The Economics of Large Language Models: Token Allocation, Fine-Tuning, and Optimal Pricing*

Dirk Bergemann† Alessandro Bonatti‡ Alex Smolin§

February 11, 2025

Abstract

We develop an economic framework to analyze the optimal pricing and product design of Large Language Models (LLM). Our framework captures several key features of LLMs: variable operational costs of processing input and output tokens; the ability to customize models through fine-tuning; and high-dimensional user heterogeneity in terms of task requirements and error sensitivity. In our model, a monopolistic seller offers multiple versions of LLMs through a menu of products. The optimal pricing structure depends on whether token allocation across tasks is contractible and whether users face scale constraints. Users with similar aggregate value-scale characteristics choose similar levels of fine-tuning and token consumption. The optimal mechanism can be implemented through menus of two-part tariffs, with higher markups for more intensive users. Our results rationalize observed industry practices such as tiered pricing based on model customization and usage levels.

*Dirk Bergemann gratefully acknowledges financial support from NSF SES 2049754 and ONR MURI. Alex Smolin gratefully acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program (grant ANR-17-EURE-0010) and through the Artificial and Natural Intelligence Toulouse Institute (ANITI).

†Department of Economics, Yale University, dirk.bergemann@yale.edu

‡MIT Sloan, bonatti@mit.edu

§Toulouse School of Economics, alexey.v.smolin@gmail.com

Keywords: Large Language Models, Optimal Pricing, Menu Pricing, Fine-Tuning.

JEL Codes: D47, D82, D83.

1 Introduction

As generative AI continues to evolve, it promises to revolutionize fields ranging from scientific research and creative industries to education and problem solving, offering unprecedented tools for human augmentation and innovation. Pricing access to generative AI tools is a complex challenge, however, as the value of computational resources and the potential economic impact of these services can vary significantly across buyers and use cases.[1]

In this paper, we develop a theoretical framework to analyze the pricing and design of Large Language Models (LLMs). We propose a framework that treats LLMs as flexible, general-purpose technologies whose value to users depends on features like variable operational costs, model size, and the potential for personalization based on user data. Within this environment, we study a monopolist's revenue-maximizing choice of service menus, reflecting a variety of contract forms that may differ in how they allocate and govern computational resources. This comprehensive approach lays the groundwork for connecting observed industry practices to core economic principles and sets the stage for analyzing how advances such as retrieval-augmented generation, chain-of-thought reasoning, and model versioning impact both platform strategy and user welfare.

Our approach to characterizing demand for LLMs is centered on rich, high-dimensional user heterogeneity: LLMs enable informed decision-making across a wide variety of tasks, and users differ in the value of completing each task accurately. The accuracy of the information provided depends on the overall quality of the LLM (e.g., the size of the neural network and the number of output tokens) and the amount of personalized data input by the user (e.g., re-training, queries, fine-tuning). The value of accuracy, however, is the buyer's private information, because it captures the diverse applications of LLMs across various users

[1]For a discussion of the practical challenges associated with LLM pricing, see https://www.wsj.com/articles/no-one-knows-how-to-price-ai-tools-f346ea8a.

and contexts. LLMs serve an expansive range of needs, from generating creative content to conducting complex analyses, each of which delivers different levels of utility depending on the user's objectives, preferences, and opportunity costs. This heterogeneity fundamentally shapes demand and hence the optimal design of LLMs, including their pricing and personalization.

We examine two contracting environments that capture different settings in which users can (or cannot) personalize the computing resources they acquire: one where the service provider has full control over how the compute is spent across tasks, and another where the user can freely allocate compute ("tokens") to various tasks. The former corresponds, for example, to contracting over numbers of servers, API calls, users with access, or differential pricing for the computational resources required for tasks like analytical work, creative writing, coding, and data processing. The latter corresponds to simpler contracts, e.g., OpenAI's pay-as-you-go pricing for models like GPT-4 and GPT-3.5.

Main Results When it is possible to contract on the entire assignment of tokens to tasks, the seller's problem ("Token Allocations") is an infinite-dimensional screening problem, which is well-known to be difficult. We are nonetheless able to make progress in two important classes of environments. The first class is the binary environment. Here the buyer's type–the entire value profile for a continuum of tasks—is binary, i.e., there are two types only (Section 4.1). We find that either the seller can extract full surplus or the one IC and one IR constraints bind (Proposition 2). Under Cobb-Douglas utility, we uncover a useful index representation of the buyer's type that allows us to identify which one is the high type.

In the second class of environments (Section 4.2), which we refer to as "value-scale" heterogeneity, the buyer's type is two-dimensional: the first dimension determines the buyer's value that she attaches to any of her ex ante homogeneous tasks; the second dimension determines the buyer's scale, i.e., the number of tasks the buyer needs to process (Proposition 4). In this case, we derive sufficient conditions under which it suffices for the seller to screen buyers along one dimension only (namely, the value dimension), i.e., the optimal menu obtains revelation of the scale dimension at no cost to the seller.

When only the total number of tokens is contractible ("Token Packages," Section 5), we leverage the tractability of the Cobb-Douglas framework to drastically simplify the problem: the buyer's type–an arbitrary function mapping each task to a marginal value of precision–is summarized by a scalar, indeed, by the same index as in Section 4.1. This is an index that allows the seller to solve a one-dimensional screening problem. The resulting optimal allocation of tokens (Proposition 5) is then measurable with respect to the buyer's index, which greatly simplifies not just the solution, but the optimal mechanism.

Finally, we connect both modeling assumptions and our results to current LLM pricing practice. In particular, we show that all of our optimal menus are implementable through menus of two-part tariffs for different kinds of tokens, which is consistent with current OpenAI pricing models (see Proposition 7 and Section 7). We also argue that the resulting menus and all our results relate to current LLM pricing practices, and we sketch how to extend our framework to accommodate heterogeneous model versions, the use of RAG, and Chain of Thought.

Related Literature Our paper contributes to the emerging and rapidly growing literature on large language models. Existing work largely examines the potential impact of LLMs in economic settings such as price competition (Fish, Gonczarowski, and Shorrer, 2024), token auctions (Duetting, Mirrokni, Paes Leme, Xu, and Zuo, 2024), and sponsored search advertising (Bergemann, Bojko, Dütting, Paes Leme, Xu, and Zuo, 2024). By contrast, our paper focuses on the details of LLM provision by a revenue-maximizing provider—an aspect that has received surprisingly little attention. For example, Mahmood (2024) studies competition among producers of generative AI who can specialize in different tasks, but restricts attention to linear token pricing.

At their core, LLMs are an information technology, so their sale relates to the sale of information as studied by Babaioff, Kleinberg, and Paes Leme (2012); Bergemann, Bonatti, and Smolin (2018); Yang (2022). However, LLMs possess distinctive features driven by their underlying technology. First, they are general-purpose technologies typically employed for multiple tasks. Second, the measurable unit of usage is tokens. Third, more precise information is achieved by combining LLM inputs, outputs, and fine-tuning. These aspects

give rise to a distinct model of revenue maximization and resulting selling mechanisms.

Methodologically, the seller's problem is a multidimensional screening problem with a multidimensional allocation and possible moral hazard on the buyer's part. Such problems are known to be challenging. Even in simpler settings—such as one-dimensional screening with moral hazard (Castro-Pires, Chade, and Swinkels, 2024) or multidimensional screening without moral hazard (Rochet and Stole, 2003; Daskalakis, Deckelbaum, and Tzamos, 2017)—tractability issues arise. Nevertheless, we show that in important environments, the problem admits a reduction and that the optimal solution can be recovered by building on the one-dimensional analysis of Mussa and Rosen (1978).

Finally, our special case of value-scale heterogeneity is reminiscent of the literature on "1.5-dimensional" mechanism design (e.g., Fiat, Goldner, Karlin, and Koutsoupias (2016); Devanur, Goldner, Saxena, Schvartzman, and Weinberg (2020)). As in that literature, our optimal mechanism is obtained by solving a continuum of separate subproblems and linking them together. Unlike most of that literature, however, these separate subproblems may lead to incentive clashes across them even when value and scale are independently distributed, thereby requiring an additional argument regarding the underlying production technology.

2 Model

A decision maker (i.e., a buyer of LLM services) faces a unit measure of tasks indexed by $i \in [0, 1]$. The buyer can decide on (i) how many input tokens to use for each task i, $x_i \geq 0$, (ii) how many output tokens to use for each task, $y_i \geq 0$, and (iii) how many fine-tuning tokens to use to improve the model's performance on *all* tasks, $z \geq 0$.

The precision of the buyer's decision on task i is given by the production function $v(x_i, y_i, z)$. We assume that v is a commonly known production function, strictly increasing, strictly concave, with all cross-partial derivatives $v_{xy}(x_i, y_i, z), v_{xz}(x_i, y_i, z)$, and $v_{yz}(x_i, y_i, z)$ strictly positive.[2] We further assume $v(0, y, z) = v(x, 0, z) = 0$, $v_x(0, y, z) = v_y(x, 0, z) = +\infty$, and $v_x(+\infty, y, z) = v_y(x, +\infty, z) = v_z(x, y, +\infty) = 0$. These properties are different from scaling laws for training LLMs but capture the fundamental assumption that time and

[2]For constants, subscripts refer to variables or tasks; for functions, they indicate partial derivatives.

computing resources spent on a single task exhibit diminishing marginal returns.[3]

Our leading functional-form example is a Cobb-Douglas production function,

$$v(x, y, z) = x^{\alpha} y^{\beta} (b + z)^{\gamma}, \tag{1}$$

where $\alpha, \beta, \gamma > 0$ are *sensitivity parameters* with $\alpha + \beta + \gamma < 1$, and $b > 0$ is a *base level*.

The buyer's tasks are potentially heterogeneous. Their marginal values are captured by a buyer *type* $\boldsymbol{w} = (w_i)_{i \in [0,1]}$, the willingness to pay for each distinct task i:[4]

$$\boldsymbol{w} : [0, 1] \to \mathbb{R}_+, \tag{2}$$

so that using a z-fine-tuned model with a profile of $(x_i)_{i \in [0,1]}$ input tokens and a profile of $(y_i)_{i \in [0,1]}$ output tokens delivers a total payoff:

$$\int_0^1 w_i v(x_i, y_i, z) di. \tag{3}$$

Buyer types are distributed according to a commonly known CDF $F_{\boldsymbol{w}}$.

An important focal case is one in which each type $\boldsymbol{w}$ is characterized by two parameters $w \in [0, 1]$ and $s \in [0, 1]$, and forms a step-function:

$$w_i = \begin{cases} w, \text{ if } i \leq s, \\ 0, \text{ if } i > s. \end{cases} \tag{4}$$

In this case, each type is effectively two dimensional, $(w, s) \in [0, 1]^2$. The first dimension determines the buyer's *value* that she attaches to any of her ex ante homogeneous tasks. The second dimension determines the buyer's *scale*, i.e., the number of tasks the buyer needs to process. We will refer to this case as a *value-scale heterogeneity*.

[3]We do not require $v(x, y, 0) = 0$ or $v_z(x, y, 0) = +\infty$ to capture the fact that even a non-fine-tuned model can be of value and that fine-tuning benefits may be bounded. We also do not necessarily require v to be bounded because the diminishing marginal value and strictly positive token costs ensure that the value never diverges.

[4]For clarity, vector-valued variables are denoted in bold.

Finally, tokens are costly to process. We assume that the marginal costs of processing tokens are constant but can vary across different token types, with the cost of input tokens being $c_x > 0$, output tokens $c_y > 0$, and fine-tuning tokens $c_z > 0$. We discuss the details of the mapping between the model and the LLM industry practices in Section 7.

3 Efficient Solution

Consider a social planner's problem who allocates the token profile $(x_i, y_i)_{i\in[0,1]}$ across tasks and fine-tunes the model with z tokens. Given any given type $\boldsymbol{w} = (w_i)_{i=0}^1$, the efficient allocation solves the problem:

$$\max_{(x_i,y_i)_{i\in[0,1]},z\geq 0} \int_0^1 \left(w_i v(x_i, y_i, z) - c_x x_i - c_y y_i\right) di - c_z z.$$

Because the marginal added value at zero is infinite, every task with $w_i > 0$ is allocated some input and output tokens. Their allocation satisfies the first-order optimality conditions:

$$w_i v'_x(x_i, y_i, z) = c_x, \tag{5}$$

$$w_i v'_y(x_i, y_i, z) = c_y. \tag{6}$$

Whether it is efficient or not to fine-tune the model depends on the aggregate added value. If the optimal fine-tuning level is strictly positive, $z > 0$, then:

$$\int_0^1 w_i v'_z(x_i, y_i, z) di = c_z. \tag{7}$$

In contrast, at a corner solution with $z = 0$:

$$\int_0^1 w_i v'_z(x_i^0, y_i^0, 0) di < c_z, \tag{8}$$

where (x_i^0, y_i^0) solve the first order conditions (5)-(6) at $z = 0$.

A more explicit characterization is obtained in the case of Cobb-Douglas production

function. In that case, the problem can be rewritten as:

$$\max_{x_i, y_i, z \geq 0} \int_0^1 \left(w_i x_i^\alpha y_i^\beta (b+z)^\gamma - c_x x_i - c_y y_i \right) di - c_z z, \tag{9}$$

and can be solved in closed form. We present the explicit characterization in the Appendix, as part of the proof of Proposition 1. Notably, the following CES aggregation index θ defined as:

$$\theta \triangleq \left(\int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} di \right)^{1-\alpha-\beta}, \tag{10}$$

as well as the threshold

$$\hat{\theta} \triangleq b^{1-\alpha-\beta-\gamma} \left(\frac{c_x}{\alpha} \right)^\alpha \left(\frac{c_y}{\beta} \right)^\beta \left(\frac{c_z}{\gamma} \right)^{1-\alpha-\beta}, \tag{11}$$

turn out to be important.

Proposition 1 (Efficient Allocation). *Under Cobb-Douglas production function and efficient allocation, all types with the same θ consume the same number of fine-tuning tokens, consume the same total amount of input and output tokens, and obtain the same total payoff. The optimal number of input and output tokens for task i is proportional to $w_i^{\frac{1}{1-\alpha-\beta}}$. Types fine-tune if and only if $\theta > \hat{\theta}$.*

Therefore, we will refer to θ as a *representative type.* In the value-scale setting, each type (w, s) corresponds to a representative type:

$$\theta = \left(s w^{\frac{1}{1-\alpha-\beta}} \right)^{1-\alpha-\beta} = w\, s^{1-\alpha-\beta}. \tag{12}$$

4 Menus of Token Allocations

We now study the revenue-maximization problem of a monopolist LLM provider. The type $\boldsymbol{w}$ is distributed according to a CDF F. The buyer knows their type, while the seller does not. The players' payoffs are quasilinear in transfers. The seller designs a direct menu that

specifies a token allocation across different tasks:

$$((x_i(\boldsymbol{w}), y_i(\boldsymbol{w}))_{i\in[0,1]}, z(\boldsymbol{w}), T(\boldsymbol{w}))_{\boldsymbol{w}}. \tag{13}$$

The allocation is contractible, which means the buyer has no freedom to reallocate tokens across tasks. This also means the seller's problem is an optimal mechanism design problem with infinitely dimensional private information. Such problems are known to be intractable in full generality. In the following sections, we obtain complete characterizations in two special settings: the case of two types (Section 4.1) and the case of value-scale heterogeneity (Section 4.2).

4.1 Binary Types

Let there be only two types, $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$, which occur with prior probabilities f_1 and f_2 respectively. In this case, the optimal mechanism depends on what happens if the seller attempts to extract the first-best level of surplus, that is, if she offers a menu containing the efficient amounts of tokens for each type with prices equal to their respective added values. If this menu is incentive compatible, then it is clearly optimal, and we associate label H with the type with higher payment. If this menu is not incentive compatible, then we associate label H with the type whose incentive constraint is violated and label L with the other type.

We then obtain a precise characterization for our payoff structure.

Proposition 2 (Binary Types). *In the optimal menu, either (i) the seller extracts full surplus, or (ii) the token allocation is efficient with respect to virtual types* $\varphi(\boldsymbol{w_H}) = \boldsymbol{w_H}$ *and* $\varphi(\boldsymbol{w_L}) = \boldsymbol{w_L} - \frac{f_H}{f_L}(\boldsymbol{w_H} - \boldsymbol{w_L})$.

Proof. A recent result by Haghpanah and Siegel (2024) on the structure of binding constraints for two buyer types in general screening problems establishes that, in the optimal menu: either (i) the seller extracts full surplus; or (ii) the incentive constraint of type $\boldsymbol{w_H}$ and the individual rationality constraint of type $\boldsymbol{w_L}$ bind. In either case, type $\boldsymbol{w_H}$ is served the efficient allocation.

The rest of the proof builds on this result. Denote by $\boldsymbol{q} = (q_i)_{i=0}^1$ the profile of "qualities" delivered to each task, $q_i \triangleq v(x_i, y_i, z)$, and denote by $C(\boldsymbol{q})$ the minimal total cost to generate

a given profile $\boldsymbol{q}$:

$$C(\boldsymbol{q}) \triangleq \min_{x_i,y_i,z\geq 0} \int_0^1 c_x x_i + c_y y_i + c_z z_i \; di,$$
$$\text{s.t. } v(x_i, y_i, z) = q_i, \quad \forall\, i \in [0,1].$$

If the seller cannot extract full surplus, then Haghpanah and Siegel (2024) show that her problem can be written as

$$\max_{\boldsymbol{q_L},\boldsymbol{q_H},T_L,T_H} f_L(T_L - C(\boldsymbol{q_L})) + f_H(T_H - C(\boldsymbol{q_H}))$$
$$\text{s.t. } \int_0^1 w_{Hi} q_{Hi} di - T_H = \int_0^1 w_{Hi} q_{Li} di - T_L,$$
$$\int_0^1 w_{Li} q_{Li} di - T_L = 0.$$

Solving for transfers from the constraints, the problem can be restated as:

$$\max_{\boldsymbol{q_L},\boldsymbol{q_H}} f_L \left(\int_0^1 \left(w_{Li} - \frac{f_H}{f_L}(w_{Hi} - w_{Li}) \right) q_{Li} di - C(\boldsymbol{q_L}) \right) + f_H \left(\int_0^1 w_{Hi} q_{Hi} di - C(\boldsymbol{q_H}) \right).$$

The result follows. □

Note that even though the virtual types in Proposition 2 follow the standard formula of the single-dimensional case, each type there is infinite-dimensional.

To complete the characterization, it is left to determine for any given two types $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ which type is $\boldsymbol{w_H}$, that is which type's constraint, if any, is violated under an attempt of full surplus extraction. If $\boldsymbol{w_2}$ dominates $\boldsymbol{w_1}$ for every task, then it is clearly $\boldsymbol{w_H}$; alternatively, the types are horizontally differentiated and the distinction is less clear. However, in the case of a Cobb-Douglas production function, the ranking is simple and coincides with the ranking of their representative types.

Proposition 3 (Two Types. Cobb-Douglas Case.). *In the case of Cobb-Douglas production function, type* $\boldsymbol{w_H}$ *is the one with higher aggregation index,* $\theta_H \geq \theta_L$*, or equivalently:*

$$\int_0^1 w_{i,H}^{\frac{1}{1-\alpha-\beta}} di \geq \int_0^1 w_{i,L}^{\frac{1}{1-\alpha-\beta}} di.$$

Furthermore, the optimal menu extracts full surplus if and only if

$$\int_0^1 (w_{i,H} - w_{i,L}) w_{i,H}^{\frac{\alpha+\beta}{1-\alpha-\beta}} di \leq 0.$$

One might wonder whether the correspondence of the incentive order and the efficiency order extends beyond binary types. Our next section shows that this is not the case. In the case of value and scale heterogeneity, the incentive constraints that bind at the optimal mechanism do not form a one-dimensional structure. Instead, they form an infinite collection of one-dimensional segments.

4.2 Value-Scale Heterogeneity

We can obtain a richer characterization in the case of value-scale heterogeneity,

$$w_i = \begin{cases} w, \text{ if } i \leq s, \\ 0, \text{ if } i > s, \end{cases} \tag{14}$$

in which w and s are independently distributed according to CDFs F_w and F_s with F_w featuring increasing virtual values. We will argue that in this case, under additional assumptions, the binding incentive constraints are those within each scale, and the seller is able to generate the same revenue as if the scale were observable.

Consider the problem in which the scale is commonly known to be s. The buyer's utility from any given item on the menu (13) is

$$w \int_{i=0}^{s} v(x_i, y_i, z) di - T.$$

It follows that for any reported w the seller should optimize token allocation to deliver a

promised level of (total) quality q,

$$w\,q - T,$$

with the minimal cost function $C(q,s)$ of delivering a given quality being:

$$C(q,s) = \min_{(x_i,y_i)_{i\in[0,1]},z\geq 0} \int_{i=0}^{s} (c_x x_i + c_y y_i)di + c_z z \tag{15}$$

$$\text{s.t.} \int_{i=0}^{s} v(x_i, y_i, z)di = q. \tag{16}$$

Since v is strictly concave, the solution to this problem is achieved by allocating the input and output tokens uniformly across the s tasks. The problem can be equivalently stated as:

$$C(q,s) = \min_{x,y,z\geq 0} s\,c_x x + s\,c_y y + c_z z \tag{17}$$

$$\text{s.t. } s\,v(x,y,z) = q. \tag{18}$$

The resulting cost function $C(q,s)$ satisfies the following properties.

Lemma 1 (Cost Function).

1. *$C(q,s)$ is strictly increasing and strictly convex in q with $C_q(0,s) = 0$.*
2. *$C(q,s)$ is strictly decreasing in s.*
3. *$C(q,s)$ is submodular, i.e., $C_q(q,s)$ is decreasing in s for all q.*

Proof. The first two statements follow directly from our assumptions on the production function $v(\cdot)$ in Section 2. To establish the third result, write the Lagrangean for the cost minimization problem:

$$L = s\,c_x x + s\,c_y y + c_z z + \lambda(q - s\,v(x,y,z)).$$

We wish to show that the partial derivative of the expenditure function $C_q(q,s)$ is decreasing in s. To do so, it suffices to show that the multiplier λ is decreasing in s for a fixed q. Consider

the necessary first-order conditions:

$$s(c_x - \lambda v_x(x, y, z)) = s(c_y - \lambda v_y(x, y, z)) = 0 \text{ and } c_z - \lambda s v_z(x, y, z) \leq 0.$$

Now suppose that λ were to increase following an increase in s. This would require v_x and v_y to decrease. Because v is strictly concave and all its cross-partial derivatives are strictly positive, this implies that both x and y must strictly increase. This means v itself must increase, which contradicts the equality constraint (18).

□

As such, the seller's problem for any given s is analogous to Mussa and Rosen (1978). Since w and s are independently distributed, we can define the virtual value to be $\varphi(w, s) = \varphi(w)$ where:

$$\varphi(w) \triangleq w - \frac{1 - F_w(w)}{f_w(w)}. \tag{19}$$

Since $\varphi(w)$ is increasing, all w with $\varphi(w) \leq 0$ are excluded and all other w receive the quality level $q(w, s)$ that solves:

$$\varphi(w) = C_q(q(w, s), s). \tag{20}$$

The corresponding optimal transfers are

$$T(w, s) = w\, q(w, s) - \int_0^w q(k, s) dk. \tag{21}$$

Lemma 1 shows that it is cheaper to generate an extra unit of (total) quality when you have more tasks. This property is intuitive given that the returns on each task are diminishing. Thus, the optimal quality (and hence the buyer's rent) increase in scale for any given value w.

Assumption 1 (Bounded Rent Increase). For all w, s, the function $q(w, s)$ defined in (20) satisfies $\int_0^w sq_s(k, s)dk \leq wq(w, s)$.

Assumption 1 requires that the buyer's rent does not grow too quickly and, specifically, that the marginal increase of buyer rent $u(w, s)$ from having an additional task is smaller than the average equilibrium value generated by LLM across existing tasks.

Proposition 4 (Optimal Menu of Token Allocations). *Under Assumption 1, an optimal menu is*

$$((x_i(w, s), y_i(w, s))_{i\in[0,1]}, z(w, s), T(w, s))_{(w,s)},$$

where for each (w, s), $(x_i(w, s), y_i(w, s), z(w, s))$ are cost-minimizing tokens from (15) that deliver quality $q(w, s)$ as defined in (20), and $T(w, s)$ as defined in (21).

Proof. If each type reports truthfully, then the menu attains the profits of the observable-scale benchmark and is thus optimal.

If type (w, s) deviates to $(w, \tilde{s})$ with $\tilde{s} \leq s$, then, under the proposed menu, she obtains exactly the same payoff as type $(w, \tilde{s})$, because she processes the same number of tasks with the same willingness to pay for quality. By Lemma 1, $C_q(q, s)$ is decreasing in s for all q, and thus $q(w, s)$ is increasing in s for all w. Therefore, the rents accrued by type (w, s) under truth-telling,

$$U(w, s) = \int_0^w q(k, s)dk,$$

are increasing in s for all w. Therefore, (w, s) does not want to deviate to $(w, \tilde{s})$ with $\tilde{s} \leq s$. Furthermore, by incentive compatibility within a given $\tilde{s}$, $(w, \tilde{s})$ does not want to deviate to $(\tilde{w}, \tilde{s})$. Therefore, (w, s) does not want to deviate to any $(\tilde{w}, \tilde{s})$ with $\tilde{s} \leq s$.

If type (w, s) deviates to $(\tilde{w}, \tilde{s})$ with $\tilde{s} > s$, then she obtains gross utility $wq(\tilde{w}, \tilde{s})\, s/\tilde{s}$ and pays the transfer $T(\tilde{w}, \tilde{s})$. Therefore, the optimal double deviation strategy for a misreporting type solves

$$\max_{\tilde{w}} \left[wq(\tilde{w}, \tilde{s})\frac{s}{\tilde{s}} - \tilde{w}q(\tilde{w}, \tilde{s}) + \int_0^{\tilde{w}} q(k, \tilde{s})dk \right].$$

Because the mechanism incentivizes truthful reporting by any $\tilde{s}$-truthtelling type, including the type $(ws/\tilde{s}, \tilde{s})$, it follows that

$$\tilde{w}^* = \frac{ws}{\tilde{s}} \text{ and hence } U(w; s, \tilde{s}) = \int_0^{\frac{ws}{\tilde{s}}} q(k, \tilde{s}) dk.$$

The condition that discourages local deviations to $\tilde{s} > s$ is:

$$s \int_0^w q_s(k, s) dk \leq w q(w, s),$$

for all (w, s), which is precisely Assumption 1.

Furthermore, observe that for $\tilde{s} > s$:

$$\begin{aligned} U_{\tilde{s}}(w; s, \tilde{s}) &= \int_0^{\frac{ws}{\tilde{s}}} q_s(k, \tilde{s}) dk - \frac{ws}{\tilde{s}^2} q(\frac{ws}{\tilde{s}}, \tilde{s}) \\ &= \frac{1}{\tilde{s}} \left(\tilde{s} \int_0^{\frac{ws}{\tilde{s}}} q_s(k, \tilde{s}) dk - \frac{ws}{\tilde{s}} q(\frac{ws}{\tilde{s}}, \tilde{s}) \right) \leq 0, \end{aligned}$$

where the inequality holds by Assumption 1 evaluated at $(w, s) \to (\frac{sw}{\tilde{s}}, \tilde{s})$. Therefore, the upward global deviations in $\tilde{s}$ are suboptimal and the result follows.

□

We now illustrate our results in an example with uniformly distributed types and Cobb-Douglas production function. We then present the characterization for a more general uniform symmetric setting in the Appendix.

Example 1 (Uniform Distribution). Let w and s be independently and uniformly distributed on $[0, 1]$, with $\alpha = \beta = \gamma = 1/4$, $c_x = c_y = c_z = 1/8$, and $b = 1$. We implement the optimal allocation scale-by-scale. Under the optimal menu, types (w, s) with $w < 1/2$ are

excluded. Types $w \in [1/2, \hat{w}(s)]$, with $\hat{w}(s) = \frac{1}{2}(1 + \frac{1}{2\sqrt{s}})$ do not fine-tune and consume:

$$\begin{aligned} q(w,s) &= 2s(2w-1), \\ x(w,s) &= y(w,s) = 4s^2(2w-1)^2, \\ z(w,s) &= 0, \\ T(w,s) &= s(2w^2 - 1/2). \end{aligned}$$

Types $w > \hat{w}(s)$ fine-tune and consume:

$$\begin{aligned} q(w,s) &= 8s^2(2w-1)^3, \\ x(w,s) &= y(s,w) = 16s^{\frac{1}{3}}(2w-1)^4, \\ z(w,s) &= 16s^{\frac{10}{3}}(2w-1)^4 - 1, \\ T(w,s) &= s^2(2w-1)^3(1+6w) - \frac{1}{16}. \end{aligned}$$

The resulting revenue is $R^* = \frac{139}{480} \simeq 0.290$ and profits are $\Pi^* = \frac{97}{960} \simeq 0.101$. $\Diamond$

5 Menus of Token Packages

In many cases, it is only feasible or practical to contract only on total number of tokens used, rather than on their allocation across different tasks. To address that case, we allow the seller to contract only on the total numbers of input, output, and fine-tuning tokens, i.e., to sell token packages. The corresponding menu is:

$$(X(\boldsymbol{w}), Y(\boldsymbol{w}), Z(\boldsymbol{w}), t(\boldsymbol{w}))_{\boldsymbol{w}}, \tag{22}$$

where X, Y, and Z are the total number of input, output, and fine-tuning tokens sold, so that upon purchase, the buyer can freely redistribute input and output tokens across tasks.

The seller's problem is an optimal mechanism design problem with infinitely dimensional private information and moral hazard. Such problems are known to be intractable. However, we show that the problem drastically simplifies in the case of Cobb-Douglas production

function. In that case, the buyer-optimal token allocation across tasks for any given total number of tokens (X, Y, Z) is:

$$U(X,Y,Z) = \max_{(x_i,y_i)_{i\in[0,1]}} \int_0^1 w_i x_i^\alpha y_i^\beta (b+Z)^\gamma di,$$
$$\text{s.t.} \int_0^1 x_i di = X, \int_0^1 y_i di = Y.$$

Lemma 2 (Buyer-Optimal Utility). *For any $X, Y, Z \geq 0$, $U(X,Y,Z) = \theta X^\alpha Y^\beta (b+Z)^\gamma$, where $\theta \in [0,1]$ is the representative type (10).*

By Lemma 2, the seller problem is equivalent to the buyer having utility $U(X,Y,Z) = \theta X^\alpha Y^\beta (b+Z)^\gamma$. Therefore, the infinite-dimensional seller problem with adverse selection and moral hazard reduces to a single-dimensional version of Mussa and Rosen (1978) with the effective type being θ and the relevant quality parameter being:

$$Q(X,Y,Z) = X^\alpha Y^\beta (b+Z)^\gamma, \tag{23}$$

with the associated production costs (cf. the definition of $C(q,s)$ in (18)):

$$C(Q) \triangleq \min_{X,Y,Z\geq 0} c_x X + c_y Y + c_z Z \tag{24}$$
$$\text{s.t. } X^\alpha Y^\beta (b+Z)^\gamma = Q. \tag{25}$$

The costs $C(Q)$ are strictly increasing and strictly convex with $C'(0) = 0$. Denote by F_θ the prior distribution of θ, and the virtual type by $\varphi(\theta) \triangleq \theta - \frac{1-F_\theta(\theta)}{f_\theta(\theta)}$. Assume that $\varphi(\theta)$ is increasing. The analysis of Mussa and Rosen (1978) applies, so that all θ with $MR(\theta) \leq 0$ are excluded and for all other θ, the optimal allocation is uniquely pinned down by:

$$\varphi(\theta) = C'(Q(\theta)). \tag{26}$$

Denote the corresponding optimal transfers by

$$T(\theta) = \theta\, Q(\theta) - \int_0^\theta Q(k)dk. \tag{27}$$

Proposition 5 (Optimal Menu of Token Packages). *The optimal menu when only the total number of tokens is contractible and $\varphi(\theta)$ is increasing is given by*

$$(X(\theta), Y(\theta), Z(\theta), T(\theta))_\theta, \tag{28}$$

where $(X(\theta), Y(\theta), Z(\theta))$ are cost-minimizing tokens from (24) that deliver quality $Q = Q(\theta)$ as defined in (26), and $T(\theta)$ is defined in (27). All types $\boldsymbol{w}$ corresponding to the same θ pick the same item.

Example 1 (Continued). In the uniform example, the representative type is $\theta = ws^{1/2}$ which is distributed according to PDF $f_\theta(\theta) = 2(1-\theta)$ with $\varphi(\theta) = \frac{3\theta-1}{2}$ and $\hat{\theta} = 2/3$. Under the optimal menu, types $\theta < 1/3$ are excluded. Types $\theta \in [1/3, 2/3]$ do not fine-tune and consume:

$$\begin{aligned} Q(\theta) &= 3\theta - 1, \\ X(\theta) &= Y(\theta) = (3\theta-1)^2, \\ Z(\theta) &= 0, \\ T(\theta) &= \frac{9\theta^2 - 1}{6}. \end{aligned}$$

Types $\theta \in (1/3 + 2/3, 1]$ fine-tune and consume:

$$\begin{aligned} Q(\theta) &= (3\theta-1)^3, \\ X(\theta) &= Y(\theta) = (3\theta-1)^4, \\ Z(\theta) &= (3\theta-1)^4 - 1, \\ T(\theta) &= -\frac{1}{12} + \frac{1+9\theta}{12}(3\theta-1)^3. \end{aligned}$$

The resulting revenue is $R^* = \frac{139}{540} \simeq 0.257$ and the resulting profits are $\Pi^* = \frac{97}{1080} \simeq 0.090$. Thus, the inability to contract on token allocations

costs the seller around 10% of the profits attainable with menus of contractible token allocations.

Figure 1 illustrates the exclusion and fine-tuning regions under the optimal contracts

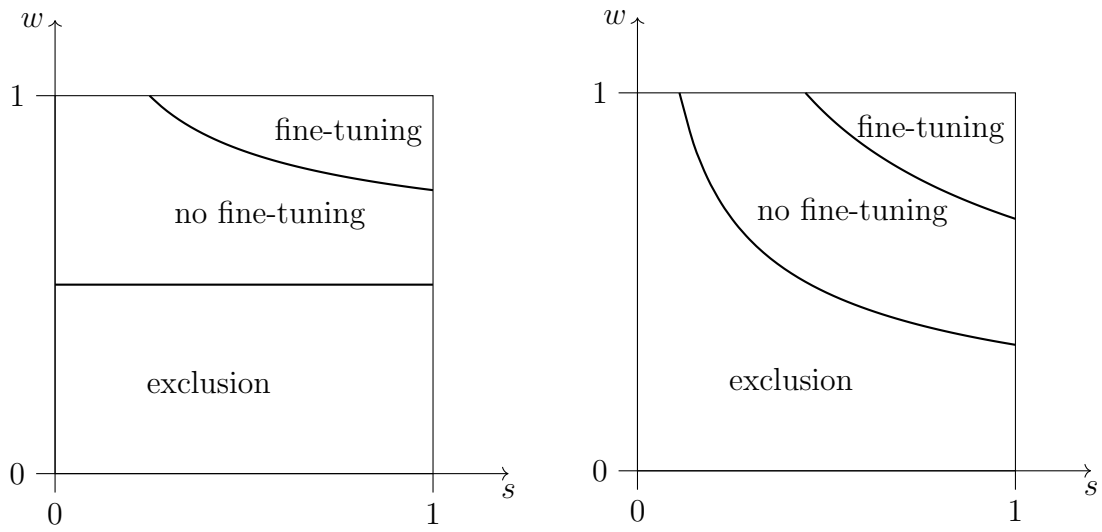


Figure 1: Optimal fine-tuning. $w, s \sim U[0,1]$, $\alpha = \beta = \gamma = 1/4$, $c_x = c_y = c_z = 1/8$, $b = 1$.

with contractible token allocations and contractible token packages. ◊

6 Two-Part Tariff Implementation

In this section, we show that the optimal mechanisms derived in the previous sections can be implemented via menus of two-part tariffs with capped and non-capped number of tasks respectively.[5] To provide a basis for a more general treatment with multiple inputs, we first present a result for a standard setting with one-dimensional private information.

6.1 Two-Part Tariff with One-Dimensional Private Information

Specifically, consider a variation of the nonlinear tariff problem in Maskin and Riley (1984), with type $\theta \in [0,1]$, increasing virtual type $\varphi(\theta) = \theta - \frac{1-F(\theta)}{f(\theta)}$, multidimensional input $x \in \mathbb{R}^J_+$, linear input cost function $\sum_{j=1}^J c_j x_j$, and the non-linear valuation function:

$$\theta v(x_1, \ldots, x_J) - T,$$

[5]The efficient solution can be trivially implemented via marginal cost pricing.

with strictly concave and increasing v. Define the optimal cost function per quality as

$$C(q) \triangleq \min_{x_1,\ldots,x_J \geq 0} \sum_{j=1}^{J} c_j x_j,$$
$$\text{s.t. } v(x_1,\ldots,x_J) = q.$$

Because the production function v is strictly increasing and strictly concave, $C(q)$ is strictly increasing and strictly convex. Therefore, this framework is equivalent to a standard Mussa-Rosen setting with cost function $C(q)$ and a linear valuation function

$$\theta q - C(q).$$

Define a type-dependent *markup* $m(\theta)$ as:

$$m(\theta) \triangleq \frac{\theta}{\varphi(\theta)}. \tag{29}$$

Lemma 3. *Let $(x_1(\theta),\ldots,x_J(\theta),T(\theta))_{\theta\in[0,1]}$ be an optimal direct mechanism with the corresponding $q(\theta) = v(x_1(\theta),\ldots,x_J(\theta))$. Then, the following menu of two-part tariffs (indirectly) implements the optimal mechanism:*

$$(p_1(\theta),\ldots,p_J(\theta),p_0(\theta))_{\theta\in[0,1]},$$

where $p_j(\theta)$ is the linear price for input x_j and $p_0(\theta)$ is the upfront payment equal to:

$$p_j(\theta) = m(\theta)c_j,$$
$$p_0(\theta) = T(\theta) - m(\theta)C(q(\theta)).$$

Proof. Faced with a given item $(p_1,\ldots,p_J,p_0)$, a problem of buyer of type w can be written as quality-maximization-payment-minimization:

$$\max_{q\geq 0} \theta q - P(q),$$

where

$$P(q) \triangleq \min_{x_1,\ldots,x_J \geq 0} \sum_{j=1}^{J} p_j x_j,$$

$$\text{s.t. } v(x) = q.$$

Under the suggested pricing, $\sum_{j=1}^{J} p_j x_j \equiv m(\theta) \sum_{j=1}^{J} c_j x_j$; thus a buyer-optimal allocation of inputs that achieves any given level q is efficient, and $P(q) = m(\theta)C(q)$.

The rest of the argument is standard—see, for example, Tirole (1988, pp. 154-157). A buyer-optimal $q(\theta)$ is determined by the first-order condition:

$$\theta = P'(q) = m(\theta)C'(q) = \frac{\theta}{\varphi(\theta)} C'(q),$$

and thus the buyer-optimal level of quality satisfies the optimality condition:

$$\varphi(\theta) = C'(q).$$

The implied transfer schedule $T(q(\theta))$ is concave and thus the markup $m(\theta)$, which is constant across inputs, implements the desired quality schedule $q(\theta)$ with each type θ consuming an optimal amount of inputs $x_1(\theta), \ldots, x_J(\theta)$ and paying the optimal total transfer $T(\theta)$. □

6.2 Optimal Token Allocations under Value-Scale Heterogeneity

We now apply the above reasoning to the case of fully contractible token allocations in the value-scale setting and show that we can implement the optimal mechanism of the contractible-task benchmark via a menu of two-part tariffs with *task caps*. This menu takes the form:

$$(p_x(w,s), p_y(w,s), p_z(w,s), p_0(w,s), \overline{s}(w,s))_{(w,s)}, \tag{30}$$

where upon selecting a given item from the menu, the buyer pays an upfront payment $p_0(w,s)$ and can freely purchase the input, output, and fine-tuning tokens at linear prices

$p_x(w,s), p_y(w,s), p_z(w,s)$ respectively for up to $\overline{s}(w,s)$ tasks. (Note that the menu does not impose caps on tokens.)

In this case, the relevant heterogeneity is the heterogeneity in value w, and a relevant markup is

$$m(w) \triangleq \frac{w}{\varphi(w)} = \frac{w}{w - \frac{1-F_w(w)}{f_w(w)}}.$$

If F_w has a monotone hazard rate, then the markup $m(w)$ is decreasing in w.

Assumption 2 (Bounded Rent Increase). For all w, s, the allocation $q(w,s)$ defined in (20) satisfies $\int_0^w q_s(k,s)dk \leq -\frac{w}{\varphi(w)} C_s'(q(w,s),s)$.

Assumption 2 admits the following interpretation. Observe that $q(w;s)$ is a buyer-optimal quality chosen by type w at per-token prices $p_x(w) = m(w)c_x, p_y(w) = m(w)c_y, p_z(w) = m(w)c_z$, and therefore the right-hand side of the assumption's condition is $-P_s'(q,s)$. As scale increases, the optimal quality increases and the payment for producing any given quality decreases. Assumption 2 requires that for any type (w,s), the marginal rent increase should never exceed the marginal saving in payment for that type producing her optimal quality level.

Proposition 6. *Under Assumption 2, a menu (30) of two-part tariffs with:*

$$p_j(w,s) = m(w)c_j, \quad j = x,y,z, \tag{31}$$

$$p_0(w,s) = T(w,s) - m(w)C(q(w,s),s), \tag{32}$$

$$\overline{s}(w,s) = s, \tag{33}$$

where $C(q(w,s),s)$ is as defined in (15), $q(w,s)$ and $T(w,s)$ are as defined in (20) and (21), implements the optimal allocation and raises the optimal profits.

Proof. If buyer type (w,s) faces token prices p_x, p_y, p_z, her optimal token allocation solves

$$\max_{(x_i,y_i)_{i\in[0,s]}, z\geq 0} \int_{i=0}^{s} (w\, v(x_i,y_i,z) - p_x x_i - p_y y_i)di - p_z z.$$

As all tasks are symmetric, she will be purchasing the same number of input and output tokens per task so the problem can be rewritten as:

$$\max_{x,y,z\geq 0} s(w\, v(x,y,z) - p_x x - p_y y) - p_z z.$$

Equivalently, the problem can be rewritten as quality-maximization-payment-minimization:

$$\begin{aligned} &\max_{q\geq 0} w\, q - P(q,s), \\ P(q,s) \triangleq &\min_{x,y,z\geq 0} s\, p_x x + s\, p_y y + p_z z, \\ &s.t.\ v(x,y,z) = q/s. \end{aligned}$$

Under the pricing of Proposition 6, $P(q,s) = m(w)C(q,s)$ where $C(q,s)$ is the efficient cost of supplying quality q as defined in (15). Thus, if reporting truthfully, each type consumes an efficient (and thus optimal) amount of tokens and pays the optimal total transfer.

Deviations to $(\tilde{w}, s)$, i.e., within the true scale, are not profitable by the arguments of Lemma 3.

Consider deviations to $(\tilde{w}, \tilde{s})$ with $\tilde{s} < s$. Upon this deviation type (w,s) obtains the exact same payoff as type $(w,\tilde{s})$. Thus, it is necessary and sufficient that the on-path rents of type (w,s) are increasing in s, and these rents are increasing.

Consider deviations to $(\tilde{w}, \tilde{s})$ with $\tilde{s} > s$. Because optimal markups $m(w)$ do not depend on s, the necessary and sufficient condition is that upfront payments $p_0(w,s)$ are increasing in s for all w. By construction,

$$p_0(w,s) = w\, q(w,s) - \int_0^w q(k,s)dk - \frac{w}{\varphi(w)} C(q(w,s),s), \tag{34}$$

and therefore,

$$\begin{aligned} \frac{d\, p_0(w,s)}{ds} &= w\, q_s(w,s) - \int_0^w q_s(k,s)dk - \frac{w}{\varphi(w)}(C_q(q(w,s),s)q_s(w,s) + C_s'(q(w,s),s), &(35)\\ &= -\int_0^w q_s(z,s)dz - m(w)C_s'(q(w,s),s). &(36) \end{aligned}$$

where for the second line we used the optimality condition $C_q(q(w,s),s) = \varphi(w)$.

Therefore, the necessary and sufficient condition for upward deviations to not be profitable is precisely the condition of Assumption 2. The result follows. □

Example 1 (Continued). Continuing with the uniform example, if full allocation is contractible, the two-part tariff implementation of the optimal mechanism is, for $w \in [1/2, 1]$:

$$p_x(w,s) = p_y(w,s) = p_z(w,s) = \frac{w}{8(2w-1)},$$
$$p_0(w,s) = \begin{cases} s(w - \frac{1}{2}), & \text{if } w \in [\frac{1}{2}, \hat{w}(s)], \\ s^2(2w-1)^3 - \frac{1}{16}, & \text{if } w \in (\hat{w}(s), 1]. \end{cases}$$

Notice, that $p_0(w,s)$ is increasing in s for all $w \in [1/2, 1]$ and thus Assumption 2 is satisfied. ◊

Remark 1 (Ranking of Assumptions). Assumption 2 is than Assumption 1 because it allows to implement the scale-by-scale optimal allocation by a more permissive mechanism. Indeed, it holds that:

$$-\frac{C_s'(q(w,s),s)}{\varphi(w)} \leq \frac{q(w,s)}{s}, \tag{37}$$

where $\varphi(q(w,s),s) = C_q(q(w,s),s)$. From the definition (15) of $C(q,s)$, and the Envelope Theorem, if $\lambda(q,s)$ is the Lagrange multiplier of the quality constraint, then:

$$C_q(q,s)q + C_s'(q,s)s = \lambda q + (c_x x(q,s) + c_y y(q,s) - \lambda(q,s)v(x(q,s), y(q,s), z(q,s)))s \tag{38}$$
$$= (c_x x(q,s) + c_y y(q,s))s \geq 0. \tag{39}$$

6.3 Optimal Token Packages

In this section, we move to the case of token packages and show that, very generally, we can implement the optimal menu of packages via a menu of two-part tariffs without task caps.

This menu takes the form:

$$(p_x(\theta), p_y(\theta), p_z(\theta), p_0(\theta))_\theta, \tag{40}$$

where upon selecting a given item from the menu, the buyer pays an upfront payment $p_0(\theta)$ and can freely purchase the input, output, and fine-tuning tokens at linear prices $p_x(\theta), p_y(\theta), p_z(\theta)$ respectively across as many tasks as the buyer likes. Recall that the relevant type in the mechanism design benchmark was the representative type θ as defined in (10). Define a markup $m(\theta)$ as

$$m(\theta) \triangleq \frac{\theta}{\varphi(\theta)} = \frac{\theta}{\theta - \frac{1-F_\theta(\theta)}{f_\theta(\theta)}},$$

and note that it is decreasing in θ as long as F_θ has a monotone hazard rate.

Proposition 7. *If $\varphi(\theta)$ is increasing, then a menu of two-part tariffs (40) with:*

$$\begin{aligned} p_x(\theta) &= m(\theta)c_x, \\ p_y(\theta) &= m(\theta)c_y, \\ p_z(\theta) &= m(\theta)c_z, \\ p_0(\theta) &= T(\theta) - m(\theta)C(Q(\theta)). \end{aligned}$$

where $C(Q)$ is as defined in (24), $Q(\theta)$ is as defined in (26), and $T(\theta)$ is as defined in (27), implements an optimal allocation and raises the optimal profits of the token package setting. All types (w, s) corresponding to the same θ pick the same item.

Proof. When type (w, s) takes any item (p_x, p_y, p_z, p_0) her optimal token allocation problem is equivalent to the efficient allocation problem (9) with prices taking the roles of costs. Therefore, all types (w, s) that correspond to the same θ have the same utility from any possible two-part tariff and can be treated as a single type. Under the proposed pricing, $P(Q(\theta)) = m(\theta)Q(\theta)$ and the buyer-optimal allocation of tokens is efficient. The optimality of the proposed menu then follows directly from Lemma 3. □

Example 1 (Continued). Continuing with the uniform example, if only token packages are contractible, the two-part tariff implementation of the optimal mechanism is, for $\theta \in [1/3, 1]$:

$$p_x(\theta) = p_y(\theta) = p_z(\theta) = \frac{\theta}{4(3\theta - 1)},$$
$$p_0(\theta) = \begin{cases} \frac{3\theta-1}{6}, \text{ if } \theta \in [\frac{1}{3}, \frac{2}{3}], \\ \frac{1}{12(3\theta-1)} + \frac{1}{12}(3\theta - 1)^3, \text{ if } \theta \in [\frac{2}{3}, 1]. \end{cases}$$

$\Diamond$

7 Mapping the Model to Practice

In this section, we provide details on how to link the features of our model the current design of LLM models. Then we discuss how the optimal pricing policy derived here map to current industry pricing practices.

7.1 Interpretation of the Model Features

Base level, b This variable captures the quality of baseline, non-fine-tuned model. The non-fine-tuned model can be viewed as a foundational model. The quality of the model depends on its size, i.e., the number of parameters[6], model architecture, the amount and quality of data used in training, and the details of training procedure. Larger size generally leads to more knowledgeable and accurate models, but to achieve the best results, a larger model must be accompanied by larger training data (Kaplan et al. (2020)). In our baseline model, we study a scenario in which the basic training has already been completed and resulted in the base level b. As such, we primarily focus on the inference-time compute rather than on the train-time compute.[7]

[6]The details of top models are not publicly disclosed, but their sizes are estimated to reach a few billions, https://codingscape.com/blog/most-powerful-llms-large-language-models

[7]For more details on the distinction between train-time compute and test-time compute, see https://arxiv.org/pdf/2104.03113.

Fine-tuning tokens, z_i This variable captures the amount of tokens used to fine-tune the foundational model. fine-tuning is similar to the original training in that it directly changes the model's parameter values (but not its size or architecture) but is much smaller in scale. It is done on a dataset directly relevant to some class of tasks, e.g., the dataset of labeled x-ray scans or the dataset of example code or conversations, and can be thought of as instilling some specific knowledge into the model. The ability of foundational LLMs to be successfully fine-tuned in a variety of specialized settings is an instance of transfer learning.

Input tokens, x_i This variable captures the amount of tokens provided by the user for a given task. A larger number of input tokens increases the quality of predictions. First, it can provide more context for the problem at hand with more detailed problem explanation leading to a more tailored and appropriate response.[8] Second, and relatedly, it can provide the necessary data for the model to act upon. For instance, a popular technique is retrieval-augmented generation (RAG), which based on the user's prompt fetches the right part of a given database and effectively appends it to the user's prompt.

Output tokens, y_i This variable captures the amount of tokens generated by the model for a given task. A larger number of output tokens increases the quality of predictions. First, it enables the model to communicate more details and nuances of the solution to the user. Second, and importantly, it enables the model to engage in an internal dialog with itself along a chain-of-thought (CoT) computation. The inner dialog seems to allow the model to implement more sophisticated algorithmic tasks, at the expense of generating a lot of intermediate steps, often hidden from the user. This computation is qualitatively distinct from the training of a foundation model, still not fully understood, and represents a frontier avenue for future progress in the performance of LLMs.[9]

Thus, input, output, and fine-tuning tokens constitute distinct categories of inputs into the production function. They are neither perfect substitutes nor perfect complements, but

[8] In practice, not only the length but also the phrasing of the context might affect the model's performance. The choice of the right phrasing is referred to as "prompt engineering."

[9] Much recent LLM progress is generated via this channel, see for example https://arcprize.org/blog/oai-o3-pub-breakthrough.

all of them contribute to the higher quality of predictions. This observation motivates our use of Cobb-Douglas function production as a leading example.

Costs, $c_x, c_y, c_z > 0$ Processing tokens is costly because it requires runs and updates of a given neural network, which in turn require energy

and hardware.[10] As tokens of the same kind are treated symmetrically from the cost-computational perspective, we assume the marginal costs are constant across tasks.[11] However, because input, output, and fine-tuning tokens enter different computational routines, we allow their marginal costs to differ.

7.2 Pricing Used in Practice

In practice, LLMs are priced based on fine-tuning, as well as input and output tokens, with pricing strategies evolving continually. For concreteness, consider the current public API pricing strategy by one of the leaders in the field, OpenAI (https://openai.com/api/pricing/, see also Figure 2).[12,13] As in our model, the pricing is based on input, output, and fine-tuning tokens and has the following qualitative features. First, more recent and larger models are priced higher. Second, the pricing is linear with dependencies across different token types. Specifically, the input, output, and fine-tuning tokens are priced linearly, but the prices of input and output tokens depend on whether the model has been fine-tuned or not: the prices are higher for the fine-tuned model than for the baseline model. Third, it is possible to specify "cached" tokens that are to be pre-appended to each prompt, and these tokens are priced lower than regular input tokens. Fourth, it is possible to have the tasks treated in batch with slower processing at lower prices. Finally, input tokens are priced lower than output tokens with the price ratio being constant across all regimes.

[10] As of January 2025, NVIDIA, which is the major producer of AI chips, has a market capitalization of 3.288 Trillion USD, making it the second most valuable company in the world.

[11] Note that fine-tuning should not affect marginal inference costs as it does not alter the model size.

[12] The details of privately negotiated deals are not publicly disclosed. However, we expect them to feature similar structure with an addition of scale discounts.

[13] Similar pricing strategies are employed by other big players in the field, such as Anthropic (https://www.anthropic.com/pricing) and Google (https://ai.google.dev/pricing).

Model	Pricing	Pricing with Batch API*
gpt-4o-2024-08-06	$3.750 / 1M input tokens	$1.875 / 1M input tokens
	$1.875 / 1M cached** input tokens	
	$15.000 / 1M output tokens	$7.500 / 1M output tokens
	$25.000 / 1M training tokens	
	. . .	
gpt-4o-2024-08-06	$2.50 / 1M input tokens	$1.25 / 1M input tokens
	$1.25 / 1M cached** input tokens	
	$10.00 / 1M output tokens	$5.00 / 1M output tokens

Figure 2: OpenAI API pricing. Top: Pricing of a fine-tuned model. Bottom: Pricing of a baseline model. Accessed: January 2025.

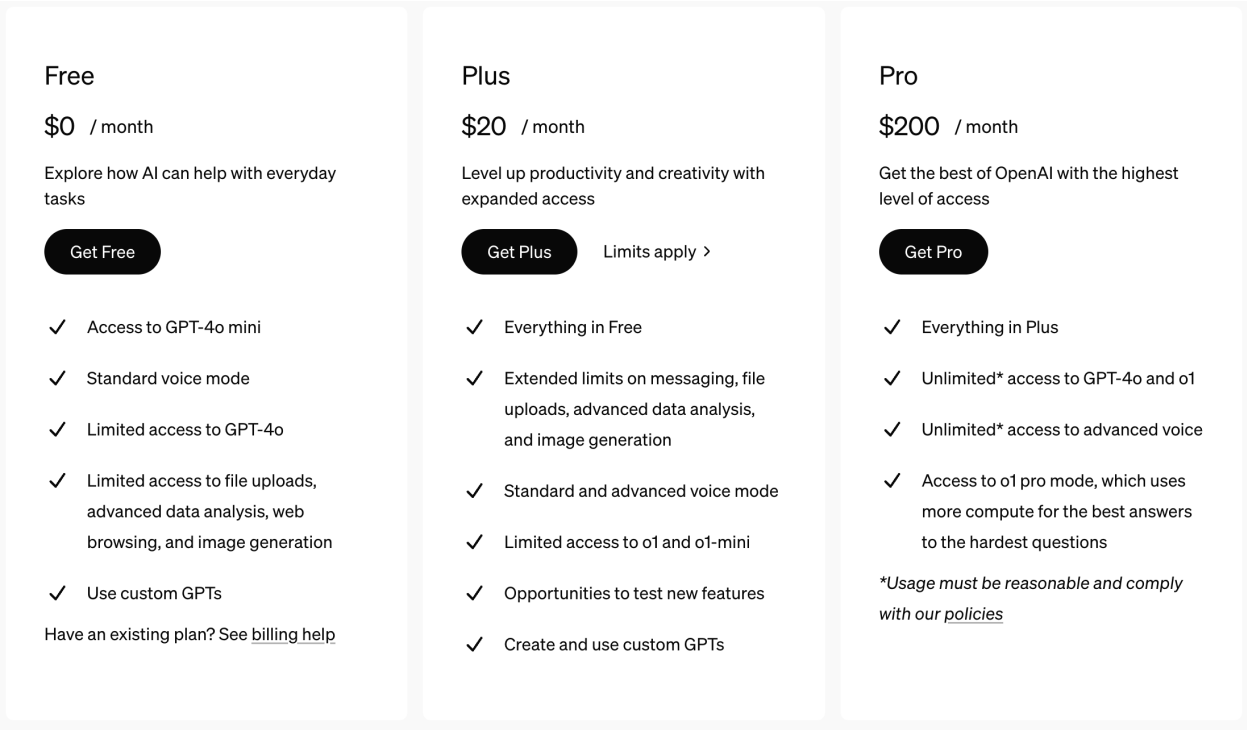


Figure 3: OpenAI ChatGPT pricing. Accessed: January 2025.

OpenAI's consumer pricing follows a subscription scheme rather than a purely usage-based model, with a tier structure that shares several notable features (https://openai.com/chatgpt/pricing/; see also Figure 3). First, higher tiers grant access to better models. Second, they increase the number of requests allowed. Third, they permit the creation of custom models, which can be viewed as a form of fine-tuning.

In both the API and ChatGPT pricing, OpenAI charges for both customization and the quality of predictions, and the pricing menu is nonlinear. In equilibrium, different buyers would select a model, a level of personalization, and a usage level. Consistent with this, our solution also features a rich menu in which higher levels of fine-tuning and usage command higher prices.

8 Discussion and Next Steps

In this paper, we developed a basic model of LLM revenue maximization. This model can be extended in multiple directions, which we discuss below.

First, we assumed that the seller has a single base model that she offers to the buyer (i.e., b is fixed). In practice, base models can differ in multiple dimensions. Historically, one of the major differentiators has been model size. One way to incorporate this heterogeneity is to let b be one of the design choices, with larger b corresponding to a larger model size. As larger models require more compute to operate, it would then be natural to assume that the token processing costs scale up with b as some functions $c_x(b), c_y(b), c_z(b)$. The resulting seller's problem would then build on our analysis but additionally include the optimal choice of b for each item in the menu.

More recently, models are often differentiated not by their size but by the amount of inference-time compute they enable the user to perform. A prominent example is the latest line of LLMs from OpenAI, including the models o1 and o3.[14] This type of differentiation is already embedded within our current model: one can think of different amounts of output tokens per task as representing different levels of CoT reasoning, resulting in different

[14] https://www.economist.com/business/2025/01/20/openais-latest-model-will-change-the-economics-of-software.

processing costs and inference quality. As such, our analysis in the fully contractible token allocation setting directly covers this kind of differentiation.

Second, we assumed that all tasks are homogeneous in their use of input, output, and fine-tuning tokens, and differ only vertically in how valuable a given task is to the buyer. It is possible that, in practice, some tasks are more input-intensive, some are more output-intensive, and some are more fine-tuning-intensive. Thus, a natural extension would be to allow the Cobb-Douglas parameters α_i, β_i, and γ_i to vary across tasks.

Third, we assumed that all buyer types have fine-tuning data readily available. In practice, some buyer types may face constraints on the amount of data they can use (e.g., $z < \overline{z}$).

Fourth, within the value-scale setting and fully contractible token allocations, we assumed that scale s and values w are independently distributed. Given additional assumptions on technology, this allowed the seller to achieve the benchmark of observable scale. However, one can imagine that value and scale can be correlated—e.g., larger firms might have higher value for the quality of their decisions—in which case the separate implementability may fail. It would be interesting to study how such a correlation would affect the optimal design in those cases.

Appendix

Proof of Proposition 1

The solution to problem (9) can be derived as follows. For any fixed $z \geq 0$, an optimal allocation of input and output tokens across tasks, as well as the corresponding value per task are:

$$x_i^*(z) = \left(\frac{\alpha w_i (b+z)^\gamma}{c_x}\right)^{\frac{1}{1-\alpha-\beta}} \left(\frac{\beta c_x}{\alpha c_y}\right)^{\frac{\beta}{1-\alpha-\beta}},$$
$$y_i^*(z) = \left(\frac{\beta w_i (b+z)^\gamma}{c_y}\right)^{\frac{1}{1-\alpha-\beta}} \left(\frac{\alpha c_y}{\beta c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}},$$
$$U_i^*(z) = (1-\alpha-\beta)\,(b+z)^{\frac{\gamma}{1-\alpha-\beta}} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta}} w_i^{\frac{1}{1-\alpha-\beta}}.$$

The corresponding total amount of input and output tokens, as well as the total value net of fine-tuning costs, are

$$X^*(z) = \left(\frac{\alpha(b+z)^\gamma}{c_x}\right)^{\frac{1}{1-\alpha-\beta}} \left(\frac{\beta c_x}{\alpha c_y}\right)^{\frac{\beta}{1-\alpha-\beta}} \int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} di, \tag{41}$$
$$Y^*(z) = \left(\frac{\beta(b+z)^\gamma}{c_y}\right)^{\frac{1}{1-\alpha-\beta}} \left(\frac{\alpha c_y}{\beta c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}} \int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} di, \tag{42}$$
$$U^*(z) = (1-\alpha-\beta)(b+z)^{\frac{\gamma}{1-\alpha-\beta}} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta}} \int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} di. \tag{43}$$

Notably, for any amount of fine-tuning $z \geq 0$, $X^*(z), Y^*(z), U^*(z)$ for type $w = (w_i)_{i=0}^1$ are equal to those for type $w = (\theta)_{i=0}^1$ who attaches to all tasks the same value θ defined as

$$\theta = \left(\int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} di\right)^{1-\alpha-\beta}. \tag{44}$$

The optimal amount of fine-tuning admits a corner solution if θ is low and an interior solution if θ is high. The threshold type is:

$$\hat{\theta} = b^{1-\alpha-\beta-\gamma} \left(\frac{c_x}{\alpha}\right)^\alpha \left(\frac{c_y}{\beta}\right)^\beta \left(\frac{c_z}{\gamma}\right)^{1-\alpha-\beta}. \tag{45}$$

If $\theta \leq \hat{\theta}$, then

$$x_i^* = w_i^{\frac{1}{1-\alpha-\beta}} \frac{\alpha}{c_x} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta}} b^{\frac{\gamma}{1-\alpha-\beta}}, \tag{46}$$

$$y_i^* = w_i^{\frac{1}{1-\alpha-\beta}} \frac{\beta}{c_y} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta}} b^{\frac{\gamma}{1-\alpha-\beta}}, \tag{47}$$

$$z^* = 0. \tag{48}$$

If $\theta > \hat{\theta}$, then

$$x_i^* = w_i^{\frac{1}{1-\alpha-\beta}} \theta^{\frac{\gamma}{(1-\alpha-\beta)(1-\alpha-\beta-\gamma)}} \frac{\alpha}{c_x} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta-\gamma}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta-\gamma}} \left(\frac{\gamma}{c_z}\right)^{\frac{\gamma}{1-\alpha-\beta-\gamma}}, \tag{49}$$

$$y_i^* = w_i^{\frac{1}{1-\alpha-\beta}} \theta^{\frac{\gamma}{(1-\alpha-\beta)(1-\alpha-\beta-\gamma)}} \frac{\beta}{c_y} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta-\gamma}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta-\gamma}} \left(\frac{\gamma}{c_z}\right)^{\frac{\gamma}{1-\alpha-\beta-\gamma}}, \tag{50}$$

$$z^* = \theta^{\frac{1}{1-\alpha-\beta-\gamma}} \frac{\gamma}{c_z} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta-\gamma}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta-\gamma}} \left(\frac{\gamma}{c_z}\right)^{\frac{\gamma}{1-\alpha-\beta-\gamma}} - b. \tag{51}$$

The optimal token allocation is continuous in θ.

Proof of Proposition 3

Consider a mechanism that attempts full surplus extraction. The efficiency characterization in Proposition 1 implies that under this mechanism type $\boldsymbol{w}$ if reporting type $\tilde{\boldsymbol{w}}$ obtains added value as follows:

if $\tilde{\theta} < \hat{\theta}$,

$$u(\boldsymbol{w}, \tilde{\boldsymbol{w}}) = \int_0^1 \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta}} b^{\frac{\gamma}{1-\alpha-\beta}} \; w_i \tilde{w}_i^{\frac{1}{1-\alpha-\beta}-1} di, \tag{52}$$

if $\tilde{\theta} \geq \hat{\theta}$,

$$u(\boldsymbol{w}, \tilde{\boldsymbol{w}}) = \int_0^1 \tilde{\theta}^{\frac{\gamma}{(1-\alpha-\beta)(1-\alpha-\beta-\gamma)}} \left(\frac{\alpha}{c_x}\right)^{\frac{\alpha}{1-\alpha-\beta-\gamma}} \left(\frac{\beta}{c_y}\right)^{\frac{\beta}{1-\alpha-\beta-\gamma}} \left(\frac{\gamma}{c_z}\right)^{\frac{\gamma}{1-\alpha-\beta-\gamma}} w_i \tilde{w}_i^{\frac{1}{1-\alpha-\beta}-1} di. \tag{53}$$

Because under truth-telling each type obtains zero rents, the corresponding payment is $T(\tilde{\boldsymbol{w}}) = u(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{w}})$ and the incentive constraint is violated if and only if:

$$\int_0^1 w_i \tilde{w}_i^{\frac{1}{1-\alpha-\beta}-1} - \tilde{w}_i^{\frac{1}{1-\alpha-\beta}} di > 0. \tag{54}$$

By Jensen's inequality and the concavity of a log function:

$$w_i \tilde{w}_i^{\frac{1}{1-\alpha-\beta}-1} = w_i^{\frac{1-\alpha-\beta}{1-\alpha-\beta}} \tilde{w}_i^{\frac{\alpha+\beta}{1-\alpha-\beta}} \leq (1-\alpha-\beta) w_i^{\frac{1}{1-\alpha-\beta}} + (\alpha+\beta) \tilde{w}_i^{\frac{1}{1-\alpha-\beta}}.$$

Therefore, inequality (54) implies

$$\int_0^1 \tilde{w}_i^{\frac{1}{1-\alpha-\beta}} - w_i^{\frac{1}{1-\alpha-\beta}} di > 0,$$

and the incentive violation is possible only if $\tilde{\theta} > \theta$. The result follows.

Proof of Lemma 2

The objective function is concave and the constraints are linear. Lagrangian approach applies. Assigning Lagrange multipliers λ and γ to the budget constraints, the optimal first-order conditions are

$$w_i \alpha x_i^{\alpha-1} y_i^{\beta} (b+Z)^{\gamma} = \lambda,$$
$$w_i \beta x_i^{\alpha} y_i^{\beta-1} (b+Z)^{\gamma} = \gamma,$$

and thus x_i/y_i is constant across i. The budget constraints imply that

$$y_i = x_i \frac{X}{Y},$$

and the first-order condition on x_i can be rewritten as:

$$w_i \alpha x_i^{\alpha+\beta-1} \left(\frac{X}{Y}\right)^{\beta} (b+Z)^{\gamma} = \lambda.$$

Thus, $w_i x_i^{\alpha+\beta-1}$ is constant across i. The budget constraint on x_i then implies that

$$x_i = \frac{w_i^{\frac{1}{1-\alpha-\beta}}}{\int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} di} X,$$

and thus,

$$U(X,Y,Z) = \left(\int_0^1 w_i^{\frac{1}{1-\alpha-\beta}} di\right)^{1-\alpha-\beta} X^\alpha Y^\beta (b+Z)^\gamma = \theta X^\alpha Y^\beta (b+Z)^\gamma, \tag{55}$$

where $\theta \in [0,1]$ is the representative type as defined in (10)

Cobb-Douglas Cost Minimization

In this section, we consider a prototypical cost-minimization problem for a Cobb-Douglas value function:

$$C(q) = \min_{x,y,z} c_x x + c_y y + c_z z \tag{56}$$

$$s.t.\ x^\alpha y^\beta z^\gamma = q \tag{57}$$

$$x \geq 0, y \geq 0, z \geq b, \tag{58}$$

where $\alpha + \beta + \gamma \leq 1$. Define:

$$\hat{q} = b^{\alpha+\beta+\gamma} \left(\frac{\alpha}{c_x}\right)^\alpha \left(\frac{\beta}{c_y}\right)^\beta \left(\frac{c_z}{\gamma}\right)^{\alpha+\beta}.$$

Lemma 4. *The solution to (56) is:*

$$C(q) = \begin{cases} (\alpha+\beta)\left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} \left(\frac{1}{b}\right)^{\frac{\gamma}{\alpha+\beta}} q^{\frac{1}{\alpha+\beta}} + c_z b, & \text{if } q < \hat{q}, \\ (\alpha+\beta+\gamma)\left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}} \left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} q^{\frac{1}{\alpha+\beta+\gamma}}, & \text{if } q \geq \hat{q}. \end{cases}$$

$C(q)$ is strictly convex and differentiable everywhere. In particular,

$$C'(\hat{q}) = b^{1-\alpha-\beta-\gamma} \left(\frac{c_x}{\alpha}\right)^\alpha \left(\frac{c_y}{\beta}\right)^\beta \left(\frac{c_z}{\gamma}\right)^{1-\alpha-\beta}.$$

If $q < \hat{q}$, then the optimal token allocation is:

$$
\begin{aligned}
x^*(q) &= \frac{\alpha}{c_x}\left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}}\left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta}}\left(\frac{1}{b}\right)^{\frac{\gamma}{\alpha+\beta}} q^{\frac{1}{\alpha+\beta}},\\
y^*(q) &= \frac{\beta}{c_y}\left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}}\left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta}}\left(\frac{1}{b}\right)^{\frac{\gamma}{\alpha+\beta}} q^{\frac{1}{\alpha+\beta}},\\
z^*(q) &= b.
\end{aligned}
$$

If $q \geq \hat{q}$, then the optimal token allocation is:

$$
\begin{aligned}
x^*(q) &= \frac{\alpha}{c_x}\left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}}\left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}}\left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} q^{\frac{1}{\alpha+\beta+\gamma}},\\
y^*(q) &= \frac{\beta}{c_y}\left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}}\left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}}\left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} q^{\frac{1}{\alpha+\beta+\gamma}},\\
z^*(q) &= \frac{\gamma}{c_z}\left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}}\left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}}\left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} q^{\frac{1}{\alpha+\beta+\gamma}}.
\end{aligned}
$$

The optimal token allocation is continuous in q.

Proof. Proof is straightforward and omitted. □

Corollary 1 (Cost Function for Fully Contractible Tokens). *Consider the problem*

$$C(q,s) = \min_{x,y,z\geq 0} s\,c_x x + s\,c_y y + c_z z \tag{59}$$

$$s.t.\ v(x,y,z) = q/s. \tag{60}$$

The optimal solution is as follows:

$$
C(q,s) = \begin{cases}
\left(\frac{1}{s}\right)^{\frac{1-\alpha-\beta}{\alpha+\beta}}(\alpha+\beta)\left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}}\left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta}}\left(\frac{1}{b}\right)^{\frac{\gamma}{\alpha+\beta}} q^{\frac{1}{\alpha+\beta}}, & \text{if } q < \hat{q}(s),\\
\left(\frac{1}{s}\right)^{\frac{1-\alpha-\beta}{\alpha+\beta+\gamma}}(\alpha+\beta+\gamma)\left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}}\left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}}\left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} q^{\frac{1}{\alpha+\beta+\gamma}} - c_z b, & \text{if } q \geq \hat{q}(s),
\end{cases}
$$

where

$$\hat{q}(s) = s^{1-\alpha-\beta}\,b^{\alpha+\beta+\gamma}\left(\frac{\alpha}{c_x}\right)^{\alpha}\left(\frac{\beta}{c_y}\right)^{\beta}\left(\frac{c_z}{\gamma}\right)^{\alpha+\beta}.$$

For $q < \hat{q}(s)$, no fine-tuning is optimal, whereas for $q > \hat{q}(s)$ some fine-tuning is optimal. C is strictly convex and continuously differentiable in q, and $C_q(q,s)$ is decreasing in s for all q.

The corresponding optimal token allocation is as follows:

If $q < \hat{q}(s)$:

$$\begin{aligned}
x^*(q) &= \left(\frac{1}{s}\right)^{\frac{1}{\alpha+\beta}} \frac{\alpha}{c_x} \left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} \left(\frac{1}{b}\right)^{\frac{\gamma}{\alpha+\beta}} q^{\frac{1}{\alpha+\beta}}, \\
y^*(q) &= \left(\frac{1}{s}\right)^{\frac{1}{\alpha+\beta}} \frac{\beta}{c_y} \left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} \left(\frac{1}{b}\right)^{\frac{\gamma}{\alpha+\beta}} q^{\frac{1}{\alpha+\beta}}, \\
z^*(q) &= 0.
\end{aligned}$$

If $q \geq \hat{q}(s)$:

$$\begin{aligned}
x^*(q) &= \left(\frac{1}{s}\right)^{\frac{1+\gamma}{\alpha+\beta+\gamma}} \frac{\alpha}{c_x} \left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}} \left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} q^{\frac{1}{\alpha+\beta+\gamma}}, \\
y^*(q) &= \left(\frac{1}{s}\right)^{\frac{1+\gamma}{\alpha+\beta+\gamma}} \frac{\beta}{c_y} \left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}} \left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} q^{\frac{1}{\alpha+\beta+\gamma}}, \\
z^*(q) &= \left(\frac{1}{s}\right)^{\frac{1-\alpha-\beta}{\alpha+\beta+\gamma}} \frac{\gamma}{c_z} \left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}} \left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} q^{\frac{1}{\alpha+\beta+\gamma}} - b.
\end{aligned}$$

Corollary 2 (Cost Function for Token Packages). *Consider the problem*

$$\begin{aligned}
C(Q) = \min_{X,Y,Z \geq 0} & \; c_x X + c_y Y + c_z Z \\
s.t. \; & X^\alpha Y^\beta (b+Z)^\gamma = Q.
\end{aligned}$$

The optimal solution is as follows:

$$C(Q) = \begin{cases}
(\alpha+\beta)\left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} \left(\frac{1}{b}\right)^{\frac{\gamma}{\alpha+\beta}} Q^{\frac{1}{\alpha+\beta}}, & \text{if } Q < \hat{Q}, \\
(\alpha+\beta+\gamma)\left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}} \left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} Q^{\frac{1}{\alpha+\beta+\gamma}} - c_z b, & \text{if } Q \geq \hat{Q},
\end{cases}$$

where

$$\hat{Q} = b^{\alpha+\beta+\gamma} \left(\frac{\alpha}{c_x}\right)^{\alpha} \left(\frac{\beta}{c_y}\right)^{\beta} \left(\frac{c_z}{\gamma}\right)^{\alpha+\beta}.$$

For $Q < \hat{Q}$, no fine-tuning is optimal, whereas for $Q > \hat{Q}$ some fine-tuning is optimal. Specifically, the optimal token allocation is as follows:
If $Q < \hat{Q}$:

$$\begin{aligned}
x^*(Q) &= \frac{\alpha}{c_x} \left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} \left(\frac{1}{b}\right)^{\frac{\gamma}{\alpha+\beta}} Q^{\frac{1}{\alpha+\beta}}, \\
y^*(Q) &= \frac{\beta}{c_y} \left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} \left(\frac{1}{b}\right)^{\frac{\gamma}{\alpha+\beta}} Q^{\frac{1}{\alpha+\beta}}, \\
z^*(Q) &= 0.
\end{aligned}$$

If $Q \geq \hat{Q}$:

$$\begin{aligned}
x^*(Q) &= \frac{\alpha}{c_x} \left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}} \left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} Q^{\frac{1}{\alpha+\beta+\gamma}}, \\
y^*(Q) &= \frac{\beta}{c_y} \left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}} \left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} Q^{\frac{1}{\alpha+\beta+\gamma}}, \\
z^*(Q) &= \frac{\gamma}{c_z} \left(\frac{c_x}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta+\gamma}} \left(\frac{c_y}{\beta}\right)^{\frac{\beta}{\alpha+\beta+\gamma}} \left(\frac{c_z}{\gamma}\right)^{\frac{\gamma}{\alpha+\beta+\gamma}} Q^{\frac{1}{\alpha+\beta+\gamma}} - b.
\end{aligned}$$

Uniform Symmetric Setting

Consider the generalization of the uniform example presented in the main text body: let w and s be independently and uniformly distributed on $[0, 1]$, $\alpha = \beta = \gamma \triangleq \rho < 1/3$, $c_x = c_y = c_z = c$, and $b = 1$. The main example corresponded to $\rho = 1/4$ and $c = 1/8$.

Menus of Token Allocations If token allocations are fully contractible, by Proposition 4, we can implement optimal allocation scale-by-scale with the relevant functions being:

$$\varphi(w) = 2w - 1,$$
$$C'(q) = \begin{cases} \frac{c}{\rho}\left(\frac{1}{s}\right)^{\frac{1-2\rho}{2\rho}} q^{\frac{1-2\rho}{2\rho}}, & \text{if } q < s^{1-2\rho}, \\ \frac{c}{\rho}\left(\frac{1}{s}\right)^{\frac{1-2\rho}{3\rho}} q^{\frac{1-3\rho}{3\rho}}, & \text{if } q \geq s^{1-2\rho}. \end{cases}$$

In the optimal menu, types (w, s) with $w < 1/2$ are excluded. Types (w, s) with $w \in [1/2, \hat{w}(s)]$, where $\hat{w}(s) = \frac{1}{2}(1 + \frac{c}{\rho}\left(\frac{1}{s}\right)^{1-2\rho})$, do not fine-tune, and consume:

$$q(w,s) = s\left(\frac{(2w-1)\rho}{c}\right)^{\frac{2\rho}{1-2\rho}},$$
$$x(w,s) = y(w,s) = s^{\frac{1}{2\rho}}\left(\frac{(2w-1)\rho}{c}\right)^{\frac{1}{1-2\rho}},$$
$$z(w,s) = 0,$$
$$T(w,s) = s\left(\frac{(2w-1)\rho}{c}\right)^{\frac{2\rho}{1-2\rho}}\left(\frac{1}{2} + (2w-1)\rho\right).$$

Types (w, s) with $w > \hat{w}(s)$ fine-tune and consume:

$$q(w,s) = s^{\frac{1-2\rho}{1-3\rho}}\left(\frac{(2w-1)\rho}{c}\right)^{\frac{3\rho}{1-3\rho}},$$
$$x(w,s) = y(w,s) = s^{\frac{1-3\rho+3\rho^2}{3\rho(1-3\rho)}}\left(\frac{(2w-1)\rho}{c}\right)^{\frac{1}{1-3\rho}},$$
$$z(w,s) = s^{\frac{1-6\rho^2}{3\rho(1-3\rho)}}\left(\frac{(2w-1)\rho}{c}\right)^{\frac{1}{1-3\rho}} - 1,$$
$$T(w,s) = -\frac{c}{2} + s^{\frac{1-2\rho}{1-3\rho}}\left(\frac{(2w-1)\rho}{c}\right)^{\frac{3\rho}{1-3\rho}}\left(\frac{1}{2} + \frac{3(2w-1)\rho}{2}\right).$$

The two-part tariff implementation of the optimal mechanism is, for $w \in [1/2, 1]$:

$$p_x(w,s) = p_y(w,s) = p_z(w,s) = \frac{w}{2w-1}c,$$

$$p_0(w,s) = \begin{cases} s\left((\frac{1}{2}+(2w-1)\rho)\left(\frac{(2w-1)\rho}{c}\right)^{\frac{2\rho}{1-2\rho}} - \frac{2wc}{2w-1}\left(\frac{(2w-1)\rho}{c}\right)^{\frac{1}{1-2\rho}}\right), \text{ if } w \in [\frac{1}{2}, \hat{w}(s)], \\ -\frac{c}{2} + s^{\frac{1-2\rho}{1-3\rho}}\left((\frac{1}{2}+\frac{3(2w-1)\rho}{2})\left(\frac{(2w-1)\rho}{c}\right)^{\frac{3\rho}{1-3\rho}} - \frac{3wc}{2w-1}\left(\frac{(2w-1)\rho}{c}\right)^{\frac{1}{1-3\rho}}\right), \text{ if } w \in (\hat{w}(s), 1]. \end{cases}$$

Menus of Token Packages If only a total number of tokens is contractible, then the relevant type is $\theta = ws^{1-2\rho}$ which is distributed according to PDF $f_\theta(\theta) = \frac{1}{2\rho}(1-\theta^{\frac{2\rho}{1-2\rho}})$ with the relevant functions being:

$$\varphi(\theta) = 2\,\frac{\theta+\rho-(1-\rho)\theta^{\frac{1}{1-2\rho}}}{1-\theta^{\frac{2\rho}{1-2\rho}}},$$

$$C'(Q) = \begin{cases} \frac{c}{\rho}Q^{\frac{1-2\rho}{2\rho}}, \text{ if } Q < 1, \\ \frac{c}{\rho}Q^{\frac{1-3\rho}{3\rho}}, \text{ if } Q \geq 1. \end{cases}$$

In general, solving for the fine-tuning threshold $\hat{\theta}$, which satisfies $\varphi(\theta) = C'(1)$ is intractable, so we directly focus on the case $\rho = 1/4$, in which

$$\varphi(\theta) = \frac{3\theta-1}{2},$$

and $\hat{\theta} = 1/3 + 8c/3$. In this case, in the optimal menu, types $\theta < 1/3$ are excluded. Types $\theta \in [1/3, \hat{\theta}]$ do not fine-tune and consume:

$$Q(\theta) = \frac{3\theta-1}{8c},$$

$$X(\theta) = Y(\theta) = \left(\frac{3\theta-1}{8c}\right)^2,$$

$$Z(\theta) = 0,$$

$$T(\theta) = \frac{9\theta^2-1}{48c}.$$

Types $\theta \in (\hat{\theta}, 1]$ fine-tune and consume:

$$
\begin{aligned}
Q(\theta) &= \left(\frac{3\theta - 1}{8c}\right)^3, \\
X(\theta) = Y(\theta) &= \left(\frac{3\theta - 1}{8c}\right)^4, \\
Z(\theta) &= \left(\frac{3\theta - 1}{8c}\right)^4 - 1, \\
T(\theta) &= -\frac{2}{3}c + \frac{1 + 9\theta}{12}\left(\frac{3\theta - 1}{8c}\right)^3.
\end{aligned}
$$

The two-part tariff implementation of the optimal mechanism is, for $\theta \in [1/3, 1]$:

$$
\begin{aligned}
m(\theta) &= \frac{2\theta}{3\theta - 1}, \\
p_0(\theta) &= \begin{cases} \frac{1}{6}\frac{3\theta-1}{8c}, \text{ if } \theta \in [\frac{1}{3}, \hat{\theta}], \\ \frac{2c}{3(3\theta-1)} + \frac{1}{12}\left(\frac{3\theta-1}{8c}\right)^3, \text{ if } \theta \in (\hat{\theta}, 1]. \end{cases}
\end{aligned}
$$

References

Babaioff, M., R. Kleinberg, and R. Paes Leme (2012): "Optimal Mechanisms for Selling Information," in *Proceedings of the 13th ACM Conference on Electronic Commerce*, 92–109.

Bergemann, D., M. Bojko, P. Dütting, R. Paes Leme, H. Xu, and S. Zuo (2024): "Data-Driven Mechanism Design: Jointly Eliciting Preferences and Information," *arXiv preprint arXiv:2412.16132*.

Bergemann, D., A. Bonatti, and A. Smolin (2018): "The Design and Price of Information," *American Economic Review*, 108, 1–48.

Castro-Pires, H., H. Chade, and J. Swinkels (2024): "Disentangling Moral Hazard and Adverse Selection," *American Economic Review*, 114, 1–37.

Daskalakis, C., A. Deckelbaum, and C. Tzamos (2017): "Strong Duality for a Multiple-Good Monopolist," *Econometrica*, 85, 735–767.

Devanur, N. R., K. Goldner, R. R. Saxena, A. Schvartzman, and S. M. Weinberg (2020): "Optimal Mechanism Design for Single-Minded Agents," in *Proceedings of the 21st ACM Conference on Economics and Computation*, 193–256.

Duetting, P., V. Mirrokni, R. Paes Leme, H. Xu, and S. Zuo (2024): "Mechanism design for large language models," in *Proceedings of the ACM on Web Conference 2024*, 144–155.

Fiat, A., K. Goldner, A. R. Karlin, and E. Koutsoupias (2016): "The Fedex Problem," in *Proceedings of the 2016 ACM Conference on Economics and Computation*, 21–22.

Fish, S., Y. A. Gonczarowski, and R. I. Shorrer (2024): "Algorithmic Collusion by Large Language Models," *arXiv preprint arXiv:2404.00806*.

Haghpanah, N. and R. Siegel (2024): "Screening Two Types," Tech. rep., Penn State.

Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020): "Scaling Laws for Neural Language Models," *arXiv preprint arXiv:2001.08361*.

Mahmood, R. (2024): "Pricing and Competition for Generative AI," *arXiv preprint arXiv:2411.02661*.

Maskin, E. and J. Riley (1984): "Monopoly with Incomplete Information," *RAND Journal of Economics*, 15, 171–196.

Mussa, M. and S. Rosen (1978): "Monopoly and Product Quality," *Journal of Economic Theory*, 18, 301–317.

Rochet, J.-C. and L. A. Stole (2003): "The Economics of Multidimensional Screening," *Econometric Society Monographs*, 35, 150–197.

Tirole, J. (1988): *The Theory of Industrial Organization*, Cambridge: MIT Press.

Yang, K. H. (2022): "Selling Consumer Data for Profit: Optimal Market-Segmentation Design and Its Consequences," *American Economic Review*, 112, 1364–1393.