The Price of Progress

Anonymous Author(s)
Affiliation
Address
email

Abstract

Language models have seen enormous progress on advanced benchmarks in recent years. However, high performance on these benchmarks requires exceedingly large computational resources. Therefore, it is difficult to get an accurate picture of the progress of practical capabilities. Here, we try to address this issue by measuring trends in benchmark price-performance. We find that the price for a given level of performance has decreased considerably for GPQA-Diamond, while our analysis of SWE-Bench Verified remains uncertain. In the process, we collect a large public dataset of benchmark prices over time. We use this data to look at trends in the price of benchmarking, which, despite trends in price-performance, have remained flat or increased, often to unexpectedly high levels. Finally, we argue that focusing on benchmark scores alone, in disregard of resource constraints, has led to a warped view of progress. Hence, 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.

1 Introduction

A critical dimension of the real-world impact of AI systems is cost, which is often ignored in discourse around evaluations. Current costs for state-of-the-art benchmark evaluation can be 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 [Ord, 2025]. At the same time, articles acclaim the huge decrease in prices for state-of-the-art LLMs [Appenzeller, 2024]. For instance, Cottier et al. [2025] finds LLM token prices controlling for task performance may be decreasing by factors of 10-1,000x per year. Here we examine each of these developments and their effect on measuring progress in AI capabilities. First, we measure benchmark progress, controlling for cost. This number serves as an important proxy of AI inference capabilities, in particular, algorithmic progress in inference, which is an important factor in models of AI capability growth [AI Futures Project, 2025]. This number also helps us predict when the currently expensive benchmark capabilities will be much more widely distributed. Unlike earlier estimates, we take into account the number of tokens as well as token costs. This adjustment is particularly important for for reasoning models, which can have lower per-token cost but a larger cost overall due to much greater token generation. We try to examine price performance trend across other benchmarks like SWE-bench Verified, but our estimates of progress in these domains remain uncertain. We use data from the Epoch AI Benchmark hub [Epoch AI, 2024] and collect historical snapshots of Artificial Analysis prices. Combining this data, we construct one of the most comprehensive datasets to date on AI benchmarking costs (linked in Appendix A).

Using our dataset of benchmark prices, we then examine trends in the price of benchmarking. Finally, we argue for more resource data transparency in evaluations. Without this data, it is hard to identify what AI developments are driving forward practical performance and what developments are simply using more compute resources.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 How Fast Is Benchmark Price Performance Improving?

As mentioned earlier, improvements in price performance are an important factor in measuring AI progress. Our initial results point to steep increases in efficiency relative to GPQA-Diamond. To measure this trend, 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 September 2025 and contains token level price data from proprietary and open-source mode 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 chose the cheapest available provider to better control for the price/latency tradeoff (see Erdil [2025]). Nevertheless, sometimes the lowest input prices are on different providers from the lowest output prices. However, in the vast majority of cases they are on the same provider or they are very comparable to the price that would exist if we had chosen the input and output costs on one provider. Second, we collect benchmark data from Epoch AI's benchmarking hub [Epoch AI, 2024]. This data includes the model's performance on the benchmark as well as the number of inputs and outputs, and cached tokens used. We then estimate the benchmark price by multiplying the input and output tokens by the lowest input and output token prices. If the price of running the benchmark changes in our dataset, we include this as a separate point in our dataset. Therefore, there may be many points for a given model if its price changes frequently. In our largest GPQA-Diamond dataset, we have 95 price datapoints with 61 unique models. While our SWE-bench Verified dataset has 13 datapoints with 12 unique models. For more information on our data collection procedure, see Appendix B.

2.1 Regression Approach

We use a fit as described in Eq. (1). i indexes models, GPQA-Diamond measures benchmark performance of model i (we also model SWE-bench Verified performance), and t is a linear time trend. ϵ_i is an i.i.d. error term. The coefficient of interest is β_2, which measures the rate of log price changes, conditional on benchmark performance. The results of this regression are given in Table 1. We chose this fit based on research showing benchmark performance increasing with the logarithm of compute [Zhang and Chen, 2024]. Owen [2024] also showed benchmark performance increasing with the logarithm of training compute. 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 linear in log inference compute/price when non-saturated. We filter GPQA-Diamond to have scores from 25 to 85 percent to identify this non-saturated region (Random guessing on GPQA-Diamond would yield 25 percent). We filter SWE-bench Verified for values above 2 percent. In addition, we run a regression on models filtered to only be on the Pareto-frontier. That is, only fitting models that were either more accurate or less expensive than previous models. We find that the trend in benchmark efficiency on the Pareto-frontier are much larger than the fit for models overall. Interestingly, we find that prices for some models have increased. This is generally due to cheap platforms no longer supporting legacy models. Since, we believe these kinds of prices increases don't represent price performance decreases, we don't include these in our analysis. In addition, we perform our regression on only open-source models. Open source models can theoretically be run by anyone, and therefore, we expect that they will not be priced at a large markup relative to the necessary GPU resources to run them. Hence, we believe that the trend in open source models more accurately measures technical progress rather than economic effects like increased market pressure. Open-source models also might not be on the technical frontier. However, we think their performance seems to parallel frontier models with a few month lag [Denain, 2024]. Almost all the models tested by Epoch on SWE-bench Verified are closed source, so measuring the extent of technical progress in this domain is much more challenging. The results of our different fits are shown in Table 1.

Overall, we find GPQA-Diamond benchmark price-performance has increased dramatically with estimates in most groups in the range of 10x. However, for SWE-bench Verified, available data is mostly inconclusive and could be consistent with both increasing and decreasing price performance.

$$\log(\text{Benchmark Price}_i) = \beta_0 + \beta_1 \text{GPQA-D}_i + \beta_2 t + \epsilon_i \tag{1}$$

Table 1: Rate of price change across several different benchmarks. Regressions include either all models or only the models that improve in accuracy or price (Pareto Restricted). A separate analysis of only open source models was possible with GPQA-Diamond (GPQA-D). Decrease factors < 1 represent increases.

Grouping	**Year Decrease Factor**	**90% CI**	**n**	R^2
Open Source Models GPQA-D	2.552	[1.210, 5.383]	57	0.4754
Open Source Models GPQA-D (Pareto Restricted)	9.776	[4.632, 21.914]	29	0.6566
All License GPQA-D	8.018	[3.757, 16.837]	91	0.443
All License GPQA-D (Pareto Restricted)	20.187	[5.962, 68.360]	34	0.6075
All License SWE-bench V	0.6807	[0.071, 6.531]	13	0.1865
All License (Pareto Restricted) SWE-bench V	.854	[0.019, 38.02]	7	0.6784

2.2 Binning Method

In addition to the full regression approach above, we can also bin models by their benchmark score and regress on each bin (Figure 1). For each bin, we find the record-setting lowest-priced models over time to get a sense of frontier capability progress. We notice, similar to Cottier et al. [2025], that there are slightly faster trends at the higher quality frontier. This may imply that there could be different economic or technical factors driving cost reduction in higher capability models. This method lends itself well to a simple graphical depiction. Yet, some bins are clearly non-linear, leading to highly unstable fits. However, the $(20\% - 40\%)$ and $(60\% - 80\%)$ are broadly consistent with the results in our regression study.

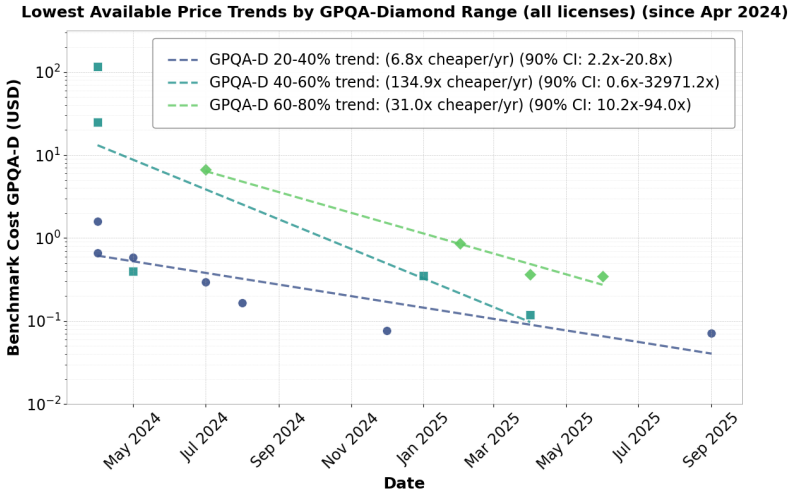

Figure 1: Graph of benchmark price vs time for models within a fixed GPQA-Diamond range. We don't have a good fit for models in the $40\% - 60\%$ range but include it here for consistency.

3 Trends in the price of Evaluations

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 price for a given level of performance outlined in the previous section, overall benchmark prices have stayed constant or increased at a moderate rate (see Figs. 2 and 3). We see less dramatic but similar trends in SWE-bench Verified. Alarmingly, the cost of running SWE-bench Verified on some models has risen to thousands of dollars. This price can be taxing for small academic

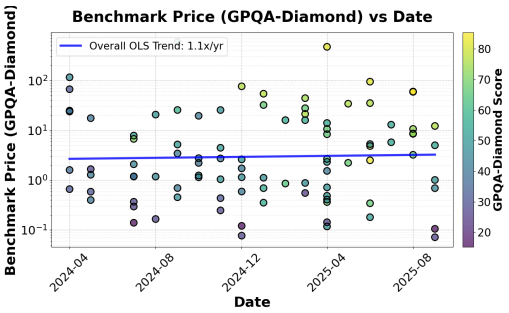

Figure 2: Graph of price to run GPQA-Diamond benchmark over time. Prices based on Epoch-AI benchmark data and Artificial Analysis Prices. Overall, benchmark prices have stayed constant despite falls in model price performance.

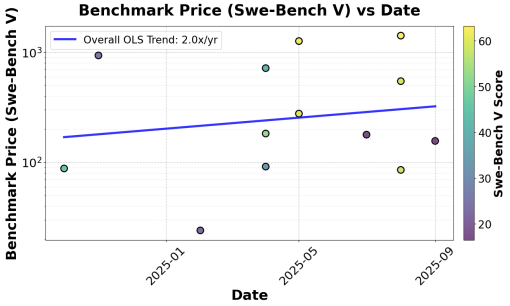

Figure 3: Graph of price to run SWE-bench Verified over time. Prices based on Epoch-AI benchmark data and Artificial Analysis Prices. Similar to Fig 2, benchmark prices have increased. In addition, the price to run SWE-bench Verified for some models is now in the thousands of dollars.

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 context and agential benchmarks, which are becoming prohibitively expensive. For instance, ∞BENCH spent 5,000 dollars evaluating 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 share some information with the public but keep other information, like benchmark costs and historical prices, either hidden or not easily accessible.

4 Recommendations for Future Evaluations

At first glance, our analysis seems to point in different directions. The price of certain abilities has dropped precipitously, while others remain uncertain. In addition, the benchmarking costs of models have remained constant or increased. Overall, we believe that focusing on increasing benchmarking scores alone has led the community to a distorted picture of AI progress. AIs with much better benchmark scores but with even greater resource demands represent a smaller leap in practical performance.

We call for evaluators and researchers to publish 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? Without data on the resources used, evaluations give us a much less clear picture of current practical capabilities and the future of AI.

References

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.

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.

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

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

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.

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.

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 is available here: The Price of Progress Project: https://anonymous.4open.science/r/The-Price-of-Progress-122B/

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 x20 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—we use 5m cache write prices. Deepseek's prices cache hit and misses. Since we do not know the hit-miss ratio, we decide not to include the model in our SWE-bench measurements. In addition, 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 where underspecified, i.e., did not mention which version of a model was used in these cases—we did not include this data. 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.