Unlocking the Potential of Generative AI: The Importance of Funding Hardware for GPT Startups

Short Report

YITEC CO LTD: Nguyen Ngoc Anh

Table of Contents

Introduction

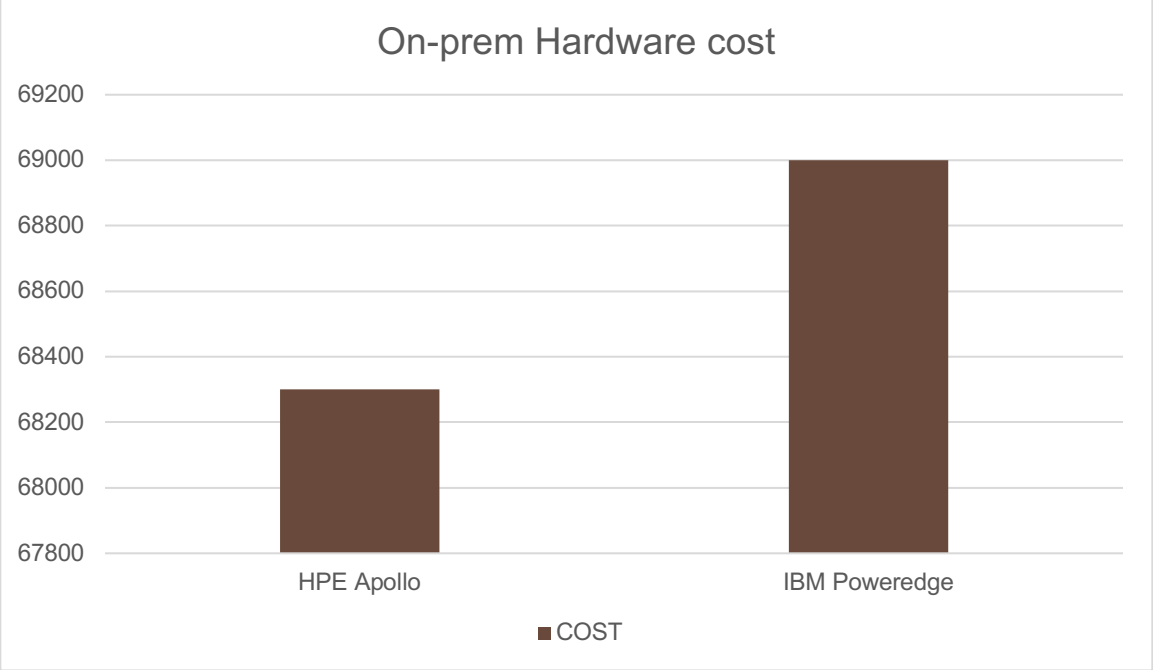
AI is revolutionizing the world, with applications ranging from healthcare to economics to entertainment. The creation of Generative Pre-trained Transformer (GPT) models, which can read, analyze, and synthesize human-like text, audio, and images, is one of the most promising and interesting fields of AI. GPT models have been employed in various applications, ranging from language translation to content creation, and can potentially transform many industries.

However, developing and training GPT models is an expensive and time-consuming procedure. Because of the enormous computing costs and infrastructure required to train these models, AI startups working on GPT technology have major barriers to entry. Despite the huge potential of GPT models, many entrepreneurs need help to get the requisite funds to purchase the appropriate hardware, limiting their capacity to innovate and scale their technology.

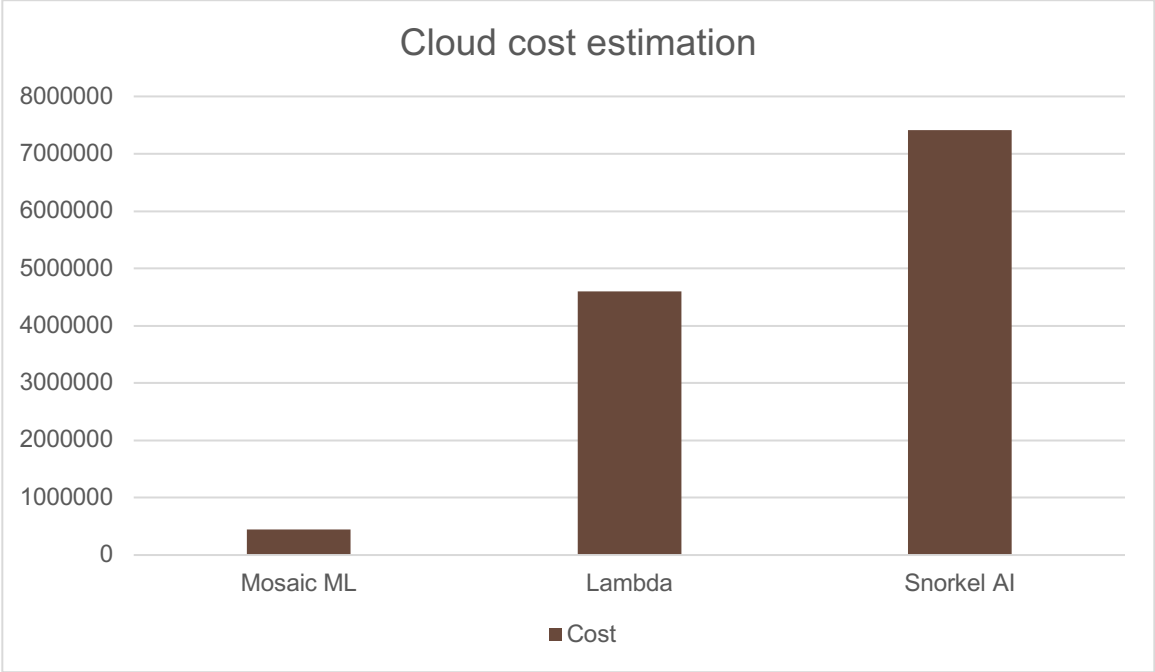
In this report, we will look at the challenges that AI startups dealing with GPT technology face and the role of hardware in developing and training GPT models. We will investigate the costs of training GPT models as well as the limitations of internal hardware funding. We will also assess the various funding options available to GPT startups and discuss why investors should fund hardware for these businesses. Finally, we'll go over the potential ROI for investors as well as the long-term advantages of investing in GPT technology. By the end of this report, we hope to have provided a thorough understanding of the challenges that GPT startups face and the importance of external funding in assisting them in realizing the full potential of this transformative technology.

Training GPT model is expensive

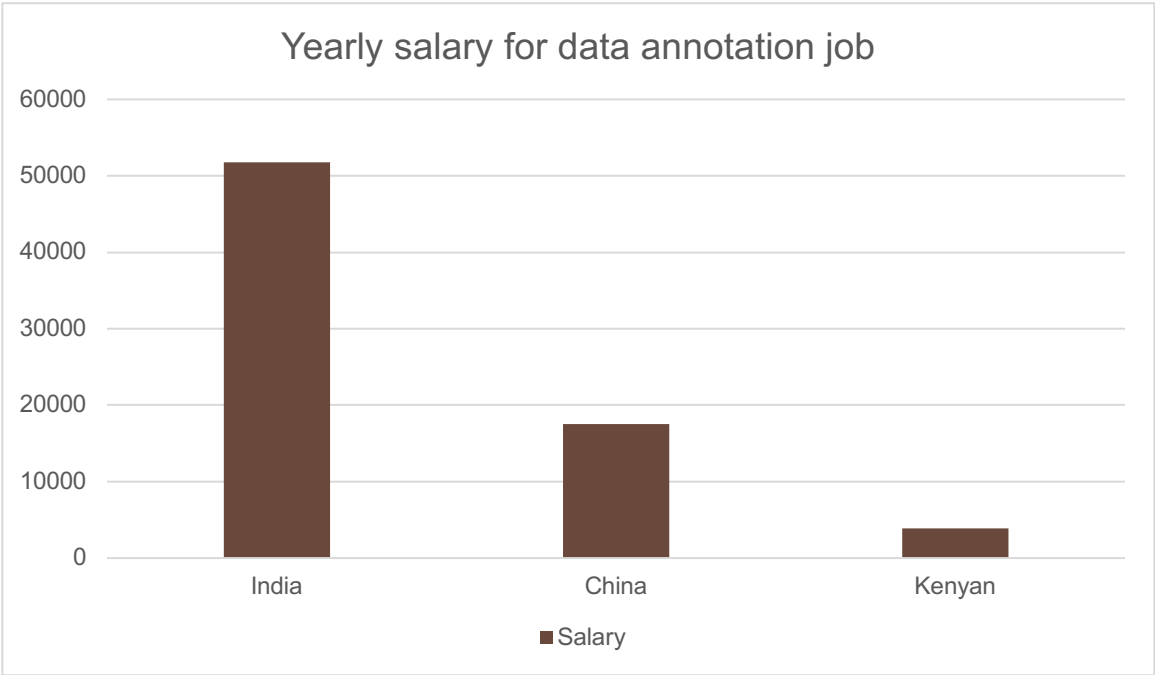
Training GPT models, like any other AI technique, demands the use of data, engineers, cloud computing resources, and hardware. It can also vary greatly based on the model's size. Here is an estimated cost for the resources required to train a GPT model:



Hardware resources: The cost of hardware varies according to the type and quantity of equipment used. For example, an HPE Apollo sx40 2SFF 4x32GB V100 SMX2 GPU CTO [1] costs $68301 for one year [2], while an IBM PowerEdge R740 supporting up to three V100 GPUs for PCIe is priced at $69,000 [3].

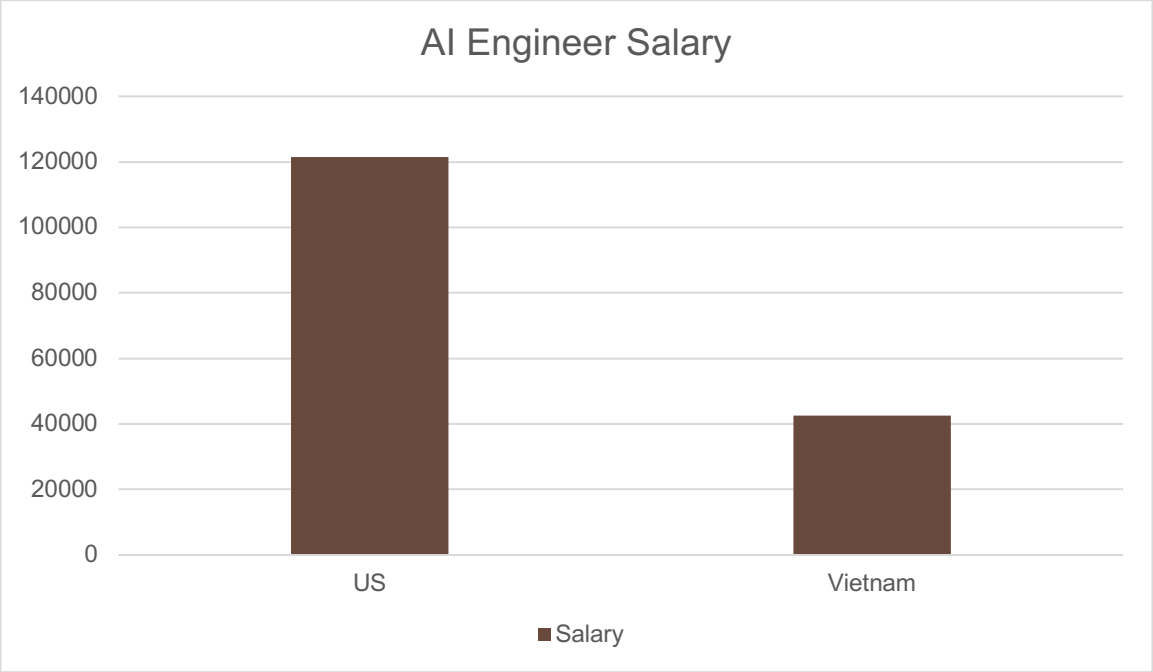


Cloud computing resources: The price of cloud computing services varies greatly depending on the provider, the amount of processing power required, and the duration of the training. Mosaic ML estimates that it costs about $450K to train a model that reaches GPT-3 quality [4]. And Lambda estimates a training cost of at least $4.6 million [5]. Finally, Snorkel AI estimates the total cost to create the model to be around $7,418 million, including fine-tuning [6].



Data resources: The cost of acquiring or producing training data varies greatly depending on the type and quantity of data required. Some startups may be able to get publically available datasets, whilst others may be required to develop their own data through crowdsourcing or other ways. For example, OpenAI hired Kenyan laborers for $2 per hour [7], although the overall

amount spent on data labeling in the ChatGPT project is unknown. In India, data annotation companies such as iMerit pay their employees an average of 2,200 local rupees (26.95 USD) per day [8] , while the average salary for a Data Analyst in China is ¥122084 (17,541.45 USD) per year [9].



Expertise: The cost of hiring a team with expertise in AI, natural language processing, and deep learning can vary depending on the location and experience of the team members. For example, the average salary of an AI engineer in the US is $127,491 per year, according to Glassdoor [7]. Lead AI engineers earn an average salary of $170,265 per year in the US [8] and 85M VND in Vietnam [9]

Time and energy: The time and energy required to train a GPT model might be difficult to calculate, but it can be substantial. Training a cutting-edge GPT model like GPT-3, for example, could take many weeks or even months, during which time the hardware and infrastructure required to operate the model will consume energy and incur expenditures.

In summary, depending on the size and complexity of the model, as well as the resources required to complete the project, the cost of designing and training a GPT model can range from hundreds of thousands to several million dollars.

The importance of Hardware and Cloud Computing

Cloud computing and hardware both play essential roles in GPT model training. GPT models take a lot of computational power to train, and both cloud computing and specialist hardware can offer it.

Cloud computing offers a scalable and adaptable alternative for GPT model training. Startups can avoid the high upfront expenses of purchasing and maintaining specialized hardware by employing cloud-based computing services, and they can scale their computing capabilities up or down as needed. Cloud computing services also give users access to a diverse set of hardware configurations, such as dedicated GPUs or TPUs designed for machine learning tasks.

A GPU (Graphics Processing Unit) is a specialized processor that performs complex mathematical calculations required for graphics rendering and other compute-intensive applications. GPUs are utilized in a variety of applications, such as gaming, scientific research, and machine learning. A TPU (Tensor Processing Unit) is a specialized processor designed by Google to accelerate machine learning workloads such as large-scale neural network training. TPUs are designed to execute matrix operations, which are commonly employed in machine learning techniques. For machine learning tasks, TPUs can give significant performance advantages over CPUs and GPUs, especially for large-scale models. TPUs are available as cloud-based services on Google Cloud as well as dedicated hardware for on-premises data centers.

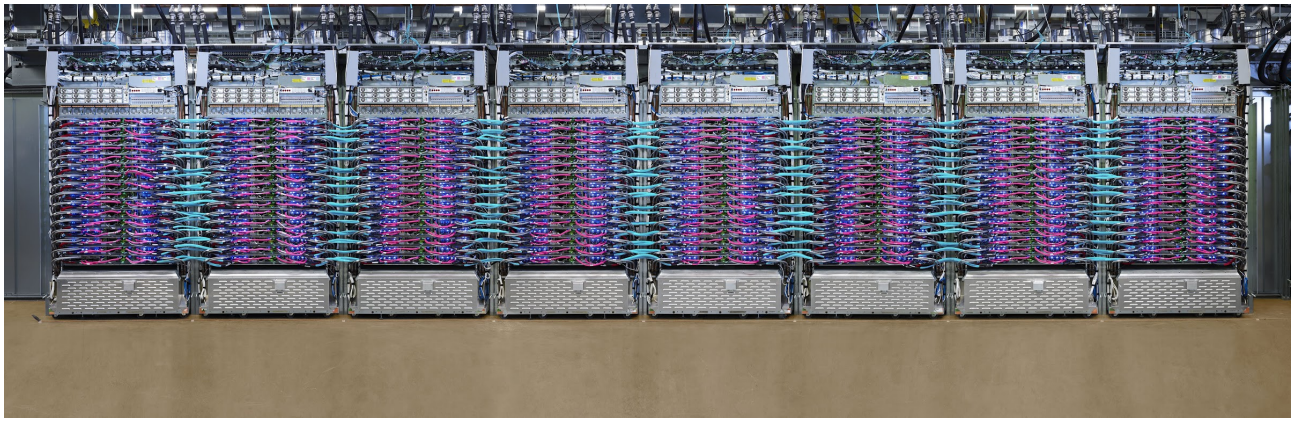


Figure 1: Google Cloud TPU

GPUs or TPUs, for example, can dramatically speed up the training process for GPT models. GPUs can execute parallel operations on huge datasets, which can significantly reduce training durations when compared to ordinary CPUs [13]. TPUs are primarily intended for machine learning workloads and can deliver even higher performance benefits, especially for large-scale models.

GPT models can also benefit from cloud computing and specialized hardware in terms of scalability and efficiency. Cloud computing enables AI companies to rapidly deploy and grow compute resources as needed, without the need for upfront hardware investments. Specialized hardware allows for the creation of larger and more complicated GPT models, resulting in more accurate and sophisticated outcomes.

However, both cloud computing and specialized hardware can also be expensive. As we analyzed how much it would cost to train GPT models in the previous section, computing costs to build this AI technology could range from a hundred thousand to millions of dollars. Cloud computing costs can quickly add up for large-scale training, and the cost of specialized hardware can be prohibitive for startups with limited resources. This is why external funding is crucial to help startups acquire the necessary resources to train their GPT models and compete with established players.

Funding options

Acquiring hardware for GPT companies might be problematic, especially given the high costs of building and testing these models. The following are some of the funding options available to entrepreneurs looking to fund their hardware:

Internal funding: Startups may choose to fund their hardware internally, either from operational revenues or by reallocating capital from other sections of the firm. This can give companies more control over their resources, but it may not be enough to cover the hefty expenses of developing and training GPT models.

Grants: To support their hardware demands, startups may be eligible for government grants or research money. Grants can provide non-dilutive capital to businesses, which means they do not have to give up equity in exchange for funding. Grants, on the other hand, can be difficult to secure and may come with strings attached, such as restrictions on how the funds can be used or a requirement to partner with research institutions.

External finance: Startups may seek external money to meet their hardware demands, such as venture capital, angel investors, or other kinds of funding. External funding can provide businesses with the funds they need to construct and train their GPT models, as well as access to seasoned investors who can offer advice and help. External investment, on the other hand, frequently comes with a loss of ownership and control, and companies may be required to fulfill particular milestones or targets in order to get additional money.

Cloud provider programs: Several cloud providers, like Google for Startups Cloud Program, IBM's Startup with IBM program, and AWS Activate for Startups, Founders, and Entrepreneurs, give free credits to startups. These initiatives offer up to $120,000 in free cloud credits and other resources to assist companies in getting started with cloud computing. These credits can be used by startups to cover the price of hardware, including the specific hardware required for GPT model training. OpenAI, for example, obtained free cloud credits from Microsoft to aid in the training of its GPT models. This solution can provide startups with a low-cost way to get the hardware resources needed to create and train their GPT models.

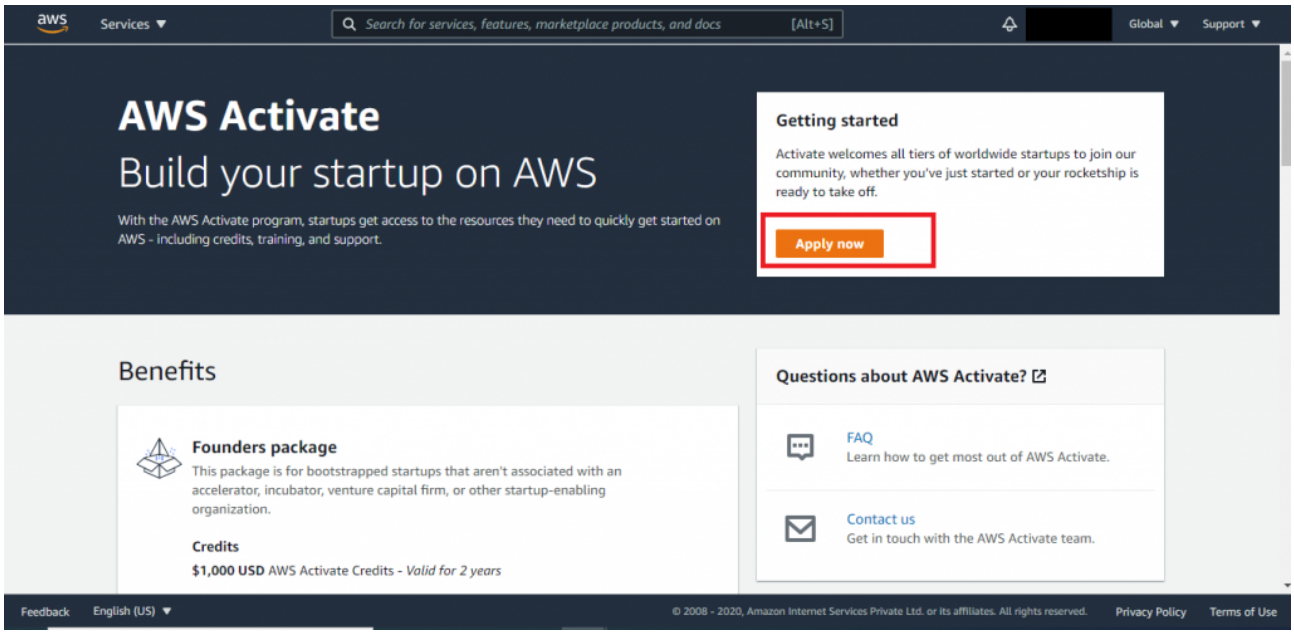


Figure 2: AWS Activate registration screen

Each of these funding methods has advantages and downsides, and the optimal solution will rely on the startup's individual needs and resources. Internal finance can give entrepreneurs more control over their resources, but it may not be enough to cover the expensive expenses of developing and training GPT models. Grants can provide non-dilutive cash, but they can be difficult to secure and come with strings attached. External investment can provide businesses with the capital they need to flourish, but it sometimes comes with a loss of ownership and control. Cloud provider programs can provide free cloud credits and other resources to help businesses get started, which can be a cost-effective approach to obtaining the necessary physical resources.

Entrepreneurs will need to carefully consider their alternatives in order to determine which funding source or mix of sources will best support their hardware needs and assist them in scaling their GPT technology. However, the necessity for outside finance is obvious, as the high cost of developing and training GPT models makes it difficult for startups to compete with established businesses without major investment.

Why funding generative AI startups

A recent Grand View Research analysis shows that the global generative AI market is estimated to reach USD 109.37 billion by 2030, growing at a 34.6 percent CAGR from 2022 to 2030. Acumen Research also predicted a CAGR of around 34.4% (See Fig.3). The growing need for generative AI applications is predicted to be driven by the growing demand to update workflows across industries.

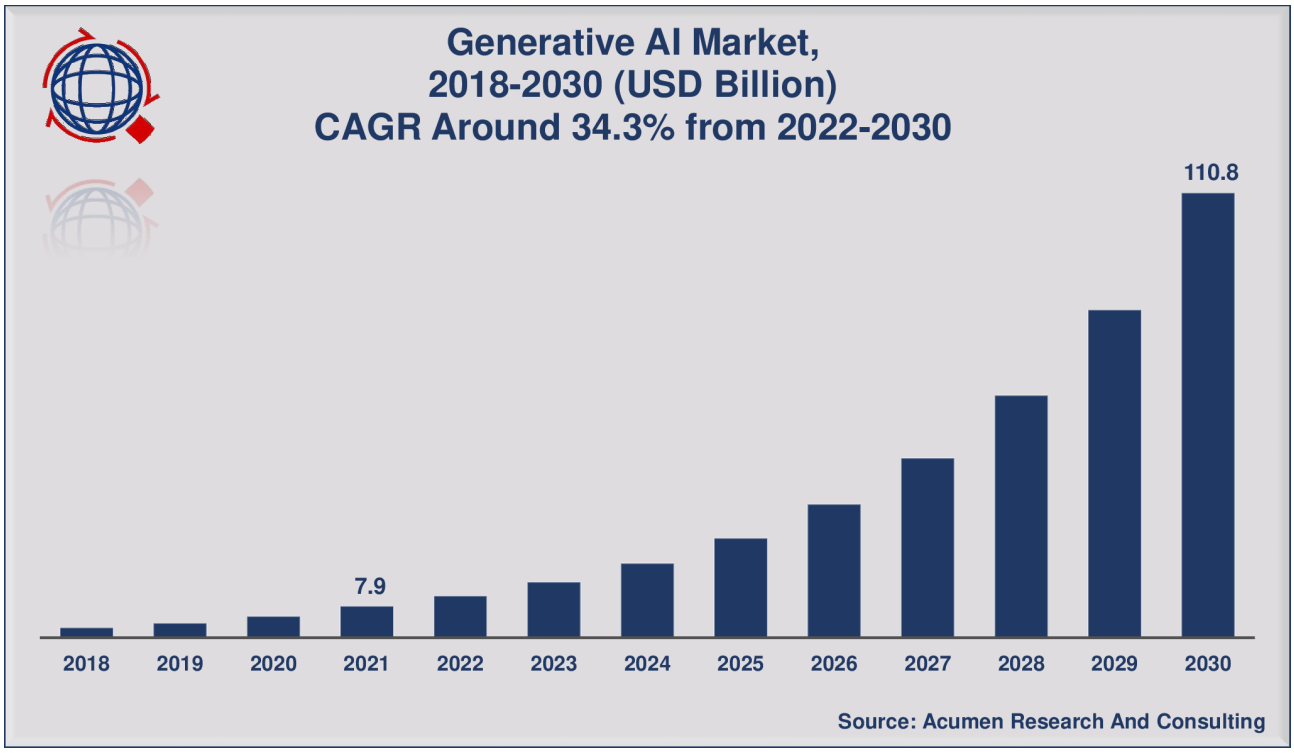


Figure 3: Generative AI Market

The COVID-19 pandemic has had a positive influence on the generative AI business, with numerous organizations deploying Machine Learning (ML) and Artificial Intelligence (AI) to combat the outbreak. During the pandemic, market players such as Microsoft, IBM, Google, and Amazon Web Services saw an increase in sales of AI-based technologies. The rapid expansion of digital platforms has also facilitated the use of generative AI applications. Amazon Web Services, for example, recently announced the inclusion of a new generative AI algorithm, Autoregressive Convolutional Neural Network (AR-CNN), to its AWS DeepComposer portfolio of products, enabling developers to digitally produce music.

Market participants in the generative AI space provide solutions for a wide range of applications, including text-to-image, image-to-image, and super-resolution. These industry players are also working on generative AI technology for enhanced picture resolution, face aging, and video resolution. Tesla, for example, is creating autonomous algorithms based on data from car sensors and neural networks that have been trained to recognize objects and perform semantic segmentation.

Investing in hardware for GPT firms might position investors to gain from the development of the generative AI market's growth and potential income. The potential ROI for investors that fund hardware for GPT firms is substantial, especially as the technology evolves and expands into other industries. Microsoft has invested $10 billion in OpenAI, the firm behind ChatGPT [14], and will receive roughly half of OpenAI's financial returns until a predetermined cap is reached. According to Gartner analyst Jason Wong, AI-generated content could be utilized for more than just producing emails or essays by 2024 [15]. Microsoft intends to bring ChatGPT functionality to Azure as well as announce the broad availability of its Azure OpenAI Service, which has been available to a select group of users since its initial release. The investment might also propel Microsoft to the forefront of

artificial intelligence, paving the path for the company to embed ChatGPT into some of its major apps, such as Word, PowerPoint, and Excel [16].

As the market expands, the need for more advanced GPT models will become more critical. Significant investment in hardware and infrastructure will be required to enable the development and training of these models. Startups that can acquire hardware investment will be better positioned to compete in this industry and capitalize on future growth prospects.

Investors that invest in hardware for GPT firms can benefit from GPT technology's long-term worth. As the technology spreads across businesses, it has the potential to disrupt multiple industries and produce enormous income. Investors can position themselves to gain from this expansion and potentially make big returns on their investment by investing in hardware for GPT firms.

Conclusion

In conclusion, this report has highlighted the substantial expenses associated with developing and training GPT models, as well as the necessity of hardware in this process. We've talked about the numerous funding alternatives for companies, such as internal funding, grants, external investment, and cloud provider programs. We also discussed the possible ROI and long-term benefits of investing in hardware for GPT businesses, such as the rise of the generative AI market and the possibility for large revenue in a variety of industries.

Investing in hardware for GPT startups is an important step toward realizing the full potential of this game-changing technology. Investors can help entrepreneurs compete in a crowded industry and capitalize on the potential growth opportunities of generative AI by assisting them in the development and training of powerful GPT models.

We advise investors to consider supporting hardware for GPT startups and assisting these businesses in realizing the full potential of this technology. The benefits of investing in hardware for GPT startups are substantial, and by assisting these businesses, investors can position themselves to gain from the development and disruption that this technology is projected to bring to numerous industries in the future years. The time has come to invest in hardware for GPT businesses and be a part of generative AI's future.

Works Cited

[1] NVIDIA, "How to Buy NVIDIA Virtual GPU Solutions," 2023. [Online]. Available: https://www.nvidia.com/en-us/data-center/v100.

[2] R. Kumar, "NVIDIA DGX Station Upgraded to Tesla V100," 28 Octorber 2017. [Online]. Available: https://www.servethehome.com/nvidia-dgx-station-upgraded-tesla-v100/.

[3] NVIDIA, "NVIDIA HGX AI Supercomputer," 2023. [Online]. Available: https://www.nvidia.com/en-us/data-center/hgx/. [Accessed 2023].

[4] L. L. Abhinav Venigalla, "Mosaic ML estimates that it costs about $450K to train a model that reaches GPT-3 quality[1], while Reddit users estimate the cost at $12 million and $4.6 million[2]. Next Platform estimates the cost of training GPT-70B at $35.71 per 1 million parameters[," 29 September 2022. [Online]. Available: https://www.mosaicml.com/blog/gpt-3-quality-for-500k#:~:text=The%20bottom%20line%3A%20it%20costs,10x%20less%20than%20people%20think.. [Accessed 2023].

[5] C. Li, "OpenAI's GPT-3 Language Model: A Technical Overview," 3 June 2020. [Online]. Available: https://lambdalabs.com/blog/demystifying-gpt-3#:~:text=The%20cost%20of%20AI%20is%20increasing%20exponentially.%20Training%20GPT-3%20would%20cost%20over%20%244.6M%20using%20a%20Tesla%20V100%20cloud%20instance.. [Accessed 2023].

[6] "Better not bigger: How to get GPT-3 quality at 0.1% the cost," 17 November 2022. [Online]. Available: https://snorkel.ai/better-not-bigger-how-to-get-gpt-3-quality-at-0-1-the-cost. [Accessed 2023].

[7] B. Perrigo, "OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic," 18 Jan 2023. [Online]. Available: https://time.com/6247678/openai-chatgpt-kenya-workers. [Accessed 19 Feb 2023].

[8] Financial Times, "AI’s new workforce: the data-labelling industry spreads globally," [Online]. Available: https://www.ft.com/content/56dde36c-aa40-11e9-984c-fac8325aaa04.

[9] Payscale, "Average Data Analyst Salary in China," 2023. [Online]. Available: https://www.payscale.com/research/CN/Job=Data_Analyst/Salary. [Accessed Feb 2023].

[10] Glassdoor, "How much does an AI Engineer make?," 2023. [Online]. Available: https://www.glassdoor.com/Salaries/ai-engineer-salary-SRCH_KO0,11.htm. [Accessed 19 Feb 2023].

[11] Salary.com, "Lead AI Engineer Salary in the United States," 2023. [Online]. Available: https://www.salary.com/research/salary/benchmark/lead-ai-engineer-salary. [Accessed 2023].

[12] Glassdoor, "How much does a Machine Learning Engineer make in Ho Chi Minh, Vietnam?," 2023. [Online]. Available: https://www.glassdoor.com/Salaries/ho-chi-minh-city-machine-learning-engineer-salary-SRCH_IL.0,16_IM1746_KO17,42.htm. [Accessed 2023].

[13] T. . Baji, "GPU: the biggest key processor for AI and parallel processing," , 2017. [Online]. Available: https://spiedigitallibrary.org/conference-proceedings-of-spie/10454/1/gpu--the-biggest-key-processor-for-ai-and-parallel/10.1117/12.2279088.full. [Accessed 19 2 2023].

[14] D. Bass, "It’s raining money for ChatGPT company OpenAI as Microsoft officially throws down a $10 billion investment," 23 Jan 2023. [Online]. Available: https://fortune.com/2023/01/23/microsoft-investing-10-billion-open-ai-chatgpt.

[15] S. Ortiz, 23 Jan 2023. [Online]. Available: Microsoft just made a huge investment in ChatGPT maker OpenAI. Here's why. [Accessed 2023].

[16] C. B. Samantha Murphy Kelly, "Microsoft confirms it’s investing billions in the creator of ChatGPT," 24 Jan 2023. [Online]. Available: https://www.cnn.com/2023/01/23/tech/microsoft-invests-chatgpt-openai/index.html. [Accessed 2023].

Contact

YITEC CO LTD

contact@yitec.group

Address: Room 2304, Leadvisor Tower, 643 Phạm Văn Đồng, Cổ Nhuế, Bắc Từ Liêm, Hà Nội 100000, Việt Nam

Tel: +84 24 7109 9234

Nguyen Ngoc Anh

anh.nguyen@yitec.group