Estimating the Scale of Closed-Source Large Language Models from API Pricing*

Giorgio Premoli[1,2,†], Elena Foà[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

Proprietary LLMs do not usually disclose their architectural details, making it difficult to estimate their parameter count. This paper explores whether public API prices can provide information about model scale. We analyze the relationship between parameter count and input costs for models with known parameters. The results show a positive correlation between input cost and parameter count, with yearly correlations increasing from 2024 to 2026. Power-law regressions in log-log space provide a moderate fit and suggest that input cost scales sub-linearly with model size. The fitted relationships are then used to estimate the parameter counts of proprietary models, including uncertainty intervals. The results suggest that API pricing can be used as a cost-based proxy for the order of magnitude of model size, although estimates remain uncertain and are reliable only within the range covered by the calibration data.

Keywords

Large Language Model, LLM, Power Law, Cost per token, Parameter Estimation, Proprietary Models

1. Introduction

The advent of Large Language Models (LLMs) has revolutionized the field of artificial intelligence, leading to intensifying interest in why they behave in such a manner. The scaling laws [1] that govern these models demonstrate how their performance and computational capacity and requirements are intrinsically linked and, consequently, could be predictable as a function of parameter count and data size [2]. This is of greater interest given the division of the current scenario. On the one hand, open-source models have publicly known weights and architectures, including parameter counts; on the other hand, companies and foundations behind proprietary models remain confidential concerning their neural network architectures behind their models [3]. Users and researchers can interact with them through commercial APIs, incurring distinct monetary costs for input token processing, prompt submission, and output token generation, which could be proportional to their execution costs [3].

It is now well-known that computational costs, hardware bottlenecks, and the throughput of an LLM scale non-linearly depending on the model size, as well as the input and output lengths [4]. Although the behavior of LLMs during real-world generation continues to be studied to evaluate their uncertainty and reliability [5, 6], most of the literature has thus far focused on comparing API usage costs for private adaptation tasks [3] or optimizing infrastructural performance assuming known parameter sizes [4].

While inference cost dynamics and scaling laws are well-documented for open models [2], there remains a profound gap in our ability to structurally quantify closed-source models [7]. Currently, there is no publicly established mathematical relationship between the input/output pricing ratio of commercial LLM APIs and the actual size of the underlying model. This is driven by the fact that many

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

leading providers do not disclose the relevant architectural details, such as total or active parameter counts [8]. Previous studies have highlighted the cost disparity between closed APIs and open-source alternatives, often demonstrating that open models offer greater computational efficiencies [3], and have modeled computational throughput as a function of input and output sizes [4]. However, no study has yet attempted to close the loop: utilizing the public economic cost of APIs as a quantitative proxy to reverse-engineer and derive the architectural parameter count of a closed model.

This study aims to bridge this gap, demonstrating that it is possible to reliably and empirically estimate the number of parameters of non-open-source LLMs by analyzing the correlation between the input cost and the output cost imposed by API providers. Our fundamental hypothesis is that API prices are not purely arbitrary business metrics but inherently reflect the underlying infrastructural costs and hardware bottlenecks [4]. Because the computational load of the prefill phase (input processing) and the auto-regressive generation phase (output) are physically constrained by the model's parameter count [6, 4], we postulate the existence of a correlation function capable of linking the input cost to the order of magnitude of the LLM parameters.

2. Data collection

The data were collected from LLM Stats[1], a comprehensive dataset comprising approximately 276 models along with their number of parameters and input and output costs, when available. The data are updated as of April 20, 2026, and an outlier, Gemma 3n E4B Instructed, was removed. For the ensuing analyses, input cost was prioritized over output cost to ensure a more rigorous evaluation of reasoning models, where tokens and business logic could have a noise effect, as we have seen in the preliminary analysis. Nevertheless, Pearson correlation between input and output costs per million tokens was preliminarily calculated, yielding a value of 0.86, indicating that they are strongly correlated.

3. Results and discussion

3.1. Relationship between parameters number and input cost

The Pearson correlation between parameter count n and input costs c_{in} was assessed, yielding a value of $r = 0.45$, which preliminary analysis showed was approximately the same for output cost c_{out}. The log-log visualization of the models on an $n - c_{in}$ plane, visible in Figure 1, shows that they appear to be distributed linearly. Consequently, power-law fitting was performed to see if the model fits the distributions [9], even if this typically requires a larger number of observations or a thorough understanding of the underlying generative model [10].

The fitted power-law model shows a $R^2 = 0.46$ compared with the actual data, validates the suitability of this mathematical function for modeling the relationship between n and c_{in}. We repeated the analyses accounting for temporal factor, which is relevant given the rate of changing of this technology [7]. To address this, we recalculated the correlations on an annual basis, yielding the results in Table 1.

Year	$r(n, c_{\text{in}})$	$r(n, c_{\text{out}})$
2024	0.38	0.39
2025	0.55	0.49
2026	0.80	0.65

Table 1
Correlation between parameter count n, input cost c_{in} and output costs c_{out} by year.

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

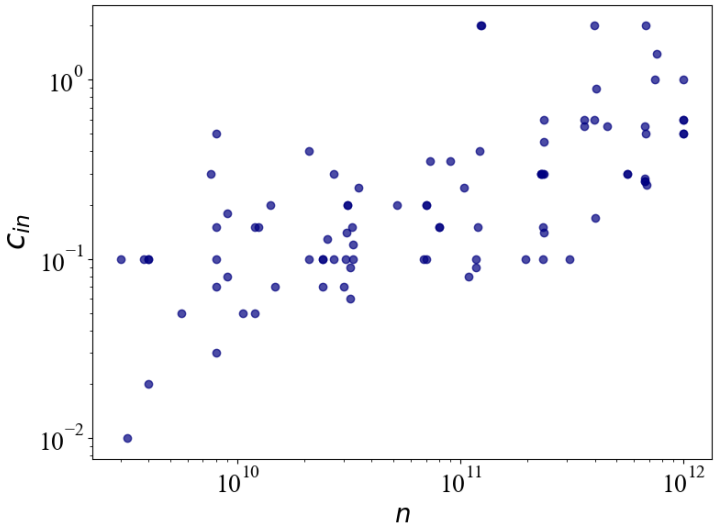

Figure 1: Visual representation of model input cost c_{in} and parameter count n of models M with a log-log scale.

The increasing yearly correlation between parameter count and API pricing suggests that, over time, commercial API costs have become progressively more aligned with the underlying scale of the models. In particular, the sharp increase in $r(n, c_{\text{in}})$ from 0.38 in 2024 to 0.80 in 2026 indicates that input pricing has become increasingly informative of model size, even in a linear way. This supports the hypothesis that API prices are not only economic metrics but may increasingly encode model scale. Nevertheless, this evidence should be interpreted cautiously, as prices could also entail market maturation, pricing standardization, or changes in the composition of available models.

Given that, on a log-log plane $n - c_{in}$ the data appear also linearly when divided per year (see Figure 2), a power law fitting has also been repeated for every single year, and the residuals observed, as in Figure 3.

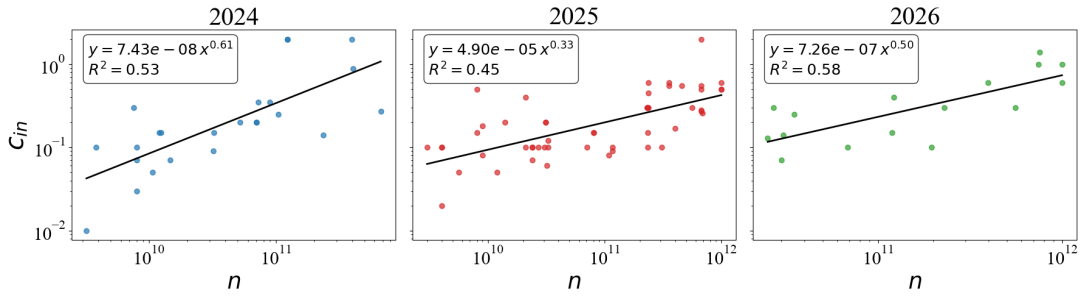

Figure 2: Application of power law regression model for each year between input cost c_{in} and parameter count n.

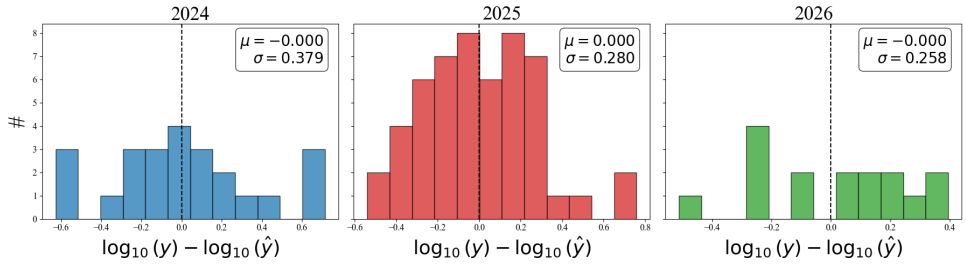

Figure 3: Residuals from the yearly log-log regressions of parameter count against input price c_{in}.

The yearly log-log regressions show that the relationship between parameter count and input price remains consistently compatible with a power-law form across the three temporal subsamples. The fitted models explain a moderate fraction of the variance, with $R^2 = 0.53$ in 2024, $R^2 = 0.45$ in 2025, and $R^2 = 0.58$ in 2026. Since the residuals are approximately normally distributed, the deviations from the fitted power laws do not appear to reflect a strong systematic misspecification, supporting the suitability of the log-log linear model. Interestingly, the estimated exponents remain below one in all years, suggesting a sub-linear scaling of input costs with parameter count. This indicates that API prices increase with model size, but less than proportionally. Furthermore, the lower exponent and lower R^2 observed in 2025 may point to a more heterogeneous pricing regime, while the stronger association observed in 2026 suggests a progressive alignment between commercial API prices and the infrastructural constraints associated with model scale.

3.2. Inference on parameters number of proprietary models

The three previously derived power laws were applied to infer the parameter counts for non-open-source models by an interpolation. The estimation was performed taking into account not only the expected estimated number of parameters $\hat{n}_i$ for a proprietary model M_i, but also its uncertainty $\Delta\hat{n}_i$.

Model name	Organization	Date	c_{in}	c_{out}	$\hat{n}_i$	$\hat{n}_i - \Delta\hat{n}_i$	$\hat{n}_i + \Delta\hat{n}_i$
Claude Opus 4.6	Anthropic	01/02/2026	$ 5.00	$ 25.00	45.87T	14.03T	150.00T
Claude Opus 4.7	Anthropic	01/04/2026	$ 5.00	$ 25.00	45.87T	14.03T	150.00T
Claude Sonnet 4.6	Anthropic	01/02/2026	$ 3.00	$ 15.00	16.53T	5.06T	54.06T
GPT-5.2	OpenAI	01/12/2025	$ 1.75	$ 14.00	74.43T	10.45T	529.96T
GPT-5.2 Pro	OpenAI	01/12/2025	$ 21.00	$ 168.00	144559.70T	20302.08T	1029328.08T
GPT-5.4	OpenAI	01/03/2026	$ 2.50	$ 15.00	11.48T	3.51T	37.55T
Gemini 3 Flash	Google	01/12/2025	$ 0.50	$ 3.00	1.64T	229.8B	11.65T
Gemini 3.1 Pro	Google	01/02/2026	$ 2.50	$ 15.00	11.48T	3.51T	37.55T

Table 2
Estimated parameter counts ($\hat{n}_i$) and uncertainty bounds for proprietary models, inferred from API input and output costs. Higher API prices generally correlate with larger inferred parameter counts, though estimates are subject to wide, asymmetric uncertainty intervals.

The results in Table 2 indicate that models with higher API input prices tend to be associated with larger inferred parameter counts. In particular, Claude Opus 4.6, Claude Opus 4.7, GPT-5.2, and GPT-5.2 Pro are placed in the upper region of the estimated scale, whereas Gemini 3 Flash is assigned a substantially lower parameter count. Models that share similar input prices, such as GPT-5.4 and Gemini 3.1 Pro, produce comparable estimates, as expected from the cost-based inversion procedure. However, the uncertainty intervals are wide and asymmetric, reflecting the fact that the estimation is performed through an inverse power-law model in log-log space. Therefore, the estimates should not be interpreted as precise measurements, but as plausible orders of magnitude conditional on the fitted relationship between API pricing and parameter count. Table 2 presents a sample of the performed estimation and the complete dataset is available in the following repository: https://osf.io/w76xy/overview.

The case of GPT-5.2 Pro is particularly informative. Given its extremely high input price, the inferred parameter count is several orders of magnitude larger than the other estimates. This is likely inaccurate; we present this instance to highlight that the validity of the results is constrained to the range where sufficient data are available. GPT-5.2 Pro was the only model so expensive, and the related estimation should be interpreted as evidence that the model lies outside the reliable interpolation regime of the fitted power law. This suggests that for very high-priced models, API pricing may include additional components beyond parameter-dependent inference costs, such as premium positioning.

4. Conclusions

This study provides preliminary evidence that API input prices can be used as an empirical proxy for estimating the order of magnitude of the parameter count of closed-source LLMs. The observed correlations, especially their increase over time, suggest that commercial pricing has become progressively more aligned with the model scale. The yearly power-law fits further support the existence of a stable, although noisy, relationship between input cost and parameter count. However, the inferred values should be interpreted as cost-implied estimates rather than direct measurements of model size. The large uncertainty intervals and the extreme case of GPT-5.2 Pro show that the method is reliable only within the range covered by the calibration data. Future work should extend the dataset and include performance variables.

References

[1] G. B. West, B. J. Enquist, The origin of universal scaling laws in, Scaling in biology (2000) 87.

[2] S. Bergsma, N. Dey, G. Gosal, G. Gray, D. Soboleva, J. Hestness, Power lines: Scaling laws for weight decay and batch size in llm pre-training, arXiv preprint arXiv:2505.13738 (2025).

[3] V. Hanke, T. Blanchard, F. Boenisch, I. E. Olatunji, M. Backes, A. Dziedzic, Open llms are necessary for current private adaptations and outperform their closed alternatives, Advances in Neural Information Processing Systems 37 (2024) 1220–1250.

[4] K. Ray, N. M. Gonzalez, B. Wassermann, R. Tzoref-Brill, D. H. Lorenz, Statistical modeling and uncertainty estimation of llm inference systems, arXiv preprint arXiv:2505.09319 (2025).

[5] H. Ma, J. Chen, G. Wang, C. Zhang, Estimating llm uncertainty with logits, arXiv e-prints (2025) arXiv–2502.

[6] Y. F. Bakman, D. N. Yaldiz, S. Kang, T. Zhang, B. Buyukates, S. Avestimehr, S. P. Karimireddy, Reconsidering llm uncertainty estimation methods in the wild, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 29531–29556.

[7] F. Bertolotti, L. Mari, An llm-based delphi study to predict genai evolution, arXiv preprint arXiv:2502.21092 (2025).

[8] D. Bergemann, A. Bonatti, A. Smolin, The economics of large language models: Token allocation, fine-tuning, and optimal pricing, arXiv preprint arXiv:2502.07736 (2025).

[9] N. Sardana, J. Portes, S. Doubov, J. Frankle, Beyond chinchilla-optimal: Accounting for inference in language model scaling laws, arXiv preprint arXiv:2401.00448 (2023).

[10] S. Roman, F. Bertolotti, A master equation for power laws, Royal Society open science 9 (2022).