An Efficiency Assessment of the Artificial Intelligence Market: Exploring the Limits

Yaroslav Kuzminov
Academic Supervisor, kouzminov@hse.ru

Ekaterina Kruchinskaia
Senior Lecturer, Department of Higher Mathematics, ekruchinskaya@hse.ru

National Research University Higher School of Economics (HSE University), 20, Myasnitkaya str., Moscow 101000, Russian Federation

Abstract

The development of Artificial Intelligence (AI) is transforming corporate strategies and enhancing business operational efficiency. This study analyzes the relative efficiency of the Generative AI (GenAI) market, considering the market size of chips, servers, and data center infrastructure (hardware), and comparing them with the market size of AI solutions (software). The study hypothesizes that the AI market is catching up despite its rapid development compared to the component market: the parity between component sales and revenue from AI products has not yet been achieved. The capital expenditures of technology giants on the creation of AI infrastructure have significantly increased and it may require decades to achieve a balance.

To assess the efficiency of the AI market, the Data Envelopment Analysis (DEA) methodology is applied, considering "inputs" (the market size of hardware components) and "outputs" (the market size of AI software solutions). The results of the DEA analysis of the GenAI market dynamics from 2016 to 2024 reveal non-linear development, starting in 2021, with a trend reversal and a decrease in efficiency indicators, which confirms the hypothesis of the catching-up nature of AI software technologies compared to the hardware component market. The fluctuations in efficiency began three years after the deployment of the first large language models, indicating their significance for the demand for hardware, but not yet demonstrating sufficient returns in the form of comparable growth of the AI solutions market. The novelty of this study lies in the application of DEA analysis for a comprehensive assessment of the AI market considering what facilitates the identification of an imbalance between the 'soft' (technological solutions) and 'hard' (components) markets, and the potential for the more efficient exploration and use of generative models.

Keywords: artificial intelligence; generative AI; DEA analysis; AI efficiency; AI investments; semiconductors market; AI market; AI Investments; large language models; AI economic impact

Citation: Kuzminov Y., Kruchinskaia E. (2025) Efficiency Assessment of the Artificial Intelligence Market: Exploring the Limits. *Foresight and STI Governance*, 19(4), pp. 6–16. https://doi.org/10.17323/fstig.2025.29079

Introduction

The adoption of artificial intelligence (AI) across a broad range of domains, from everyday life to key business processes, has marked the biggest shift in the structure of the global economy in recent decades. Evolving from a highly specialised field into a ubiquitous corporate strategy tool, AI significantly impacts operational efficiency and shapes the macroeconomic development trajectory (WEF, 2025). Market players expect these breakthrough solutions to deliver industrial returns faster, and more efficiently than humans. It's this hope for autonomous profit generation that largely underlies the projected growth rates for the segment. A key factor fuelling investor expectations is the potential to automate a wide range of processes, and reduce dependence on expensive human labour, which large language models (LLMs) are supposed to have. Beyond the corporate and industrial sectors, this type of model (also known as generative AI, or GenAI), has also attracted a lot of attention in academia (Bang et al., 2023; Zhao et al., 2023). Companies have begun to actively adopt ready-made solutions, and develop own LLM-based applications to deal with specific problems (Fu et al., 2024; Laskar et al., 2023; Mayer et al., 2025).

In our previous paper (Kuzminov, Kruchinskaya, 2024) the AI market was analysed at the micro-level by assessing various LLMs' actual performance in the economy, law, and education. The results revealed that in 2024 even the most advanced models were unable to accomplish relevant tasks autonomously (error rate in excess of 50%). Subsequent evaluations show a certain improvement due to the release of new models (such as GPT-5, DeepSeek, Grok), but next-generation LLMs still have unacceptably high error rates (Cai et al., 2025; Georgiou, 2025). The inability to operate autonomously continues to hinder the application of AI agents as human replacements in most areas, except initial customer contact and basic information gathering.

At the same time a certain contradiction arises: it's the development of agent-based AI that is seen as the way to apply GenAI autonomously and highly efficiently. It is envisioned that such systems will process user requests on their own, acting as independent intelligent agents, leading to increased organisational productivity and reduced transaction and labour costs. This paradigm implies AI's ability to independently set goals, plan actions, and solve problems, while interacting with the external environment without direct human intervention, potentially automating complex processes in dynamically changing conditions.

However, developing strategies for integrating AI into companies' business processes wouldn't be enough for professional AI implementation. Significant investments in three areas are required: components, computation (initial and subsequent model training, fine-tuning, and testing), and hiring professionals skilled in relevant core fields (e.g., medicine), and in AI as such. The reason is that even benchmarking reveals that complete, unsupervised delegation is impossible. The most common errors and complaints about such systems are the latter's hallucinations, selective use of available information, and the inability to maintain a stable core of professional knowledge, all of which are due to the specific nature of LLMs. Investments are mainly channelled in developing models whose architecture is designed to imitate communication, rather than professional activity. Unsuccessful (because of such solutions' limitations)[1] attempts to capitalise on them lead not to abandoning such strategies, but rescheduling the positive expectations to the next year, then the year after that, etc. It's the combined effect of inflated expectations from AI shared by a wide range of market players, and the constant supply of new, ever-more-highly performing solutions.

Underestimating this factor in forecasting, and failing to compare the costs with expected revenues creates both practical (risk of an investment bubble), and scientific problems (insufficient understanding and inadequate methodologies). The purpose of this study is to analyse the relative GenAI market performance by comparing chip, server, and data centre hardware sales with the growth of the AI solutions (software) sector. Building on the critical positions of other scholars (Hajikhani, Cole, 2024), we evaluate the practical potential of such solutions and their financial returns, taking into account the amount of investments in the three areas mentioned above.

To assess the prospects, and the returns on technological investments in agent-based AI in a specific industry or business, in addition to benchmarks one must also consider implementation costs. The AI market, and expectations of it are analysed by comparing inputs (hardware costs) and outputs (software revenues). The underlying hypothesis is that the software market has a catch-up nature, and does not yet cover companies' component costs needed to support system development and integration into business processes. The following sections present a literature review describing the current and projected trends in the global AI market, and the authors' assessment of its relative performance.

[1] According to a recent MIT study, 95% of organisations get zero return on their investments in AI technologies (Challapally et al., 2025).

Literature review: Trends in, and prospects of the global AI market

The global AI market is showing a steady and rapid growth. According to Statista Market Insights, by the end of 2025 it will reach 244.22 billion USD, and by 2031, due to the compound annual growth rate (CAGR) of 26.60%, will exceed 1 trillion USD (Figure 1).

Segment data analysis (Figure 2) reveals significant growth across all sub-sectors. However, the machine learning (ML) and natural language processing (NLP) markets have a significantly higher base and growth rate than autonomous and sensor systems, computer vision, and AI robotics.

This disparity indicates the faster proliferation of algorithmic advances focused on data analysis and information processing, due to the opportunities to use them via familiar interfaces (phone and computer screens). At the same time, technologies which require tight integration with the physical world and complex hardware solutions are growing more slowly, due to higher entry barriers and problems with implementing them in real-world processes. In fact, this reflects the current trend: applications accessible to millions of users via standard interfaces are developing faster than solutions which require physical interaction with the real world.

Alternative forecasts by Grand View Research suggest an even higher AI market growth rate, at 36.6% (CAGR) in 2023-2030. According to this estimate, by 2030 the market will reach approximately 1.7 trillion USD. The discrepancy in projected growth rates may be due to the differences in analysis methodology, segment coverage, and forecasting models.

This trend is observed in almost all developed and emerging economies. Table 1 presents the projected GenAI market growth rate for 2023-2031 in the top 10 countries by this indicator.

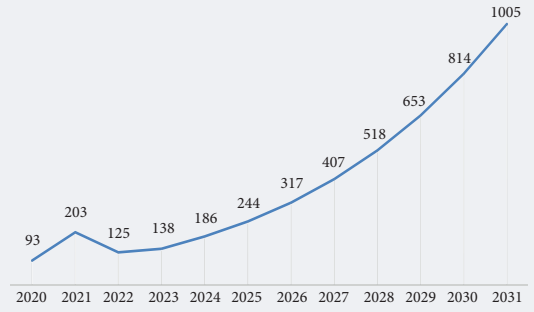
Throughout the forecast period, the US will remain the leader in the GenAI market. In 2025, the US market (73.98 billion USD) will exceed the Japanese, German, Indian, UK, French, Canadian, Russian, and Italian ones combined. This reflects the US ecosystem's maturity, and extensive investment in technology development. By 2031, the US market will reach 309.7 billion USD, leaving the competitors far behind.

China confidently comes second, significantly behind the US but ahead of all other countries. In 2025 the Chinese market will be worth 46.53 billion USD, or almost 1.6 times less than the US one. However, its growth rate is particularly noteworthy: by 2031 the market will be more than 7.5 times bigger than in 2023. It indicates the enormous potential of the Chinese market, and the possibility of narrowing the gap with the US.

In 2025 Japan, Germany, India, the UK and France make the middle group: their markets range from 6.64 billion (France) to 10.15 billion USD (Japan). Canada, Russia and Italy are at the end of the list, with markets ranging from 4.8 billion (Italy) to 5.20 billion USD (Canada).

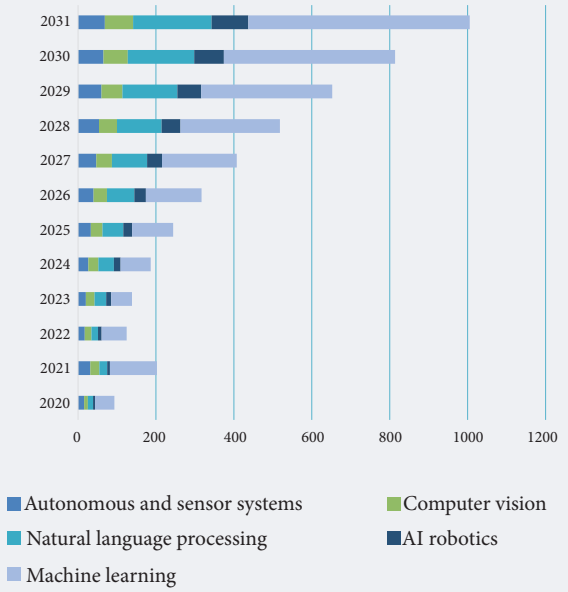
All countries in Table 1 show positive GenAI market growth, which stresses the global proliferation of the technology. Investments are also growing: in 2024, the US attracted 109.1 billion USD in pri-

Figure 1. **Overall AI market growth until 2031** (*billion USD*)



Source: Statista Market Insights, https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide?currency=USD, accessed 24.09.2024.

Figure 2. **AI market growth by industry, until 2031** (***billion USD***)



Source: Statista Market Insights, https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide?currency=USD, accessed 24.09.2024

Table 1. **Top 10 GenAI markets (*Billion USD*)**

Country	2023	2024	2025	2026	2027	2028	2029	2030	2031
US	39.74	55.45	73.98	96.76	124.91	159.26	200.80	250.58	309.70
China	25.61	35.15	46.53	60.70	78.23	99.72	125.78	157.04	194.19
Japan	5.96	7.85	10.15	13.09	16.77	21.29	26.78	33.37	41.19
Germany	5.84	7.70	9.96	12.85	16.45	20.89	26.29	32.76	40.46
India	4.58	6.05	7.84	10.13	12.98	16.48	20.75	25.86	31.94
UK	4.53	5.94	7.68	9.90	12.68	16.10	20.26	25.26	31.21
France	3.93	5.15	6.64	8.55	10.94	13.89	17.47	21.77	26.88
Canada	3.09	4.04	5.20	6.70	8.57	10.87	13.68	17.05	21.06
Russia	2.98	3.91	5.04	6.49	8.31	10.55	13.27	16.54	20.42
Italy	2.83	3.72	4.80	6.18	7.92	10.05	12.64	15.74	19.44

Source: Statista Market Insights. https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide?currency=USD, accessed 24.09.2024.

vate investment in AI, significantly ahead of other countries. To compare, US investments exceeded Chinese ones (9.3 billion USD) almost 12 times, and the UK ones (4.5 billion USD) 24 times. The GenAI sector is showing particularly rapid growth, having globally attracted 33.9 billion USD (18.7% more than in 2023) (Stanford University, 2025).

Major investments like these must yield economic returns, or the risk of the "philosopher's stone" scenario described earlier (Kuzminov, Kruchinskaya, 2024) would arise. Effective monetisation of these investments is critical to maintaining the sector's growth, and preventing speculative bubbles (Acemoglu, 2025).

In 2024 investments in agent-based AI reached 1.1 billion USD, while job openings in the industry in 2023-2024 grew by 985% (Yee et al., 2025). This trend reflects the transition from simple algorithms to complex autonomous systems which require continuous data analysis and processing for real-time decision-making.[2] Agent-based AI is applied in finance, sales, education, and medicine (Raza et al., 2025).

In the *financial sector*, generative models (which serve as the basis of agent-based AI) are increasingly applied to automate various operations. Studies of performance, growth history, and technical aspects of applying language models in finance have focused on model training data, fine-tuning methods, and specialised datasets (Lee et al., 2024). Other studies (Li et al., 2023; Zhao et al., 2024) analysed current approaches to implementing AI tools in finance, including adjusting models to specific tasks; researchers note the continued dissemination of LLMs. This trend confirms the focus on automation and "smartification" of financial procedures. Decision support systems for industry professionals are proposed, emphasising the need to select models best-suited for specific tasks.

In *e-commerce*, LLMs provide innovative solutions for improving customer experience and optimising business processes (Ren et al., 2024). A study on

Table 2. **GenAI and semiconductor market revenues (*billion USD*)**

Product	2023	2024	2025	2026	2027	2028	2029	2030
GenAI	20.47	37.87	66.89	107.91	160.83	223.94	294.23	368.12
All semiconductor products	1218.35	1213.92	1381.19	1535.82	1675.69	1804.03	1969.52	2166.49
Out of that:								
AI chips	34.44	48.58	66.28	90.89	122.27	160.66	205.37	254.61
Discrete semiconductors	529.03	504.23	542.96	580.84	616.72	651.73	685.45	717.49
Optoelectronics	231.42	212.88	237.30	261.90	284.97	306.46	327.33	347.79
Microchips	406.32	403.68	487.26	552.38	600.09	632.15	697.40	791.89
Sensors and actuators	51.58	44.54	47.39	49.82	51.65	53.02	53.96	54.69

Sources:
1. Statista Market Insights, https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide?currency=USD, accessed 24.09.2024.
2. Statista Market Insights, https://www.statista.com/outlook/tmo/semiconductors/worldwide?currency=USD, accessed 24.09.2024.

[2] https://www.statista.com/study/38609/artificial-intelligence-ai-statista-dossier/, accessed 17.09.2025.

LLM application in e-commerce recommendation systems (Hu et al., 2024) offers a comprehensive analysis of recent advances and promising development directions. The authors emphasise the need for a deep understanding of LLM-based recommendation systems in the context of their rapid evolution.

In *medicine*, LLMs are applied to process text data, increase healthcare awareness, support clinical practices, and improve doctor-patient communication (Huang et al., 2024). In *education*, language models are widely used to optimise student-teacher interaction, improve teacher feedback, and empower students. A study on the application of various LLM technologies and tools in education (Wang et al., 2024) covers both students and teachers. Another study (Xu et al., 2024) analyses LLMs' contribution to improving teaching methodology and models, and the integration of language models into the educational process.

Our hypothesis about the catch-up nature of AI market growth is partially confirmed by the slowing growth rate of the GenAI segment in the medium term. According to Statista Market Insights, after rapid growth in 2021-2024 (39.70%, 56.58%, 69.97%, and 85.00% year-on-year, respectively), a decline is predicted: 76.64% in 2025, 61.31% in 2026, and then to 20.09% by 2031.[3] Such dynamics may indicate saturation of certain AI segments, increased competition, and growing ROI pressure.

The data in Table 2 allows to expand the analysis of AI economic feasibility focusing on the ratio of hardware costs (AI chips) to technology companies' potential profits in the GenAI market.

An analysis of the AI market in 2020-2022 reveals a mixed picture, due to the lack of generally accepted methodologies for delineating its segments. In particular, there are problems with defining the boundaries between the AI chip segment, the broader semiconductor segment, and the GenAI market. Still, the available data allows to identify certain trends.

In 2020-2021 AI chip revenues (18.01 billion and 24.87 billion USD, respectively) exceeded those from GenAI (11 billion and 18 billion USD, respectively). However, this data must be interpreted with caution: at early stages of AI development tech companies heavily invested in specialised hardware, among other reasons in anticipation of the expected growth in demand for AI solutions. Thus capital expenditures on the infrastructure required to develop and train generative models outpaced the actual monetisation of technological solutions. This fact highlights GenAI's dependence on the availability and performance of relevant hardware.

Since 2022 the gap between GenAI and AI chip revenues narrowed (37 billion and 34.78 billion USD, respectively), which reflects the growing interest in generative models but does not eliminate GenAI's fundamental dependence on hardware. However, in the long-term this ratio may change. According to current estimates, by 2030 the AI chip market will reach 332.77 billion USD, or approximately 34% of GenAI's projected value (981 billion USD). This forecast is conditional due to the high uncertainty associated with rapid technological advancement and market growth. However, even in this scenario a large share of GenAI revenue will be allocated to cover hardware costs, which may limit the profit margins of companies specialising in the development and implementation of GenAI solutions. Furthermore, the monetisation of GenAI is currently primarily based on consumer and entertainment applications, while industrial implementation (which produces direct economic benefits) is only just beginning.

Analysis of the entire semiconductor product market provides a broader perspective; by 2030, its revenues are projected at 1,285.13 billion USD, or 1.3 times higher than those generated by GenAI. This forecast highlights the fundamental role the semiconductor industry (and hardware in general) play in the AI development. In the medium term, the industry will maintain a dominant market share compared to GenAI. This may influence AI companies' investment strategies and business models, encouraging integration with hardware manufacturers and development of specialised solutions for specific machine learning tasks.

At the same time, it's important to consider the growing computing needs associated with the development of AI. While there are obvious benefits of increased performance and expanded functionality, research and company reports don't always fully reflect the exponential growth in demand for computing power. To train complex AI models, businesses require increasingly powerful systems. The smaller and more efficient chips become, the more expensive they become to manufacture, which affects development companies' survival and market concentration. The need for constant performance increases creates a financial burden for AI market players, and may hinder broad adoption of relevant solutions.

The second cost item of AI implementation (along with hardware) is computing. The cost of training AI models has significantly increased in 2024.

[3] https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide?currency=USD, accessed 18.09.2025.

Training LLMs such as Gemini costs between 30-191 million USD. The technical costs of creating ChatGPT-4 are in the range of 41-78 million USD,[4] which is supported by OpenAI CEO Sam Altman's statement that the development cost have exceeded 100 million USD. To compare, previous-generation models' training costs were much lower: in 2020 ChatGPT-3 cost 2-4 million USD, and in 2022 PaLM required 3-12 million USD (for computing resources only).

According to McKinsey (Nossfinger et al., 2025), by 2030 global demand for data centre capacity will nearly triple, reaching 220 GW (vs 82 GW in 2025). 70% of this growth is caused by the resource intensity of AI systems, which require 5.2 trillion USD in investment. Despite potential increases in AI performance (e.g. reduced DeepSeek training costs), model development will keep this demand at a high level. The forecast (see Fig. 3) shows a dominant growth of AI usage (up to 156 GW in 2030) compared to other needs (up to 64 GW). The AI market's demand for data centre capacity confirms the need for significant investments, since the benefits of algorithm optimisation will be offset by the scaling of technology adoption. This trend is reflected in the growing demand for specialised chips, and the expansion of data centre infrastructure.

Finally, the third significant factor driving investment activity is skilled labour. In the US, the country with the most highly developed AI market, in 2023 employment in the technology sector reached 9.6 million - an increase of 1.2%, or over 117,000 new positions, compared to 2022 (CompTIA, 2024). The median salary in the technology sector in 2022 was 104,556 USD, more than double the national average. In some states it was 89,544 USD, or 1.3 times higher than the state average. The information and communications services and custom software development sector has remained the main driver of employment since 2018 (50% of the growth), followed by the software sector (29%). For 2024, the most rapid growth in demand was projected for data scientists (5.5%), cybersecurity (5.1%), and software engineers (4.8%). Steady employment growth and high salaries in technology areas, particularly AI-related ones, reflect a growing shortage of talent, and pose a critical barrier to scaling AI solutions.

The problems described above are already addressed in the literature. The AutoDroid (Wen et al., 2024) and WebAgent (Gur et al., 2024) frameworks designed to automate operations in mobile and web environments demonstrate the difficulties in achieving optimal response speed and accuracy. The issues include limited computing resources, the need to process large volumes of data in real time, and ensure high accuracy of results. Both these frameworks reflect the challenges of efficiently distributing the computational load between a centralised server and periphery devices, and the need to optimise LLMs for resource-constrained devices.

Some authors also note a clear limit to AI capabilities (Chang et al., 2024; Biderman et al., 2024). Deploying LLMs in real-world industrial scenarios requires testing them not only across a wide range of benchmarks, but also across real business operations throughout the entire workflow cycle. This necessitates hiring highly-paid experts with relevant skills. And results obtained in laboratory conditions often do not reflect LLMs' actual performance in complex production scenarios. Research on such systems' effectiveness in the field remains insufficient. Empirical AI testing in real-world processes requires applied research, and the development of metrics to assess the returns on investment in AI.

According to Morgan Stanley forecasts, in 2025-2028 companies will have to invest in chips, servers, and data centre infrastructure at least 2.9 trillion USD, and by 2030 these costs will exceed 3 trillion USD. Meanwhile Sequoia[5] (2023) estimates that to recoup the investments, revenues from GenAI solutions must exceed 600 billion USD,[6] compared to the current market of 196.6 billion USD.[7] The 400 billion dollar gap is due to the fact that in 2025 tech giants' capital expenditures reached 40% of the revenue, when just ten years ago they did not exceed 5-15%. Alphabet, Microsoft, and Amazon intend to spend in 2025 nearly 400 billion USD, primarily on AI infrastructure. Investment payback will take decades. Even Nvidia, the most successful of them all, brings in just 14.51 billion USD a year. The UK AI start-up Stability AI, the creator of Stable Diffusion, is considering selling the business amid mounting investor pressure given the modest 1 billion USD annual revenue.

The GenAI market's readiness for such investments, and estimates of their returns are discussed in critical studies of the AI market. E.g. an analysis of the macroeconomic impact of technology development (Acemoglu, 2025) shows that despite automation and improved performance, the overall

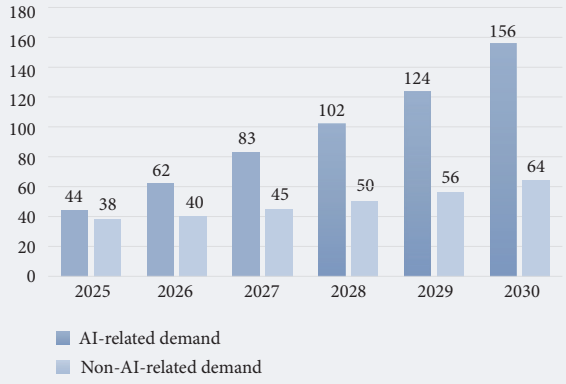
[4] https://www.statista.com/chart/33114/estimated-cost-of-training-selected-ai-models/, accessed 22.08.2025.

[5] https://habr.com/ru/companies/onlinepatent/articles/843730/, accessed 19.09.2025.

[6] https://www.tomshardware.com/tech-industry/artificial-intelligence/ai-industry-needs-to-earn-dollar600-billion-per-year-to-pay-for-massive-hardware-spend-fears-of-an-ai-bubble-intensify-in-wake-of-sequoia-report, accessed 07.09.2025.

[7] https://www.forbes.com/sites/petercohan/2024/07/09/ai-industry-may-never-earn-back-its-150-billion-nvidia-chip-spend/, accessed 19.09.2025.

Figure 3. **Forecasted global demand for data centre capacity, by type (AI vs Non-AI), under the current growth rate scenario (GW)**



Source: (Nossfinger et al., 2025).

increase in total factor productivity over a decade could amount only to 0.71%. If AI is applied to deal with more simple tasks, this estimate could drop to 0.55%. The impact on GDP will be moderate: 0.9-1.8% over 10 years, depending on the scale of investment. Furthermore, there is a risk of increased inequality: the benefits of AI may be distributed across economic sectors unevenly.

Accurate analysis requires understanding the technological limitations associated with dependence on training data, and the difficulty of adapting to new domains. Training and fine-tuning LLMs is highly resource-intensive, which limits smaller organisations' and researchers' access to these technologies. Dependence on the quality and diversity of training data poses challenges in understanding specific contexts, and generating relevant answers on insufficiently mastered topics (Saleh et al., 2025). Adapting models to specialised tasks requires significant resources, and does not guarantee optimal results (Deng et al., 2023). These barriers highlight the importance of data-sparse learning methods, and developing domain-specific models.

Finally, the ethical implications of applying AI must be considered. Effective collaboration of humans and models remains a key challenge, since AI is expected to augment human potential without making critical decisions on its own (Bender et al., 2021; Trott et al., 2023). Furthermore, maintaining LLMs' relevance and accuracy requires regular updates. This implies a need for a robust infrastructure to manage model development, seamless integration of new data sources, and ongoing fact-checking, which involves significant expenditures on human resources.

The observed dichotomy between considerable AI costs ("inputs") and insufficient returns from its application ("outputs") reveals a gap in understanding the current state of the market and its prospects. This discrepancy raises a number of research questions that require detailed study. A statistical tool for estimating the AI market bubble and tracking its growth is needed, along with models to identify input/output imbalances. A solution is offered by the Data Envelopment Analysis (DEA) model, which is applied in the next section of our study.

Methodology

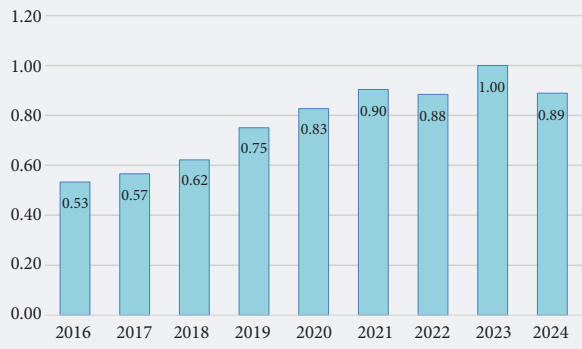
We assessed GenAI market's relative performance using DEA analysis: a nonparametric method for measuring the efficiency of decision-making units (DMU) based on multiple input and output parameters. For the dynamic and unpredictable AI market this approach is particularly relevant, as it allows to identify trends and patterns unattainable using alternative methods. In our study, DMUs reflect the market state over various years. Each DMU is evaluated relative to the efficiency threshold set by the best performers in the sample. Aleskerov and Petrushchenko (2016), Liu et al. (2021), Gładysz et al. (2024), and Shi et al. (2024) note that DEA is a powerful tool for assessing performance in a situation of data constraints. The method allows to identify inefficient DMUs and potential improvements.

To analyse the GenAI market growth in 2016-2024, two constant returns-to-scale models were applied (Charnes-Cooper-Rhodes, CCR). In both of them, DMUs represent years reflecting market state in the corresponding periods. Each model contains one input and one output, and uses an output-oriented approach. I.e. the model seeks to increase output at a fixed level of input. Specifically, it determines how much generative model owners' revenues (outputs) can be increased at a given level of component manufacturers' revenue (inputs).

Three key AI investment areas were identified earlier: chips, computing power, and skilled labour. For the purposes of this study, investments in chips were considered as the most capital-intensive component, potentially leading to the emergence of an AI market bubble. A comprehensive assessment of all three factors will be continued in future publications.

Model 1 uses absolute revenue values in US dollars. The input parameter is the total revenue of AI components (hardware) manufacturers, reflecting the resources invested in the hardware basis of AI systems. The output parameter is the total revenue of AI models (software) owners, which reflects the monetisation (commercial success) of these models' application (Table 3). Model 1 assesses how effectively investments in hardware production are

Figure 4. **Relative annual performance of hardware investments in the GenAI market in 2016–2024**



Source:calculated by the authors.

transformed into profits for AI technology owners, in absolute terms.

To test for robustness in *Model 2*, the above indicators were adjusted for global GDP by analysing the total "inputs" and "outputs" revenues as GDP proportions (Table 4). Relative indicators offset the effect of macroeconomic factors such as inflation and exchange rate fluctuations, ensuring a valid comparison of GenAI market performance across different economic cycles.

Revenue was chosen as the primary performance indicator for several reasons. First, it's the most readily available and transparent financial indicator, published by companies in open sources, allowing for comparative analysis across different periods and businesses. Second, revenue reflects products' and services' true market value. While investments are no less important, they are not always directly linked to current profitability.

The selection of companies in the "inputs" and "outputs" groups was based on their value generating role in the GenAI market, and the availability of financial data. "Inputs" companies (Nvidia, IBM, TSMC, ASML, Broadcom, AMD, Intel, Micron, and Qualcomm) make hardware needed to develop and deploy AI models. Nvidia, AMD, and Intel lead the production of graphics (GPUs) and central processing units (CPUs), which serve as the computing basis for AI training. TSMC and ASML make the technological foundation of semiconductor manufacturing. Broadcom, Micron, and Qualcomm specialise in networking equipment, memory, and other components. Together, they form the resource basis of the AI market.

"Outputs" firms (Sony, OpenAI, Google DeepMind, Microsoft, Alibaba, Baidu, Amazon, Alphabet, Apple) develop and own AI models and platforms which provide access to them, and monetise AI. OpenAI develops the GPT family, Google DeepMind creates AlphaGo and AI systems, and Microsoft integrates AI into its products and services. Alibaba, Baidu, Amazon, and Apple apply AI to optimise the provision of services, and develop their own AI solutions. Sony creates AI technologies for use in entertainment and robotics. These software developers' revenues reflect the value of technologies created in the AI market.

Using generalised AI market metrics as input and output indicators would be inappropriate, since it would lead to duplication and distortion of results. Generalised indicators include revenues from both hardware manufacturers and software developers, which wouldn't allow to assess the efficiency of resource transfer from one sector to another (from semiconductors to technology). The study aims to independently analyse each sector, and assess their relationship.

For each DMU (year), the following linear programming (LP) problem was solved:

- n - total duration of the series (years)
- i - DMU index (year), i = 1, ..., n (where N = 9, 2016, ..., 2024)
- x_i - input value for DMU_i
- y_i - output value for DMU_i
- u - output weight
- v - input weight
- θ - maximised DMU efficiency

Objective function (maximisation):

$$max\ \theta = (u * yi) / (v * xi)$$

The following constraints were imposed to ensure that the performance of any DMU (year) does not exceed 1, and the input and output weights remain non-negative:

$$(u * yj) / (v * xj) \leq 1, for\ all\ \ j = 1,...,n;\ u,v \geq 0$$

In both models, the θ value obtained by solving the LP problem shows the relative performance of the AI market,[8] i.e. the conversion of hardware into software in a given year. A θ value close to 1 indicates high efficiency: in the relevant year the generative model market was relatively successful in converting hardware investments into commercial gains from applying AI models. A θ value less than 1 indicates that the market generated suboptimal

[8] Serves as the basis of the "AI performance" concept in this paper.

Table 3. **Model 1 parameters**

Year	Total revenue of input companies (*billion USD*)	Total revenue of output companies (*billion USD*)
2024	437.1	2031.3
2023	353.3	1846.8
2022	378.2	1748.3
2021	337.4	1594.1
2020	286.2	1237.6
2019	267.4	1049.1
2018	287.1	933.4
2017	257.4	762.0
2016	234.6	653.9

Source: calculated by the authors.

revenue from AI models at the same level of hardware investment.

Empirical results

A comparison of Model 1 (absolute values) and Model 2 (GDP-adjusted) data reveals a high consistency of the performance estimates. The analysis results are not significantly affected by macroeconomic fluctuations, since they reflect the internal GenAI market growth processes. Adjusting for global GDP did not lead to a significant change in the relative estimates, strengthening the reliability of the conclusions.

The dynamics of the GenAI market performance in 2016-2024 (Figure 4) reflects a nonlinear growth pattern, with a trend reversal in 2021. Between 2016 and 2021, indicators grew, due to the development and commercialisation of generative models; after 2021 they declined, reaching a local minimum in 2022. In 2023 an intense, but short-lived surge in market activity occurred, which did not continue. In 2024 the downward performance trend returned, with the value falling to the 2022 level.

The dynamics described above confirm the hypothesis of a catch-up nature of AI technology development in relation to the hardware market. Performance fluctuations begin three years after the deployment of the first LLMs, indicating this technology's importance for hardware demand. Increased demand for hardware is primarily caused by the development of LLMs. However, even counting giant companies whose revenues are not entirely dependent on AI, the performance fluctuations indicate that returns on hardware investments have limits, while the level of generative models' adoption and application remain insufficient. Therefore the software market has not yet fully realised its potential compared to the hardware market.

Conclusions

The results of a DEA analysis of the AI market performance in 2016-2024 suggest GenAI's prospects to increase technological capacity and revenue should be assessed cautiously. The revealed performance trend, with a surge in 2023 and a decline in 2024, indicates that the GenAI market's contribution to the growth of related industries, particularly the components market, outpaces the market's ability to generate comparable added value. This conclusion is consistent with critical assessments of the AI's macroeconomic impact (Acemoglu, 2025), whose authors express doubts about the possibility of significant increases in total factor productivity and impact on GDP, even with appreciable investments.

Comparing these findings with Statista Market Insights data on the slowing growth of the GenAI industry supports the argument that the market may be reaching saturation, and a cost-benefit assessment is necessary, taking into account AI's limitations (Chang et al., 2024; Biderman et al., 2024). Full implementation of GenAI's potential, and sustainable profitability growth require not only technological development, but also more efficient application of these technologies, and careful planning of resources allocated to improve the hardware basis.

Table 4. **Model 2 parameters**

Year	Total revenue of input companies (*billion USD*)	Total revenue of output companies (*billion USD*)	Global GDP (*billion USD*)	Share in Global GDP, inputs (%)	Share in Global GDP, outputs (%)
2024	437.1	2031.3	197428.1	0.22	1.03
2023	353.3	1846.8	185886.7	0.19	0.99
2022	378.2	1748.3	174155	0.22	1.00
2021	337.4	1594.2	156376.5	0.22	1.02
2020	286.2	1237.6	139883.6	0.20	0.88
2019	267.4	1049.1	140300.1	0.19	0.75
2018	287.1	933.4	132441.6	0.22	0.70
2017	257.5	762.0	124428	0.21	0.61
2016	234.6	653.9	117844	0.20	0.55

Source: calculated by the authors.

References

Acemoglu D. (2025) The simple macroeconomics of AI. *Economic Policy*, 40(121), 13–58. https://doi.org/10.1093/epolic/eiae042

Aleskerov F., Petrushchenko V. (2016) DEA by sequential exclusion of alternatives in heterogeneous samples. *International Journal of Information Technology & Decision Making,* 15(1), 5–22. https://doi.org/10.1142/S021962201550042X

Bang Y., Cahyawijaya S., Lee N., Dai W., Su D., Wilie B., Lovenia H., Ji Z., Yu T., Chung W., Do Q.V., Xu Y., Fung P. (2023) *A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity* (arXiv preprint 2302.04023). https://doi.org/10.18653/v1/2023.ijcnlp-main.45

Bender E.M., Gebru T., McMillan-Major A. Shmitchell S. (2021) "On the dangers of stochastic parrots: can language models be too big? In: *FAccT 2021 – Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York: Association for Computing Machinery, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Biderman S., Schoelkopf H., Sutawika L., Gao L., Tow J., Abbasi B., Aji A.F., Ammanamanchi P.S., Black S., Clive J., DiPofi A., Etxaniz J., Fattori B., Forde J.Z., Foster Ch., Jaiswal M., Lee W.Y., Li H., Lovering Ch., Muennighoff N., Pavlick E., Phang J., Skowron A., Tan S., Tang X., Wang K.A., Winata G.I., Yvon F., Zou A. (2024) *Lessons from the trenches on reproducible evaluation of language models* (arXiv Preprint 2405.14782). https://doi.org/10.48550/arXiv.2405.14782

Cai Z., Wang Y., Sun Q., Wang R., Gu C., Yin W., Lin Z., Yang Z., Wei C., Shi X., Deng K., Han X., Chen Z., Li J., Fan X., Deng H., Lu L., Li B., Liu Z., Wang Q., Lin D., Yang L. (2025) *Has GPT-5 Achieved Spatial Intelligence? An Empirical Study* (arXiv Preprint 2508.13142v1). https://doi.org/10.48550/arXiv.2508.13142

Challapally A., Pease C., Raskar R., Chari P. (2025) The GenAI Divide: State of AI Iin Business 2025, Cambridge, MA: MIT.

Chang Y., Wang X., Wang J., Wu Y., Yang L., Zhu K., Chen H., Yi X., Wang C., Wang Y., Ye W., Zhang Y., Chang Y., Yu P.S., Yang Q., Xie X. (2024) A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1–45. https://doi.org/10.48550/arXiv.2307.03109

CompTIA (2024) *State of the Tech Workforce*, Downers Grove, IL: The Computing Technology Industry Association (CompTIA).

Deng Y., Xia C. S., Peng H., Yang C., Zhang L. (2023) Large language models are zero-shot Fuzzers: fuzzing deep-learning libraries via large language models. In: *ISSTA 2023 – Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis,* New York: Association for Computing Machinery, pp. 423–435. https://doi.org/10.1145/3597926.3598067

Fu Y., Weng Z. (2024) Navigating the ethical terrain of AI in education: A systematic review on framing responsible human-centered AI practices. *Computers and Education Artificial Intelligence*, 7(1), 100306. https://doi.org/10.1016/j.caeai.2024.100306

Georgiou G.P. (2025) *Capabilities of GPT-5 across critical domains: Is it the next breakthrough?* (arXiv preprint 2508.19259). https://doi.org/10.48550/arXiv.2508.1925

Gładysz B., Despotis D., Kuchta D. (2024) Application of data envelopment analysis to IT project evaluation, with special emphasis on the choice of inputs and outputs in the context of the organization in question. *Journal of Information and Telecommunication*, 8(3), 301–314. https://doi.org/10.1080/24751839.2023.2286764

Gur I., Furuta H., Huang A. V., Safdari M., Matsuo Y., Eck D., Faust A. (2024) *A real-world WebAgent with planning, long context understanding, and program synthesis* (arXiv preprint 2307.12856). https://doi.org/10.48550/arXiv.2307.12856

Hajikhani A., Cole C. (2024) A critical review of large language models: Sensitivity, bias, and the path toward specialized AI. *Quantitative Science Studies*, 5(3), 736–756. https://doi.org/10.1162/qss_a_00310

Hu X., Xu Z., Ling Z., Jin Z., Du S. (2024) *Emerging Synergies Between Large Language Models and Machine Learning in Ecommerce Recommendations* (arXiv preprint 2403.02760). https://doi.org/10.48550/arXiv.2403.02760

Huang Y., Tang K., Chen M.A. (2024) *Comprehensive Survey on Evaluating Large Language Model Applications in the Medical Industry* (arXiv preprint 2404.15777). https://doi.org/10.48550/arXiv.2404.15777

Kouzminov Y., Kruchinskaia E. (2024) The Evaluation of GenAI Capabilities to Implement Professional Tasks. *Foresight and STI Governance,* 18(4), 67–76. https://doi.org/10.17323/2500-2597.2024.4.67.76

Laskar M.T.R., Fu X.-Y., Chen C., Bhushan T.N.S. (2023) Building real-world meeting summarization systems using large language models: A practical perspective. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Singapore: Association for Computational Linguistics, pp. 343–352. https://doi.org/10.18653/v1/2023.emnlp-industry.33

Lee J., Stevens N., Han S.C., Song M. (2024) *A survey of large language models in finance (finllms)* (arXiv preprint 2402.02315). https://doi.org/10.1007/s00521-024-10495-6

Li Y., Wang S., Ding H., Chen H. (2023) Large language models in finance: A survey. In: *Proceedings of the 4th ACM International Conference on AI in Finance*, New York: Association for Computing Machinery, pp. 374–382. https://doi.org/10.1145/3604237.3626869

Liu J., Gong Y., Zhu J., Titah R. (2021) Information technology and performance: Integrating data envelopment analysis and configurational approach. *Journal of the Operational Research Society*, 73(6), 1278–1293. https://doi.org/10.1080/01605682.2021.1907238

Mayer H., Yee L., Chui M., Roberts R. (2025a) *Superagency in the Workplace. Empowering People to Unlock AI's Full Potential*, New York: McKinsey & Company.

Noffsinger J., Patel M., Sachdeva P. (2025c) The Cost of Compute: A $7 Trillion Race to Scale Data Centers, New York: McKinsey & Company.

Raza M., Jahangir Z., Riaz M.B., Saeed M.J., Sattar M.A. (2025) Industrial applications of largelanguage models. *Scientific Reports*, 15, 13755. https://doi.org/10.1038/s41598-025-98483-1

Ren Q., Jiang Z., Cao J., Li C., Liu Y., Huo S., He T., Chen Y. (2024) *A survey on fairness of large language models in e-commerce: progress, application, and challenge* (arXiv preprint 2405.13025). https://doi.org/10.48550/arXiv.2405.13025

Saleh Y., Abu Talib M., Nasir Q., Dakalbab F. (2025) Evaluating large language models: a systematic review of efficiency, applications, and future directions. *Frontiers in Computer Science*, 7, 1523699. https://doi.org/10.3389/fcomp.2025.1523699

Shi J., Mei J., Zhu L., Wang Y. (2024) Estimating the Innovation Efficiency of the Artificial Intelligence Industry in China Based on the Three-Stage DEA Model. *IEEE Transactions on Engineering Management*, 71, 9217–9228. https://doi.org/10.1109/TEM.2023.3323292

Stanford University (2025) *Artificial Intelligence Index Report 2025,* Stanford, CA: The Stanford Institute for Human-Centered Artificial Intelligence.

Trott S., Jones C., Chang T., Michaelov J., Bergen B. (2023) Do large language models know what humans know? *Cognitive Science*, 47, 13309. https://doi.org/10.1111/COGS.13309

Wang S., Xu T., Li H., Zhang C., Liang J., Tang J., Yu P.S., Wen Q. (2024) *Large language models for education: A survey and outlook* (arXiv preprint 2403.18105). https://doi.org/10.48550/arxiv.2403.18105

Wen H., Li Y., Liu G., Zhao S., Yu T., Li T. J.-J., Jiang S., Liu Y., Zhang Y., Liu Y. (2024) *AutoDroid: LLM-powered Task Automation in Android*. Paper presented at the ACM MobiCom'24 International Conference on Mobile Computing and Networking, September 30 – October 4, 2024. https://doi.org/10.1145/3636534.3649379

WEF (2025) *The Future of Jobs Report 2025*, Geneva: World Economic Forum.

Xu H., Gan W., Qi Z., Wu J., Yu P.S. (2024) *Large Language Models for Education: A Survey* (arXiv preprint 2405.13001). https://doi.org/10.48550/arXiv.2405.13001

Yee L., Chui M., Roberts R., Smit S. (2025) *Technology Trends Outlook 2025,* New York: McKinsey & Company.

Zhao H., Liu Z., Wu Z., Li Y., Yang T., Shu P., Xu S., Dai H., Zhao L., Jiang H., Pan Y., Chen J., Zhou Y., Mai G., Liu N., Liu T. (2024) *Revolutionizing finance with LLMs: An overview of applications and insights* (arXiv preprint 2401.11641). https://doi.org/10.48550/arXiv.2401.11641

Zhao W.X., Zhou K., Li J., Tang T., Wang X., Hou Y., Min Y., Zhang B., Zhang J., Dong Z., Du Y., Yang C., Chen Y., Chen Z., Jiang J., Ren R., Li Y., Tang X., Liu Z., Liu P., Nie J.Y., Wen J.R. (2023) *A Survey of Large Language Models* (arXiv preprint 2303.18223). https://doi.org/10.48550/arXiv.2303.18223