The Price of Progress

Algorithmic Efficiency and the Falling Cost of AI Inference

Hans Gundlach[*]
MIT CSAIL, MIT FutureTech
hansgund@mit.edu

Jayson Lynch
MIT CSAIL, MIT FutureTech
jaysonl@mit.edu

Matthias Mertens
MIT Sloan, MIT FutureTech
mmertens@mit.edu

Neil Thompson[*]
MIT CSAIL, MIT FutureTech
neil_t@mit.edu

Abstract

Language models have seen enormous progress on advanced benchmarks in recent years, but much of this progress has only been possible by using more costly models. Benchmarks may therefore present a warped picture of progress in practical capabilities *per dollar*. To remedy this, we use data from Artificial Analysis and Epoch AI to form the largest dataset of current and historical prices to run benchmarks to date. We find that the price for a given level of benchmark performance has decreased remarkably fast, around $5\times$ to $10\times$ per year, for frontier models on knowledge, reasoning, math, and software engineering benchmarks. These reductions in the cost of AI inference are due to economic forces, hardware efficiency improvements, and algorithmic efficiency improvements. Isolating out open models to control for competition effects and dividing by hardware price declines, we estimate that algorithmic efficiency progress is around $3\times$ per year. Finally, we recommend that evaluators both publicize and take into account the price of benchmarking as an essential part of measuring the real-world impact of AI. [2]

1 Introduction

A critical dimension of the real-world impact of language models (and AI systems in general) is cost, which is often ignored in discourses around evaluations and AI performance. Popular blog posts have ignited discussions on potentially large declines in prices for accessing a given level of (high) LLM performance [Appenzeller, 2024]. At the same time, Cottier et al. [2025] finds that, controlling for benchmark performance, LLM token prices may be decreasing by factors of 10–1,000× per year, depending on the performance level. On the other hand, Erol et al. [2025] finds that the cost-of-pass on benchmarks like MATH 500 and AIME 2024 has gone down by 24.5× and 3.23× per year, respectively. Understanding these price trends is key for many issues, such as predicting the cost efficiency of models versus labor-based work, or democratizing access to state-of-the-art AI capabilities.[3]

In this study, we carefully examine how the price to run LLMs has changed for a given LLM performance level. We use data from the Epoch AI Benchmark Hub [Epoch AI, 2024] and collect

[*] Corresponding authors. hansgund@mit.edu, neil_t@mit.edu

[2] This paper was accepted to the NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling (https://sites.google.com/view/llm-eval-workshop).

[3] Documenting changes in quality-adjusted AI prices is also relevant to economic science, as these data can be helpful for understanding substitution elasticities between labor and AI inputs.

historical snapshots of Artificial Analysis prices using the Internet Archive. Combining these data, we construct the largest dataset of AI benchmarking prices we know of to date (linked in Appendix A). This dataset encompasses knowledge and reasoning benchmarks like GPQA-Diamond (GPQA-D), mathematics benchmarks like OTIS Mock AIME 2024–2025 (AIME), and software engineering benchmarks like SWE-bench Verified (SWE-V).

Using our data, we estimate regression models that recover an exponentially decreasing price trend for a given LLM performance level. This quantity serves as an important proxy for changing AI inference efficiency. We employ two approaches. In the first, we run a pooled regression of model prices on an exponential trend while controlling for model performance with a regression term. In the second, we bin the data and run regressions separately for models with similar benchmark performance levels. Both approaches yield qualitatively consistent results.

We find that overall progress is closer to 10× than 1,000×. Our estimates are in general lower than those of Cottier et al. [2025] and closer to the AIME 2024 cost-of-pass estimates from Erol et al. [2025]. We attribute our lower estimates to three main factors. First, we use a benchmark- rather than token-level approach, which compensates for the increased number of tokens used in recent models. This is particularly important for reasoning models, which can have a lower per-token cost but a larger overall cost due to generating more tokens. Second, we examine the period from April 2024 to November 2025, where we have high-quality price data. Cottier et al. [2025] and related work examine price trends since April 2023 and late 2022, respectively. Therefore, our data do not capture possible price changes during the very initial stages of a new technology. Finally, we include an almost an order of magnitude more models per benchmark in our fit than Cottier et al. [2025].

In contrast to previous estimates of inference efficiency progress, we measure both the overall trend and separate trends for open-weight models. Since open-weight models can, in principle, be run by anyone, they are subject to greater market competition than closed-weight models. As a result, we argue that reductions in open-weight model inference costs are more likely to reflect fundamental technical innovation rather than purely economic changes. Using the open-weight model trend and dividing by the hardware price progress, we arrive at an estimate for algorithmic progress in AI inference. This measures the reduction in computational operations needed for a given problem over time [Ho et al., 2024] and is an important factor in models of AI capability growth [AI Futures Project, 2025]. In particular, if we cannot grow computational resources indefinitely, then inference efficiency gains represent the limit for the growth of AI labor and the democratization of AI.

After establishing our core findings on declining LLM prices, we use our dataset of benchmark prices to additionally examine trends in the price of running benchmarks. It has been reported that current costs for state-of-the-art benchmark evaluation can reach thousands of dollars. For instance, OpenAI was able to attain breakthrough performance on the ARC-AGI benchmark; however, achieving this score reportedly cost 3,000 dollars just to run the model [Ord, 2025]. Despite the trends we document in price-performance, the cost of running benchmarks in our dataset has remained flat or increased, often to unexpectedly high levels. This suggests that declining per-unit price-performance has been offset—and in some cases overcompensated—by the larger models and greater reasoning demand required to reach higher performance levels.

Finally, we argue that evaluations should be more transparent about computational resource usage. Without this data, it is difficult to identify which AI developments are driving meaningful performance improvements and which are simply leveraging more computing resources.

2 How Fast Is Benchmark price-performance Improving?

To measure the trend in benchmark price-performance controlling for quality, we run a variety of regressions on a dataset of benchmarking costs. We first describe our data collection before we turn to our regression results.

2.1 Data

We collect data on input and output token prices over time by gathering Internet Archive data from Artificial Analysis. This data ranges from April 2024 to October 2025 and contains token-level price data from proprietary and open-source model inference providers (OpenAI, Deepinfra, Cerebras, etc). We collect exclusively the lowest input and output token prices (the full price data across all providers is not available on the Internet Archive). We focus on the cheapest available provider to better control for the price/latency tradeoff (see Erdil [2025]). In some cases, the lowest input and output prices were offered by different providers. However, in the vast majority of cases, the lowest input and output price are offered by the same provider, and when they are not, the resulting costs differences were negligible. Interestingly, in our dataset, we sometimes observe price increases for some models. This is generally due to cheap platforms no longer supporting legacy models. Since we believe that these kinds of price increases do not represent price-performance decreases, we do not include them in our analysis. In addition, we remove datapoints with 0 token cost as these are not reflective of actual inference prices.

We supplement our dataset with data from Epoch AI's benchmarking hub [Epoch AI, 2024]. This data includes models' benchmark performance as well as the number of inputs, outputs, reasoning, and cached tokens used to run the benchmark. We compute the benchmark price by multiplying the corresponding tokens by their corresponding prices sourced from Artificial Analysis.

Prices of running a benchmark may change over time and we treat any such changes as separate data points. As a result, there may be many points for a given model if its price changes frequently. To illustrate: for GPQA-Diamond, which is our largest benchmark sample, we have 138 price data points with 93 unique models; our smallest sample is SWE-bench Verified, which has 21 data points with 19 unique models. Appendix B provides more details on our data collection procedure.

2.2 General Regression Approach

Regression specification. Our first approach involves running the following regression:

$$\log(\text{Benchmark Price}_{it}) = \beta_0 + \beta_1 \text{logit}(\text{Performance}_i) + \beta_2 t + \epsilon_{it}, \tag{1}$$

where i indexes models, Performance_i measures benchmark performance of model i, $\text{Benchmark Price}_{it}$ is the cost of running model i on the chosen benchmark, t is time, and ϵ_{it} is an i.i.d. error term. The coefficient of interest in Eq. (1) is β_2, which measures the rate of log price changes, conditional on benchmark performance. We run separate regressions for each benchmark individually, where Performance_i measures GPQA-D, SWE-V, or AIME benchmark scores. Note that we apply a logit-transformation to our performance measure, that is $\text{logit}(Y) = \ln\left(\frac{Y}{1-Y}\right)$, where Y is a reported benchmark score. We choose this transformation for several reasons. First, benchmark performance is bounded between 0 and 1 and increases with the logarithm of inference compute in the non-saturated region [Zhang and Chen, 2024]. Second, Owen [2024] and Ruan et al. [2024] found that benchmark performance increases logistically with training compute. And since parameters are roughly proportional to inference compute and the square root of training compute [Villalobos and Atkinson, 2023], we infer that benchmark performance is logistic in log inference compute/price.

Regression samples. In our main specification, we estimate the regression on the Pareto frontier. Specifically, we filter out models that are Pareto dominated by earlier models that have a better benchmark score and a lower price to run a given benchmark. This yields a more relevant sample, as economically-optimizing users will typically choose the cheapest model for a given performance level. In additional analyses, we also run the regression for all available models and compare the results.

Finally, we also run separate regressions for open-weight models. Open-weight models can, in principle, be run and modified by anyone. We therefore expect that they will not be priced at a significant markup relative to the necessary GPU resources to run them. As a result, we argue that the trend in open-weight models more accurately measures technical progress rather than economic effects such as increased market pressure, and that the difference between our results for open-weight and proprietary models is (to some extent) informative about non-technological drivers of price changes, such as competition.

One potential concern is that open-weight models might not lie on the overall technical frontier. However, their performance appears to parallel frontier models with a lag of a few months [Denain,

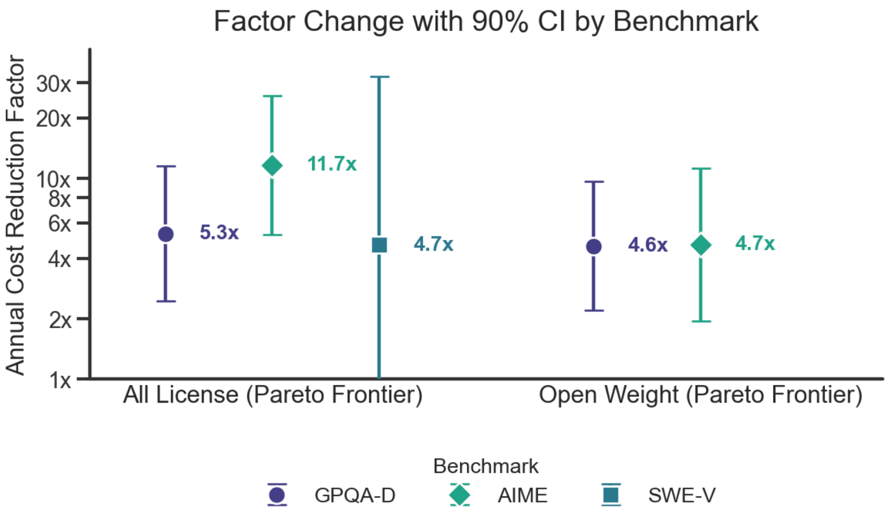

Figure 1: Annual factor change in price controlling for performance. We restrict our analysis to 2024-2025 models on the accuracy price pareto-frontier, and we present separate analyses for all models and only open weight models. Note, we do not have enough open-source data on SWE-V to report it 4. See Table 1 and Table 2 for more information.

2024], so the general technical trend should be approximately similar.[4] The results of our different fits are shown in Fig. 1.

2.3 Main Results: Frontier Models Are Getting 5-10 Times Cheaper Each Year

Results for our main specification (Pareto-frontier models) are shown in Fig 1. We focus first on our main sample. Overall, GPQA-Diamond and MOCK AIME benchmark price-performance has increased dramatically with prices reducing annually by a factor of $5 - 10\times$ for models on the cost-performance Pareto-frontier. This is similar to trends in the cost-of-pass for some math benchmarks in Erol et al. [2025].[5] For SWE-bench Verified, we have less data, and our estimates feature much larger confidence bands. Nonetheless, the average reduction rate looks similar to our estimates for GPQA-D and AIME.

Tables 1 and 2 report results for all models in our dataset (i.e., including models not on the Pareto frontier). The price-performance decreases are smaller by almost a factor of 2. We think that this likely reflects that our Pareto frontier is defined as optimizing a particular set of benchmarks for cost, but in the real world models are optimized for many different goals (latency, performance for other capabilities, etc.). As such, we do not assume that overall progress is necessarily slower, just that other model builders are optimizing for other goals.

To get a better sense of the role of algorithmic progress (i.e., technical non-hardware progress) [Ho et al., 2024], we estimate β_2 (the price decline) for open-weight models only. Additionally, we divide the estimated price trends by the decline in hardware prices as reported in Rahman [2024]. After these adjustments, the remaining price reduction factor is around **$3\times$ per year**, which can be interpreted as the contribution of algorithmic progress to declining price. Our estimates for algorithmic progress are more similar to experimental measures of energy efficiency gains like [Saad-Falcon et al., 2025] which find $3.1\times$ gains from 2023-2025 controlling hardware. See Table 2 for more details on our measurement.

[4] Almost all the models tested by Epoch AI on SWE-bench Verified are closed-source, so we cannot measure the trend in open-source price-performance for SWE-bench Verified.

[5] Our findings are also consistent with evidence in Fischl-Lanzoni et al. [2025] who document how algorithmic progress helps to make models of a given quality level smaller but without studying the resulting price implications.

2.4 Results by Performance Bins

In addition to the pooled regression approach above, we also bin models by their benchmark score and estimate Eq. (1) within each bin. Specifically, for each bin we identify, at each point in time, the lowest-priced models, thereby constructing the frontier of capability progress at a given quality level. We then run the regression on the subsample of these frontier models. Fig. 2 depicts the resulting binned trends for all models, while Fig. 3 shows the corresponding results for open-weight models.

Similar to Cottier et al. [2025], we observe faster trends at the higher-quality frontier. For instance, in Fig. 2, models in the highest GPQA-D bin declined in price by $31\times$ per year, whereas models in the lowest bin declined by only $1.7\times$ per year. This pattern may indicate that different economic or technical factors drive cost reductions in higher-capability models (e.g., greater use of distillation or Mixture-of-Experts (MoE) architectures).

In addition, we find that the closed-weight model trend is slightly faster than the open-weight model trend. This is particularly pronounced for closed-weight models in the 40%–60% group, where we see a sudden drop in price that is not mirrored in open-weight models, hinting at non-technical competitive effects.

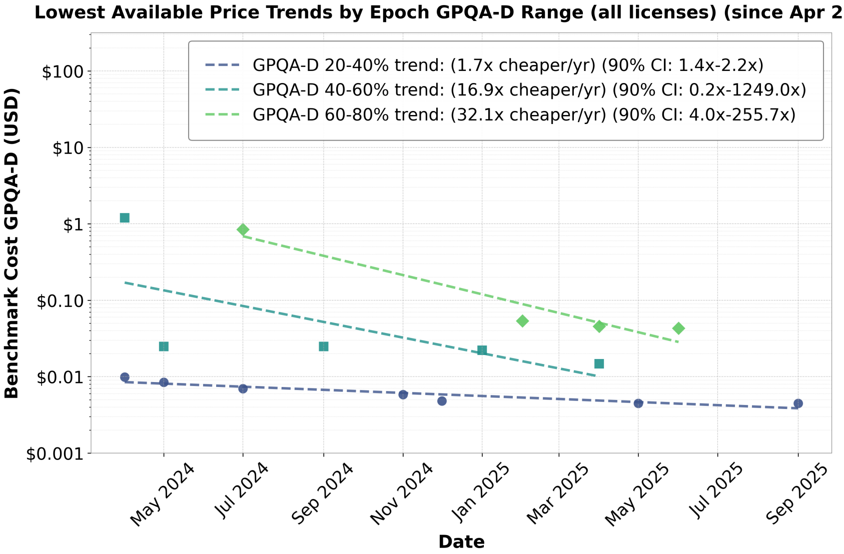

Figure 2: Graph of benchmark price vs time for all models within a fixed GPQA-Diamond range. We don't have a good fit for models in the $40\% - 60\%$ range, but include it here for consistency. We suspect that the large drop in overall price in this range is due to increased market competition.

3 Inference Costs Are Falling but Benchmarking Costs are Increasing

We close our paper by extending our analysis to benchmark costs. Benchmarking is a fundamental part of AI research and development. Large benchmarking costs pose a challenge to maintaining a healthy evaluation ecosystem. We take the previous price estimates of benchmarking and look at the overall trend regardless of performance. We find that, despite the dramatic fall in the price for a given performance level previously discussed, benchmarking costs have stayed constant or increased (see Figs. 4 and 5). We believe that this reflects the demand for much higher quality models, which are much larger and use much longer reasoning traces.

For SWE-bench Verified, average evaluation prices have remained roughly constant. However, beneath the stable overall trend lies considerable model-specific variation in evaluation prices: The cost of running SWE-bench Verified on some models has risen to thousands of dollars. Overall, increasing evaluation prices can be taxing for small academic research groups, and this problem is compounded if many different models must be run or if multiple benchmarks are involved in a study. Our analysis does not take into account new, longer contexts and agential benchmarks, which are becoming prohibitively expensive. For instance, ∞BENCH spent 5,000 dollars evaluating

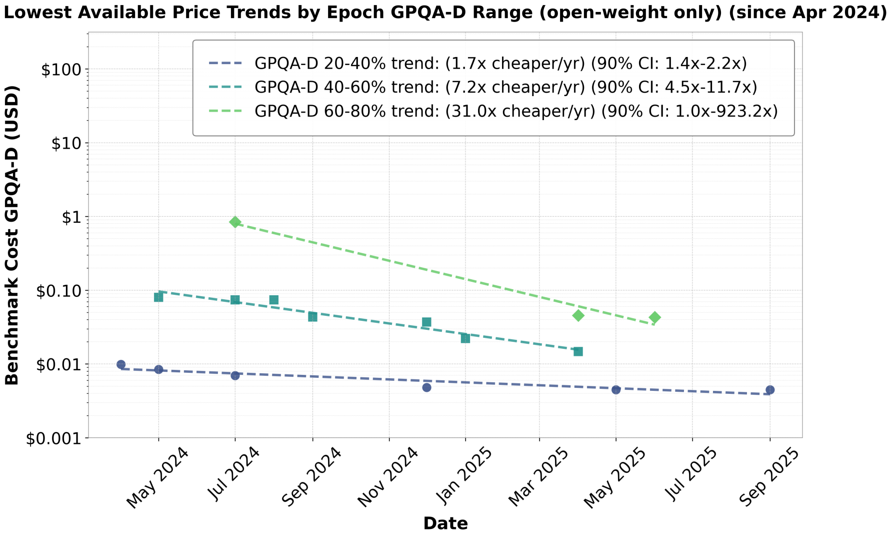

Figure 3: Graph of benchmark price vs time for open weight models within a fixed GPQA-Diamond range. The price for higher quality models is decreasing faster than for lower quality models.

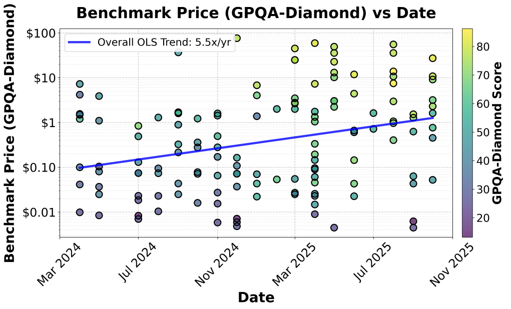

Figure 4: Price to run GPQA-Diamond benchmark. Prices based on Epoch-AI benchmark data and Artificial Analysis Prices. Overall, benchmark prices in our dataset have increased despite a dramatic fall in model price-performance.

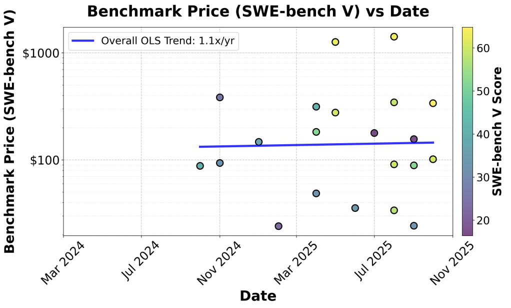

Figure 5: Price to run SWE-bench Verified. Prices based on Epoch-AI benchmark data and Artificial Analysis Prices. Similar to Fig 4, benchmark prices have increased. In addition, the price to run SWE-bench Verified for some models is now in the thousands of dollars.

long-context abilities on GPT-4 [Zhang et al., 2024]. Such costs make independent and academic benchmarking prohibitive and have led to the rise of private firms like Artificial Analysis, which have the capital necessary to do cutting-edge evaluations on many models. These companies are valuable resources but only share some information with the public and keep other information, like benchmark costs and historical prices, either hidden or not easily accessible.

4 Conclusions and Recommendations for Future Evaluations

At first glance, our analysis seems to point in different directions. The overall price for accessing a given level of LLM performance has dropped significantly, by 5× to 10× per year, although still substantially less than the reported 1000× upper bound by Cottier et al. [2025]. However, like Cottier et al. [2025], we find much larger price declines for higher-performance models—almost 32× per year. For the least performant models, by contrast, we see much smaller price declines, around 1.7× per year, close to the estimates for energy efficiency improvements in AI models overall [Saad-Falcon et al., 2025]. In terms of benchmarks, progress looks similar across GPQA-D and AIME, but the data for SWE-V is so limited that our confidence bounds are large enough to be consistent with there having been no progress at all.

Despite these large price decreases, the benchmarking costs of models in our dataset have remained constant or increased as the demand for quality has risen, leading to larger models and longer reasoning.

Taken together, these findings suggest that focusing on benchmark scores alone does not provide a full view of the nuanced nature of AI progress. A model that substantially improves benchmark performance but only by consuming dramatically more computational resources represents a smaller advance in practical capability than headline scores suggest. Hence, we must take into account price if we want to understand progress.

We therefore call for evaluators and researchers to publish price and resource data on evaluations more widely, so that the evaluation community can examine capability progress in light of real-world economic constraints. Given the limited data available, it is hard to know what real-world progress has been made in SWE-bench or in any benchmark. Has math price-performance increased? Have agentic tool-using agents become more efficient? Are we closer to an AI software developer that can efficiently replace humans? Without data on the resources used, evaluations provide a much less clear picture of current practical capabilities and the future of AI.

Acknowledgements

We thank Zachary Brown for insightful comments on this manuscript.

References

Guido Appenzeller. Welcome to LLMflation — LLM inference cost is going down fast. https://a16z.com/llmflation-llm-inference-cost/, November 2024. a16z (Andreessen Horowitz) blog post.

Ben Cottier, Ben Snodin, David Owen, and Tom Adamczewski. LLM inference prices have fallen rapidly but unequally across tasks, 2025. URL https://epoch.ai/data-insights/llm-inference-price-trends. Accessed: 2025-09-03.

Mehmet Hamza Erol, Batu El, Mirac Suzgun, Mert Yuksekgonul, and James Zou. Cost-of-pass: An economic framework for evaluating language models. *arXiv preprint arXiv:2504.13359*, 2025.

Epoch AI. "AI Benchmarking Hub", 11 2024. URL https://epoch.ai/benchmarks. Accessed: 2025-09-02.

Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan C Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. Algorithmic progress in language models. *Advances in Neural Information Processing Systems*, 37:58245–58283, 2024.

AI Futures Project. About — AI 2027, August 2025. URL https://ai-2027.com/about. Accessed 2025-09-03.

Toby Ord. Inference scaling and the log-x chart. https://www.tobyord.com/writing/inference-scaling-and-the-log-x-chart, January 2025. Accessed: 2025-09-01.

Ege Erdil. Inference economics of language models, 2025. URL https://arxiv.org/abs/2506.04645.

Hugh Zhang and Celia Chen. Test-Time Compute Scaling Laws. https://github.com/hughbzhang/o1_inference_scaling_laws, 2024.

David Owen. How predictable is language model benchmark performance? *arXiv preprint arXiv:2401.04757*, 2024.

Yangjun Ruan, Chris J Maddison, and Tatsunori B Hashimoto. Observational scaling laws and the predictability of langauge model performance. *Advances in Neural Information Processing Systems*, 37:15841–15892, 2024.

Pablo Villalobos and David Atkinson. Trading Off Compute in Training and Inference, 2023. URL https://epoch.ai/blog/trading-off-compute-in-training-and-inference. Accessed: 2025-09-02.

Jean-Stanislas Denain. Models with downloadable weights currently lag behind the top-performing models, 2024. URL https://epoch.ai/data-insights/open-vs-closed-model-performance. Accessed: 2025-09-04.

Natalia Fischl-Lanzoni, Matthias Mertens, and Neil Thompson. Is There a Secret Sauce in Large Language Model Development? *Mimeo*, 2025.

Robi Rahman. Performance per dollar improves around 30 URL https://epoch.ai/data-insights/price-performance-hardware. Accessed: 2025-11-07.

Jon Saad-Falcon, Avanika Narayan, Hakki Orhun Akengin, J Griffin, Herumb Shandilya, Adrian Gamarra Lafuente, Medhya Goel, Rebecca Joseph, Shlok Natarajan, Etash Kumar Guha, et al. Intelligence per watt: Measuring intelligence efficiency of local ai. *arXiv preprint arXiv:2511.07885*, 2025.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. Infinity Bench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*, 2024.

Andrey Fradkin. Demand for LLMs: Descriptive Evidence on Substitution, Market Expansion, and Multihoming. *arXiv preprint arXiv:2504.15440*, 2025.

A Data

Our dataset of benchmark prices, along with the code used for analysis, is available here: The Price of Progress Project: https://github.com/hansgundlach/Algorithmic_Progress_Inference

B Details on Dataset Selection, and Preprocessing

Some models have input and output costs of 0 dollars on Artificial Analysis. We do not include these models in our dataset. We believe these are generally company promotional offers. The Internet Archive has only limited data on some models. We collect all information accessible (many Internet Archive pages failed to load). However, if data is logged 6 months apart this could lead to irregular progress estimates.

Epoch benchmarks also include cached tokens. We do not include these for our GPQA-Diamond benchmark cost estimates as the number of cached tokens is generally 20x smaller (as well as around 10x cheaper) than either input or output tokens, and artificial analysis does not have cache token prices. However, for SWE-bench Verified, cached tokens constitute a significant portion of the cost. Therefore, we use proprietary vendors' current cache tokens prices. Vendors generally have a variety of cache token prices for Anthropic models—we use 5m cache write prices. For Deepseek models, we use cache token prices from the official API. Additionally, we do not have historical data on cache token prices. However, proprietary models generally do not change either input or output tokens for a given model over time. For instance, we find only one instance of this in our dataset. This is also mentioned by Fradkin [2025]. Sometimes, Epoch benchmark reports were underspecified, i.e., did not mention which version of a model was used in these cases—we did not include this data. In general, we did not include data in our analysis where we could not match the Epoch model benchmark card with the model name on Artificial Analysis.

Finally, Epoch includes multiple versions of some models, in particular Claude 3.7 with different reasoning levels. We include these as separate models in our estimates.

For many models Epoch uses multiple runs (8-16) of a given benchmark i.e runs the benchmark multiple times. We are only able to see the total token numbers so we normalize all benchmarks to 1 iteration by dividing tokens by number of runs.

C Data Tables

Table 1: Rate of price change across several different benchmarks using general regression approach. Regressions include either all models or only the models that improve in accuracy or price (Pareto Restricted). A separate analysis of only open weight models was possible with GPQA-Diamond (GPQA-D) and OTIS-MOCK AIME 2024-2025 (AIME). Decrease factors < 1 represent increases.

Benchmark	Restriction	Annual Reduction Factor	90% CI	n	R^2
GPQA-D	Pareto Restricted All License	5.315	[2.449, 11.534]	53	0.8304
	Pareto Restricted Open Weight	4.602	[2.195, 9.648]	35	0.7272
	All License (no restriction)	3.769	[2.085, 6.814]	135	0.6571
	Open Weight (no restriction)	1.214	[0.607, 2.426]	72	0.5166
AIME	Pareto Restricted All License	11.664	[5.250, 25.911]	42	0.7846
	Pareto Restricted Open Weight	4.680	[1.943, 11.273]	30	0.7539
	All License (no restriction)	6.988	[3.687, 13.243]	109	0.6287
	Open Weight (no restriction)	2.661	[1.373, 5.157]	55	0.7087
SWE-V	Pareto Restricted All License	4.675	[0.680, 32.156]	13	0.6281
	Pareto Restricted Open Weight	—	[—, —]	—	—
	All License (no restriction)	1.603	[0.373, 6.881]	21	0.1628
	Open Weight (no restriction)	—	[—, —]	—	—

C.1 Adjusting For GPU price-performance Gains

Trends in benchmark price-performance are also influenced by hardware performance trends. If we want to isolate the component due purely to algorithmic advances, we have to divide the annual factor decrease by the annual hardware price efficiency gain. Here we use our general regression approach and estimates from Rahman [2024], which finds that for a fixed performance level costs have dropped by 30% a year.

Table 2: Annual reduction factor (hardware-adjusted) and 90% CI (hardware-adjusted).

Benchmark	Restriction	Annual reduction factor (hardware-adjusted)	90% CI (hardware-adjusted)	n	R^2
GPQA-D	Pareto Restricted All License	3.720	[1.714, 8.074]	53	0.8304
	Pareto Restricted Open Weight	3.221	[1.536, 6.754]	35	0.7272
	All License (no restriction)	2.638	[1.459, 4.770]	135	0.6571
	Open Weight (no restriction)	0.850	[0.425, 1.698]	72	0.5166
AIME	Pareto Restricted All License	8.165	[3.675, 18.138]	42	0.7846
	Pareto Restricted Open Weight	3.276	[1.360, 7.891]	30	0.7539
	All License (no restriction)	4.891	[2.581, 9.270]	109	0.6287
	Open Weight (no restriction)	1.863	[0.961, 3.610]	55	0.7087
SWE-V	Pareto Restricted All License	3.273	[0.476, 22.509]	13	0.6281
	Pareto Restricted Open Weight	—	[—, —]	—	—
	All License (no restriction)	1.122	[0.261, 4.817]	21	0.1628
	Open Weight (no restriction)	—	[—, —]	—	—