The Declining Cost of Intelligence: Temporal and Economic Trends in LLM Performance on GPQA*

Elena Foà[1,2,†], Giorgio Premoli[1,2,†] and Francesco Bertolotti[1,3,*,†]

[1]Intelligence, Complexity, and Technology Lab (ICT Lab), University Cattaneo – LIUC - 21053 Castellanza, Italy
[2]School of Industrial Engineering, University Cattaneo – LIUC - 21053 Castellanza, Italy
[3]Università Cattolica di Milano, Department of Philosophy, L. Gemelli 1 - 20123 Milano, Italy

Abstract

This study examines the temporal evolution and economic efficiency of Large Language Models on the GPQA benchmark from 209 models released between 2023 and April 2026. Our analysis reveals a logarithmic progression in GPQA performance over time, with SOTA models approaching benchmark saturation. However, substantial heterogeneity persists among models, indicating that high performance remains challenging for some architectures or configurations. Economic efficiency analysis, measured as the score-to-output-cost ratio, demonstrates marked improvement from 2024 to 2025 and 2026, suggesting that comparable performance levels are being achieved at progressively lower computational costs. Provider-level analysis identified Google, Mistral, and Meituan as efficiency leaders, while model-level rankings revealed that mid-sized architectures such as Gemma family, GLM-4.7-Flash, and Mistral Small 4, achieve near-frontier performance at significantly reduced cost. These findings show the trend of the declining cost of intelligence, and how optimization enables cost-effective competition with frontier-scale models.

Keywords

LLM, large language model, benchmark, time evolution, GPQA, cost of intelligence, GenAI,

1. Introduction

The rapid advancement of Large Language Models (LLMs) has generated the necessity for evaluation frameworks capable of assessing sophisticated reasoning capabilities[1, 2]. The Graduate-Level Google-Proof Q&A (GPQA) benchmark, developed by researchers at the Align Research Group, New York University [3], represents a significant contribution to this evaluation landscape. Published on November 20, 2023, GPQA addresses the growing need for assessment tools that can differentiate among increasingly capable models operating at the frontier of artificial intelligence performance. The GPQA benchmark comprises 448 multiple-choice questions designed to evaluate graduate-level scientific reasoning across three domains: biology, physics, and chemistry. Each question presents four response options, expertly crafted to appear plausible, thereby requiring models to present an understanding rather than simple pattern matching. The difficulty calibration of GPQA questions is intentionally set at PhD-level complexity, positioning the benchmark as a measure of advanced cognitive skills rather than mere knowledge retrieval.

As LLM development accelerates and model capabilities converge toward theoretical performance ceilings on established benchmarks, understanding the relationship between performance and deployment cost becomes increasingly critical for both researchers and practitioners. Previous studies have focused primarily on absolute performance metrics, the economic dimension of model deployment—particularly the cost per unit of capability—remains underexplored [4]. This study addresses this gap by investigating two primary research questions: (1) How has GPQA performance

Ital-IA 2026: CINI National Conference on Artificial Intelligence, June 18–19, 2026, Rome, Italy
*Corresponding author.
†These authors contributed equally.
el16.foa@stud.liuc.it (E. Foà); gi06.premoli@stud.liuc.it (G. Premoli); fbertolotti@liuc.it (F. Bertolotti)
0000-0003-1274-9628 (F. Bertolotti)

evolved temporally across the model population? (2) What is the relationship between benchmark performance and economic efficiency, measured by the score-to-cost ratio?[5]

This work shows and discusses the behavior of GPQA over time: as benchmark saturation approaches and marginal performance gains require disproportionate computational resources, the optimization landscape shifts from pure capability maximization toward multi-objective optimization, balancing performance, cost, and inference efficiency [6]. This trend is highlighted by the progressive decrease in the ratio between mean benchmark performance and economic efficiency, which declined from 1.62 in 2024 to 1.17 in 2025 and 0.64 in 2026, up to April.

2. Material and Methods

2.1. Data Collection

Data were taken from the LLM Stats[1], a publicly accessible aggregator of benchmark results compiled from official provider announcements and technical blog posts. The dataset was collected on April 20, 2026, encompassing all models with publicly available GPQA scores released through that date. The GPQA leaderboard on LLM STATS contained 209 model entries at the time of extraction. For each model, the following attributes were collected: model name, provider organization, announcement date, parameter count (where disclosed), and pricing per million tokens, for both input and output. No filtering criteria were applied to the dataset, and all available entries were retained for analysis to preserve the full distributional characteristics of the model population.

2.2. Temporal Trend Analysis

Benchmark performances were visualized on a scatter plot with announcement date on the x-axis and normalized GPQA score on the y-axis. To observe the temporal trend while mitigating sensitivity to individual outliers, a rolling mean was computed using a fixed window of k = 40 days, which was selected as it provided stable trend estimation without excessive smoothing. To highlight frontier performance progression, a cumulative maximum curve was constructed by selecting, at each time point, the highest-scoring model released to date. This approach traces the state-of-the-art boundary (SOTA), effectively filtering out lower-performing releases and focusing exclusively on technological ceiling advancement.

2.3. Economic Efficiency Metrics

Economic efficiency η_i of a model M_i is quantified via as the ratio

$$\eta_i = \frac{C_{out,i}}{GPQA_i^n} \tag{1}$$

where $GPQA_i^n$ is the normalized GPQA score, adopted rather than the raw score to ensure comparability across models evaluated on heterogeneous benchmarks scales, where raw scores may span ranges beyond the standard 0-1 interval, and $C_{out,i}$ the output cost, selected rather than input cost due to the asymmetric cost structure of reasoning-intensive models, which generate substantially more output tokens during chain-of-thought inference [7]. This metric enables direct comparison across heterogeneous model architectures, including those employing inference-time compute techniques [4][8].

Temporal trends in efficiency were assessed by partitioning the dataset by each year y, starting from 2024, and computing $E_y[\eta]$ [USD/(token*score)] within each cohort. Provider-level and model-level rankings were generated by aggregating efficiency metrics across all releases from each organization and for each specific model variant.

[1] https://llm-stats.com/

3. Results

3.1. Temporal Evolution of GPQA Performance

The temporal distribution in Figure 1 shows a logarithmic progression in GPQA performance over time, with recent models approaching benchmark saturation. The rolling mean curve demonstrated consistent upward trajectory from initial releases in late 2023 through April 2026, with the rate of improvement decelerating as scores approached theoretical maximum performance.

Despite the overall upward trend, substantial heterogeneity persisted across the model population, as evidenced by the scatter plot dispersion. A considerable proportion of recently released models continue to achieve relatively low GPQA scores, indicating that high performance on this benchmark remains challenging for non-specialized or smaller-scale models. This observation suggests that while the SOTA boundary has advanced significantly, the broader model ecosystem has not uniformly converged toward saturation-level performance.

In the figure, a black dashed line highlights the SOTA progression, inclusive of the GPT-4o model released in May 2024. Conversely, the light blue cumulative curve represents the SOTA trajectory when this specific model, that outperformed the rest of the ecosystem for more than 9 months is excluded.

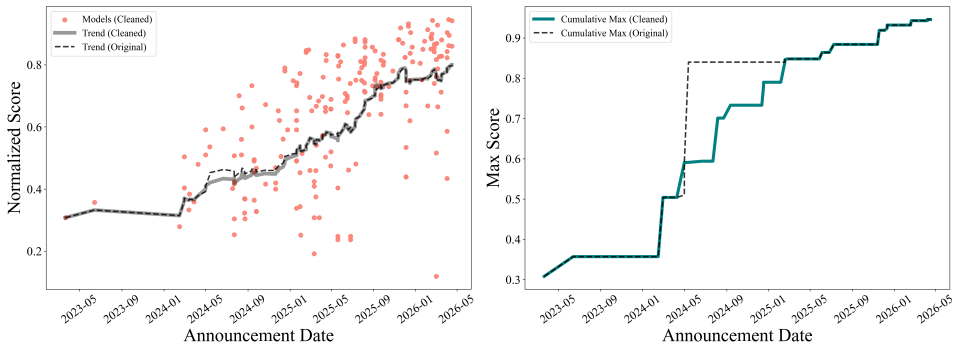

Figure 1: GPQA rolling mean (k=40) and trend, comparison with and without outlier

3.2. Economic Efficiency Distribution

The distribution of economic efficiency η_i across all models M_i shows that high-performance GPQA scores remain economically expensive to achieve for most models (Figure 2). The frequency histogram of score-to-output-cost ratios presents a right-skewed distribution, with the majority of models exhibiting low η. Only a small subset of models achieved high GPQA scores and favorable cost efficiency [9]. The heatmap on the right complements this analysis by plotting GPQA score versus output cost, with color intensity representing release recency—darker points correspond to more recent models. This temporal encoding shows that newer releases increasingly occupy the high-performance, low-cost position on a Paretian frontiers that gradually moves toward the upper-left corner, confirming the progressive efficiency gains observed across annual cohorts.

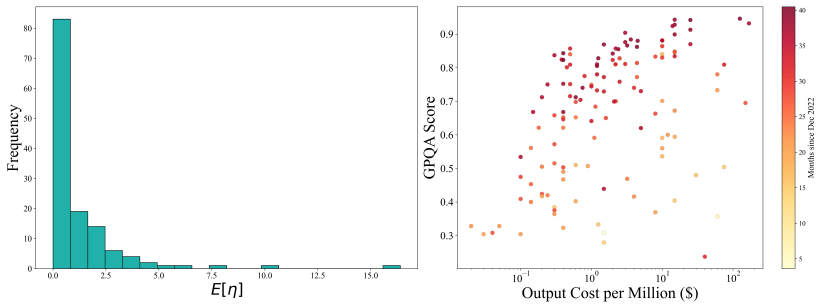

Figure 2: Economic efficiency distribution. Left: histogram of $E_y[\eta]$ (USD/(token*score) values. Right: GPQA score vs. output cost with temporal color encoding (darker = more recent).

3.3. Temporal Trends in Economic Efficiency

The yearly comparison shows a progressive improvement in mean efficiency $E_y[\eta]$ [USD/(token*score)]: 1.62 in 2024, 1.17 in 2025, and 0.64 in 2026 (through April). Despite the persistence of models with low efficiency scores across all years, the 2025 and 2026 distributions exhibited rightward shifts relative to 2024, indicating that newer models increasingly achieve comparable performance levels at reduced output costs (Figure 3).

This trend suggests that GPQA performance, and more generally a certain level of intelligence, is transitioning toward a commodity capability, where architectural optimizations and inference improvements enable cost-effective achievement of scores that previously required frontier-scale computational resources. The rightward distributional shift of η_i implies that the marginal cost of intelligence, as measured by this benchmark, is declining over time.

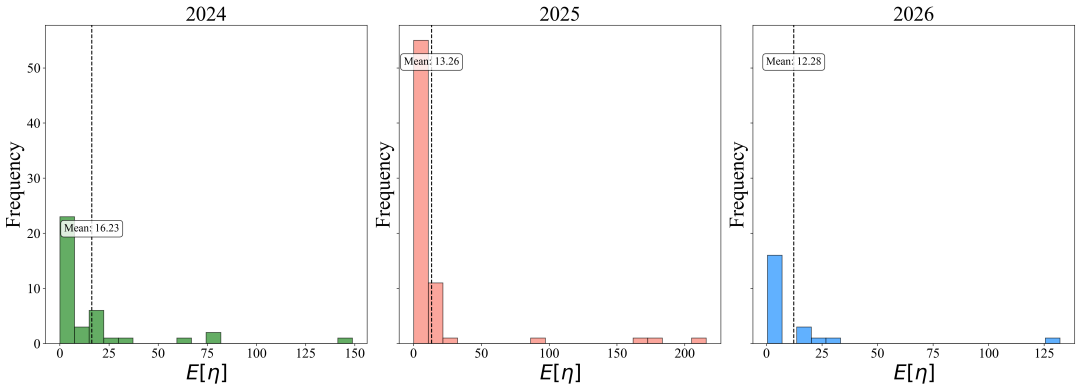

Figure 3: Economic efficiency trends 2024-2026. Histograms of $E_y[\eta]$ distributions showing rightward shift.

Mean efficiency values across years are illustrated in Figure 4, which traces the observed trend from 2024 through April 2026 via a magenta line, connecting the three annual $E_y[\eta]$ [USD/(token*score)] data points. The dashed segment extending to the 2026 endpoint represents a within-year forecast, projecting the expected mean efficiency once the full set of 2026 model releases will be accounted for. Forward-looking estimates for 2027 and 2028—depicted as salmon-colored markers—were derived by extrapolating the observed 2024–2026 trajectory, deliberately excluding the 2026 within-year projection to avoid compounding forecast uncertainty. If the current trend persists, $E_y[\eta]$ [(USD/(token*score)] is expected to decline to approximately 9.97 by 2027 and 7.99 by 2028, suggesting a continued, systematic reduction in the marginal cost of benchmark-level intelligence.

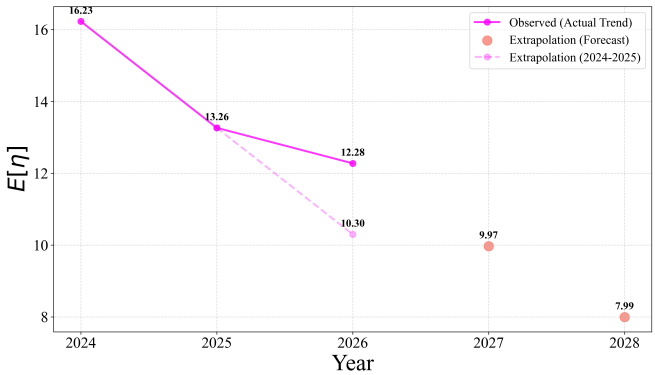

Figure 4: Mean efficiency trends and forward projections. Observed annual values (magenta line), 2026 within-year forecast (dashed), and extrapolated estimates for 2027-2028 based on 2024-2026 trajectory.

3.4. Provider-Level and Model-Level Efficiency Rankings

Provider-level analysis for 2026 releases revealed marked heterogeneity in cost-efficiency strategies despite comparable performance levels across organizations [10]. Google occupied the top position in the efficiency ranking, followed by Mistral and Meituan, suggesting the adoption of architectures or inference strategies optimized for cost-performance balance. Conversely, Alibaba, OpenAI, and Moonshot AI exhibited higher costs for comparable performance levels, while Anthropic and Qwen occupied the lower positions in the efficiency distribution.

This provider-level heterogeneity reflects divergent strategic priorities. Marginal improvements in GPQA scores—particularly the final 2-3 percentage points approaching the benchmark ceiling—require disproportionately large computational resources. Providers pursuing SOTA performance employ inference-time compute scaling techniques, generating thousands of intermediate reasoning tokens to produce a single final answer. This approach yields minimal performance gains at substantial cost inflation, driven by the proximity to the theoretical benchmark maximum, but could be explained by the necessity to obtain the lead in this and other benchmark for financial reasons.

Model-level analysis further illuminated efficiency heterogeneity within the ecosystem (Figure 5). Models from the Gemma family occupied the top efficiency positions, achieving $E[\eta] > 2$. At the opposite extreme, four Claude variants exhibited $\eta < 0.25$. Specifically, Gemma 4, GLM-4.7-Flash, and Mistral Small 4 achieved $\eta < 1.5$, attaining GPQA scores in the 80-85% range—proximate to frontier model performance—at fractional costs compared to "Opus" or "Pro" tier offerings. Observing that $\eta \in [2.10, 0)$ suggests that GPQA has become a metric to get how close moderately-sized and optimized models are to frontier-scale models [11].

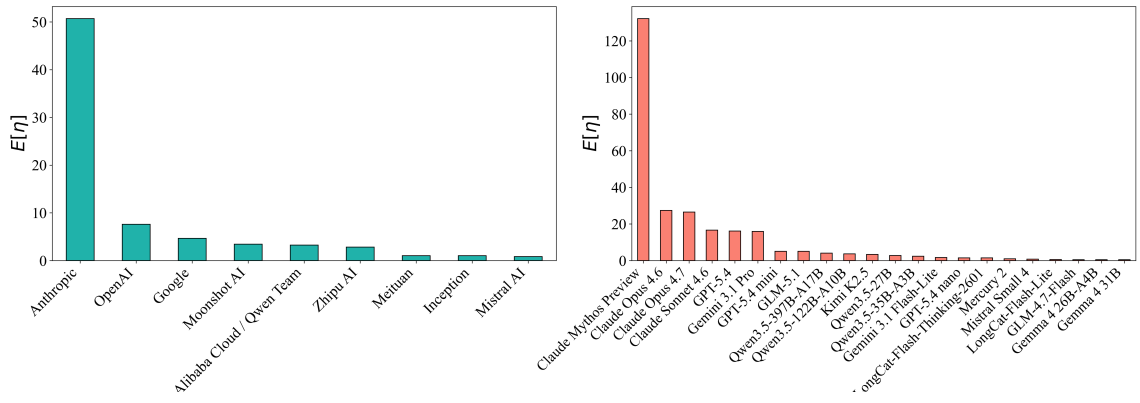

Figure 5: Provider and model-level efficiency rankings for 2026 releases.

4. Discussions and conclusions

The results indicate that GPQA saturation is not yet complete for non-SOTA models. This suggests that the benchmark still provides meaningful discriminatory power across a broad range of current systems. In particular, its relevance is likely to persist as newly released models become progressively smaller and more resource-efficient. Under this scenario, GPQA can continue to serve as a useful testbed for assessing advanced reasoning capabilities.

In addition, the results show a marked decline in the marginal cost of intelligence. There is no clear reason to expect this trend to reverse; rather, the available evidence suggests that it may be accelerating. This is especially relevant because, above a certain capability threshold, these models can be applied to cognitively complex tasks, such as coordinating with one another.

These results also provide an indication of what may be expected in the near future. Even without major technological disruptions, it appears plausible that the cost of intelligence will continue to decline

over the next one or two years. In this scenario, models available within a few years could provide, on average, similar capabilities at costs not achievable today. This would further expand the range of tasks for which advanced models can be used in practice.

This work has some limitations that should be considered when interpreting the results. First, the analysis focuses on a single benchmark, and therefore does not capture the full range of model capabilities. Second, it does not explicitly account for architectural differences among models, such as mixture-of-experts designs or reasoning-oriented systems. The cost-based approach partly reduces this issue, since it does not rely directly on parameter counts. Future work could extend the analysis by including these model-specific features more explicitly.

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM transactions on intelligent systems and technology 15 (2024) 1–45.

[2] M. Shanahan, Talking about large language models, Communications of the ACM 67 (2024) 68–79.

[3] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, S. R. Bowman, Gpqa: A graduate-level google-proof q&a benchmark, in: First conference on language modeling, 2024.

[4] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, R. McHardy, Challenges and applications of large language models, arXiv preprint arXiv:2307.10169 (2023).

[5] A. Aryan, A. K. Nain, A. McMahon, L. A. Meyer, H. S. Sahota, The costly dilemma: generalization, evaluation and cost-optimal deployment of large language models, arXiv preprint arXiv:2308.08061 (2023).

[6] E. T. Katz, Correct answers do not supervise intelligence (????).

[7] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, V. Gadepally, From words to watts: Benchmarking the energy costs of large language model inference, in: 2023 IEEE high performance extreme computing conference (HPEC), IEEE, 2023, pp. 1–9.

[8] S. Shashidhar, A. Chinta, V. Sahai, Z. Wang, H. Ji, Democratizing llms: An exploration of cost-performance trade-offs in self-refined open-source models, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 9070–9084.

[9] H.-Y. Chen, Process supervision via verbal critique improves reasoning in large language models, arXiv preprint arXiv:2604.21611 (2026).

[10] E. A. Khalafyan, Better thinking or a bigger model? thinking–answering shuffles with qwen3 on gpqa, Mathematical Problems of Computer Science 64 (2025) 17–28.

[11] A. X. Tian, R. Zhang, J. Tang, Y. M. Cho, X. Li, Q. Yi, J. Wang, Z. Zhang, D. Qi, Z. Li, et al., Beyond the strongest llm: Multi-turn multi-agent orchestration vs. single llms on benchmarks, arXiv preprint arXiv:2509.23537 (2025).