

LLM Time-Saving and Demand Theory

Tom Cunningham

2025-11-19

Many thanks to Elsie Jang for discussion.

Summary

I anticipate that we're going to be arguing a lot about LLM-speedups & task substitution over the next year.

- Suppose an LLM speeds you up by a factor β on tasks that take up share s of your work (pre-LLM), what's the overall efficiency gain?
- How does the answer change if you measure the time-share s , *after* you adjust to use LLMs?

Quick checklist:

1. The total **output gain** will be between $\frac{1}{(1-s)+s/\beta}$ (if perfect complements/Amdahl) and β (perfect substitutes).
2. If the time-savings are tiny then Hulten's theorem applies: $\Delta \ln y \approx s \ln \beta$ (and for tiny changes, time-saved $\approx \Delta \ln y$ too).
3. If the time-savings are small then you can use elasticity of substitution (there's a closed-form expression but it's ugly).
4. If the time-savings are large then you should use the entire area under the demand curve (i.e. not reasonable to assume constant elasticity). Though in demand theory this relies on income effects being small, & I think a sufficient condition is that the expenditure share is small. When the expenditure share is large I'm not confident how to think about this.
5. Applying Amdahl's law ex-ante will under-estimate productivity improvements; applying it ex-post will over-estimate productivity improvements.

	10% saving on 50%	50% saving on 10%	80% saving on 10%
$\varepsilon = 0$ (perfect complements/Amdahl)	5.0%	5.0%	8.0%
$\varepsilon = 1/2$ (complements)	5.1%	5.8%	10.7%
$\varepsilon = 1$ (Cobb-Douglas/Hulten)	5.1%	6.7%	14.9%
$\varepsilon \rightarrow \infty$ (perfect substitutes)	10%	50%	80%

Time savings using *ex-ante* time shares

	10% saving on 50%	50% saving on 10%	80% saving on 10%
$\varepsilon = 0$ (perfect complements/Amdahl)	5.3%	9.1%	28.6%
$\varepsilon = 1/2$ (complements)	5.2%	7.8%	20.8%
$\varepsilon = 1$ (Cobb-Douglas/Hulten)	5.1%	6.7%	14.9%

	10% saving on 50%	50% saving on 10%	80% saving on 10%
$\varepsilon \rightarrow \infty$ (perfect substitutes)	N/A	N/A	N/A

Time savings using *ex-post* time shares:

How these are calculated:

- “**X% saving**” means task-2 productivity increases so that time-per-unit falls by factor $(1 - X)$, i.e., $\beta = 1/(1 - X)$. So 10% saving $\rightarrow \beta = 1.11$; 50% saving $\rightarrow \beta = 2$; 80% saving $\rightarrow \beta = 5$.
- “**Y% of work**” means the time share on task 2 is $s = Y$.
- **Output gain** from the CES formula:

$$\frac{y'}{y} = ((1 - s_0) + s_0 \beta^{\varepsilon-1})^{1/(\varepsilon-1)}$$

where s_0 is the *ex-ante* share.

- **Time savings** $= 1 - (y'/y)^{-1}$, i.e., the fraction of time saved to produce the same output.
- **Table 1:** The column header specifies the ex-ante share s_0 directly. Compute the output gain and convert to time savings.
- **Table 2:** The column header specifies the ex-post share s_1 . First back out the implied ex-ante share using:

$$\frac{s_0}{1 - s_0} = \frac{s_1}{1 - s_1} \cdot \beta^{1-\varepsilon}$$

Then compute the true output gain using s_0 .

- **Perfect substitutes (Table 2):** After any productivity improvement, you reallocate entirely to the better task, so the ex-post share is always 100%. Specifying it as 10% or 50% is inconsistent with optimization—hence N/A.

An analogy: if we can turn lead into gold, what benefit do I get? Using ex ante expenditure: my expenditure share on gold is ~0%, so my benefit is very small.

Using ex post expenditure: if gold is cheap then I’ll start buying gold cutlery. Suppose I spend \$1K on gold/year, then it makes it look like I’m getting value worth \$100K/year, which is clearly wrong.

I think the resolution is that gold has high substitutability with other goods (steel, bronze), and so demand is highly elastic. But that substitutability only appears when prices are low, so just estimating a CES function I think would get this wrong.

There are some nice crisp results from economics that apply here. I discuss some related literature below.

Loose ends:

- I give bounds on aggregate time-savings with a CES model below, but I’m not sure whether you might not get wider bounds if you relax the CES assumption, e.g. with non-homotheticities so there are income effects.

Model

We set up a two-task CES production problem and derive the optimal time split, the implied output, and the response to productivity changes, with limits for common special cases.

Practical implications (at a glance)

Let $s \equiv t_2^*$ denote the optimal time share on task 2 (and $1 - s = t_1^*$). Express all effects as log-changes $\Delta \ln y^* = \ln(y^{*'}/y^*)$ when task-2 productivity moves from A_2 to $A_2' = \beta A_2$. The last column plugs in $s = 0.1$ and $\beta = 2$.

Case	Output effect ($\Delta \ln y^*$)	Intuition	Example $\Delta \ln y^*$ ($s = 0.1, \beta = 2$)
General finite change	$\frac{1}{\varepsilon - 1} \ln((1 - s) + s \beta^{\varepsilon - 1})$	CES-weighted average of the shock	$\frac{1}{\varepsilon - 1} \ln(0.9 + 0.1 \times 2^{\varepsilon - 1})$ (depends on ε)
Perfect substitutes ($\varepsilon \rightarrow \infty$)	$\ln \beta$	All time moves to the better task	≈ 0.69
Cobb–Douglas ($\varepsilon = 1$)	$s \ln \beta$	Log-linear weighting by the task share	≈ 0.069
Perfect complements ($\varepsilon \rightarrow 0$)	$-\ln((1 - s) + s/\beta)$	Bottlenecked by the slow task	≈ 0.051
Infinitesimal change (Hulten)	$s d \ln A_2$	Percent gain equals time share on improved task	$0.1 \times \ln 2 \approx 0.069$

Setup and parameters

- Time endowment is 1; choose $t_1 \in [0, 1]$ and $t_2 = 1 - t_1$.
- Productivities: $A_1 > 0$ for task 1, $A_2 > 0$ for task 2.
- Taste weight: $\alpha \in (0, 1)$ on task 1.
- Substitution parameter: $\varepsilon > 0$; take $\varepsilon \neq 1$ for the algebra and then send $\varepsilon \rightarrow 1$ for the Cobb–Douglas limit.
- Output aggregator (CES):

$$y(t_1, t_2) = \left(\alpha (A_1 t_1)^{\frac{\varepsilon - 1}{\varepsilon}} + (1 - \alpha) (A_2 t_2)^{\frac{\varepsilon - 1}{\varepsilon}} \right)^{\frac{\varepsilon}{\varepsilon - 1}}.$$

Assumptions

1. Feasible set: $t_1 \in [0, 1], t_2 = 1 - t_1$.
2. Parameters satisfy $A_i > 0$ and $\alpha \in (0, 1)$.
3. Decision problem: choose t_1 to maximise $y(t_1, 1 - t_1)$.

Proposition 1 (optimal time split). The interior optimum is

$$t_1^* = \frac{1}{1 + \left(\frac{1 - \alpha}{\alpha}\right)^\varepsilon \left(\frac{A_2}{A_1}\right)^{\varepsilon - 1}}, \quad t_2^* = 1 - t_1^*.$$

Proof (explicit)

1. Write the Lagrangian $\mathcal{L} = y(t_1, t_2) + \lambda(1 - t_1 - t_2)$ with y as above.
2. First-order conditions (interior): $\partial \mathcal{L} / \partial t_1 = 0$ and $\partial \mathcal{L} / \partial t_2 = 0$ give

$$\lambda = \alpha A_1^{\frac{\varepsilon - 1}{\varepsilon}} t_1^{-\frac{1}{\varepsilon}} y^{\frac{1}{\varepsilon}} = (1 - \alpha) A_2^{\frac{\varepsilon - 1}{\varepsilon}} t_2^{-\frac{1}{\varepsilon}} y^{\frac{1}{\varepsilon}}.$$

3. Cancel $y^{\frac{1}{\varepsilon}}$ and rearrange to obtain $\frac{t_2}{t_1} = \left(\frac{1 - \alpha}{\alpha}\right)^\varepsilon \left(\frac{A_2}{A_1}\right)^{\varepsilon - 1}$.
4. Impose $t_1 + t_2 = 1$ and solve for t_1^* ; set $t_2^* = 1 - t_1^*$.
5. The interior solution is valid for $\varepsilon > 0$ with finite A_i ; only the perfect-substitutes limit $\varepsilon \rightarrow \infty$ or $A_i \rightarrow 0$ forces a corner.

Proposition 2 (indirect output). At t_1^*, t_2^* the output is

$$y^* = \left(\alpha^\varepsilon A_1^{\varepsilon - 1} + (1 - \alpha)^\varepsilon A_2^{\varepsilon - 1} \right)^{\frac{1}{\varepsilon - 1}}.$$

Proof (explicit)

1. Substitute t_1^*, t_2^* from Proposition 1 into $y(t_1, t_2)$.
2. Factor out $\alpha^\varepsilon A_1^{\varepsilon - 1} + (1 - \alpha)^\varepsilon A_2^{\varepsilon - 1}$ inside the braces; the exponent $\frac{\varepsilon}{\varepsilon - 1}$ collapses to the stated form.

Proposition 3 (infinitesimal productivity change). Holding A_1 fixed, a small change in A_2 satisfies

$$\frac{dy^*}{y^*} = t_2^* \frac{dA_2}{A_2}.$$

Proof (explicit)

1. Take $\log y^* = \frac{1}{\varepsilon-1} \log (\alpha^\varepsilon A_1^{\varepsilon-1} + (1-\alpha)^\varepsilon A_2^{\varepsilon-1})$.
2. Differentiate with respect to $\log A_2$:

$$\frac{dy^*}{y^*} = \frac{(1-\alpha)^\varepsilon A_2^{\varepsilon-1}}{\alpha^\varepsilon A_1^{\varepsilon-1} + (1-\alpha)^\varepsilon A_2^{\varepsilon-1}} \cdot \frac{dA_2}{A_2}.$$

3. The fraction equals t_2^* from Proposition 1, so the result follows.

Proposition 4 (finite productivity change on task 2). If $A_2' = \beta A_2$ with $\beta > 0$, then

$$\frac{y^{*'}}{y^*} = (t_1^* + (1-t_1^*)\beta^{\varepsilon-1})^{\frac{1}{\varepsilon-1}}.$$

Proof (explicit)

1. Replace A_2 by βA_2 in y^* from Proposition 2:

$$y^{*'} = \left(\alpha^\varepsilon A_1^{\varepsilon-1} + (1-\alpha)^\varepsilon (\beta A_2)^{\varepsilon-1} \right)^{\frac{1}{\varepsilon-1}}.$$

2. Factor out the old level y^* to form a ratio; the remaining weights inside the braces are t_1^* and $t_2^* = 1 - t_1^*$, giving the stated expression.

Proposition 5 (canonical limits). Take limits of Proposition 4:

- Cobb–Douglas ($\varepsilon \rightarrow 1$): $\frac{y^{*'}}{y^*} \rightarrow \beta^{1-\alpha}$ and t_i^* is unchanged.
- Perfect complements ($\varepsilon \rightarrow 0$): $\frac{y^{*'}}{y^*} \rightarrow \frac{1}{t_1^* + t_2^*/\beta}$.
- Perfect substitutes ($\varepsilon \rightarrow \infty$): $\frac{y^{*'}}{y^*} \rightarrow \beta$ with $t_2^* \rightarrow 1$ if $\beta A_2 > A_1$.

Proof sketch For $\varepsilon \rightarrow 1$ apply L'Hôpital to the CES form. For $\varepsilon \rightarrow 0$ the CES aggregator converges to $\min\{A_1 t_1, A_2 t_2\}$. For $\varepsilon \rightarrow \infty$ it converges to $\max\{A_1 t_1, A_2 t_2\}$. Substitute these limits into Proposition 4 and simplify.

Example Applications

Anthropic (2025) They estimate Claude saves time by 80% on certain tasks, and apply Hulten's theorem. But Hulten's theorem only applies for *small* efficiency changes, so I believe this effectively assumes Cobb–Douglas substitution (i.e. constant time-shares on these tasks).

Becker et al. (2025) They estimate time-savings on tasks using LLMs. In this case the subjects mostly were *not* using AI, but in followup studies they *will* be using AI. This makes it hard to think about interpreting uplift studies over time, insofar as AI causes them to change the task distribution. It would be nice to have a good clear language here.

Related Theory

Index numbers and the “which share do you weight by?” problem Caves, Christensen, and Diewert (1982) formalize why productivity (or “quantity”) measurement becomes an index-number problem as soon as *shares move endogenously*. Their key point is that fixed-weight indices answer the wrong question when agents substitute:

- A Laspeyres-style measure (base-period weights) is biased toward *understating* gains when a shock makes you reallocate toward the improved input/task, because it freezes the old bundle.
- A Paasche-style measure (end-period weights) is biased toward *overstating* gains if you implicitly treat the new bundle as if it had always been purchased at the old relative prices/technologies.
- The object you actually want is a path (Divisia) integral of share-weighted growth rates; “superlative” discrete indices (e.g., Fisher/Törnqvist) are designed to approximate that integral well.

Connection to our setup: our two-task CES makes this completely explicit. Proposition 3 is the Divisia/Hulten-style statement:

$$d \ln y^* = t_2^* d \ln A_2,$$

so the “right weight” is the *current* optimal time share t_2^* , not a fixed pre/post share. For a finite β shock, the exact change is the integral of that weight along the adjustment path, which collapses to the closed form in Proposition 4. This is exactly the sense in which “ex-ante Amdahl/Hulten” (freeze t_2 at baseline) understates and “ex-post Hulten” (freeze t_2 at the endpoint) can overstate.

Consumer surplus / welfare for large “price” (time-cost) changes Willig (1976) defends using Marshallian consumer surplus as a welfare proxy: the area under the (Marshallian) demand curve is close to the Hicksian objects (EV/CV) when income effects are small, and Willig provides conditions/bounds under which the approximation error is limited.

Hausman (1981) shows how to compute exact welfare measures (EV/CV, deadweight loss) from an estimated demand curve by imposing integrability (i.e., that the demand actually comes from some underlying utility/expenditure function).

Deaton and Muellbauer (1980) provides a workhorse *integrable* demand system (AIDS) that flexibly captures income effects and substitution patterns while keeping welfare calculations coherent.

Connection to our setup: if you reinterpret the model in dual form, “task outputs” ($x_i = A_i t_i$) are the “goods,” and the time budget implies prices ($p_i = 1/A_i$) (time per unit of task output). An LLM that raises A_2 by β is literally a price drop for good 2 by a factor $1/\beta$. For small task shares ($s = t_2^*$) (your “gold is tiny in the budget” example), the Willig logic says “income-effect-like problems are muted,” so simple surplus approximations are safer. When the effective budget share is not tiny—or when the shock is huge—the Hausman/Deaton-Muellbauer message is: you want an *integrable system* (or in our case, an explicit aggregator like CES) so that “area under a curve” corresponds to a defensible welfare change.

Economics of time allocation (time is a scarce input with shadow prices) DeSerpa (1971) treats time as a fundamental scarce resource: activities require time as well as market goods, and constraints on required time generate a shadow value of time. The key conceptual result is that “time-saving” innovations matter because they relax a binding constraint, and the value of that relaxation depends on the shadow price of time and substitution across activities.

Connection to our setup: we’re essentially running the cleanest possible DeSerpa-style problem: a single time endowment allocated across tasks, where task technologies (A_i) turn time into effective task output. The LLM shock is a direct relaxation of the time requirement for task 2. What DeSerpa adds conceptually is the interpretation: the welfare/productivity effect is not “minutes saved” per se, but “minutes saved *valued at the shadow price of time*,” and that shadow price is determined jointly with substitution across tasks—exactly what ε is doing for us.

Task substitution and computerization as task-specific technology shocks Autor, Levy, and Murnane (2003) is the empirical foundation for the modern “tasks” approach: computerization substitutes for routine tasks and complements non-routine tasks, shifting task content within occupations and generating distributional consequences (e.g., polarization). The paper’s core claim is that you cannot summarize tech change as “labor-augmenting” in the aggregate; it’s task-directed, and substitution patterns matter.

Acemoglu and Autor (2011) synthesizes and formalizes this task-based view. A central message is that the impact of a task-specific productivity shock depends on: (i) which tasks are affected, (ii) how substitutable tasks are, and (iii) how the economy re-optimizes task assignment across workers/technologies.

Connection to our setup: our $A_2 \mapsto \beta A_2$ change is exactly a “task-directed” shock. Your entire “ex-ante vs ex-post task share” issue is basically the same point Autor/Acemoglu-Autor emphasize: task shares are

endogenous responses to technology, not fixed primitives. In our CES aggregator, that endogeneity is governed by ε : when ε is high, reallocation is large and the “Amdahl intuition” breaks; when ε is low, reallocation is small and Amdahl becomes accurate.

Hulten’s theorem and when first-order share-weighting breaks Hulten (1978) shows (in a competitive, CRS setting with intermediates) that a *small* productivity shock’s effect on aggregate productivity can be summarized by share-weighted sectoral TFP growth (Domar/revenue-share weights). The key takeaway is the legitimacy of first-order share weighting—but only locally.

Baqee and Farhi (2019) pushes beyond that local logic: in production networks, micro shocks can have macro consequences and nonlinearities/higher-order terms matter, especially for large shocks or when substitution patterns interact with network structure.

Baqee and Burstein (2021) highlights that welfare vs. output can diverge when income effects / demand shifts are important, and that changing expenditure patterns can be central to aggregate responses (so “just Domar weights” can miss key welfare channels).

Comin, Lashkari, and Mestieri (2021) emphasizes that over longer horizons, relative price changes and non-homotheticities induce structural change: shares drift systematically, so you should expect “task/sector shares” to be part of the response to technology, not a nuisance.

Connection to our setup: Proposition 3 is the Hulten logic in a two-task world (local, share-weighted). Proposition 4 is the “nonlinear / finite change” version where the share itself moves with β .

Amdahl’s law as the perfect-complements benchmark Amdahl’s law says the speedup from improving one component is bounded by the unimproved fraction.

Connection to our setup: that’s precisely the $\varepsilon \rightarrow 0$ (Leontief / perfect complements) limit in Proposition 5:

$$\frac{y^*'}{y^*} \rightarrow \frac{1}{(1-s) + s/\beta}.$$

<!-- **Reallocation and aggregation across units (Olley–Pakes)** Olley and Pakes (1996) shows that aggregate productivity can be decomposed into: (i) the unweighted mean productivity across firms (“within”), plus (ii) a covariance between productivity and market share (“between” / reallocation). Their empirical contribution is that changes in aggregate productivity can come as much from reallocating activity toward more productive firms as from within-firm improvements.-->

Illustrations

Indifference Curve

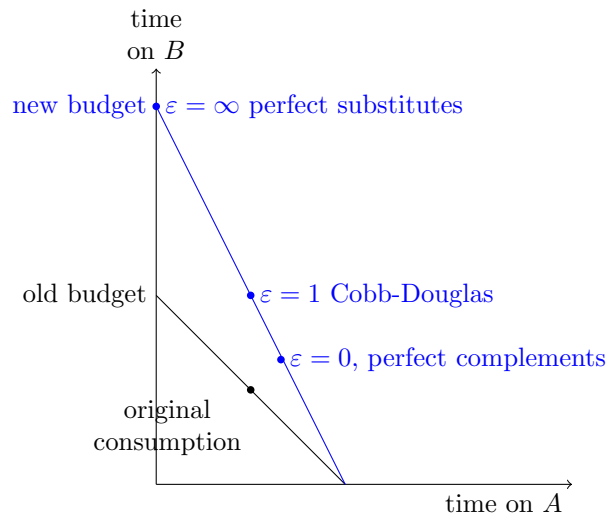


Figure 1: Budget constraint and optimal allocations under different elasticities

Demand Curve

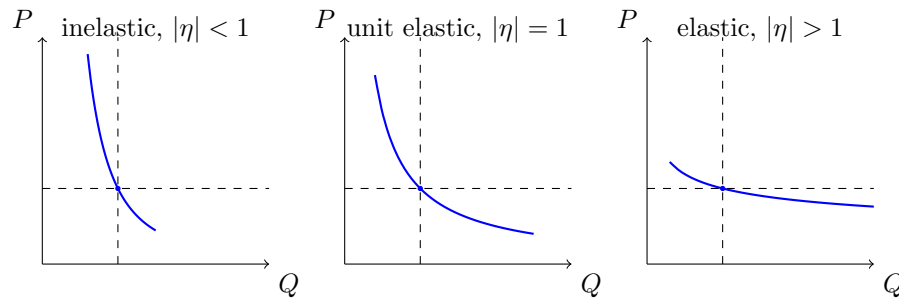


Figure 2: Demand curves with different price elasticities

- Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In, edited by David Card and Orley Ashenfelter, 4:1043–1171. *Handbook of Labor Economics*. Elsevier. [https://doi.org/https://doi.org/10.1016/S0169-7218\(11\)02410-5](https://doi.org/https://doi.org/10.1016/S0169-7218(11)02410-5).
- Anthropic. 2025. "Estimating AI Productivity Gains from Claude Conversations." 2025. <https://www.anthropic.com/research/estimating-productivity-gains>.
- Autor, David, Frank Levy, and Richard J Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *The Quarterly Journal of Economics* 118 (4): 1279–1333.
- Baqaei, David Rezza, and Ariel Burstein. 2021. "Welfare and Output with Income Effects and Demand Instability." Working Paper.
- Baqaei, David Rezza, and Emmanuel Farhi. 2019. "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem." *Econometrica* 87 (4): 1155–1206.
- Becker, Joel, Nate Rush, Elizabeth Barnes, and David Rein. 2025. "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity." <https://arxiv.org/abs/2507.09089>.
- Caves, Douglas W., Laurits R. Christensen, and W. Erwin Diewert. 1982. "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity." *Econometrica* 50 (6): 1393–1414. <https://www.jstor.org/stable/1913382>.
- Comin, Diego, Danial Lashkari, and Martín Mestieri. 2021. "Structural Change with Long-Run Income and Price Effects." Working Paper.
- Deaton, Angus, and John Muellbauer. 1980. "An Almost Ideal Demand System." *American Economic Review* 70 (3): 312–26.
- DeSerpa, Allan C. 1971. "A Theory of the Economics of Time." *The Economic Journal* 81 (324): 828–46.
- Hausman, Jerry A. 1981. "Exact Consumer's Surplus and Deadweight Loss." *American Economic Review* 71 (4): 662–76.
- Hulten, Charles R. 1978. "Growth Accounting with Intermediate Inputs." *The Review of Economic Studies* 45 (3): 511–18.
- Olley, G. Steven, and Ariel Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica* 64 (6): 1263–97.
- Willig, Robert D. 1976. "Consumer's Surplus Without Apology." *American Economic Review* 66 (4): 589–97.