The Economics of Data in the Scaling of Large Language Models

Anna Pastwa,[*] Georgios Petropoulos,[†] Dan Zhao,[‡] Martin Fleming,[§]
Brian Goehring,[¶] Kate Soule[‖] and Neil Thompson.[**]

March 2026

Abstract

Traditional scaling laws for large language models (LLMs) have focused on the optimal model size and amount of training data needed to achieve a given loss level. More recent works have also begun considering the computational costs of inference in optimal training and scaling. Yet, these analyses can still lack a degree of realism as data is typically assumed to be free, which is not always be the case. Data-related costs can significantly alter what optimal scaling can realistically be. As the race towards ever better models intensifies, acquiring and preparing newer or higher quality training data is expensive but necessary to improve performance and maintain a competitive edge. In this paper, we re-derive scaling laws to incorporate training data costs under several cost projections/trends and focus on minimizing economic costs. We characterize the resulting optimal expansion paths and derive comparative statics showing how changes in marginal data costs and inference usage affect the optimal allocation between parameters and training tokens. Compute-optimal scaling rules arise as a special knife-edge case. Whenever data or inference costs are economically relevant, the compute-optimal and economic optima diverge. Numerical experiments calibrated to contemporary LLM training and deployment regimes demonstrate that ignoring data and inference costs can lead to substantial misallocation and large excess costs. Our framework provides a tractable economic generalization of existing scaling laws and offers new guidance for efficient LLM development under realistic resource constraints.

Keywords: Large Language Models, Scaling Laws, Economic Optimization.
JEL Classifications: D24, O33, L86.

[*]MIT. E-mail: apastwa@mit.edu.
[†]University of Southern California. E-mail: georgios@marshall.usc.edu.
[‡]New York University. E-mail: dz1158@nyu.edu.
[§]MIT. E-mail: marti264@mit.edu.
[¶]IBM. E-mail: goehring@us.ibm.com.
[‖]IBM. E-mail: Kate.Soule@ibm.com.
[**]MIT. E-mail: neil_t@mit.edu

1 Introduction

As large language models (LLMs) continue to grow in capability, understanding the true cost of their development is crucial. While existing scaling laws primarily focus on compute costs, our research introduces a more comprehensive framework that incorporates the growing costs of acquiring, processing, and managing data. Despite the enormous influence of LLM scaling laws, the guidance they offer to model builders neglects the often significant costs of inference and data. Guidance on the "best" ways to scale these models or to characterize the limits of resource-performance trade-offs has become essential given the explosive growth of LLMs and other foundation models, fueled by improvements in algorithms, data and compute. Past works have attempted to capture and model the interplay and relative significance of these factors in the form of neural scaling laws that illustrate how model performance or quality scales with data size, model size, and compute (see Section 2). However, two factors stand out given the prevalence of how LLMs are used and trained today. The first is the demand for inference. As LLMs are increasingly used in products and services, model choices must balance optimal training, where particular ratios of compute and data are desirable, and optimal inference. The second, often missing, factor is the cost of high-quality training data (Subramanyam et al., 2025; Chen et al., 2025). These can include the direct costs for acquiring the data as well as the indirect costs for processing it (e.g., human labor for annotation, de-duplication, quality filtering, detoxification, etc.). Since not all tokens are created equal in terms of data quality or training signal, these costs can be expressed either as effective increases in per token costs or a quality-adjusted analog. However, data cost remains relatively unexplored despite its growing importance as existing open-source datasets are exhausted (Kandpal and Raffel, 2025). This omission is particularly important as the search for new quality data heats up: just a few examples include OpenAI partnering with Axel Springer in a content licensing agreement (OpenAI, 2023b) and, more recently, with Reddit (Chapman, 2024) to secure high-quality data. Some estimates (Villalobos et al., 2022) project that the stock of high-quality language/text data will be depleted within the next two decades. The realism of these projections aside, data availability–and by extension, data costs—is becoming an ever important in the scaling and improvement of LLMs. These costs, in turn, can hold significant implications for various other considerations such as datacenter demand, resultant energy consumption, and AI's overall energy footprint among others.

Contributions. In this paper, we focus on the potentially growing cost of data and attempt to incorporate these differential data costs into the guidance of traditional scaling laws. Specifically, we provide an economic lens on scaling: organizations naturally face binding resource and budget constraints in the development and scaling of LLMs and other foundation models, incurring trade-

offs in the resource allocations made. Many existing scaling laws like Chinchilla provide guidance on what is compute-optimal rather than what is realistically or economically optimal.

We develop a theoretical model with insights and predictions that substantially enrich our understanding of AI scaling beyond existing compute–optimal rules. We illustrate that parameters and training data are not merely technical substitutes along an iso–loss curve, but become *economic substitutes* whose optimal ratio depends systematically on marginal data cost and inference demand. This perspective is absent from existing scaling–law work, which implicitly assumes that data are free and inference is negligible. Our analysis shows that these assumptions place classical compute–optimal scaling rules (such as Chinchilla) on a knife–edge.

With the scarce amount of information available on data costs (especially from large companies and AI labs), however, it is challenging to tackle data cost problems in any truly substantive way. Nonetheless, given the growing importance of data and inference, we turn to simulations and cost scenarios in an attempt to validate our theoretical predictions and to provide a sense of scale and impact as data availability/costs and inference demand vary. This way, we intend for our work to help spur further conversation, research, and transparency regarding this topic.

Our simulations, cost scenarios, and experiments confirm that when accounting for data costs, economically-optimal scaling can differ significantly from Chinchilla scaling depending on the level of inference demand and training data costs per token. Depending on the cost of training data, we also show that model/data choices implied by Chinchilla's scaling can lead to under or over-estimates of cost optimal model sizes (Figure 1). As such, compared to the costs of Chinchilla's optimal choices, an economically-optimal scaling can produce choices that reduce total costs by up to and over 90% in several regimes (Tables 3, 4, and 5).

Our primary goal is to highlight how data costs and availability affect scaling, despite the limited transparency around organizations' data budgets and sourcing strategies. Because such information is scarce, our analysis necessarily relies on simplified cost scenarios. These scenarios are designed to capture any data-related expense—whether from purchasing new data or investing in preprocessing and curation—as an effective increase in per-token cost. However, because we lack real-world benchmarks, our simulations, results, and interpretations are naturally limited by these simplifying assumptions.

2 Related Work

Scaling laws and LLMs. Earlier works (Kaplan et al., 2020; Hernandez et al., 2021; Kandpal and Raffel, 2025) in scaling laws studied how language models empirically scaled in terms of performance during training, showing how training loss can predictably scale with model/data size, compute, etc. The Chinchilla scaling laws (Hoffmann et al., 2022) found similar results, estimating

that a doubling of data size (i.e., number of training tokens) should accompany a doubling of the model size and that, for a 1 trillion parameter model under a fixed compute budget, 11x more training data than that used for GPT-3 would be necessary. Many works have since sprung up to test, build upon, or further study the implications of Chinchilla's scaling laws. Given the diversity of LLMs, works like Isik et al. (2024) have studied how scaling behavior varies in transfer learning settings for downstream tasks, how the inductive bias of model architectures influence scaling behavior (Tay et al., 2022), how scaling affects trade-offs for smaller language models (Hu et al., 2024), the potential impact of synthetic data on scaling and model collapse (Dohmatob et al., 2024), how data-constrained language models scale with repeated data Muennighoff et al. (2023), how scaling laws can be used to derive optimal model/data sizes when accounting for inference demand (Sardana and Frankle, 2023), and much more.

Data on Data Costs. Recently there is increasing recognition of data as an essential ingredient in the best-performing LLMs and similar large foundation models (OpenAI, 2023a; Oderinwale and Kazlauskas, 2025). This importance relates not only to data size, but also data quality, focusing on filtering strategies or data generation to achieve the best performance (Marion et al., 2023). Freely available datasets, such as Common Crawl, Wikipedia, and open-source code repositories, constitute the majority of training corpora, incurring minimal expenses beyond preprocessing and/or storage. Licensed datasets, including proprietary books, news archives, and academic papers, introduce moderate costs depending on exclusivity and access restrictions. High-value proprietary data, such as financial reports or legal documents, often require direct agreements, significantly increasing per-token costs. Additionally, synthetic data generated by other LLMs is increasingly used to augment training corpora (Bansal et al., 2024), offering a scalable but computationally expensive alternative.

Yet, there is little information disclosed on the cost of data from acquisition to pre-processing and curation despite the hidden costs/challenges of data ownership and usage as well as data collection and quality, which are increasingly expensive for AI companies due to a variety of factors—from lack of transparency due to secrecy and competition to the complex nature of measuring, collecting, or recording the cost of data throughout its lifecycle (Whang et al., 2023; Vipra and West, 2023). Despite this, these costs have proved important as companies like OpenAI and Google have even made strategic decisions based on the high costs and scarcity of compute and data, influencing their market strategies (Alliance and ClearML, 2023).

3 Theoretical Model and Main Predictions

The economic realities of resource constraints necessarily mean that using traditional scaling laws to find the optimal model size and data size, given a desired level of model quality/performance, is likely to provide incorrect guidance. Instead, by incorporating models of data costs, we aim to characterize alternative *expansion paths*—paths connecting optimal input combinations (data, model size) as the scale of production (improving model quality) expands (Figure 1)—than those implied by traditional scaling. Each point on the expansion path characterizes the producer's optimal (i.e., cost minimizing) input allocations given a certain level of output to produce.

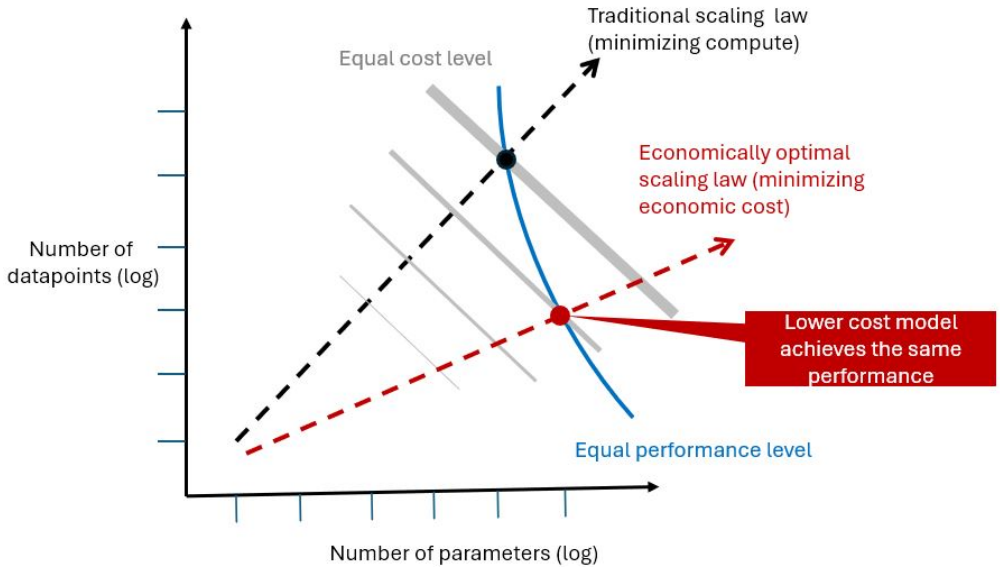


Figure 1: Economically-optimal Scaling vs. Compute-optimal Scaling. A high-level illustrative example of what we aim to characterize: the differences between traditional scaling laws and those that account for minimizing data costs. The red line indicates the expansion path for economically-aware scaling while the black line indicates the original path under Chinchilla scaling (without considering data costs). Gray lines are iso-cost curves where points along each curve represent model/data configurations of equal cost.

Our goal is to find a model of sufficient quality/performance that minimizes total development cost, with a focus on data costs, for both training and inference. We frame this as a cost minimization problem as typically found in economics: minimizing the cost of production (i.e., the total cost of inputs required for production) subject to producing a certain level of output. Here, this translates into finding the optimal size of a model and amount of data that minimizes the total cost of data and compute for both training and inference, to achieve a certain level of performance.

3.1 Environment and primitives

Let $N \in \mathbb{R}_+$ denote model parameters, $D_{\text{tr}} \in \mathbb{R}_+$ the number of training tokens, and $D_{\text{inf}} \in \mathbb{R}_+$ the total inference tokens demanded by the product. Model quality is summarized by a training loss ℓ that follows a Chinchilla-style production function:

$$\ell = \frac{A}{N^{\alpha}} + \frac{B}{D_{\text{tr}}^{\beta}} + \varepsilon, \qquad A, B > 0, \;\; \alpha, \beta \in (0,1). \tag{1}$$

A target quality level $\bar{\ell} > 0$ is externally fixed by the AI engineer.[1]

Training and inference floating point operation (FLOP) prices are $P_{\text{F,tr}}$ and $P_{\text{F,inf}}$, respectively. Let also token/parameter–to–FLOP conversion constants be κ_{tr} and κ_{inf} for training and inference, respectively. The economic cost of training data (acquisition, curation, licensing, labeling) is $C(D_{\text{tr}})$. We initially consider a general cost function which is potentially non-linear and possibly kinked.

3.2 Costs and firm's problem

Training compute scales (approximately) with the token–parameter product:

$$\text{Cost}_{\text{tr,compute}} = P_{\text{F,tr}} \, \kappa_{\text{tr}} \, N D_{\text{tr}}. \tag{2}$$

Inference compute scales with model size and total inference tokens:

$$\text{Cost}_{\text{inf,compute}} = P_{\text{F,inf}} \, \kappa_{\text{inf}} \, N D_{\text{inf}}. \tag{3}$$

Total *economic* cost at (N, D_{tr}) is

$$\boxed{\mathscr{C}(N, D_{\text{tr}}) = C(D_{\text{tr}}) + P_{\text{F,tr}} \kappa_{\text{tr}} N D_{\text{tr}} + P_{\text{F,inf}} \kappa_{\text{inf}} N D_{\text{inf}}} \tag{4}$$

Given a target loss $\bar{\ell}$, the AI engineer chooses (N, D_{tr}) to minimize total economic cost subject

[1] Our use of pre-training loss as the performance quality metric follows standard practice in scaling law literature, though we acknowledge that the mapping from loss to downstream capabilities can be nonlinear and task-dependent. The economic trade-offs we characterize—between data, parameters, and inference costs—remain structurally valid under alternative quality metrics, as our comparative statics depend only on the functional form of the production function rather than the specific interpretation of $\bar{\ell}$.

to quality:

$$\begin{aligned} \min_{N,D_{\text{tr}}} \quad & \mathscr{C}(N, D_{\text{tr}}) \\ \text{s.t.} \quad & \frac{A}{N^{\alpha}} + \frac{B}{D_{\text{tr}}^{\beta}} \leq \bar{\ell}. \end{aligned} \tag{5}$$

The Lagrangian is

$$\mathscr{L} = C(D_{\text{tr}}) + P_{\text{F,tr}} \kappa_{\text{tr}} N D_{\text{tr}} + P_{\text{F,inf}} \kappa_{\text{inf}} N D_{\text{inf}} + \lambda \left(\frac{A}{N^{\alpha}} + \frac{B}{D_{\text{tr}}^{\beta}} - \bar{\ell} \right). \tag{6}$$

First-order conditions (FOCs):

$$\partial_N : \quad P_{\text{F,tr}} \kappa_{\text{tr}} D_{\text{tr}} + P_{\text{F,inf}} \kappa_{\text{inf}} D_{\text{inf}} = \lambda \, \alpha A N^{-(\alpha+1)}, \tag{7}$$

$$\partial_{D_{\text{tr}}} : \quad C'(D_{\text{tr}}) + P_{\text{F,tr}} \kappa_{\text{tr}} N = \lambda \, \beta B D_{\text{tr}}^{-(\beta+1)}, \tag{8}$$

$$\text{constraint:} \quad \frac{A}{N^{\alpha}} + \frac{B}{D_{\text{tr}}^{\beta}} = \bar{\ell}. \tag{9}$$

Eliminating λ yields the following key trade-off condition:

$$\boxed{\frac{\alpha A}{N^{\alpha+1}} \Big/ \frac{\beta B}{D_{\text{tr}}^{\beta+1}} = \frac{P_{\text{F,tr}} \kappa_{\text{tr}} D_{\text{tr}} + P_{\text{F,inf}} \kappa_{\text{inf}} D_{\text{inf}}}{C'(D_{\text{tr}}) + P_{\text{F,tr}} \kappa_{\text{tr}} N}} \tag{10}$$

The left-hand side in equation (10) captures how much more effective are parameters than data at improving quality, at the current scale. The numerator $\frac{\alpha A}{N^{\alpha+1}}$ measures how much loss falls if you slightly increase model size. The denominator $\frac{\beta B}{D_{\text{tr}}^{\beta+1}}$ measures how much loss falls if you slightly increase training data.

The right-hand side of (10) captures how expensive is it to improve quality via parameters rather than via data. The numerator refers to the marginal cost of parameters, while the denominator to the marginal cost of data.

3.3 Comparative statics

At the optimum, quality improvements obtained via parameters and via data must deliver the same loss reduction per dollar. If parameters are relatively more effective at reducing loss than their relative cost would suggest, the optimal choice shifts toward a larger model; if data are relatively more cost-effective, the optimal choice shifts toward larger training datasets.

To see this more formally, we engage in a comparative statics analysis over (10) without considering a specific functional form of $C(\cdot)$. We only assume that for each θ in an open interval

$\Theta \subset \mathbb{R}$, the function $C(\cdot;\theta) : \mathbb{R}_+ \to \mathbb{R}_+$ is twice continuously differentiable, nondecreasing and weakly convex:

$$C'(D_{tr};\theta) \geq 0, \qquad C''(D_{tr};\theta) \geq 0 \quad \forall D_{tr} > 0.$$

We also assume a standard mild regularity condition ensuring that the solution of FOC (10) is locally unique and that the Implicit Function Theorem applies (see the Appendix for a formal exposition).

Proposition 1 (Pointwise increase in marginal data cost)**.** *Suppose that the parameter θ shifts marginal data cost* upwards pointwise*, i.e.*

$$\frac{\partial C'(D;\theta)}{\partial \theta} > 0 \quad \textit{for all } D \equiv D_{tr} > 0 \textit{ and all } \theta \in \Theta_0.$$

Then along the solution manifold $\theta \mapsto (N^*(\theta), D^*(\theta))$, we have

$$\frac{dN^*(\theta)}{d\theta} > 0, \qquad \frac{dD^*(\theta)}{d\theta} < 0.$$

Any parameter that shifts marginal data cost up pointwise induces substitution away from training tokens and toward model parameters.

Proposition 1 shows that *any* parameter that shifts the marginal data-cost schedule $C'(D_{\text{tr}};\theta)$ upward pointwise (for example a uniform increase in token price, a new licensing fee, or a reduction in effective data supply that raises marginal cost at all scales) induces an increase in optimal model size N and a decrease in optimal training tokens D_{tr}, holding target loss $\bar{\ell}$ fixed.

Proposition 2 (Inference demand reduces optimal parameter count)**.** *Under an interior optimum with differentiable $C(\cdot)$, an increase in inference tokens D_{inf} decreases the optimal parameter count and increases the optimal training tokens:*

$$\frac{\partial N^*}{\partial D_{\text{inf}}} < 0, \qquad \frac{\partial D_{tr}^*}{\partial D_{\text{inf}}} > 0.$$

Intuitively, when inference demand D_{inf} increases, each additional model parameter becomes more expensive, because it raises the cost of every future inference call. Training data, by contrast, only affects cost once, during training. Holding target quality fixed at $\bar{\ell} > 0$, the optimizer therefore shifts away from improving quality through parameters and instead relies more heavily on training data. As a result, the optimal parameter count falls and the optimal training data size rises.

Proposition 3 (Divergence from Chinchilla in economically relevant regimes)**.** *Consider two optimization problems for a fixed target loss $\bar{\ell}$:*

Compute–only (Chinchilla) problem (C).

$$\min_{N, D_{\text{tr}} > 0} \quad P_{\text{F,tr}}\, \kappa_{\text{tr}}\, N D_{\text{tr}}$$
$$s.t. \quad \frac{A}{N^{\alpha}} + \frac{B}{D_{\text{tr}}^{\beta}} = \bar{\ell}.$$

Economic scaling problem (E).

$$\min_{N, D_{\text{tr}} > 0} \quad C(D_{\text{tr}}) + P_{\text{F,tr}}\, \kappa_{\text{tr}}\, N D_{\text{tr}} + P_{\text{F,inf}}\, \kappa_{\text{inf}}\, N D_{\text{inf}}$$
$$s.t. \quad \frac{A}{N^{\alpha}} + \frac{B}{D_{\text{tr}}^{\beta}} = \bar{\ell}.$$

Let $(N^{\text{C}}, D^{\text{C}})$ denote an interior solution of (C) and (N^{*}, D^{*}) an interior solution of (E). Then:

(i) If $C'(D) \equiv 0$ on $(0, \infty)$ and $D_{\text{inf}} = 0$, the Chinchilla and economic scaling problems coincide: any interior solution of (C) is an interior solution of (E), and vice versa. In particular, their expansion paths in (N, D) space are identical.

(ii) If either $C'(D) > 0$ at D^{*} or $D_{\text{inf}} > 0$, then any pair (N, D) that simultaneously solves the first-order conditions of both (C) and (E) must satisfy the knife-edge relation

$$C'(D_{\text{tr}})\, D_{\text{tr}} = P_{\text{F,inf}}\, \kappa_{\text{inf}}\, N\, D_{\text{inf}}. \tag{11}$$

Unless (11) happens to hold at $(N^{\text{C}}, D^{\text{C}})$, the Chinchilla and economic optima differ. In particular, for generic choices of C and $D_{\text{inf}} > 0$ the Chinchilla and economic expansion paths are distinct.

Proposition 3(i) shows that classical compute–optimal scaling rules, such as Chinchilla, arise as a special case of the economic model when data are effectively free and inference usage is negligible. Proposition 3(ii) shows that once either data or inference costs are economically relevant, the Chinchilla and economic optima coincide only on a knife-edge locus defined by (11). For generic cost functions and realistic inference demand, this condition fails, and optimal scaling departs systematically from Chinchilla.

3.4 Specific Cost Scenarios

In addition to the above results, with a general cost function for training data, we also explore more specific strategies and pricing scenarios for assessing the data cost $C(D_{\text{tr}})$ which we describe

in turn below. Through these scenarios, we aim to shed light on the efficacy and implications of different data cost scenarios and their impact on scaling trajectories.

Base case: free data. If $C'(D_{\text{tr}}) = 0$ and $D_{\text{inf}} \approx 0$, then (10) collapses to the compute-only tradeoff, recovering the standard Chinchilla expansion path (Sardana and Frankle, 2023). Such frameworks implicitly presume that sufficiently high quality training data is free and infinitely available. Following Proposition 3, we refer to the solutions and scaling behavior under this case as the *Chinchilla optimal* scaling—i.e., the anticipated scaling behavior under Chinchilla-like scaling laws that treat data costs as zero and neglect inference demand. This base case serves as a reference against which we analyze other cost projections.

Uniform data cost. In the simplest scenario, data costs are constant or uniform: they neither grow nor shrink but stay at some constant per-unit cost P_{tok}. If so, then $C(D_{\text{tr}})$ is simply:

$$C(D_{\text{tr}}) = P_{\text{tok}} D_{\text{tr}} \tag{12}$$

where P_{tok} is non-zero. If $P_{\text{tok}} = 0$, then this reduces to the free data case.

Piece-wise uniform data cost. Another scenario is where sufficiently high-quality data exists but is free only up to a certain point. For example, publicly available large curated datasets like "The Pile" (Gao et al., 2020) can be used to pre-train LLMs; however, depending on specific needs and different workflows either by industry or use-case, training can require either specific bespoke datasets or ones that require additional cost to procure and process. One way to model this phenomena is a branching cost structure that depends on the amount of data involved.

Let

$$C(D_{\text{tr}}) = \begin{cases} 0, & D_{\text{tr}} \leq \bar{D}_{\text{free}}, \\ P_{\text{tok}}\,(D_{\text{tr}} - \bar{D}_{\text{free}}), & D_{\text{tr}} > \bar{D}_{\text{free}}. \end{cases} \tag{13}$$

At $D_{\text{tr}} = \bar{D}_{\text{free}}$, the subgradient is $\partial C = [0, P_{\text{tok}}]$. Optima either (i) sit at the kink with $D_{\text{tr}} = \bar{D}_{\text{free}}$ (corner) and a discrete jump in N, or (ii) lie in the paid region where (10) holds with $C'(D_{\text{tr}}) = P_{\text{tok}}$.

Piece-wise increasing data cost. Finally, we consider a scenario where the cost of data grows at an non-constant, increasing pace past a certain point, like a power-law or polynomial growth to reflect diminishing returns or increasing costs:

$$C(D_{\text{tr}}) = \begin{cases} 0, & \text{if } D_{\text{tr}} \leq \bar{D}_{\text{free}} \\ P_{\text{tok}}\left(a(D_{\text{tr}} - \bar{D}_{\text{free}})^b\right), & \text{if } D_{\text{tr}} > \bar{D}_{\text{free}} \end{cases} \tag{14}$$

where, a is a scaling coefficient and b captures the rate of increase in cost with respect to non-free data volume (if $b = 1$ cost is linearly increasing in data and if $b > 1$ increases non-linearly). We note that this polynomial form also captures our different scenarios above where $a = 0$ reduces Equation 14 to the free data base case and setting $a = 1$ and $b = 1$ corresponds to the piece-wise uniform case.

3.5 Predictions

The theoretical model delivers the following set of testable predictions about the way optimal model size and training data scale with economic primitives such as data cost, compute prices, and inference demand.

1. **Substitution under rising marginal data cost.** Whenever the marginal cost of training data increases—due to higher token prices, licensing frictions, or reduced data availability—the model predicts a systematic shift toward parameter–intensive solutions:

$$P_{\text{tok}} \uparrow \quad \Rightarrow \quad N \uparrow, \quad D_{tr} \downarrow .$$

   This is a robust consequence of the first-order condition and does not depend on functional-form assumptions. Empirically, this implies that regimes with higher data prices will exhibit larger parameter counts and shorter training sets for models targeting the same loss.

2. **Inference demand shifts optimal scale.** The model predicts that heavier inference usage decreases optimal model size and increases the optimal volume of training data:

$$D_{\text{inf}} \uparrow \quad \Rightarrow \quad N \downarrow, \quad D_{tr} \uparrow .$$

   This implies that inference-heavy deployments should favor smaller models trained on larger datasets rather than larger deployed models. Because inference demand magnifies the recurring cost of parameters but not the one-time cost of training data, quality improvements for high-traffic systems are most efficiently achieved through data expansion, refinement, and distillation, while model size should be capped early.

3. **Kinks and discontinuities at data-supply thresholds.** When data cost is piecewise (free up to a threshold and costly thereafter), the model predicts a discontinuous "jump" in the optimal allocation at the point where marginal data cost becomes positive. The expansion path bends sharply at $\bar{D}_{\text{free}}$, leading to regime switches from data-heavy to parameter-heavy configurations. In practice, we should observe distinct scaling regimes depending on whether a training run exhausts the pool of free or low-cost tokens.

4. **Amplified substitution when marginal data cost is convex.** If the marginal cost of data increases with dataset size ($C''(D_{tr}) > 0$), then the response to data scarcity becomes even stronger:

$$\frac{\partial \log(N^*/D_{tr}^*)}{\partial \log P_{\text{tok}}} \quad \text{is strictly larger when} \quad C''(D_{tr}) > 0.$$

Empirically, this predicts that organizations facing steeply rising data-curation costs (e.g., deduplication, filtering, or high-quality human annotation) will favor more parameter-intensive scaling than organizations with access to abundant, low-curvature data sources.

5. **Divergence from Chinchilla in economically relevant regimes.** The model predicts that the classical compute-optimal scaling rule (Chinchilla) is a knife-edge benchmark that applies only when both data and inference are effectively free:

$$C'(D_{tr}) = 0, \quad D_{\text{inf}} \approx 0 \quad \Rightarrow \quad (N^*, D_{tr}^*) \text{ matches Chinchilla.}$$

As soon as either constraint becomes binding, the economic optimum departs from Chinchilla, generally implying larger models and fewer tokens. Empirically, this predicts systematic deviations from Chinchilla in real-world training runs where data costs, licensing, or inference demand are non-negligible.

4 Experiments & Results

4.1 Experimental Setup

Following Sardana and Frankle (2023), we perform numerical experiments to evaluate our optimization framework using their reference value ranges. We make a number of similar assumptions for the approximation of compute cost of training and inference as well as inference input and output data. Taking values from Sardana and Frankle (2023) as guidance, we define the approximate training cost per FLOP as $C_{\text{train}}/MFU_{\text{train}}$ for training and as $C_{\text{inf}}[D_{\text{input}}/MFU_{\text{input}} + D_{\text{output}}/MFU_{\text{output}}]$ for inference where MFU stands for Model FLOPs Utilization; MFU takes values of 50% for both training and inference input but takes 1% for inference output. For compute costs of training and inference we assume, similarly, $C_{\text{train}} = 1.34 \times 10^{-18}$ dollars per FLOP and $C_{\text{inf}} = 4.9 \times 10^{-19}$ dollars per FLOP.

We design each of the experimental conditions to be a combination of three levels of data costs per token and three levels of inference demand, distinct enough to represent high (H), medium (M), and low (L) values each, as presented in Table 1 below.[2]

[2]Due to lack of publicly available and reliable data on data costs and inference demand, we look to sources like

	Low (L)	Medium (M)	High (H)
(Per) Token Cost ($)	10^{-12}	10^{-6}	10^{-2}
Inference Demand (#)	100 K	500 M	35.1 B

Table 1: The range of per token cost values (in dollars per token) and inference demand levels (# requests) used in our experiments. For tractability, we explore three ranges of values: high (H), medium (M), and low (L) that attempt to reflect sufficiently distinct ranges. K denotes thousand, M million, and B billion.

For inference demand levels we considered low condition to be associated with niche, specialized entreprise-only class of models, with 100K requests. Medium level of 500M inference requests is typically associated with consumer-facing models, powering for instance chatbots. High levels of inference requests correspond to models exposed to global, very large regional or platform specific users, for instance OpenAI's ChatGPT or China's Doubao. Although the overall lifetime of the model might necessitate higher values, we also take into account what is typical for present-day LLM deployments where version updates happen more frequently.

Moreover, we assume models are dense (no Mixture of Experts (MoE)) and we target a fixed training-loss. For the range of the model size (i.e., number of parameters N), we focus on what is possibly or potentially achievable in real-world scenarios for the loss/performance range of interest (between 10^6 and 10^{17}). We consider loss values covering a slightly wider range than the Chinchilla-scaled (Hoffmann et al., 2022), ranging from 1.8 to 2.6. For our results, we use the common pre-training loss value of 1.89 (Hoffmann et al., 2022) and the optimal model (71.69B) and data (4.40 T) sizes backed out from the Chinchilla scaling laws for said loss.

Finally, we parameterize the economic cost of acquiring/curating/licensing one training token (not API inference tokens). We design three levels by triangulating public content licensing deals, labeling/curation benchmarks, and the API token market (used as an upper bound proxy for scarce/exclusive content). For data cost scenarios with free and non-free data, we consider free tokens amount to be 600 billion which is close to the 400 billion tokens used to train GPT-like LLMs such as GPT-J/Neo from The Pile and a close lower bound for LLMs (Komatsuzaki, 2021). For the polynomial representation of piece-wise increasing data cost we assume $a = 1$ and $b = 1.2$. Calculations from this optimization takes up to a few minutes at most on an Intel Xeon CPU with 2 vCPUs (virtual CPUs) and 13GB of RAM.

Technology (2023) and AIPRM (2024) to inform our choice of data/demand values based on estimates on usage, data annotation costs, etc. Based on these estimates, realistic per unit/token data costs sit between our medium and high scenarios while inference demand can vary across all scenarios depending on the scale/popularity of the enterprise that owns the model.

4.2 Base Case: Chinchilla Scaling

Inf. Request (#)	Token Cost ($)	Total Cost ($)
100K	10^{-12}	5.06M
100K	10^{-6}	9.46M
100K	10^{-2}	44.03B
500M	10^{-12}	5.82M
500M	10^{-6}	10.22M
500M	10^{-2}	44.03B
35.1B	10^{-12}	58.39M
35.1B	10^{-6}	62.79M
35.1B	10^{-2}	44.09B

Table 2: Total costs implied from traditional Chinchilla scaling (Hoffmann et al., 2022) and Eq. 12 using the inference requests, token costs, etc. assumed in the experimental setup section.

4.3 Model Quality

We first examine how our different cost structures and scenarios for data affect scaling behavior in terms of model quality (i.e., performance in terms of training loss). Figure 2 provides an overview of these expansion paths.

Uniform cost scenario. Overall, we see that even with uniform data costs (2b, top-right), expansion paths and scaling can vary considerably depending upon inference demand and the cost of training data.

In the case of low inference demand and low token cost (Figure 2b), we see that there is virtually no difference in the expansion paths implied by Chinchilla scaling (near overlap of the two expansion paths); the costs presented by inference demand and training tokens are sufficiently negligible such that guidance under traditional Chinchilla-like scaling remains optimal. However, as we keep inference demand low and increase the token cost to medium (M) and high (H) levels, expansion paths move up and towards the left so that, to achieve the same level of model performance, the optimal choices involve scaling back the amount of data and scaling up the model's number of parameters. As shown in Figure 2b, the rate at which the model size needs to scale up to compensate in order to stay at the same loss/performance level is much larger than the amount at which data scales down. For a medium level of inference demand, as token costs increase, the optimal choice is similarly to cut back on data but scale up model size disproportionately. Lastly, for high inference demand, expansion paths for low and medium level token costs tend towards the

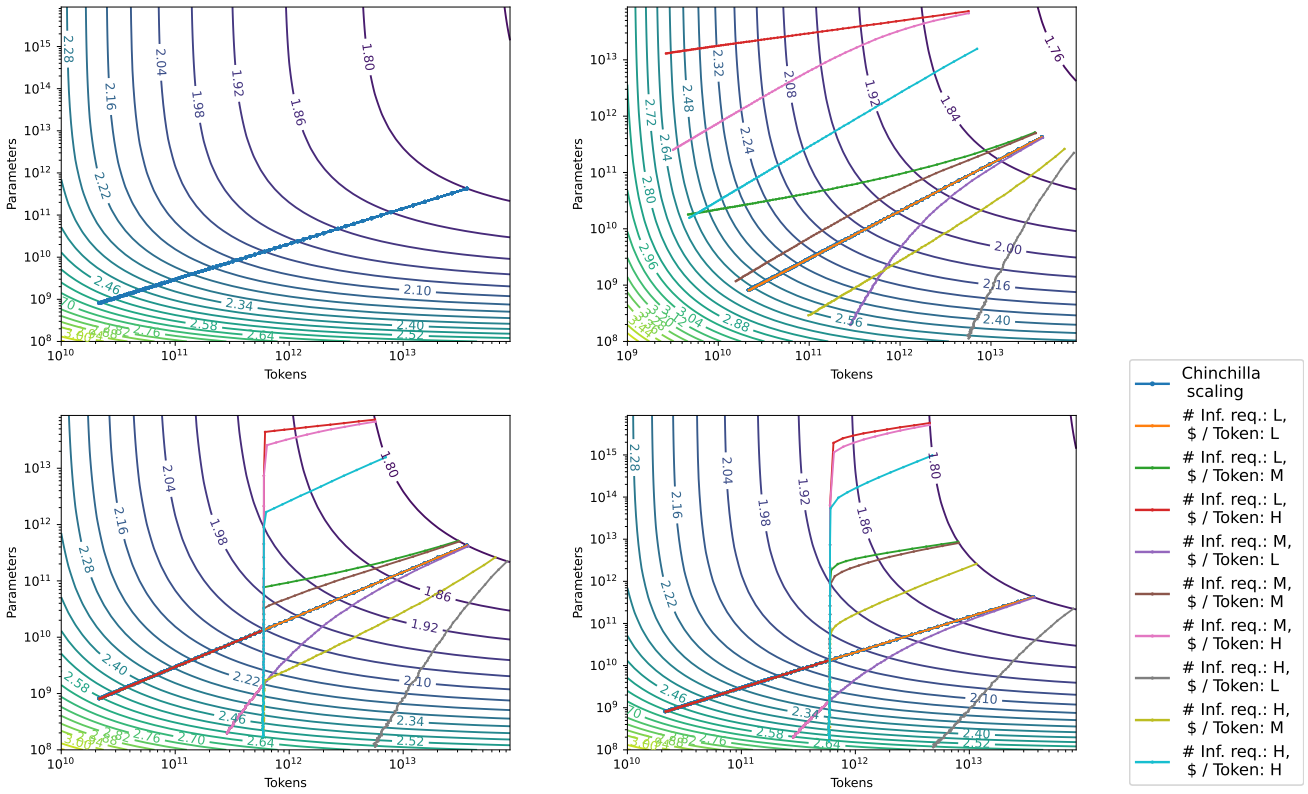


Figure 2: Optimal Expansion Paths for Model Performance. Expansion paths corresponding to the base case of Chinchilla scaling (2a, top-left), uniform data cost (2b, top-right), piece-wise uniform (2c, bottom-left), and piece-wise increasing cost (2d, bottom-right) under different amounts of inference demand (# inf. req.) and token cost ($ / Token). Contours represent iso-loss curves where every point on the same curve represents parameter and data choices corresponding to the same level of quality/performance.

bottom right of Figure 2b. Like the other cases, increasing token costs push expansion paths towards the top-left, similarly reflecting expansion paths that have larger models (more parameters) and fewer tokens to achieve the same loss.

Piece-wise cost scenarios. For piece-wise uniform costs, one notable difference (Figure 2c) is the presence of a discontinuity in several expansion paths. These behaviors are **likely** caused by discontinuities in our piece-wise cost scenario where data is free—but only up to a certain point beyond which a constant/uniform cost per token kicks in. Although increasing token costs from L to H continues to shift expansion paths upwards (i.e., less data, larger model) like before, the discontinuity suggests that certain inference demand and token cost circumstances lead to expansion paths that contain a regime where no further scaling up of data is optimal if token costs are sufficiently high or grow sufficiently quickly; instead, relative to Chinchilla's expansion path and for a given/fixed level of performance or quality, the strategy is to cut back on training data and

significantly increase the model's parameters. The piece-wise increasing cost scenario (Figure 2d) is largely similar to the piece-wise uniform cost scenario with the exception that the behavior is more pronounced and distinct at each level of inference demand and token cost level.

4.4 Model Cost

Having focused on model quality in the previous sub-section, we now turn to characterizing expansion paths in terms of total cost.

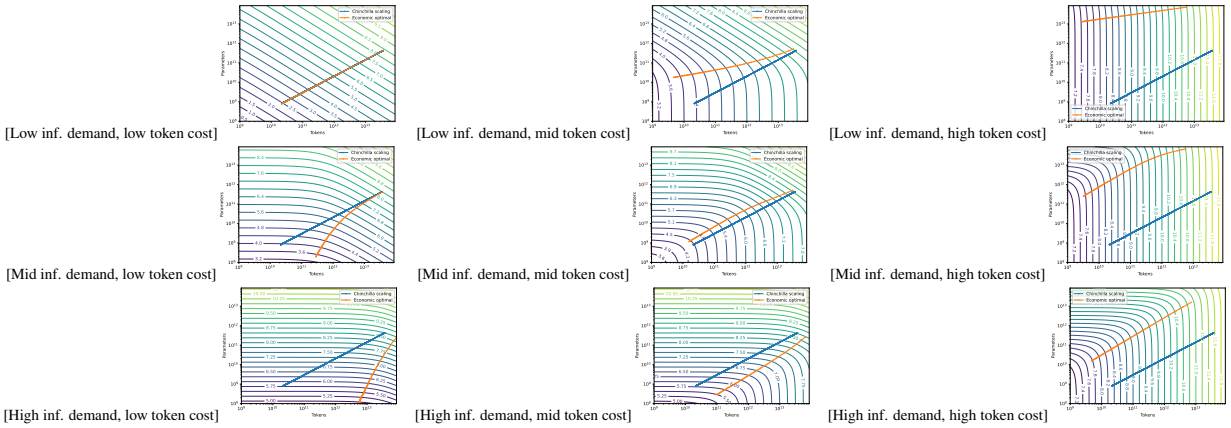


Figure 3: Expansion paths under the uniform data cost scenario in log cost space for different configurations of inference demand (# of inference requests) and token price/cost. The blue path corresponds to Chinchilla scaling and the orange to cost-aware scaling. Contours correspond to total cost in $\log_{10}$ dollars where each contour is an iso-cost curve (i.e., any point on the same contour cost the same).

Uniform cost scenario. For the uniform cost scenario where each unit of training data can be acquired at constant cost, Figure 3 illustrates the differences in cost expansion curves across different regimes of inference demand and token cost. Under the low inference demand regime (Figures 3(a-c)), we see that, as expected, sufficiently low token cost produces an expansion path that is near identical to that produced under traditional Chinchilla scaling; if token cost and inference demand are sufficiently low or negligible, then the problem simply reduces to Chinchilla scaling.

As inference increases to medium levels of demand (Figures 3(d-f)), low token costs imply an expansion path, for a given fixed level of costs, consisting of optimal choices that have more training tokens but approximately the same number of parameters within an order of magnitude (Figure 3d). However, as token cost increases, for the same total cost level, we instead see significant increases in model size under economically optimal scaling relative to Chinchilla while the amount of data between the two stays approximately the same. High levels of inference demand exhibit

Inference Request (#)	Token Price ($)	EC Params (#)	EC Tokens (#)	EC Total Cost ($)	Cost Reduction (%)
100K	10^{-12}	71.69 B	4.40 T	5.06 M	0.0%
100K	10^{-6}	138.24 B	2.55 T	8.20 M	13.31%
100K	10^{-2}	44.11 T	614.21 B	6.58 B	85.06%
500M	10^{-12}	59.89 B	5.33 T	5.75 M	1.2%
500M	10^{-6}	110.40 B	3.01 T	9.50 M	7.08%
500M	10^{-2}	25.71 T	638.92 B	6.92 B	84.27%
35.1B	10^{-12}	21.67 B	29.70 T	26.44 M	54.72%
35.1B	10^{-6}	33.68 B	11.82 T	43.25 M	31.12%
35.1B	10^{-2}	2.29 T	874.28 B	10.48 B	76.24%

Table 3: Economically-optimal (EC) scaling with *Uniform Data Costs*. We show the optimal amount of data and model size under economically-optimal scaling—as well as the total costs these choices imply—under a variety of inference request numbers and per-token prices for a 70B parameter model. Cost reduction is calculated as (CC Total Cost – EC Total Cost)/(CC Total Cost) where CC Total Cost corresponds to the total cost derived from traditional Chinchilla scaling (Table 2). Note: K denotes thousand, M million, B billion, and T trillion.

similar behaviors with the exception that only sufficiently high token costs moves the economically optimal expansion path above Chinchilla's (Figure 3i).

Table 3 summarizes these implications: generally, within each inference demand level (e.g., 100K), rising token cost results in Chinchilla significantly over-estimating the optimal amount of training tokens and total cost—with economically-optimal scaling implying cost reductions of as high as 85% in some cases. Even for low token cost, as inference demand itself rises to 35B, we see considerable increases in Chinchilla's over-estimation as well as growing cost reduction from following economically-optimal scaling instead when accounting for data costs and inference.

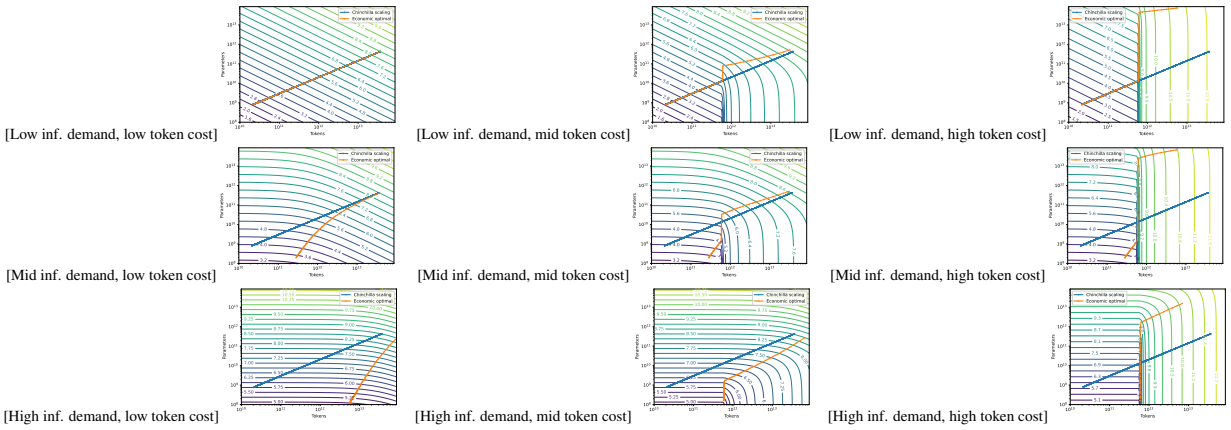


Figure 4: Expansion paths under the piece-wise uniform data cost scenario in log cost space for different configurations of inference demand (# of inference requests) and token price/cost. The blue path corresponds to Chinchilla scaling and the orange to cost-aware scaling. Contours correspond to total cost in $\log_{10}$ dollars where each contour is an iso-cost curve (i.e., points on the same contour cost the same).

Piece-wise cost scenarios. Figures 4 and 5 show cost expansion paths under piece-wise uniform data costs and piece-wise increasing data costs, respectively. We see similar discontinuities as before due to the piece-wise cost structure of our scenarios. Under the piece-wise uniform cost scenario, we see dis-continuous expansion paths across all inference demand levels—so long as token costs become sufficiently high. At those levels, past a certain amount of training tokens ($\approx 10^{12}$), we see that abrupt jumps characterizing increases in model size as data becomes more expensive; at high token costs, we see even sharper cost contours and jumps in the economically-optimal expansion path, illustrating that, for the same total cost, the optimal choice would be substitute towards significantly scaling up model size, keeping data size about the same if not slightly less. The piece-wise increasing data cost scenario tells the same story but with even more pronounced jumps and increases.

Inference Request (#)	Token Price ($)	EC Params (#)	EC Tokens (#)	EC Total Cost ($)	Cost Reduction (%)
100K	10^{-12}	71.69 B	4.40 T	5.06 M	0.0%
100K	10^{-6}	138.24 B	2.55 T	7.60 M	14.21%
100K	10^{-2}	44.11 T	614.21 B	576.36 M	98.48%
500M	10^{-12}	59.89 B	5.33 T	5.75 M	1.2%
500M	10^{-6}	110.40 B	3.01 T	8.90 M	7.52%
500M	10^{-2}	25.71 T	638.92 B	924.99 M	97.57%
35.1B	10^{-12}	21.67 B	29.70 T	26.44 M	54.72%
35.1B	10^{-6}	33.68 B	11.82 T	42.65 M	31.42%
35.1B	10^{-2}	2.29 T	874.28 B	4.48 B	88.25%

Table 4: Economically-optimal (EC) scaling with *Piece-wise Uniform Data Costs.* We show the optimal amount of data and model size under economically-optimal scaling—as well as the total costs these choices imply—under a variety of inference request numbers and per-token prices for a 70B parameter model. Cost reduction is calculated as (CC Total Cost – EC Total Cost)/(CC Total Cost) where CC Total Cost corresponds to the total cost derived from traditional Chinchilla scaling (Table 2). Note: K denotes thousand, M million, B billion, and T trillion.

Inference Request (#)	Token Price ($)	EC Params (#)	EC Tokens (#)	EC Total Cost ($)	Cost Reduction (%)
100K	10^{-12}	71.69 B	4.40 T	5.06 M	0.0%
100K	10^{-6}	3.02 T	829.71 B	83.20 M	93.36%
100K	10^{-2}	64.36 T	599.98 B	618.93 M	100.0%
500M	10^{-12}	59.89 B	5.33 T	5.75 M	1.19%
500M	10^{-6}	2.02 T	896.92 B	108.86 M	91.32%
500M	10^{-2}	64.36 T	599.98 B	1.30 B	99.99%
35.1B	10^{-12}	21.67 B	29.70 T	26.45 M	54.7%
35.1B	10^{-6}	305.05 B	1.63 T	494.74 M	62.12%
35.1B	10^{-2}	54.25 T	606.14 B	46.45 B	99.63%

Table 5: Economically-optimal (EC) scaling with *Increasing Data Costs.* We show the optimal amount of data and model size under economically-optimal scaling—as well as the total costs these choices imply—under a variety of inference request numbers and per-token prices for a 70B parameter model. Cost reduction is calculated as (CC Total Cost – EC Total Cost)/(CC Total Cost) where CC Total Cost corresponds to the total cost derived from traditional Chinchilla scaling (Table 2). Note: K denotes thousand, M million, B billion, and T trillion.

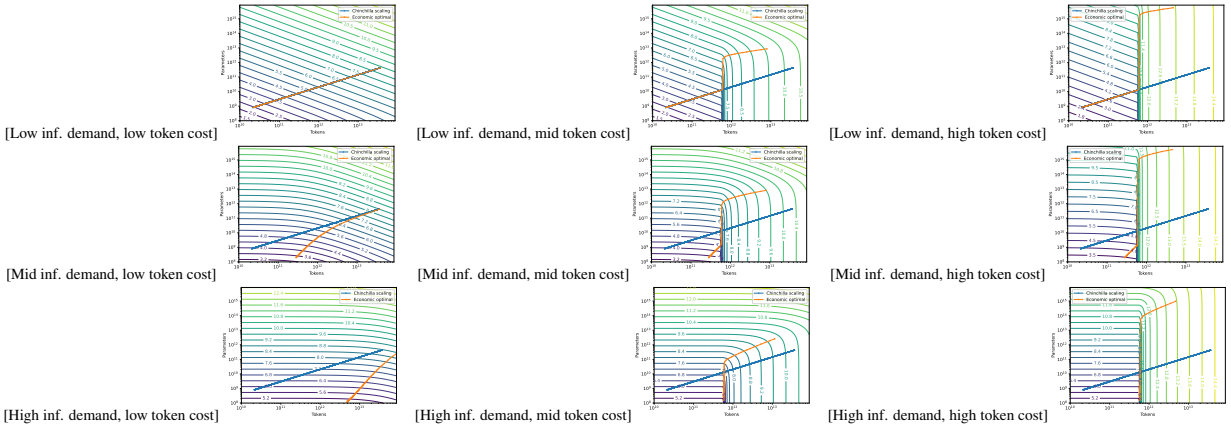


Figure 5: Expansion paths under the piece-wise increasing data cost scenario in log cost space for different configurations of inference demand (# of inference requests) and token price/cost. The blue path corresponds to Chinchilla scaling and the orange to cost-aware scaling. Contours correspond to total cost in $\log_{10}$ dollars where each contour is an iso-cost curve (i.e., any point on the same contour cost the same).

Tables 4 and 5 illustrate the differences in total cost and optimal model/data choices in more detail. Under the piece-wise cost scenarios, we see Chinchilla implying even higher over-estimates than the uniform cost scenario (Table 3) with the piece-wise increasing cost scenario showing Chinchilla's most significant over-estimates overall. In some cases, we see cost reductions of virtually $> 99\%$ if one were to follow economically-optimal scaling instead of Chinchilla scaling.

Theoretical predictions–experiment mapping. The experimental simulations provide a direct quantitative validation of the model's predictions. Proposition 1 predicts that increases in the marginal cost of training data induce substitution toward parameters and away from tokens; empirically, this appears as upward–leftward rotations of the optimal expansion path as token prices rise (Figures 2b–d; Tables 3–5). Proposition 2 shows that higher inference demand penalizes parameters because their cost is incurred repeatedly at deployment; consistent with this prediction, holding the target loss fixed, scenarios with higher inference usage optimally rely on more training data and fewer parameters (Figures 2b; Table 3). The model further predicts kinks and discrete jumps in optimal scale when data costs are non-smooth; these regime changes are visible at the free-token threshold in Figures 2c and 4. The prediction that substitution away from data is amplified when marginal data costs are convex is also consistent with our simulations. We observe that piecewise increasing data costs exhibit substantially steeper expansion paths, larger deviations from Chinchilla scaling, and the largest cost savings across all scenarios (Figures 2d and 5; Table 5). Finally, the knife-edge equivalence between economic-optimal and Chinchilla scaling arises only when data are free and inference demand is negligible, a condition under which the simulated paths coincide (Figure 2a). Taken together, the experiments confirm that the observed deviations

from Chinchilla scaling are not artifacts of calibration but arise from the structural trade-offs highlighted by the theory.

5 Discussion, Implications and Limitations

Our analysis reveals systematic departures from compute-optimal scaling when economic considerations, particularly data costs and inference demand enter the optimization problem. In this section, we discuss the implications of these findings for practitioners, market structure, and policy, while also clarifying the scope and limitations of our framework.

5.1 Implications for LLM training and deployment

Resource allocation under budget constraints. Organizations developing LLMs operate under binding resource constraints that traditional scaling laws do not address. Our framework provides actionable guidance: when data acquisition costs are non-trivial, the economically optimal strategy shifts toward parameter-intensive configurations. Tables 3–5 quantify these effects, showing that adhering to Chinchilla-style scaling can result in cost overruns exceeding 90% in high data-cost regimes. For practitioners, this suggests that the "optimal" model size depends critically on the cost structure facing the organization, not merely on available compute.

Inference-aware training decisions. Proposition 2 formalizes an intuition increasingly adopted by industry: models destined for high-volume deployment should be trained differently than research prototypes. Our comparative statics show that higher anticipated inference demand D_{inf} shifts the optimum toward larger models trained on relatively fewer tokens. This prediction aligns with recent industry practice, for instance Meta's decision to train Llama 3 8B on 15 trillion tokens (Meta AI, 2024), approximately 75 times the Chinchilla-optimal quantity for that model size, explicitly citing inference efficiency as the motivation.

Regime-dependent strategies. The discontinuities in our piece-wise cost scenarios (Figures 4–5) suggest that organizations may face discrete strategic choices rather than smooth trade-offs. When data costs exhibit threshold effects, that is free up to some quantity $\bar{D}_{\text{free}}$ then costly, optimal scaling paths can jump discontinuously. Practitioners should anticipate such regime switches when planning data acquisition strategies, particularly as freely available high-quality data becomes exhausted.

5.2 Market Structure and Competition

Our framework carries implications for competitive dynamics in the LLM market that could inspire further research.

Asymmetric data access as a barrier to entry. If marginal data costs $C'(D)$ differ across firms due to proprietary data assets, existing licensing agreements, or vertical integration with content providers, then firms face fundamentally different optimization landscapes. Proposition 1 implies that a firm with lower marginal data costs will, ceteris paribus, optimally train on more data and potentially achieve better model quality at the same total cost. This asymmetry could entrench incumbents who have amassed data advantages, suggesting that data access, not merely compute capacity, may constitute a durable competitive advantage.

Strategic over-investment in parameters. Our model assumes cost minimization, but firms may deviate strategically. A firm anticipating future data scarcity might over-invest in model parameters today, building capabilities that competitors cannot replicate once data costs rise. Similarly, firms may use large parameter counts as signals of capability to attract customers or talent, even when smaller models would be economically efficient. Extending our framework to incorporate strategic behavior represents an interesting avenue for future work.

Vertical integration incentives. The knife-edge condition in Proposition 3 (Equation 11) clarifies when data costs materially affect optimal scaling. As $C'(D)$ rises, the returns to securing cheaper data sources increase. This provides a formal rationale for the wave of content licensing agreements observed in the industry, such as OpenAI's partnerships with news organizations and Reddit Chapman (2024)as firms seek to shift their data cost curves downward.

5.3 Policy Implications for Data Governance and Access

Our results suggest that policies affecting data availability and cost can significantly influence the trajectory of LLM development. Restrictive data access regime, whether through copyright enforcement, privacy regulation, or platform terms of service, effectively raise $C'(D)$, which our model predicts will push development toward parameter-heavy, data-light configurations (Proposition 1). Policymakers should consider whether such shifts align with broader objectives for AI development, including safety, capability distribution, and environmental impact. In particular, policymakers could consider this research in designing regulations and frameworks around data governance and intellectual property rights in the AI domain. With the growing need for high-quality, diverse datasets, policies that facilitate more sustainable and equitable data sharing could

alleviate cost pressures in LLM training, making it more accessible to a wider range of organizations.

5.4 Limitations

We acknowledge several limitations that qualify the interpretation of our results.

Reliance on simulated cost parameters. In the absence of disclosed data on actual training data costs, our experiments rely on assumed cost scenarios spanning several orders of magnitude (Table 1). While we have designed these scenarios to bracket plausible real-world values based on available estimates of data annotation, licensing, and curation costs, we cannot validate that our parameter choices reflect any specific organization's cost structure. Our results should therefore be interpreted as characterizing the *sensitivity* of optimal scaling to data costs rather than as precise prescriptions.

Static optimization framework. We model the firm's problem as a one-shot optimization (Equation 5), whereas real-world LLM development is dynamic: models are iteratively improved, data pipelines evolve, and compute prices decline over time. A richer model would incorporate intertemporal optimization, option values associated with data acquisition, and learning-by-doing effects. Our static framework provides a useful baseline but does not capture these dynamic considerations.

Single-objective focus. Our optimization focuses on cost minimization subject to a quality constraint (Equation 5). In practice, organizations balance multiple objectives including safety, alignment, latency, interpretability, and robustness that may trade off against cost efficiency. A multi-objective extension of our framework could provide richer guidance for practitioners navigating these trade-offs.

6 Conclusion

As high-quality data becomes increasingly important fuel for LLM, and foundation model, development, their costs should not be neglected. Yet, the lack of transparency or record-keeping means no real data on the costs of data acquisition, processing, curation, and utilization for model development exists. Until then, simulations and projections (like our work) can help but may not provide the most effective form of grounded guidance on the resource trade-offs and costs involved. We hope this work spurs more study and transparency on the growing costs of the never-ending

search for sufficiently high-quality or diverse data especially as traditional sources grow scarce or over-used. As the field moves towards even larger and more capable models, optimizing the balance between compute and data is essential for ensuring that AI advancements remain sustainable, efficient, and accessible.

References

AIPRM. 100+ chatgpt statistics 2024, 2024. URL https://www.aiprm.com/chatgpt-statistics/. Accessed: 2025-02-10.

A. I. Alliance and ClearML. Ai infrastructure survey report. https://ai-infrastructure.org/wp-content/uploads/2023/09/AIIA-ClearML-Survey-Report-Sept-2023.pdf, September 2023. Accessed: 2024-08-16.

H. Bansal, A. Hosseini, R. Agarwal, V. Q. Tran, and M. Kazemi. Smaller, weaker, yet better: Training llm reasoners via compute-optimal sampling, 2024. URL https://arxiv.org/abs/2408.16737.

M. Chapman. Openai, reddit teaming in deal that will bring reddit's content to chatgpt, 2024. URL https://apnews.com/article/reddit-openai-chatgpt-bd2291fcc226bc737a44dbef4a31563f.

Z. Chen, S. Wang, T. Xiao, Y. Wang, S. Chen, X. Cai, J. He, and J. Wang. Revisiting scaling laws for language models: The role of data quality and training strategies. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23881–23899, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1163. URL https://aclanthology.org/2025.acl-long.1163/.

E. Dohmatob, Y. Feng, P. Yang, F. Charton, and J. Kempe. A tale of tails: Model collapse as a change of scaling laws, 2024.

L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020.

D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish. Scaling laws for transfer, 2021.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training compute-optimal large language models, 2022.

S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao, X. Zhang, Z. L. Thai, K. Zhang, C. Wang, Y. Yao, C. Zhao, J. Zhou, J. Cai, Z. Zhai, N. Ding, C. Jia, G. Zeng, D. Li, Z. Liu, and M. Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024.

B. Isik, N. Ponomareva, H. Hazimeh, D. Paparas, S. Vassilvitskii, and S. Koyejo. Scaling laws for downstream task performance of large language models, 2024.

N. Kandpal and C. Raffel. Position: The most expensive part of an llm should be its training data, 2025. URL https://arxiv.org/abs/2504.12427.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.

A. Komatsuzaki. Gpt-j-6b: 6b jax-based transformer. https://arankomatsuzaki.wordpress.com/2021/06/04/gpt-j/, June 2021.

M. Marion, A. Üstün, L. Pozzobon, A. Wang, M. Fadaee, and S. Hooker. When less is more: Investigating data pruning for pretraining llms at scale, 2023.

Meta AI. Introducing Meta Llama 3: The most capable openly available LLM to date, Apr. 2024. URL https://ai.meta.com/blog/meta-llama-3/.

N. Muennighoff, A. M. Rush, B. Barak, T. L. Scao, A. Piktus, N. Tazi, S. Pyysalo, T. Wolf, and C. Raffel. Scaling data-constrained language models, 2023.

H. Oderinwale and A. Kazlauskas. The economics of ai training data: A research agenda, 2025. URL https://arxiv.org/abs/2510.24990.

OpenAI. Gpt-4 technical report, 2023a.

OpenAI. Partnership with Axel Springer to deepen beneficial use of AI in journalism, 2023b. URL https://openai.com/index/axel-springer-partnership/.

N. Sardana and J. Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws, 2023.

A. Subramanyam, Y. Chen, and R. L. Grossman. Scaling laws revisited: Modeling the role of data quality in language model pretraining, 2025. URL https://arxiv.org/abs/2510.03313.

Y. Tay, M. Dehghani, S. Abnar, H. W. Chung, W. Fedus, J. Rao, S. Narang, V. Q. Tran, D. Yogatama, and D. Metzler. Scaling laws vs model architectures: How does inductive bias influence scaling?, 2022.

K. Technology. Data labeling services price [q3 2023 benchmark], 2023. URL https://kili-technology.com/data-labeling/data-labeling-services-price-q3-2023-benchmark. Accessed: 2025-02-10.

P. Villalobos, J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, and A. Ho. Will we run out of data? an analysis of the limits of scaling datasets in machine learning, 2022.

J. Vipra and S. M. West. Computational power and ai. https://ainowinstitute.org/publication/policy/compute-and-ai, September 2023. Accessed: 2024-08-16.

S. E. Whang, Y. Roh, H. Song, and J.-G. Lee. Data collection and quality challenges in deep learning: a data-centric ai perspective. *The VLDB Journal*, 32:791–813, 2023. doi: 10.1007/s00778-022-00775-9. URL https://link.springer.com/article/10.1007/s00778-022-00775-9.

A Proof of Proposition 1

We restate the key first-order condition in a convenient notation. Recall that at an interior optimum (N,D) the loss constraint

$$\frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} = \bar{\ell} \tag{15}$$

binds, and the N–D first-order condition can be written as

$$\frac{\alpha A/N^{\alpha+1}}{\beta B/D^{\beta+1}} = \frac{TD + ID_{\text{inf}}}{C'(D;\theta) + TN}, \tag{16}$$

where

$$T := P_{\text{F,tr}}\kappa_{\text{tr}} > 0, \qquad I := P_{\text{F,inf}}\kappa_{\text{inf}} > 0,$$

and $C(D;\theta)$ is the (possibly parameterized) data-cost function with marginal cost $C'(D;\theta)$.

For notational convenience, define

$$\Phi(N,D) := \frac{\alpha A N^{-(\alpha+1)}}{\beta B D^{-(\beta+1)}}, \qquad \Psi(N,D;\theta) := \frac{TD + ID_{\text{inf}}}{C'(D;\theta) + TN}.$$

Then the two equations determining (N^*,D^*) are

$$F_1(N,D;\theta) := \Phi(N,D) - \Psi(N,D;\theta) = 0, \tag{17}$$

$$F_2(N,D) := \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} - \bar{\ell} = 0. \tag{18}$$

We will study how (N^*,D^*) responds to changes in parameters that enter via $C'(D;\theta)$ or via D_{inf} under the following three assumptions:

1. *Loss parameters and exponents.* $A,B > 0$ and $\alpha,\beta \in (0,1)$.

2. *Data cost.* For each θ in an open interval $\Theta \subset \mathbb{R}$, the function $C(\cdot;\theta) : \mathbb{R}_+ \to \mathbb{R}_+$ is twice continuously differentiable, nondecreasing and weakly convex:

$$C'(D;\theta) \geq 0, \qquad C''(D;\theta) \geq 0 \quad \forall D > 0.$$

3. *Interior solution and regularity.* For each $\theta \in \Theta$ there exists an interior solution $(N^*(\theta), D^*(\theta))$ to (17)–(18), and the Jacobian

$$J(N, D; \theta) := \begin{pmatrix} F_{1N} & F_{1D} \\ F_{2N} & F_{2D} \end{pmatrix}$$

satisfies

$$\det J(N^*(\theta), D^*(\theta); \theta) \; > \; 0.$$

Assumption (3) is a standard "stability" or regularity condition ensuring that the solution is locally unique and that the Implicit Function Theorem applies.

Lemma 1 (Existence of differentiable policy functions)**.** *Under assumptions* (1)–(3)*, there exist open intervals* $\Theta_0 \subseteq \Theta$ *and* $\mathscr{N}, \mathscr{D} \subset \mathbb{R}_+$ *and continuously differentiable functions*

$$\theta \mapsto N^*(\theta) \in \mathscr{N}, \qquad \theta \mapsto D^*(\theta) \in \mathscr{D},$$

such that for all $\theta \in \Theta_0$ *we have* $F_1(N^*(\theta), D^*(\theta); \theta) = 0$ *and* $F_2(N^*(\theta), D^*(\theta)) = 0$*.*

Proof. By (3), at each solution (N^*, D^*) we have $\det J \neq 0$. Hence the Implicit Function Theorem applies to the system $(F_1, F_2) = (0, 0)$ in a neighborhood of (N^*, D^*, θ), which yields the desired continuously differentiable functions $N(\cdot)$ and $D(\cdot)$ on some neighborhood Θ_0 of θ. □

We differentiate $F_1(N^*(\theta), D^*(\theta); \theta) = 0$ and $F_2(N^*(\theta), D^*(\theta)) = 0$ with respect to θ. Denoting partial derivatives by subscripts we obtain the linear system

$$\begin{pmatrix} F_{1N} & F_{1D} \\ F_{2N} & F_{2D} \end{pmatrix} \begin{pmatrix} \dfrac{dN^*}{d\theta} \\ \dfrac{dD^*}{d\theta} \end{pmatrix} + \begin{pmatrix} F_{1\theta} \\ F_{2\theta} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{19}$$

By construction F_2 does not depend on θ, hence $F_{2\theta} = 0$. We therefore have

$$\begin{pmatrix} \dfrac{dN^*}{d\theta} \\ \dfrac{dD^*}{d\theta} \end{pmatrix} = -J^{-1} \begin{pmatrix} F_{1\theta} \\ 0 \end{pmatrix}.$$

We now compute $F_{1\theta}$. Recall that $F_1 = \Phi - \Psi$ and Φ does not depend on θ, so $F_{1\theta} = -\Psi_\theta$. Using

$$\Psi(N, D; \theta) = \frac{TD + ID_{\text{inf}}}{C'(D; \theta) + TN},$$

we have

$$\Psi_\theta = -\frac{(TD + ID_{\text{inf}})\,\partial_\theta C'(D;\theta)}{\big(C'(D;\theta) + TN\big)^2}.$$

By assumption, $\partial_\theta C'(D;\theta) > 0$ and the numerator $TD + ID_{\text{inf}}$ and denominator are strictly positive, so $\Psi_\theta < 0$ and thus

$$F_{1\theta} = -\Psi_\theta > 0. \tag{20}$$

The Jacobian can be written as

$$J = \begin{pmatrix} F_{1N} & F_{1D} \\ F_{2N} & F_{2D} \end{pmatrix}, \qquad \text{where,} \quad \det J > 0 \quad \text{by assumption (3).}$$

Cramer's rule applied to (19) gives

$$\frac{dN^*}{d\theta} = -\frac{F_{1\theta}F_{2D} - 0 \cdot F_{1D}}{\det J} = -\frac{F_{1\theta}F_{2D}}{\det J}, \tag{21}$$

$$\frac{dD^*}{d\theta} = -\frac{F_{1N} \cdot 0 - F_{2N}F_{1\theta}}{\det J} = \frac{F_{2N}F_{1\theta}}{\det J}. \tag{22}$$

From (18) we have

$$F_{2N} = -\alpha A N^{-(\alpha+1)} < 0, \qquad F_{2D} = -\beta B D^{-(\beta+1)} < 0.$$

Combining the signs:

$$F_{1\theta} > 0, \quad F_{2D} < 0, \quad F_{2N} < 0, \quad \det J > 0.$$

Thus, in (21) the numerator $F_{1\theta}F_{2D}$ is negative, so $-F_{1\theta}F_{2D} > 0$, and dividing by $\det J > 0$ yields

$$\frac{dN^*}{d\theta} > 0.$$

In (22), the numerator $F_{2N}F_{1\theta}$ is negative, and dividing by $\det J > 0$ gives

$$\frac{dD^*}{d\theta} < 0.$$

This establishes the claim in Proposition 1.

B Proof of Proposition 2

Differentiate $F_1(N^*, D^*; D_{\text{inf}}) = 0$ and $F_2(N^*, D^*) = 0$ with respect to D_{inf}. As above we obtain

$$\begin{pmatrix} \dfrac{dN^*}{dD_{\text{inf}}} \\ \dfrac{dD^*}{dD_{\text{inf}}} \end{pmatrix} = -J^{-1} \begin{pmatrix} F_{1,D_{\text{inf}}} \\ 0 \end{pmatrix},$$

where now J is the same Jacobian as before (and has negative determinant by assumption (3)).

Compute $F_{1,D_{\text{inf}}}$. Again $F_1 = \Phi - \Psi$, with Φ independent of D_{inf}, so

$$F_{1,D_{\text{inf}}} = -\Psi_{D_{\text{inf}}}.$$

From the definition of Ψ,

$$\Psi(N, D; D_{\text{inf}}) = \frac{TD + ID_{\text{inf}}}{C'(D;\theta) + TN},$$

we have

$$\Psi_{D_{\text{inf}}} = \frac{I}{C'(D;\theta) + TN} > 0,$$

hence

$$F_{1,D_{\text{inf}}} = -\Psi_{D_{\text{inf}}} < 0. \tag{23}$$

Applying Cramer's rule as before,

$$\frac{dN^*}{dD_{\text{inf}}} = -\frac{F_{1,D_{\text{inf}}} F_{2D} - 0 \cdot F_{1D}}{\det J} = -\frac{F_{1,D_{\text{inf}}} F_{2D}}{\det J}, \tag{24}$$

$$\frac{dD^*}{dD_{\text{inf}}} = -\frac{F_{1N} \cdot 0 - F_{2N} F_{1,D_{\text{inf}}}}{\det J} = \frac{F_{2N} F_{1,D_{\text{inf}}}}{\det J}. \tag{25}$$

From (18), as before $F_{2N} < 0$ and $F_{2D} < 0$. Combining with (23) and $\det J > 0$ we conclude:

$$F_{1,D_{\text{inf}}} < 0, \quad F_{2D} < 0 \;\Rightarrow\; F_{1,D_{\text{inf}}} F_{2D} > 0 \;\Rightarrow\; -\frac{F_{1,D_{\text{inf}}} F_{2D}}{\det J} < 0,$$

so $dN^*/dD_{\text{inf}} < 0$.

Similarly,

$$F_{2N} < 0, \quad F_{1,D_{\text{inf}}} < 0 \;\Rightarrow\; F_{2N} F_{1,D_{\text{inf}}} > 0 \;\Rightarrow\; \frac{F_{2N} F_{1,D_{\text{inf}}}}{\det J} > 0,$$

so $dD^*/dD_{\text{inf}} > 0$.

C Proof of Proposition 3

(i) Equivalence when data are free and there is no inference cost. If $C'(D) \equiv 0$ and $D_{\text{inf}} = 0$, the objective in (E) reduces to TND up to an additive constant, so (E) and (C) have the same objective and the same constraint. Their Lagrangians coincide and hence so do their first-order conditions. Therefore, their interior solution sets coincide for each $\bar{\ell}$, and the expansion paths in (N, D) space are identical.

(ii) Divergence when data or inference are economically relevant. We first characterize the first-order conditions for (C) and (E).

For the Chinchilla problem (C), the Lagrangian is

$$\mathscr{L}^{\text{C}} = TND + \lambda^{\text{C}}\Big(\frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} - \bar{\ell}\Big),$$

yielding FOCs (at an interior optimum)

$$\begin{aligned} \partial_N: &\quad TD = \lambda^{\text{C}} \alpha A N^{-(\alpha+1)}, \\ \partial_D: &\quad TN = \lambda^{\text{C}} \beta B D^{-(\beta+1)}. \end{aligned}$$

Dividing the first equation by the second, we obtain the familiar Chinchilla trade-off condition

$$\frac{\alpha A / N^{\alpha+1}}{\beta B / D^{\beta+1}} = \frac{TD}{TN} = \frac{D}{N}. \tag{26}$$

For the economic problem (E), the Lagrangian is

$$\mathscr{L}^{\text{E}} = C(D) + TND + IND_{\text{inf}} + \lambda^{\text{E}}\Big(\frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} - \bar{\ell}\Big),$$

with FOCs

$$\begin{aligned} \partial_N: &\quad TD + ID_{\text{inf}} = \lambda^{\text{E}} \alpha A N^{-(\alpha+1)}, \\ \partial_D: &\quad C'(D) + TN = \lambda^{\text{E}} \beta B D^{-(\beta+1)}. \end{aligned}$$

Dividing gives the economic trade-off condition

$$\frac{\alpha A / N^{\alpha+1}}{\beta B / D^{\beta+1}} = \frac{TD + ID_{\text{inf}}}{C'(D) + TN}. \tag{27}$$

Now suppose (N, D) is an interior pair that satisfies the FOCs of both problems, i.e. (26) and

(27) simultaneously (for the same (N,D) and some multipliers $\lambda^{\mathrm{C}},\lambda^{\mathrm{E}}$). Equating the right-hand sides of (26) and (27) yields

$$\frac{D}{N}=\frac{TD+ID_{\text{inf}}}{C'(D)+TN}.$$

Cross-multiplying and rearranging, we obtain

$$\big(TD+ID_{\text{inf}}\big)N=D\big(C'(D)+TN\big)\quad\Leftrightarrow\quad TND+IND_{\text{inf}}=C'(D)D+TND,$$

which simplifies to the knife-edge condition

$$C'(D)\,D=IND_{\text{inf}}.$$

This establishes (11).

If $C'(D)\equiv 0$ and $D_{\text{inf}}=0$, (11) holds trivially at any (N,D), and part (i) has already shown that the problems coincide. In any *economically relevant* regime with either $C'(D)>0$ at D or $D_{\text{inf}}>0$, (11) imposes a nontrivial algebraic restriction on (N,D). For a given D_{inf} and cost function $C(\cdot)$, this relation can only hold on a lower-dimensional set of (N,D) pairs (typically a curve, and often at most a discrete set when combined with the loss constraint). In particular, there is no reason for $(N^{\mathrm{C}},D^{\mathrm{C}})$—which solves (26) without regard to $C'(D)$ or D_{inf}—to satisfy (11). Absent this coincidence, the Chinchilla optimum $(N^{\mathrm{C}},D^{\mathrm{C}})$ cannot satisfy the economic FOCs, and vice versa, so the two optima differ.

Thus, as soon as $C'(D)$ is positive at the optimum or $D_{\text{inf}}>0$, the equivalence between Chinchilla and economic scaling requires the knife-edge identity (11), which is generically violated. Hence the Chinchilla and economic expansion paths are distinct for generic cost functions and inference loads.

D Formal exposition that convex data costs amplify substitution

Proposition 4 (Convex marginal data costs amplify substitution)**.** *Fix a target loss $\bar{\ell}>0$ and suppose an interior optimum $(N^*(p),D^*(p))$ solves*

$$\min_{N,D>0}\;C(D;p)+TND+IND_{\text{inf}}\quad s.t.\quad \frac{A}{N^{\alpha}}+\frac{B}{D^{\beta}}=\bar{\ell},$$

where $T:=P_{\mathrm{F,tr}}\kappa_{\mathrm{tr}}>0$ and $I:=P_{\mathrm{F,inf}}\kappa_{\mathrm{inf}}>0$. Assume $C(\cdot;p)$ is twice continuously differentiable in a neighborhood of $D^*(p)$, with $C'(D;p)\geq 0$ and $C''(D;p)\geq 0$.

Let p be a scalar that shifts the marginal data cost pointwise:

$$C'(D;p) = p\,g(D), \qquad g(D) > 0, \quad g'(D) \geq 0, \tag{28}$$

so that $C''(D;p) = p\,g'(D)$. Define the substitution index

$$S(p) \;:=\; \log\!\left(\frac{N^*(p)}{D^*(p)}\right).$$

Then, holding fixed $(A, B, \alpha, \beta, T, I, D_{\inf}, \bar{\ell})$, the local elasticity

$$\mathscr{E}_S(p) \;:=\; \frac{\partial S(p)}{\partial \log p}$$

is (weakly) increasing in the curvature $C''(D^*(p);p) = p\,g'(D^*(p))$. In particular, compared to the linear case $g'(D) \equiv 0$ (so $C'' \equiv 0$), convex marginal data costs $g'(D) > 0$ imply

$$\left|\mathscr{E}_S(p)\right| \;\geq\; \left|\mathscr{E}_S^{\text{lin}}(p)\right|,$$

with strict inequality whenever $g'(D^*(p)) > 0$ and the Jacobian regularity condition holds.

Proof. **Step 1 (FOC system).** For an interior optimum, the key first-order condition can be written as

$$\Phi(N,D) \;=\; \Psi(N,D;p), \tag{29}$$

together with the binding constraint

$$\frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} = \bar{\ell}. \tag{30}$$

Here

$$\Phi(N,D) := \frac{\alpha A N^{-(\alpha+1)}}{\beta B D^{-(\beta+1)}} = K\,\frac{D^{\beta+1}}{N^{\alpha+1}}, \qquad K := \frac{\alpha A}{\beta B} > 0,$$

and

$$\Psi(N,D;p) := \frac{TD + ID_{\inf}}{C'(D;p) + TN}.$$

Define $F_1(N,D;p) := \Phi(N,D) - \Psi(N,D;p)$ and $F_2(N,D) := \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} - \bar{\ell}$. The optimum satisfies $F_1 = 0$ and $F_2 = 0$.

Step 2 (Derivatives and Jacobian). Compute partial derivatives:

$$\Phi_N = -(\alpha+1)\frac{\Phi}{N} < 0, \qquad \Phi_D = (\beta+1)\frac{\Phi}{D} > 0.$$

For Ψ, write $Q := TD + ID_{\text{inf}} > 0$ and $M := C'(D;p) + TN > 0$, so $\Psi = Q/M$. Then

$$\Psi_N = -\frac{TQ}{M^2} < 0, \qquad \Psi_D = \frac{TM - QC''(D;p)}{M^2} = \frac{T}{M} - \frac{QC''(D;p)}{M^2}.$$

Hence Ψ_D is *strictly decreasing* in the curvature $C''(D;p)$.

Let J denote the Jacobian of (F_1, F_2) with respect to (N, D):

$$J = \begin{pmatrix} F_{1N} & F_{1D} \\ F_{2N} & F_{2D} \end{pmatrix} = \begin{pmatrix} \Phi_N - \Psi_N & \Phi_D - \Psi_D \\ -\alpha A N^{-(\alpha+1)} & -\beta B D^{-(\beta+1)} \end{pmatrix}.$$

Assume the regularity condition $\det J \neq 0$ at the optimum, so the Implicit Function Theorem applies.

Step 3 (How p enters). By (28), $C'(D;p) = pg(D)$, so

$$\frac{\partial \Psi}{\partial p} = \frac{\partial}{\partial p}\left(\frac{Q}{pg(D) + TN}\right) = -\frac{Q\,g(D)}{(pg(D) + TN)^2} < 0.$$

Since Φ does not depend on p, we have $\partial_p F_1 = -\partial_p \Psi > 0$ and $\partial_p F_2 = 0$.

Step 4 (Implicit derivatives dN/dp and dD/dp). Differentiating $F_1 = 0$ and $F_2 = 0$ w.r.t. p yields

$$J \begin{pmatrix} \frac{dN^*}{dp} \\ \frac{dD^*}{dp} \end{pmatrix} = -\begin{pmatrix} \partial_p F_1 \\ 0 \end{pmatrix}.$$

By Cramer's rule,

$$\frac{dN^*}{dp} = -\frac{1}{\det J}\begin{vmatrix} \partial_p F_1 & F_{1D} \\ 0 & F_{2D} \end{vmatrix} = -\frac{\partial_p F_1 \, F_{2D}}{\det J}, \tag{31}$$

$$\frac{dD^*}{dp} = -\frac{1}{\det J}\begin{vmatrix} F_{1N} & \partial_p F_1 \\ F_{2N} & 0 \end{vmatrix} = \frac{\partial_p F_1 \, F_{2N}}{\det J}. \tag{32}$$

Note that $F_{2N} < 0$ and $F_{2D} < 0$, while $\partial_p F_1 > 0$.

Step 5 (Substitution index and its elasticity). The substitution index is $S = \log N - \log D$, so

$$\frac{dS}{dp} = \frac{1}{N^*}\frac{dN^*}{dp} - \frac{1}{D^*}\frac{dD^*}{dp} = -\frac{\partial_p F_1}{\det J}\left(\frac{F_{2D}}{N^*} + \frac{F_{2N}}{D^*}\right). \tag{33}$$

Multiplying by p gives $\mathscr{E}_S(p) = p\,dS/dp$.

Crucially, the only place where the *curvature* C'' enters is through $F_{1D} = \Phi_D - \Psi_D$ and thus through $\det J$.

Step 6 (How curvature affects $\det J$**).** Compute

$$\det J = F_{1N}F_{2D} - F_{1D}F_{2N}.$$

Since $F_{2N} < 0$ and $F_{2D} < 0$, increasing C'' lowers Ψ_D, hence raises $F_{1D} = \Phi_D - \Psi_D$. Therefore,

$$\frac{\partial(\det J)}{\partial C''} = -\frac{\partial F_{1D}}{\partial C''}F_{2N} = -\Big(-\frac{\partial \Psi_D}{\partial C''}\Big)F_{2N} = -\Big(\frac{Q}{M^2}\Big)F_{2N} \;<\; 0,$$

because $\frac{\partial \Psi_D}{\partial C''} = -\frac{Q}{M^2} < 0$ and $F_{2N} < 0$. Thus, *higher curvature* C'' *decreases* $\det J$ locally.

Step 7 (Amplification). Inspect (33). The numerator term

$$\partial_p F_1 \left(\frac{F_{2D}}{N^*} + \frac{F_{2N}}{D^*}\right)$$

does not depend on C'' (given (N^*, D^*)), while the denominator $\det J$ is strictly decreasing in C''. Hence the magnitude of dS/dp (and therefore of $\mathscr{E}_S(p) = p\,dS/dp$) is (weakly) increasing in curvature: moving from $C'' = 0$ (linear marginal costs) to $C'' > 0$ (convex marginal costs) *amplifies* the local responsiveness of $S = \log(N^*/D^*)$ to a marginal increase in p.

Strict amplification obtains whenever $g'(D^*) > 0$ (so $C''(D^*; p) > 0$) and $\det J \neq 0$. This establishes the claim. □