Modeling the Future of AI Development: Economic Viability and Geopolitical Implications

Geopolitics of AGI
Anton Shenk
September 26, 2024

Executive Summary

This memo presents a model for estimating the profitability of future frontier AI. Our analysis aims to predict whether private companies (e.g., OpenAI, Anthropic) will continue investing in larger, more powerful models – or if state intervention may become necessary to push model scaling further. This information is crucial for policymakers considering the trajectory of AI development and its geopolitical implications.

Key Points

1. We've developed a model to estimate the profitability of frontier AI models based on various factors, including training costs, operational costs, and potential revenue.
2. An online interface allows users to input their own assumptions and explore different scenarios: AGI Profitability Simulator
3. Our case study examines the profitability of 2024 models under different assumptions – demonstrating the model's utility in predicting private AI developers' behavior.
4. We project that, by 2028, training costs for frontier models could reach approximately $20 billion – potentially necessitating state intervention for continued development.

Case Study: 2024 Models

This case study demonstrates how our model can be used to evaluate the potential behavior of private frontier AI developers – and the implications of increasing costs of frontier AI training. While the current model relies on assumptions that require further investigation, it nonetheless provides valuable insights into the economics of AI development.

Default Scenario[1]

First, we consider the profitability of a model developed today under our default assumptions:

- Markup per token: 250%
- Annual growth rate in model usage: 40%
- Initial usage: 50,000,000,000,000 tokens per year
- Estimated training cost: $337,222,404

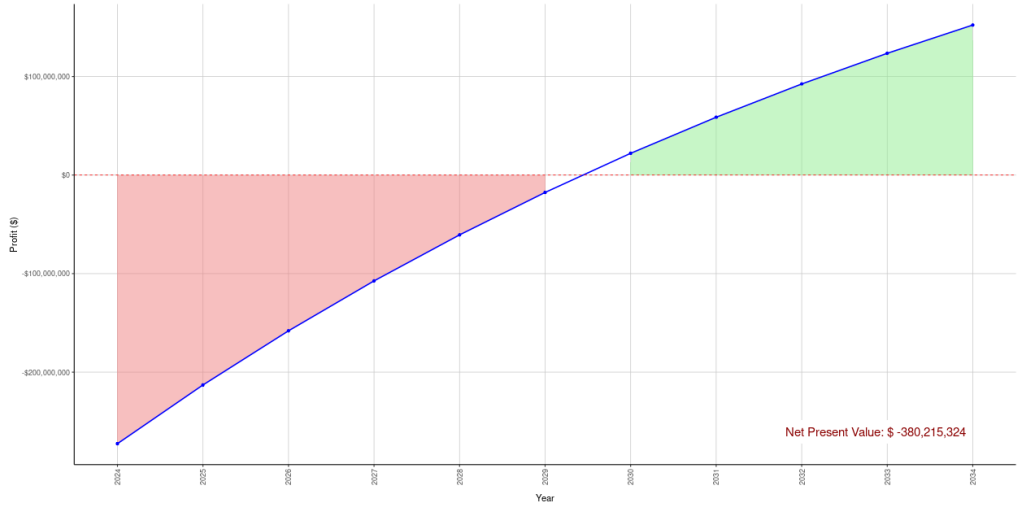
Key findings:

- Time to break even: Approximately 5 years
- Net present value: Negative

The long timeline to profitability suggests challenges for the most advanced models developed today.

[1] The methodology for determining these values is described in detail in *Appendix A*.

Figure 1: Net Present Value of a 2024 Model Under Default Assumptions



Alternative Scenarios

It's possible that models developed today may be able to command a higher markup or grow their usage at faster rates than our default assumptions suggest. This could be due to their increased capabilities, allowing them to perform more valuable work. Additionally, there may be factors that could reduce the inference cost, which our model may not capture. Such reductions would increase the profit per token that these models can generate. To account for these possibilities, we explored two alternative scenarios:

1. **Increased markup to 300% per token** – which assumes that the advanced capabilities of the model justify a higher price point (see *Figure 2*).
2. **Increased annual growth rate to 50%** – which assumes a faster adoption rate and expanding use cases for the model (see *Figure 3*).

Figure 2: Net Present Value of a 2024 Model with Markup of 300% per Token

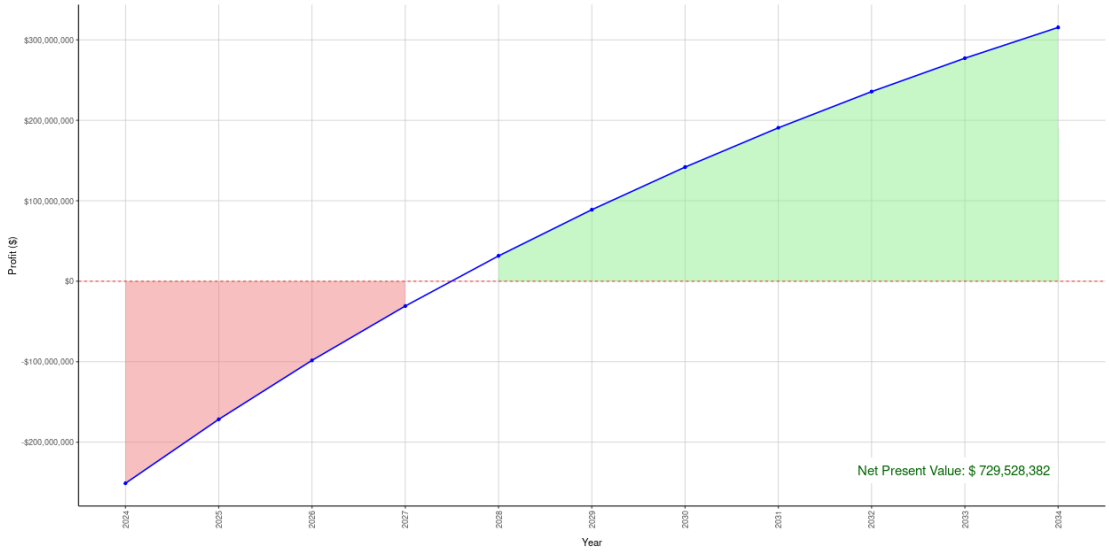
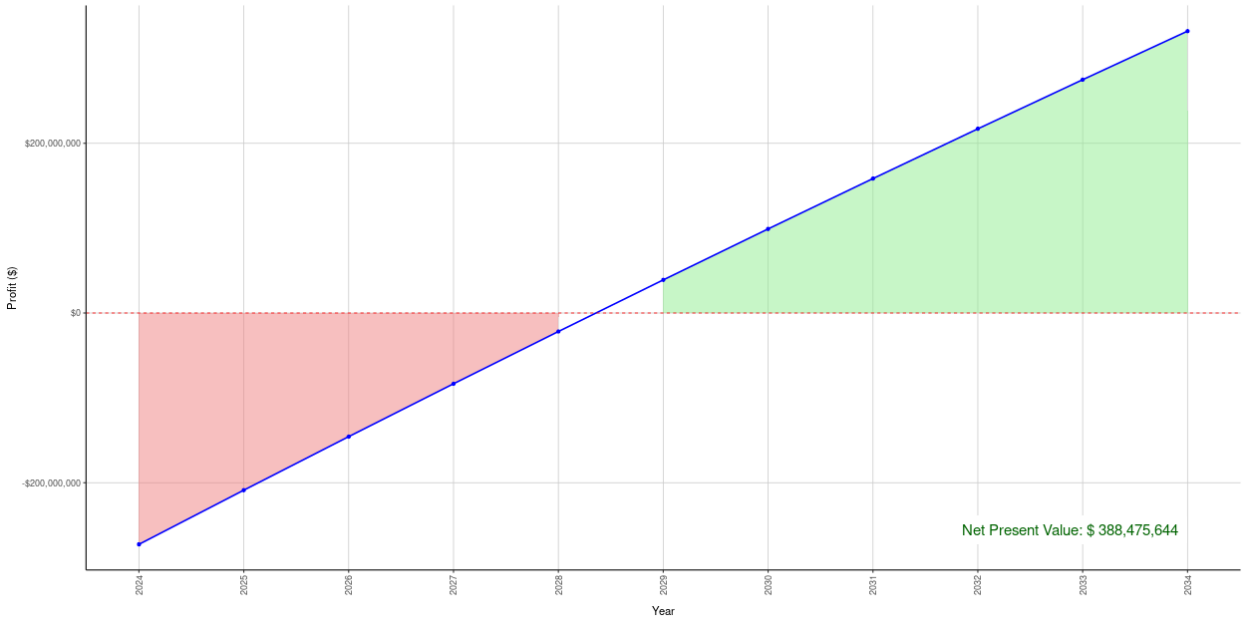


Figure 3: Net Present Value of a 2024 Model with Annual Growth Rate of 50%



In both cases:

- Break-even time reduced to 3-4 years
- Positive net present value achieved

These scenarios suggest a more viable business case for training 2024 models. The improved profitability in these cases indicates that even small changes in markup or growth rate can significantly impact the economic viability of frontier AI models. However, these alternative scenarios represent potential outcomes based on optimistic assumptions about the model's performance and adoption. They highlight the sensitivity of our projections and underscore the importance of accurately forecasting these factors.

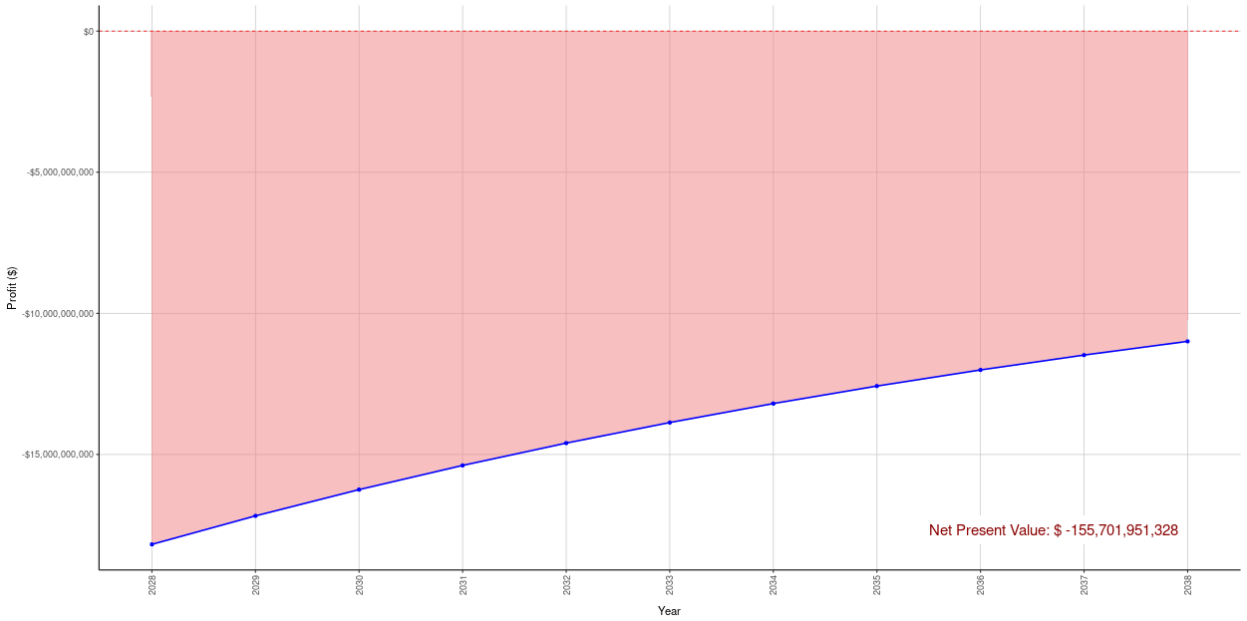
Future Projections and Implications

Looking ahead, our model projects significant challenges for the economic viability of frontier AI development. By 2028, we estimate that training costs for frontier models could reach a staggering $20 billion. Under our default assumptions, such models may not achieve profitability within a 10-year timeframe, presenting a considerable obstacle for private sector investment (see *Figure 4*). For these 2028 models to become economically viable, they will likely require a combination of factors:

- Significantly higher markup on their services,
- Substantially increased utilization rates,
- Breakthroughs in inference cost reduction.

Our analysis offers crucial insights into the economic landscape facing frontier AI developers. It outlines the challenging thresholds of profitability, cost markups, and usage growth rates that future models must achieve to be economically viable. The magnitude of these projections points to a potentially seismic shift in AI development dynamics. As costs escalate and traditional business models strain under the weight of massive investments, we may see a fundamental restructuring of how advanced AI systems are funded and developed. This could lead to new collaborative models between private enterprises and public institutions, or even necessitate increased government involvement in pushing the boundaries of AI capabilities.

Figure 4: Net Present Value of a 2028 Model Under Default Assumptions



Areas for Future Study and Limitations

Our current model, while informative, faces several limitations that warrant further investigation. One significant challenge is the limited data available for forecasting model demand and pricing. As the field of AI rapidly evolves, historical data may not accurately reflect future trends. Future analysis will need to focus on better quantifying the markup and usage rates required for profitable deployment of frontier models, taking into account emerging applications and market dynamics.

Another area of uncertainty lies in capturing the true inference costs for frontier AI developers. The AI industry has made significant efforts to lower these costs, which may result in greater effective profit per token than our model currently captures. Further research is crucial to understand these costs accurately, as even small variations can have substantial impacts on long-term profitability projections.

It's also important to consider that some frontier AI developers may make decisions based on factors beyond current model profitability. These non-economic motivations could include assumptions about future model profitability, belief in the technology's importance beyond immediate economic returns, or, in the case of state actors, national security considerations. Understanding these motivations is crucial for predicting the trajectory of AI development, especially as the costs of training frontier models continue to escalate.

A key area for future analysis is identifying the tipping point for state intervention in AI development. As training costs rise, there may come a point when it's no longer reasonable to expect economically rational private actors to continue developing frontier AI models. Determining this threshold could help predict when intervention from non-economically motivated actors, likely state entities, becomes necessary to advance AI capabilities further.

Appendix A: Methodology for Modeling the End of Industry AI Development

This appendix briefly describes the methodology used for estimating when private industry may cease to fund the development of increasingly capable AI models. For greater detail, see *Shenk 2024a* (below).

The model utilizes a net present value (NPV) approach to determine the economic viability of conducting large-scale AI training runs, factoring in fixed training costs, operational costs, revenues, and discounting future cash flows. In particular, the model aims to forecast the point at which the development of increasingly large AI models becomes economically irrational for private firms. The profit function *P* is expressed as:

$$P = \sum_{t=1}^{T} \frac{R(t) - O(t)}{(1+\delta)^t} - C$$

Where *P* is the net present value (NPV) of profits from a model deployed for *T* periods, *R(t)* is the revenue from a model at time t, O*(t)* is the operational cost of a model at time t, δ is the discount rate, and C is the fixed cost of model training.

We suppose private industry continues to invest in AI model development where $P > 0$.

Future Model Training Costs – C

Model training cost, C, is a one-time expenditure representing the compute required to train a large-scale AI model. It is primarily determined by the number of floating-point operations required for training ($tFLOP$) and the cost per FLOP ($tFLOP\_\$$) executed at a given precision with assumed advances in hardware.

$$C_{\,Year,\ Precision,\ Rate} = \frac{tFLOP_{\,Year}}{tFLOP\_\$_{\,Year,\ Precision,\ Rate}}$$

Profit Per Period

Operational Cost – O(t)

Operational costs reflect the expenses incurred from model inference. These costs depend on the number of tokens generated by a model of a given size and the efficiency of the hardware used.

$$O(t)_{\,Year,\ Precision,\ Rate,\ \ Size} = \frac{iFLOP_{\,Size}}{iFLOP\_\$_{\,Year,\ Precision,\ Rate}} \times N(t)$$

Where $iFLOP$ is the computational requirement per token generated as a function of model size, $iFLOP\_\$$ is the price-performance of inference compute, and *N(t)* is the number of tokens generated in period t.

Revenue – R(t)

Revenue is modeled as a markup on operational costs:

$$R(t) = \ (1+m)\ \times O(t) \times N(t)$$

Where *m* is the markup applied to the operational cost, reflecting the firm's ability to generate profit from selling access to a model. In our model, we use a baseline markup of 250% – as it aligns with the gross margins commonly observed in the AI and tech sectors. Major cloud providers and AI software companies often target margins in the range of 60-80%, which corresponds to a markup of approximately 150-400%. This 250% markup is a reasonable middle ground, balancing between the more conservative and aggressive pricing strategies seen in industry.

Model Usage – N(t)

Model usage, N(t), evolves over time, reflecting increasing demand for models. Our current implementation models usage as purely increasing – without factoring in decay from obsolescence.

$$N_{\,initial\,|\,year}\,(g) = N_{\,initial} \times (1+g)^{year-2024}$$

Where $N_{\,initial}$ is the amount of inference in the model's release year and g is the growth rate, representing the annual increase in model usage as it gains adoption. We base our assumptions for model usage on data shared by Sam Altman, which highlights that OpenAI generates approximately 100 billion words per day. We convert this figure to tokens (at a rate of 0.75 words per token) and scale it up to reflect annual output. Meanwhile, our default growth rate is informed by projections for the AI market and cloud services, where annual growth is expected to range from 30-50% in the near term.

Discount Factor – δ

The discount factor δ represents the firm's time preference for revenue and accounts for investment risk. It adjusts future cash flows to their present value, assuming that money today is worth more than the same amount in the future due to opportunity costs and risk. We set the default discount rate at 10%, considering a balance between long-term risk-free rates, typical equity risk premiums, and the specific uncertainties inherent in AI development. This baseline incorporates a 4-5% risk-free rate (aligned with 10-year U.S. Treasury yields as of 2024), a 5-6% equity risk premium, and a small additional premium to reflect the risks and volatility unique to long-term AI investments. Using 10% as the default discount rate ensures we account for both investment risks and the opportunity costs faced by tech firms, while remaining conservative enough for long-term projections.

Modeling When Industry Development of Increasingly Capable AI Models May End

Anton Shenk

July 23, 2024

Introduction

An essential question for understanding the geopolitical implications of A(G)I is identifying who will control AGI at the time of its development. The emergence of AGI across multiple private companies might portend a highly decentralized AGI future where it was unnecessary to concentrate vast amounts of compute, while the development of AGI in a single government-controlled "project" owning significant quantities of compute would allow for a much more centralized and controllable development of such an advanced technology. This memo presents a model to help determine who may fund the creation of AGI and potentially control it, by estimating the level of compute at which economically rational actors may cease supporting advanced AI development – and where other actors, such as states, may need to step in to push forward continued AI progress.

The Current and Future Landscape of AI Development

Currently the development of increasingly capable AI models is driven almost entirely by industry – as, for now, the capital costs required for training frontier models have exceeded the capacity of available public funding. We believe much of the private capital currently funding AI scaling is "economically rational," meaning that they are funding AI development on the belief that there will be a positive return on these investments. It is possible that, in the future, private capital may no longer be able to support increasingly large training runs because it will no longer be economically rational to do so. For example, even if the model is increasingly capable, the willingness of customers to pay for access to it may be lower than it costs to build and deploy.

If models can no longer "pay off" their underlying costs, it will no longer be rational for economic actors to develop increasingly large and capable AI. When such a threshold is reached it will be necessary for "irrational" capital, such as states seeking security or ideologically motivated actors seeking to produce increasingly capable AI, to take over the funding of AI training. This threshold could be reached before or after the development of AGI, with significant implications for what actors may fund the technology's development. We propose below a model that could estimate such a threshold.

Modeling Decisions to Invest in a Training Run

This analysis could employ simple economic modeling to analyze the decision to conduct large-scale AI training runs. Our proposed model incorporates key variables including fixed costs (training), variable costs (inference), and marginal revenue, utilizing a net present value (NPV) approach to account for the time value of money and investment risk.

Decision Criteria

Economically rational actors are assumed to invest in a training run if the expected profit (P) is positive. The breakeven point, where $P = 0$, determines the maximum economically viable model training run.

Profit Equation

The profit function is expressed as:

$$P = \sum \frac{R(t) - O(t)}{(1+\delta)^t} - C$$

In this equation, P represents the lifetime profit from deploying a model, R(t) and O(t) denote the revenue and operational costs at time *t*, δ is the discount rate accounting for the time value of money and investment

risk, and C represents the initial fixed cost of training the model. Altogether, this equation calculates the present value of future cash flows, subtracting the initial fixed cost of model training. Such a model could produce estimates of expected profits from the largest foundation model training run in any given year – given assumptions about model development decisions including:

- Model size (e.g., parameter count)
- Parameter precisions (e.g., FP16, FP32)
- Hardware improvement (rate of change in FLOP/$)
- Scale of model deployment over time
- Firm markups

Training Compute Cost Estimation

Representing the one-time compute expense of model training in a given year, C can be estimated by:

$$C_{\,Year,\,Precision,\,Rate} = \frac{tFLOP_{\,Year}}{tFLOP\_\$_{\,Year,\,Precision,\,Rate}}$$

Where:

- $C_{\,Year,\,Precision,\,Rate}$ is the estimated cost for the specified year, parameter precision, and assumed growth rate of FLOP/$.
- $tFLOP_{\,Year}$ is the predicted floating-point operations required for training in the specified year.
- $tFLOP\_\$_{\,Year,\,Precision,\,Rate}$ is the projected training floating-point operations per dollar for the specified year, parameter precision, and rate of hardware improvement.

Estimates of expected training cost for future models have already been generated as part of the earlier "Are Foundation Models a Natural Monopoly?" project. This model would leverage this earlier work to estimate the one-time compute costs of training new models. Appendix A contains a detailed methodology for estimating model training compute costs.

Operational Cost Estimation

Operational costs, $O(t)$, represent expenses associated with model inference. These costs are estimated primarily through estimating the cost per model prediction (e.g., token) of the model multiplied by the number of predictions generated during a particular period. These costs are calculated as:

$$O(t)_{\,Year,\,Precision,\,Rate,\,\,Size} = \frac{iFLOP_{\,Size}}{iFLOP\_\$_{\,Year,\,Precision,\,Rate}} \times N(t)$$

Where:

- $O(t)_{\,Year,\,Precision,\,Rate,\,Size}$ is the operational cost in period t of a model trained in a given year, parameter precision, assumed growth rate of FLOP/$, and model size.
- $iFLOP_{\,Size}$ – see Appendix B – is the predicted floating-point operations required for generating a model prediction from a model of a given size.
- $iFLOP\_\$_{\,Year,\,Precision,\,Rate}$ is the projected inference floating-point operations per dollar for the specified year, parameter precision, and rate of hardware improvement.
- $N(t)$ – detailed in Appendix B – is the number of model predictions generated in period t.

Revenue Estimation

Revenue, R(t), is modeled assuming a markup on operational costs:

$$R(t) = (1 + m) \times O(t) \times N(t)$$

Where m represents the markup percentage (e.g., 0.75 for a 75% markup). Estimation of both $R(t)$ and $O(t)$ accounts for the declining utilization of models over their lifecycle due to the introduction of newer, more advanced models, through the evolution of $N(t)$. Additional work could be done to estimate the change in markup percentage across a model's lifecycle.

Limitations and Assumptions

This model assumes a continuation of current trends in AI development and market dynamics. It does not account for potential disruptive technologies or significant shifts in the AI landscape. The revenue model simplifies complex market dynamics, model production inputs, and assumes a relatively stable competitive environment. Additionally, this model does not account for decisions to open-source AI models. Future work could refine these assumptions with more granular market data as it becomes available. However, some of this uncertainty may be unresolvable with available data, and therefore it may be appropriate to execute this analysis under multiple sets of assumptions reflecting various plausible trends in AI development.

Analysis

By integrating fixed costs, variable costs, and revenue projections, this methodology provides a structured approach to analyzing the future trajectory of AI model development from an economic perspective. The results of this analysis offer insights into the potential limits of profit-driven AI progress and the conditions under which alternative funding models may become necessary for continued progress in the field.

If properly executed, we expect that this analysis would allow for estimations of the "maximum" size of a training run that economically rational actors may undertake. Identifying the "maximum" training run will allow us to estimate to what point private, economically rational actors may drive AI development without "irrational" funding from ideological or state-based actors.

Because compute is, as of now, roughly correlated with model capabilities, identifying the "maximum" model economically rational actors may generate would allow for an estimate of what level of capability private corporations will develop models to before requiring external intervention. Model results suggesting that it would be economically rational to train very large and capable models would suggest that private companies would be significant actors in the continued development of AI without state intervention. This in turn implies that increasingly capable AI will emerge in private hands and be deployed in a decentralized fashion. This could potentially include the development of AGI, should it be shown that training a model with an amount of compute estimated to equate to AGI would be economically rational. Such results could indicate that AGI development would be undertaken by economically rational and decentralized actors without state support.

On the other hand, results showing that training increasingly capable models will quickly become economically irrational would suggest that "irrational" actors, such as states seeking to ensure their security or ideologically motivated private actors, would be required to develop increasingly capable AI and eventually AGI. This would suggest that AGI will be created by few, and potentially only one actor, and that its emergence will be centralized and highly controlled in comparison to alternative scenarios.

Appendix A. **Method for Estimating Future Model Training Costs**

This appendix outlines the methodology employed in forecasting future training costs of AI models. Our analysis hinges on two pivotal elements: first, calculating the requisite floating-point operations (FLOP) for the training of future models; second, forecasting the progression of the cost-effectiveness of computing hardware, quantified as FLOP per dollar (FLOP/$). Estimating these elements allows us to calculate the training costs of AI models under a range of hypothetical future scenarios.

Projecting Training FLOP of Future AI Models

This section outlines the methodology that projects the required floating-point operations (FLOP) for training future AI models, updating the analysis originally performed by Tamay Besiroglu, Lennart Heim, and Jaime Sevilla in their 2022 report, *Projecting Compute Trends in Machine Learning*[2]. Our approach builds upon their work, extrapolating historical trends in model training compute requirements under various assumptions – and utilizes Epoch's Database of notable machine learning systems.[3]

Rates of Model Scaling

Our projections are grounded in the assumption that the scaling hypothesis continues to hold – otherwise, training highly-capable models is not bottlenecked by capital requirements.

This scenario assumes ongoing rapid growth in the compute required for developing AI models, a trend that has been evident since the Deep Learning (DL) era began[4]. Historically, the usage of compute for training has doubled approximately every six months, a notable acceleration from the pre-DL era, where compute doubled roughly every 20 months, aligning with Moore's Law. This scenario envisions an eventual moderation of growth rates to align with Moore's Law, influenced by economic and technological constraints as suggested by Carey (2018)[5] and Lohn and Musser (2022)[6]. To model this transition, we employ three scenarios from Besiroglu et al., 2022 – Bearish, Middle of the Road, and Bullish – reflecting the potential duration that DL-era growth rates persist before reverting to Moore's Law.

Model

A Monte Carlo simulation – generating a broad spectrum of potential future outcomes – is utilized to address the inherent uncertainty in long-term forecasting. By sampling from 40 different growth rates per model and conducting 10,000 model runs, mirroring the approach used in Besiroglu et al., 2022, we achieve a statistically reliable spread of outcomes. For each model **j** and year **i**, the logarithm of compute (FLOP) required for training is given by:

$$\log C_{j,i} = \log C_{j,i-1} + \left(\frac{end - 2022}{timeline\_length - 1}\right) \cdot growth_i$$

Where $\log C_{j,i-1}$ represents the logarithm of compute (FLOP) required for training in the previous year, and $growth_i$ is a weighted growth rate based on the sampled reversion date from the distribution of reversion dates to Moore's Law. Additionally, *coef_pre_dl* represents the growth rate consistent with Moore's Law prior to the advent of deep learning, and *coef_post_dl* denotes the growth rate observed after the introduction of deep learning.

$$growth_j = (coef\_post\_dl^{\,weights[i]}) \cdot (coef\_pre\_dl^{\,1-weights[i]})$$

[2] Tamay Besiroglu, Lennart Heim, and Jaime Sevilla, *Projecting Compute Trends in Machine Learning*, Epoch, March 7, 2022.
[3] Epoch, "Epoch Database," webpage, undated.
[4] Jaime Sevilla, Lennart Heim, Anson Ho, Tamary Besirogly, Marius Hobbhahn, and Pablo Villalobos, "Compute Trends Across Three Eras of Machine Learning," *arXiv:2202.05924v2*, March 9, 2022.
[5] Ryan Carey, "Interpreting AI compute trends," AI Impacts, July 2018.
[6] Andrew Lohn and Micah Musser, "AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progreess," CSET, January 2022.

The weight for each year, $weights[i]$, is determined by a logistic-like function that smoothly transitions based on the reversion date relative to 2022. The term scaling $growth_i$ adjusts the growth rate to match the number of years covered in each step of the simulation.

Results

We assess the predicted FLOP required for training the largest foundation model each year under the continuation of the scaling hypothesis. While our projections focus on raw compute trends, it's important to acknowledge the potential influence of algorithmic progress on required training compute. Advances in algorithms could significantly increase the efficiency of compute usage – potentially requiring less FLOP for equivalent or superior outcomes compared to current models. However, quantifying the precise effect of these advancements is challenging and introduces additional uncertainties. Therefore, our current projections do not explicitly account for these algorithmic improvements. By applying the growth rates from the DL era and Pre-DL era to our Monte Carlo simulations and prior over reversion dates to Moore's Law, we provide updated projections that reflect potential future compute requirements for ML systems.

Projecting Price-Performance (FLOP/$) of Compute Hardware

This section outlines the methodology used for projecting the future price-performance of compute hardware, specifically focusing on floating-point operations per dollar (FLOP/$) for machine learning (ML) applications. Our approach utilizes a combination of regression analysis and bootstrapping techniques to estimate the annual change in FLOP/$ for different numeric precisions, particularly FP32 and FP16, due to greater data availability and relevance in the ML field. This method reproduces the work detailed in Epoch's *Trends in Machine Learning Hardware* report, which emphasizes the importance of precision-specific performance analysis in contemporary ML hardware evaluation[7].

Assumptions

Our estimation is predicated on several key assumptions:

- **Continued Trend of Numeric Precision**: The analysis primarily focuses on FP32 and FP16 precision in ML hardware, reflecting their relevance for contemporary ML applications[8]. This focus aligns with trends that favor lower precision, such as FP16, for efficiency gains in training models.
- **Data Reliability**: The data used for this analysis, sourced from the *Trends in Machine Learning Hardware* report, provides a comprehensive view of ML hardware performance from 2010 to 2023[9]. Price-performance is calculated using release prices or cloud service rates, adjusted for inflation and with assumed profit margins. Including FLOP/$ data for 40 ML accelerators, methods used for determining FLOP/$ are extensively described in Hobbhahn et al.[10]

Model

The methodology for projecting the future price-performance of machine learning hardware involves several stages. Initially, the data is prepared by converting all dates into numerical values that represent years since the dataset's start date, a step essential for facilitating time-based regression analysis. Next, regression analysis is performed for each numeric precision. This involves performing a linear regression of the logarithm of FLOP per dollar (log10(FLOP/$)) against the numeric date representation, aiming to estimate the rate of change in FLOP/$ over time. To enhance the robustness of these estimates and account for potential variability, a bootstrapping method is employed. This method entails resampling the dataset with replacement and recalculating the regression coefficients over 1,000 iterations, from which the 5th and 95th percentile estimates for FLOP/$ growth rates are derived to provide a confidence interval for the

[7] Marius Hobbhahn, Lennart Heim, and Gökçe Aydos, " Trends in Machine Learning Hardware," *Epoch*, November 9, 2023.
[8] NVIDIA, *Train with Mixed Precision: User's Guide*, DA-08617-001_v001, February 2023.
[9] Hobbhahn, Heim, and Aydos, 2023.
[10] Hobbhahn, Heim, and Aydos, 2023.

projections. Finally, using the obtained regression coefficients and the confidence intervals from bootstrapping, projections are made for the FLOP/$ for 2023 for both FP32 and FP16 precisions to serve as the base case upon which varying growth rates are extrapolated.

Estimating Training Costs of Future Foundation Models

Our methodology for estimating the training costs of future foundation models uses as inputs the two projections described above: the required FLOP for training these models and the price-performance (FLOP/$) of compute hardware. It is important to note that while our model separately estimates the required FLOP for training and the FLOP/$ of computing hardware, these variables are potentially correlated i.e., they are not strictly independent. This acknowledged limitation should be considered when evaluating the projections made by our model.

Results

The following table details the resulting projections – providing a breakdown of estimated training costs for future foundation models between 2024 and 2027.

Table A.1. Compute Training Costs of Future Foundation Models

Year	Projected Growth Rate In FLOP/$	
	Median (0.14 OOM/Year) *Compute trends persist*	**5th Percentile (0.04 OOM/Year)** *Stagnant cost improvement*
2024	$337,222,427	$424,537,883
2025	$911,220,979	$1,444,187,927
2026	$1,916,747,917	$3,824,414,886
2027	$6,233,530,230	$15,657,920,006

Note: Estimates assume use of FP16 in training.

Appendix B. **Method for Estimating Operational Cost and Model Revenue**

Projecting iFLOP of Future Foundation Models

The estimation of iFLOP for future AI models builds upon the current state-of-the-art in large language models (LLMs), which predominantly utilize decoder-only Transformer architectures. Our methodology assumes a baseline computational cost of approximately 2N FLOP per token during inference, where N represents the model's parameter count (see Kaplan et al. 2020).

$$iFLOP_{Size} = \text{Size} \times 2 \text{ FLOP per token per model parameter}$$

We propose using data compiled on Notable ML Systems by Epoch AI – and presented in Sevilla et al. (2021) – to facilitate projections of future model sizes.

Estimating Token Generation for Deployed Models

To estimate $N(t)$, the number of model predictions generated per period *t* of model deployment, we can consider two key factors:

1. **Increase in Model Usage at Release Over Time:** The initial usage of new models ($N_{initial\,|\,year}$) increases over time due to factors such as growing popularity, increased usage, and deployment of more advanced models. We assume $N_{initial\,|\,year}$ follows an exponential growth pattern due to increasing model adoption and usage. This can be modeled as:

$$N_{\,initial\,|\,year}\,(g) = N_{\,initial}(2024) \times (1+g)^{year-2024}$$

Where $N_{\,initial}(2024)$ is the initial number of tokens generated per period t at the starting year (2024), and g is the annual growth rate. To estimate $N_{\,initial}(2024)$, we could use the calculation found in Alan Thompson's GPT-3.5 + ChatGPT: An Illustrated Overview.

2. **Decay Rate**: For each model, the usage decreases over time as the model matures, users' novelty interest wanes, or more efficient models are introduced. We define the rate at which model usage decreases as d, where $0 < d < 1$. For instance, if $d = 0.95$, then each month the token generation reduces by 5%.

Together, the formula to estimate the number of predictions generated in a *period t* is:

$$N(t) = N_{\,initial\,|\,year}\,(g) \times d^{t}$$

Using this methodology, we estimate the number of predictions per period t of model deployment – considering both increasing initial model use over time and decreasing use over a given model's lifecycle.