conference on Innovations in Technology and Education, Belovo, 2025, ISBN 978-5-00137-528-9, ISBN 978-5-00137-314-8, pp. 183–188.

[6] **Hatib, I., Denev, D.** *Integrating Artificial Intelligence for Autonomous Traffic Management in Industrial Networks and Automation Systems*, Proceedings of International Scientific Conference Defense Technology Forum, National Military University "Vasil Levski", Artillery, Aircraft Defense and CIS Faculty, Shumen, 2025, ISSN 2815-4282, pp. 813–820.

[7] **Nikolova, S., Denev, D., Tsankov, Ts.** *Comparative Analysis of the Performance of Deep Learning-Based AI for Video Cameras Working with Different Databases*, XIX International Conference on Communication Electromagnetics and Medical Application CEMA, Athens, Greece, 2025, ISSN 1314-2100, pp. 59–65.

[8] **Shiller, R. J.** *Irrational Exuberance*, Princeton University Press, 2015.

[9] **Sotirov, P., Denev, D., Docheva, A.** *Automation of Low-Earth Orbit Communication Networks through Artificial Intelligence*, Proceedings of Scientific Conference "Near Space – a Common Goal", National Military University "Vasil Levski", Veliko Tarnovo, vol. 2, 2025, ISSN 2815-3510 (Print), pp. 159–166.