Winning the large language capability battle and losing the production economy

Peter Cotton

Abstract

The race to dominate machine intelligence is scored on capability benchmarks. This may be the wrong metric: in competitive equilibrium, work goes to whichever model creates the most value per token of compute, and we exhibit a closed form in which the most capable model's share of production compute tends to zero. In competitive deployment, task-level average PnL per token equals the shadow price of compute divided by the elasticity of task value with respect to effective compute; at a common shadow price it ranks productivity exactly when that elasticity declines with crowding. The model supplies no canonical task-invariant scalar quality: any aggregate PnL-per-token ranking embeds a task distribution and deployment weighting, measuring the economic niches a model occupies.

Keywords: division of labor, compute allocation, artificial intelligence, technological displacement, marginal product
JEL: D24, D41, O33

1. Introduction

The contest for leadership in large language models is treated as strategically decisive. Governments restrict exports of the chips that train frontier models, laboratories commit capital on the scale of national infrastructure, and progress is scored on capability benchmarks. Most public leaderboards primarily rank capability, sometimes alongside price and latency. Almost none ranks realized economic value net of compute expenditure. Beneath the race sits a premise: the most capable model captures the work, and with it the economic value of machine intelligence.

The premise fails in equilibrium. When compute is itinerant, free to move between prediction tasks and to run any available algorithm, and returns to

Email address: peter.cotton@gsmc.ai (Peter Cotton)

compute on each task diminish with crowding, each subcontractible task goes to whichever model creates the most value per token of compute. Capability carries no allocation value beyond its contribution to that quotient: we exhibit a closed-form equilibrium in which a model that outscores every rival on every benchmark has production compute share tending to zero as the division of labor refines. The observable that measures the intensive economic fitness driving this allocation is profit-and-loss (PnL) per token in a competitive prediction market, which pays each entry its marginal contribution at the state it faced. Task-level average PnL per token under sole supply equals the shadow price of compute divided by the elasticity of task value with respect to effective compute; at a common shadow price it ranks productivity when that elasticity declines with crowding, and no comparable theorem covers an arbitrary lifetime quotient.

The mechanism is old. Smith (1776) tied the division of labour to the extent of the market; Ricardo (1817) showed the absolutely better party does not get all the work; Becker and Murphy (1992) bound specialization by coordination costs; Acemoglu and Autor (2011) model production as a task continuum with technologies assigned by comparative advantage. Cotton (2022) supplies the vision this note formalizes: analytical tasks naturally evolve into radically low-cost prediction supply chains. We import the classical skeleton, replace labor with compute, and add the feature specific to prediction: alpha decays with crowding, so returns to effort on a task are concave.

2. The economy

Work is a unit mass $\Omega = [0, 1]$ carrying Lebesgue measure μ. At date t it is partitioned into finitely many subcontractible tasks $\mathcal{P}_t = \{B_1, \ldots, B_{n_t}\}$, and partitions refine: every cell of $\mathcal{P}_t$ is a union of cells of $\mathcal{P}_{t+1}$. Finitely many models $m \in \mathcal{M}_t$ are available, with $\mathcal{M}_t \subseteq \mathcal{M}_{t+1}$. Each model m has a feasibility set $A_m \subseteq \Omega$, the work it can perform, itself a union of cells of $\mathcal{P}_t$ whenever $m \in \mathcal{M}_t$, and a measurable productivity $\theta_m : A_m \to (0, \infty)$, constant on cells at date t; write $\theta_m(B)$ for its value on a cell $B \subseteq A_m$, and require $q_{mB} = 0$ otherwise.

A *token* is a standardized unit of compute cost, not a vendor text token, and PnL is expected trading gain under a fixed capital and risk protocol. Without both normalizations, PnL per token is not comparable across models.

Let $q_{mB} \geq 0$ be the density of tokens of model m on task B. A token of model m contributes $\theta_m(B)$ units of effective effort, so effort on B is

$z_B = \sum_m \theta_m(B)\, q_{mB}$, and the task pays $\mu(B)\,\phi(z_B)$, where ϕ is continuously differentiable on $[0,\infty)$ with $\phi(0) = 0$, $\phi' > 0$, and ϕ' strictly decreasing. The concavity says marginal alpha declines with crowding: prediction opportunities exhibit diminishing returns to compute. Assume $K > 0$, every cell has positive measure, and every cell is feasible for some model. Total value is

$$V(q) = \sum_{B \in \mathcal{P}_t} \mu(B)\,\phi\Big(\sum_m \theta_m(B)\, q_{mB}\Big), \qquad \sum_B \mu(B) \sum_m q_{mB} = K. \tag{1}$$

Compute is *itinerant*: at zero switching cost a token may move to any task and run any available model. At the margin, a token of model m on task B earns $r_{mB} = \theta_m(B)\,\phi'(z_B)$. An equilibrium is an allocation and a return λ with $r_{mB} = \lambda$ wherever $q_{mB} > 0$ and $r_{mB} \le \lambda$ elsewhere. In English: no token can raise its PnL by moving. Existence and efficiency are immediate because the routing game has the concave potential V over the compute simplex, and λ is the shadow price of compute.

Proposition 1 (Displacement). *If $\theta_{m^*}(B) > \theta_m(B)$ for every $m \neq m^*$, then in any equilibrium supplying B, only m^* runs on B.*

Proof. If some $m \neq m^*$ had $q_{mB} > 0$ then

$$\theta_m(B)\,\phi'(z_B) = \lambda \qquad \text{and} \qquad \theta_{m^*}(B)\,\phi'(z_B) > \lambda,$$

so a token improves by switching to m^* on the same task. □

Selection therefore runs on the upper envelope $\theta^*(B) = \max_m \theta_m(B)$, the maximum over models feasible on B, and an inferior technology receives zero execution share, not a small one. Displacement is this stark because value aggregates through the scalar z_B: model outputs are perfect substitutes after productivity scaling. Models with complementary errors, whose ensemble improves on either alone, fall outside Proposition 1 and can survive being individually inferior. Take $\phi(z) = \log(1+z)$. Active tasks satisfy $\theta^*(B)/(1+\theta^*(B)q_B) = \lambda$, so

$$q_B = \left(\frac{1}{\lambda} - \frac{1}{\theta^*(B)}\right)_+, \qquad K = \sum_B \mu(B)\left(\frac{1}{\lambda} - \frac{1}{\theta^*(B)}\right)_+, \tag{2}$$

the water-filling allocation. The right side is strictly decreasing in λ, so λ and the aggregate task intensities q_B are unique; the model-level allocation is unique when the frontier model on every active task is unique.

3. A more capable model exits

Let b_m be a model's success probability, c_m its compute cost per unit of work, and v any increasing conversion of accuracy into economic value, so that $\theta_m = v(b_m)/c_m$. The example needs only

$$b_G > b_S \qquad \text{and} \qquad \frac{v(b_G)}{c_G} < \frac{v(b_S)}{c_S}\,: \tag{3}$$

more capable, less productive.

Take $v(b) = b$. A generalist G succeeds with probability 0.99 on every task at $c_G = 4$ tokens per unit of work, so $\theta_G = 0.2475$. Specialist S_k succeeds with probability 0.95 at one token, with feasibility set $A_{S_k} = I_k$ defined below, so $\theta_S = 0.95$ there; the generalist has $A_G = \Omega$. At date t, expose tasks $I_k = (2^{-k}, 2^{-k+1}]$ for $k \le t$, leaving a residual $R_t = [0, 2^{-t}]$ that only G can serve. By Proposition 1 each exposed I_k goes entirely to S_k. The generalist keeps the residual, of measure $u_t = 2^{-t}$.

Compute share vanishes too, not just task measure. With $K = 4$, clearing (2) across the two active regions gives $\lambda_t = [4 + (1 - u_t)/0.95 + u_t/0.2475]^{-1}$ and

$$Q_{G,t} = u_t \left(\frac{1}{\lambda_t} - \frac{1}{0.2475} \right) = u_t \left(1.0122 + 2.9878\, u_t\right) \longrightarrow 0. \tag{4}$$

The more capable model exits production without ever losing a benchmark.

The example generalizes. Let $E_t = \{x : \theta_G(x) \ge \max_{m \in \mathcal{M}_t,\, x \in A_m} \theta_m(x)\}$; the sets decrease, and by Proposition 1 the generalist executes only on E_t. If specialists eventually dominate almost everywhere and equilibrium token intensities are bounded by $\bar{q}$, then $Q_{G,t} \le \bar{q}\, \mu(E_t) \to 0$. Conversely, at any fixed date, unique frontier status on a positive-measure, strictly active set guarantees a positive compute share; a share bounded away from zero through time requires uniform measure and margin conditions. Membership on a null or inactive set supports none, and a tie leaves the share indeterminate: some equilibria assign G positive compute, others none. These are allocation results conditional on which technologies arrive. They do not predict arrival.

4. What PnL per token measures

Marginal PnL per token is not an intrinsic quality of a model. It factorizes as $r_{mB} = \theta_m(B)\, \phi'(z_B)$: the first factor is the model, the second is the state of the market. What is intrinsic, given the task, is $\theta_m(B)$, and at a matched

task and congestion state the market factor is common, so ratios of marginal PnL across models are ratios of θ.

A competitive prediction market pays exactly this margin. Under a market scoring rule each entry is paid its improvement over the standing forecast, and the payments telescope (Hanson, 2007). Telescoping gives additivity of payments; to identify them with the production technology, let $p(z)$ be the standing forecast after effective effort z and *define* $\phi(z) = \mathbb{E}[S(p(z), Y)] - \mathbb{E}[S(p(0), Y)]$, the expected score improvement that effort has purchased. These payments are funded by a principal who procures the forecast, so ϕ is market-compensated information output rather than a welfare measure; the argument needs only these private returns. An increment of effort $dz = \theta\, dq$ then earns $d\Pi = \phi'(z)\, dz = \theta\, \phi'(\theta q)\, dq$ for a model working a task alone, and payments integrate to $\Pi(q) = \int_0^q \theta\, \phi'(\theta s)\, ds = \phi(\theta q)$; sole supply is the case Proposition 1 delivers on strict frontiers, and with interleaved suppliers aggregate payments still telescope but individual attribution is path dependent. Every dollar is a marginal payment at the state it faced; the form $\phi(z) = \log(1 + z)$ is a convenient parametrization of this map, not a consequence of the scoring rule.

Relative to the shadow price of compute, that average is an inverse elasticity.

Lemma 1 (PnL per token is an inverse elasticity). *Let* $\varepsilon_\phi(z) = z\, \phi'(z)/\phi(z)$. *On any active task with productivity* θ *at shadow price* λ,

$$\frac{AP(\theta; \lambda)}{\lambda} = \frac{1}{\varepsilon_\phi(z)}, \qquad \theta\, \phi'(z) = \lambda, \tag{5}$$

and $z(\theta, \lambda)$ *is increasing in* θ. *Hence AP ranks* θ *at a common* λ *if and only if* ε_ϕ *decreases on the relevant range.*

Proof. $AP = \phi(z)/q = \theta\, \phi(z)/z$ and $\lambda = \theta\, \phi'(z)$ give $AP/\lambda = \phi(z)/(z\, \phi'(z))$, and $z = (\phi')^{-1}(\lambda/\theta)$ increases in θ. □

Corollary 1 (Log production ranks). *For* $\phi(z) = \log(1 + z)$, $\varepsilon_\phi(z) = z/\big((1+z)\log(1+z)\big)$ *is strictly decreasing, since* $\log(1+z) < z$; *hence AP is strictly increasing in* θ, *explicitly* $AP = \lambda\, g(\theta/\lambda)$ *with* $g(r) = r \log r/(r-1)$ *and* $g'(r) \propto r - 1 - \log r > 0$.

Concavity alone does not deliver the ranking. At the Inada boundary, admitting an infinite right derivative at zero, the power family $\phi(z) = z^\alpha$ has constant elasticity, so $AP/\lambda = 1/\alpha$ for every θ: the statistic carries no

ranking information at all. Within the smooth class above, take $\phi(z) = z + a(1 - e^{-z})$ with $a > 0$: its elasticity tends to one as $z \to 0$ and as $z \to \infty$ yet, by strict concavity, lies below one at every positive z, so it increases somewhere, and on that range competitive average PnL reverses the productivity ranking. Whether the elasticity of alpha with respect to compute declines with crowding is therefore the empirical condition on which the interpretation of PnL per token rests.

Remark 1 (Elastic compute pins $\lambda = 1$). The λ-conditioning matters only when compute is capacity constrained. If compute is elastically supplied at price c per token, entry gives $\theta\,\phi'(\theta q) = c$; since a token is already a standardized dollar of compute cost, normalize $c = 1$. Then $\lambda = 1$ across unconstrained competitive environments and $AP(\theta) = g(\theta)$ is directly comparable. Deflation is needed only where compute is scarce or its real price moves.

In English: under log production the margin is the same λ for every survivor, but the inframarginal rents are monotone in θ. The λ-conditioning still binds: mixing markets with different shadow prices can invert the quotient even under competitive deployment, and the repair is to deflate, since AP/λ depends on θ/λ alone. Raw PnL per token is a nominal return, and the shadow value of compute is its deflator.

Administered deployment can invert the ranking within one environment, since it need not satisfy the common-λ conditions: $\theta_A = 2$ forced to scale $q_A = 100$ averages $\log 201/100 \approx 0.053$ per token against 0.693 for $\theta_B = 1$ at $q_B = 1$. A laboratory routing all traffic to its own generalist administers exactly this experiment.

Three statistics must therefore be kept apart. The sole-supplier task-level average product obeys Lemma 1. The incremental return $\Delta\text{PnL}/\Delta T \approx \theta_m(B)\,\phi'(z_B)$, taken in small increments at matched task and congestion, estimates the marginal product directly and is the right statistic under randomized or interleaved deployment. An arbitrary lifetime quotient is a deployment-weighted average of marginal returns across tasks, congestion states, and scarcity regimes, and no inverse-elasticity theorem applies to it without further structure.

One attempt at a realized-PnL ranking, `modelrank.net`, currently divides prediction-market PnL by vendor-metered tokens under randomized routing, with open positions marked to market. Read through the model, that is a randomized experiment estimating average incremental trading contribution over an assigned distribution of tasks and states: the randomization

serves causal identification, and is not itself the itinerant equilibrium. The theory's version of the statistic denominates in dollar compute cost, estimates ΔPnL per dollar at matched task and state, deflates by the contemporaneous shadow price, and presumes a common, or suitably normalized, task-response ϕ across the tasks compared. Nor is there a canonical task-invariant scalar underneath: a model serving many tasks reports a token-weighted portfolio of $\lambda\, g(\theta_m(B)/\lambda)$, so every aggregate ranking embeds a task distribution and deployment weighting, measuring the economic niches a model occupies. Capability benchmarks measure b alone, while equilibrium prices b jointly with compute cost and task value through $\theta = v(b)/c$.

5. Discussion

Capable large models need not disappear. Their productivity advantage may move up the production hierarchy, to decomposition, routing, verification, and the hard residual, whenever their value per token on those tasks sits on the frontier; in the present model that statement is an application of Proposition 1 rather than a theorem, since decomposition here is free and exogenous. For strategy, the question is therefore not who owns the most capable model but who owns the productivity frontier on an economically material part of the limiting division of labor. PnL per token is the intensive margin: value per token on the niches occupied. Production share also requires the extensive margin, the economic mass of those niches; a model with $\theta = 100$ on a niche of measure 10^{-9} combines spectacular PnL per token with negligible share. Commercial survival requires more still. Under log production an active task, $\theta > \lambda$, yields rent $R(\theta, \lambda) = \log(\theta/\lambda) - 1 + \lambda/\theta$ per unit of task mass above the shadow opportunity cost of its compute, and an inactive task yields zero; with compute elastically supplied at unit cash cost (Remark 1), $\lambda = 1$ and $R(\theta) = \log\theta - 1 + 1/\theta$ is operating rent above actual compute cost. It is total rent across owned tasks, not average PnL per token, that must cover fixed maintenance and development cost. The chain runs: task productivity, PnL per token, allocation, total rents, commercial survival. This note proves the middle of the chain; the entry model closing it, rents against fixed cost, is left to future work.

References

Acemoglu, D., Autor, D., 2011. Skills, tasks and technologies: Implications for employment and earnings, in: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, Vol. 4B. Elsevier, pp. 1043–1171.

Becker, G.S., Murphy, K.M., 1992. The division of labor, coordination costs, and knowledge. Quarterly Journal of Economics 107, 1137–1160.

Cotton, P., 2022. Microprediction: Building an Open AI Network. MIT Press, Cambridge, MA.

Hanson, R., 2007. Logarithmic market scoring rules for modular combinatorial information aggregation. Journal of Prediction Markets 1, 3–15.

Ricardo, D., 1817. On the Principles of Political Economy and Taxation. John Murray, London.

Smith, A., 1776. An Inquiry into the Nature and Causes of the Wealth of Nations. W. Strahan and T. Cadell, London.