LLM Time-Saving and Demand Theory

Tom Cunningham (METR)*

2026-01-28

Suppose an LLM speeds you up by factor β on tasks that are a share s of your time. What is your overall increase in output for a given time? Should we expect it to be $s \times (\beta - 1)$, or larger or smaller?

How does the answer change if you measure the time-share s before vs after you start using LLMs?

If we can observe the time-shares both before and after LLMs, does that help with estimating the overall efficiency gain?

Economic theory has crisp answers to these questions. These questions are all equivalent to classic economic questions of how people change expenditure in response to changes in prices. Below I give a cheat sheet, a lookup table, derivations, and some brief survey of different relevant literatures (there are surprisingly many related literatures).

My guess is that LLMs are mostly substitutes. LLMs let me complete 16 hours of work in an 8 hour day, but the LLM is mostly accelerating me on things that I wouldn't otherwise spend my time on (e.g. fact checking, literature reviews, visualizations), meaning they are substitutes, and so my effective productivity lift is much lower than a doubling of time.

Cheat Sheet

Assume people spend their time rationally. We will assume people allocate time between sped-up and non-sped-up tasks rationally, trying to maximize their overall output.

The output gain will be between $\frac{1}{(1-s)+s/\beta}$ and β. We can put upper and lower bounds on the effect on aggregate output (for a fixed time input):

1. If the tasks are perfect substitutes: β.
2. If the tasks are perfect complements: $\frac{1}{(1-s)+s/\beta}$

The percent gain is simple in two cases. We can apply the small-change approximation $y'/y \approx 1 + s(\beta - 1)$ if either of two cases holds:

1. Most tasks are affected by the speedup ($s \simeq 1$)
2. The productivity increase is small ($\beta \simeq 1$) (AKA Hulten's theorem)

In these two cases we can estimate the aggregate productivity improvement without knowing the degree of substitutability between tasks.

If the time-savings are somewhat small then you can use elasticity of substitution. If you know the elasticity of substitution between sped-up tasks and other tasks, ε, then we can write an expression for the aggregate increase:

$$\frac{y'}{y} = \left((1-s) + s\,\beta^{\varepsilon-1}\right)^{1/(\varepsilon-1)}$$

If the time-savings are large, use the area under the demand curve. If the time-savings are large then it's more dangerous to assume a constant elasticity. Instead we ideally want to trace out the entire demand curve (i.e. how time-allocated to a task changes as the efficiency increases), and the speed-up will be proportional to the area under the demand curve.

[*] with many thanks to Elsie Jang

Using pre-LLM shares will under-estimate value for substitutes (and over-estimate for complements). If using the simple $s \times (\beta - 1)$ estimate, then using pre-LLM shares will under-estimate productivity improvements when tasks are substitutes, while using post-LLM time-shares will over-estimate; the direction flips when tasks are complements.

Observing pre-LLM and post-LLM shares helps. If you observe the time-share both pre-LLM and post-LLM then you can back out the elasticity of substitution, and thus the aggregate efficiency improvement. Graphically, if we observe the change in budget constraint, and change in consumption point, we can infer substitutability, and therefore aggregate productivity improvement.

Applications

Estimating productivity improvements from query-level time-savings. Anthropic (2025) samples a range of tasks from Claude chatbot logs and estimates the time required for each task both with and without AI assistance.

My understanding of their calculation:

1. Claude is used for around 10% of tasks, $s = 0.10$ (using the pre-LLM distribution of tasks).
2. When Claude is used, time-required falls by 80%, $\beta = 5$.
3. Therefore the total time-saving is around 20% ($\simeq 5^{0.1}$).

However as we discussed above, Hulten's theorem only applies for *small* efficiency changes, but these are large changes (80%), so this approximation will only be accurate assuming Cobb-Douglas substitution, i.e. that time-shares are constant.

==my guess: people are doing tasks they wouldn't otherwise do.==

Estimating time-savings in an RCT. Becker et al. (2025) report an RCT, where software engineers first choose tasks, then get assigned to either with-AI or without-AI conditions. In this case the subjects mostly were *not* using AI, but in follow-up 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.

Lookup Tables

Output increase using *ex-ante* time shares

	1.1X speedup on 50%	2X speedup on 10%	5X speedup on 10%
$\varepsilon = 0$ (perfect complements/Amdahl)	4.8%	5.3%	8.7%
$\varepsilon = 1/2$ (complements)	4.8%	6.1%	12.0%
$\varepsilon = 1$ (Cobb-Douglas/Hulten)	4.9%	7.2%	17.5%
$\varepsilon = 2$ (substitutes)	5.0%	10.0%	40.0%
$\varepsilon \to \infty$ (perfect substitutes)	10.0%	100.0%	400.0%

Output increase using *ex-post* time shares:

	1.1X speedup on 50%	2X speedup on 10%	5X speedup on 10%
$\varepsilon = 0$ (perfect complements/Amdahl)	5.0%	10.0%	40.0%
$\varepsilon = 1/2$ (complements)	4.9%	8.5%	26.2%
$\varepsilon = 1$ (Cobb-Douglas/Hulten)	4.9%	7.2%	17.5%
$\varepsilon = 2$ (substitutes)	4.8%	5.3%	8.7%
$\varepsilon \to \infty$ (perfect substitutes)	N/A	N/A	N/A

How these are calculated:

- **Output gain** from the CES formula:

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

  where s_0 is the *ex-ante* share. The reported numbers are $(y'/y - 1) \times 100\%$.
- **Table 1:** The column header specifies the ex-ante share s_0 directly. Compute the output gain and report the percent increase.
- **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.

Other Points

An analogy: we're turning lead into gold. Suppose I invent a technology to turn lead into gold, so that the price of gold falls by 99%. I'd like to quantify my welfare increase in terms of equivalent income. I could apply a simple share-weighted price-change rule in two ways:

1. If I use *ex ante* expenditure the effect looks small: my expenditure share on gold is ~0%, so even a 100% price decrease has a negligible effect on my effective income.

2. If I use *ex post* expenditure the effect looks large: if gold is sufficiently cheap then I'll start buying many things which are made of gold. Suppose I now spend 1% of my income on gold, then it looks like the price reduction has doubled my effective income, because at the original price of gold I would've had to have twice as much income to afford my new basket.

This discrepancy between *ex ante* and *ex post* values arises because of the high substituability between gold and other materials (steel, bronze). However I'm not confident that we would see that elasticity at *current* prices, it would only occur when gold's price gets sufficiently low, meaning that estimating a CES function wouldn't be sufficient to give a good estimate of the value. To get a realistic estimate we need to map elasticities at different prices, i.e. draw the entire demand function.

Sensitivity to CES. I give bounds on aggregate time-savings with a CES model below, but I'm not sure whether you'd get wider bounds if you relax the CES assumption, e.g. account for second-order effects.

Non-homotheticities. In demand theory there can be significant effects from non-homotheticities.

Tasks are fake. (…)

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 $\Delta \ln y^* = \ln\left(y^{*\prime}/y^*\right)$ 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	$\frac{1}{\varepsilon - 1} \ln\left((1 - s) + s\,\beta^{\varepsilon-1}\right)$	CES-weighted average of the shock	$\frac{1}{\varepsilon - 1} \ln\left(0.9 + 0.1 \times 2^{\varepsilon-1}\right)$ (depends on ε)

Case	Output effect ($\Delta \ln y^*$)	Intuition	Example $\Delta \ln y^*$ ($s = 0.1,\ \beta = 2$)
Perfect substitutes ($\varepsilon \to \infty$)	$\ln \beta$	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 ($\varepsilon \to 0$)	$-\ln\big((1-s) + s/\beta\big)$	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 $\varepsilon \to 1$ 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

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

Proof (structured).

Given: $y(t_1, t_2) = \left(\alpha (A_1 t_1)^{\frac{\varepsilon-1}{\varepsilon}} + (1-\alpha)(A_2 t_2)^{\frac{\varepsilon-1}{\varepsilon}}\right)^{\frac{\varepsilon}{\varepsilon-1}}$ with $t_1 + t_2 = 1$, $t_1, t_2 \in (0, 1)$, $\alpha \in (0, 1)$, $A_1, A_2 > 0$, $\varepsilon > 0$, $\varepsilon \neq 1$.

To show: At an interior optimum, t_1^* and t_2^* have the stated closed form.

1. Define $\rho \equiv \frac{\varepsilon-1}{\varepsilon}$. Since the map $u \mapsto u^{\varepsilon/(\varepsilon-1)}$ is strictly increasing on $u > 0$, maximizing $y(t_1, t_2)$ is equivalent to maximizing the inside of the CES bracket:

$$g(t_1, t_2) \equiv \alpha (A_1 t_1)^{\rho} + (1-\alpha)(A_2 t_2)^{\rho}.$$

2. Form the Lagrangian for the constrained maximization of g:

$$\mathscr{L} = g(t_1, t_2) + \lambda(1 - t_1 - t_2).$$

3. Compute the interior first-order conditions:
   1. $\frac{\partial \mathscr{L}}{\partial t_1} = \alpha \rho A_1^{\rho} t_1^{\rho-1} - \lambda = 0.$
   2. $\frac{\partial \mathscr{L}}{\partial t_2} = (1-\alpha) \rho A_2^{\rho} t_2^{\rho-1} - \lambda = 0.$
4. Eliminate λ by equating the two expressions:

$$\alpha \rho A_1^{\rho} t_1^{\rho-1} = (1-\alpha) \rho A_2^{\rho} t_2^{\rho-1}.$$

5. Cancel $\rho \neq 0$ and rearrange:

$$\left(\frac{t_2}{t_1}\right)^{\rho-1} = \frac{\alpha A_1^{\rho}}{(1-\alpha) A_2^{\rho}}.$$

6. Substitute $\rho - 1 = -\frac{1}{\varepsilon}$ and invert both sides to solve for t_2/t_1:

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

7. Let $K \equiv \left(\frac{1-\alpha}{\alpha}\right)^{\varepsilon} \left(\frac{A_2}{A_1}\right)^{\varepsilon-1}$. Then $t_2 = Kt_1$ and the constraint $t_1 + t_2 = 1$ becomes $t_1(1+K) = 1$, hence

$$t_1^* = \frac{1}{1+K}, \qquad t_2^* = 1 - t_1^*.$$

8. This completes the derivation of the interior critical point. □

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 (structured).

Given: The setup above, and that (t_1^*, t_2^*) is an interior optimum.

To show: $y^* = y(t_1^*, t_2^*)$ equals the stated closed form.

1. Define $\rho \equiv \frac{\varepsilon-1}{\varepsilon}$ and write $y = g^{1/\rho}$ where

$$g(t_1, t_2) \equiv \alpha(A_1 t_1)^{\rho} + (1-\alpha)(A_2 t_2)^{\rho}.$$

2. Consider the Lagrangian $\mathscr{L} = g(t_1, t_2) + \lambda(1 - t_1 - t_2)$ from Proposition 1, and let λ^* denote its multiplier at (t_1^*, t_2^*).
3. From the (interior) first-order condition for t_1:

$$\lambda^* = \alpha \rho A_1^{\rho} (t_1^*)^{\rho-1}.$$

4. Multiply both sides of the previous equality by t_1^*:

$$\lambda^* t_1^* = \alpha \rho (A_1 t_1^*)^{\rho}.$$

5. Analogously, from the first-order condition for t_2:

$$\lambda^* t_2^* = (1-\alpha) \rho (A_2 t_2^*)^{\rho}.$$

6. Add the two equalities from Steps 4 and 5, and use $t_1^* + t_2^* = 1$:
    1. Left-hand side: $\lambda^*(t_1^* + t_2^*) = \lambda^*$.
    2. Right-hand side: $\rho\left[\alpha(A_1 t_1^*)^{\rho} + (1-\alpha)(A_2 t_2^*)^{\rho}\right] = \rho\, g(t_1^*, t_2^*)$. Therefore,

$$\lambda^* = \rho\, g(t_1^*, t_2^*).$$

7. Since $y^* = g(t_1^*, t_2^*)^{1/\rho}$, Step 6 implies

$$\lambda^* = \rho\, (y^*)^{\rho}.$$

8. Return to the first-order condition for t_1 in Step 3 and substitute $\lambda^* = \rho(y^*)^{\rho}$:

$$\rho(y^*)^{\rho} = \alpha \rho A_1^{\rho} (t_1^*)^{\rho-1}.$$

9. Cancel ρ and solve for t_1^*:
    1. $(y^*)^{\rho} = \alpha A_1^{\rho} (t_1^*)^{\rho-1}$.
    2. Multiply both sides by t_1^*:

$$(y^*)^{\rho} t_1^* = \alpha (A_1 t_1^*)^{\rho}.$$

    3. Raise both sides to the power ε (using $\rho\varepsilon = \varepsilon - 1$):

$$(y^*)^{\varepsilon-1} t_1^{*\varepsilon} = \alpha^{\varepsilon} A_1^{\varepsilon-1} t_1^{*\varepsilon-1}.$$

4. Cancel $t_1^{*\varepsilon-1} > 0$ to get

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

10. By the same argument for t_2,

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

11. Use $t_1^* + t_2^* = 1$ and substitute the expressions from Steps 9–10:

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

12. Rearranging gives

$$(y^*)^{\varepsilon-1} = \alpha^\varepsilon A_1^{\varepsilon-1} + (1-\alpha)^\varepsilon A_2^{\varepsilon-1},$$

hence

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

13. □

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 (structured).

Given: $y^* = \left(\alpha^\varepsilon A_1^{\varepsilon-1} + (1-\alpha)^\varepsilon A_2^{\varepsilon-1}\right)^{\frac{1}{\varepsilon-1}}$ from Proposition 2, with A_1 fixed.

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

1. Define

$$S \equiv \alpha^\varepsilon A_1^{\varepsilon-1} + (1-\alpha)^\varepsilon A_2^{\varepsilon-1}.$$

Then Proposition 2 says $y^* = S^{1/(\varepsilon-1)}$.

2. Take logs:

$$\ln y^* = \frac{1}{\varepsilon-1} \ln S.$$

3. Differentiate both sides (holding A_1 fixed):

$$\frac{dy^*}{y^*} = \frac{1}{\varepsilon-1} \frac{dS}{S}.$$

4. Compute dS:
    1. The A_1 term is constant by assumption.
    2. For the A_2 term: $d\big(A_2^{\varepsilon-1}\big) = (\varepsilon-1)A_2^{\varepsilon-2}\,dA_2$. Therefore,

$$dS = (1-\alpha)^\varepsilon(\varepsilon-1)A_2^{\varepsilon-2}\,dA_2.$$

5. Substitute Step 4 into Step 3 and simplify:

$$\frac{dy^*}{y^*} = \frac{(1-\alpha)^\varepsilon A_2^{\varepsilon-2}}{S}\,dA_2 = \frac{(1-\alpha)^\varepsilon A_2^{\varepsilon-1}}{S} \cdot \frac{dA_2}{A_2}.$$

6. Identify the fraction as t_2^*:
    1. From Proposition 2, $(y^*)^{\varepsilon-1} = S$.
    2. From Step 10 in the proof of Proposition 2, $t_2^* = \frac{(1-\alpha)^\varepsilon A_2^{\varepsilon-1}}{(y^*)^{\varepsilon-1}}$.
    3. Combining these, $t_2^* = \frac{(1-\alpha)^\varepsilon A_2^{\varepsilon-1}}{S}$.

7. Substitute Step 6 into Step 5 to get
$$\frac{dy^*}{y^*} = t_2^* \frac{dA_2}{A_2}.$$
8. □

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

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

Proof (structured).

Given: Proposition 2, and a shock $A_2' = \beta A_2$ with $\beta > 0$.

To show: $\dfrac{y^{*'}}{y^*} = \left(t_1^* + (1 - t_1^*)\beta^{\varepsilon-1}\right)^{\frac{1}{\varepsilon-1}}$.

1. Define the “inside” term
$$S \equiv \alpha^\varepsilon A_1^{\varepsilon-1} + (1-\alpha)^\varepsilon A_2^{\varepsilon-1},$$
so that Proposition 2 gives $y^* = S^{1/(\varepsilon-1)}$.
2. After the shock $A_2' = \beta A_2$, the corresponding term is
$$S' \equiv \alpha^\varepsilon A_1^{\varepsilon-1} + (1-\alpha)^\varepsilon (A_2')^{\varepsilon-1} = \alpha^\varepsilon A_1^{\varepsilon-1} + (1-\alpha)^\varepsilon (\beta A_2)^{\varepsilon-1}.$$
3. Apply Proposition 2 to the post-shock economy:
$$y^{*'} = (S')^{1/(\varepsilon-1)}.$$
4. Form the ratio:
$$\frac{y^{*'}}{y^*} = \left(\frac{S'}{S}\right)^{\frac{1}{\varepsilon-1}}.$$
5. Rewrite S'/S by factoring out S:
$$\frac{S'}{S} = \frac{\alpha^\varepsilon A_1^{\varepsilon-1}}{S} + \frac{(1-\alpha)^\varepsilon A_2^{\varepsilon-1}}{S}\,\beta^{\varepsilon-1}.$$
6. Identify the two fractions as optimal shares:
    1. From Steps 9–10 in the proof of Proposition 2 and $(y^*)^{\varepsilon-1} = S$, we have
    $$t_1^* = \frac{\alpha^\varepsilon A_1^{\varepsilon-1}}{S}, \qquad t_2^* = \frac{(1-\alpha)^\varepsilon A_2^{\varepsilon-1}}{S}.$$
    2. Therefore, $\frac{S'}{S} = t_1^* + t_2^*\,\beta^{\varepsilon-1}$.
7. Substitute Step 6 into Step 4 and use $t_2^* = 1 - t_1^*$:
$$\frac{y^{*'}}{y^*} = \left(t_1^* + (1 - t_1^*)\beta^{\varepsilon-1}\right)^{\frac{1}{\varepsilon-1}}.$$
8. □

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

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

Proof (structured).

Given: Proposition 4 gives, for each $\varepsilon \neq 1$,

$$\frac{y^{*\prime}}{y^*} = \left(t_1^*(\varepsilon) + t_2^*(\varepsilon)\,\beta^{\varepsilon-1}\right)^{\frac{1}{\varepsilon-1}}, \qquad t_2^*(\varepsilon) = 1 - t_1^*(\varepsilon).$$

To show: The three canonical limits stated in the proposition.

1. **Cobb–Douglas limit ($\varepsilon \to 1$).**
   1. Let $\rho \equiv \frac{\varepsilon-1}{\varepsilon}$. Then $\varepsilon \to 1$ implies $\rho \to 0$.
   2. For any $x > 0$, $x^\rho = e^{\rho \ln x}$, so as $\rho \to 0$ we have the first-order expansion $x^\rho = 1 + \rho \ln x + o(\rho)$.
   3. Apply this to $g(t_1, t_2) = \alpha(A_1 t_1)^\rho + (1-\alpha)(A_2 t_2)^\rho$:
   $$g(t_1, t_2) = \alpha(1 + \rho \ln(A_1 t_1)) + (1-\alpha)(1 + \rho \ln(A_2 t_2)) + o(\rho).$$
   4. Simplify:
   $$g(t_1, t_2) = 1 + \rho(\alpha \ln(A_1 t_1) + (1-\alpha)\ln(A_2 t_2)) + o(\rho).$$
   5. Since $y = g^{1/\rho}$, take logs and use $\ln(1+u) = u + o(u)$ as $u \to 0$:
   $$\ln y = \frac{1}{\rho} \ln g \to \alpha \ln(A_1 t_1) + (1-\alpha)\ln(A_2 t_2).$$
   6. Exponentiate to get the Cobb–Douglas limit aggregator:
   $$y \to (A_1 t_1)^\alpha (A_2 t_2)^{1-\alpha}.$$
   7. In that Cobb–Douglas problem, maximize $(A_1 t_1)^\alpha (A_2(1-t_1))^{1-\alpha}$ over $t_1 \in (0,1)$. Taking logs, the objective is $\alpha \ln t_1 + (1-\alpha)\ln(1-t_1)$ plus constants, so the FOC is
   $$\frac{\alpha}{t_1} - \frac{1-\alpha}{1-t_1} = 0,$$
   which solves to $t_1^* = \alpha$, $t_2^* = 1-\alpha$ (i.e. time shares are unchanged by A_1, A_2).
   8. Under $A_2' = \beta A_2$, the Cobb–Douglas optimum output scales by $\beta^{1-\alpha}$, hence $\frac{y^{*\prime}}{y^*} \to \beta^{1-\alpha}$.
2. **Perfect-complements limit ($\varepsilon \to 0$).**
   1. In Proposition 4, take $\varepsilon \to 0$. Then $\varepsilon - 1 \to -1$ and $\beta^{\varepsilon-1} \to \beta^{-1}$.
   2. Therefore, provided $t_i^*(\varepsilon)$ has a well-defined limit as $\varepsilon \to 0$ (which it does in this two-task CES), we obtain
   $$\frac{y^{*\prime}}{y^*} = \left(t_1^*(\varepsilon) + t_2^*(\varepsilon)\,\beta^{\varepsilon-1}\right)^{\frac{1}{\varepsilon-1}} \to \left(t_1^* + t_2^*\beta^{-1}\right)^{-1} = \frac{1}{t_1^* + t_2^*/\beta}.$$
3. **Perfect-substitutes limit ($\varepsilon \to \infty$).**
   1. The CES aggregator converges to a max aggregator as $\varepsilon \to \infty$ (equivalently $\rho \to 1$):
   $$y(t_1, t_2) \to \max\{A_1 t_1,\ A_2 t_2\}.$$
   2. In the max problem, the optimum allocates all time to the task with the higher A_i (up to ties). Therefore after the shock, if $\beta A_2 > A_1$ then the optimal allocation satisfies $t_2^* \to 1$.
   3. In that case, the optimized output scales by β, so $\frac{y^{*\prime}}{y^*} \to \beta$.
4. This proves the three limit statements. □

Related Theory

I don't consider myself an expert on these literatures, take this survey at your own risk.

The index number problem. There's an old literature on calculating an aggregate price index in a way that accounts for substitutability between different goods. The same theory can be applied to change in goods-prices or task-productivities, see Caves, Christensen, and Diewert (1982).

- The Laspeyres index uses base-period weights, and will understate gains when a shock makes you reallocate toward the lower-price good.
- The Paasche index uses end-period weights, and will overstate gains when a shock makes you reallocate toward the lower-price good.
- A Divisia index is a path integral of share-weighted growth rates. Discrete indices (e.g., Fisher/Törnqvist) are designed to approximate that integral.

Consumer surplus / welfare for large price changes. The classic consumer surplus measure from a price change is the area under the demand curve, however this measures the Marshallian consumer surplus, which will approximate the welfare-relevant Hicksian surplus only when there are negligible income effects. Willig (1976) gives approximation bounds.

- EV = Equivalent Variation, the change in income that would have the same welfare effect as the price change.
- CV = Compensating Variation, the change in income which could *accompany* the price change and restore your utility.
- P = Cost of the new bundle at old prices
- L = Cost of the old bundle at new prices

For a price increase we can show that

$$P \leq EV \leq CV \leq L.$$

If utility is homothetic then you can summarize the entire price vector with a single price index, and the ratio of EV to CV is the ratio of price indices.

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 standard integrable demand system (AIDS) that flexibly captures income and substitution patterns.

Economics of time allocation (time is a scarce input with shadow prices) DeSerpa (1971) is a classic reference on time allocation, and time-saving innovations as relaxing the budget constraint.

Task substitution and computerization as task-specific technology shocks Autor, Levy, and Murnane (2003) gives 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). You cannot summarize tech change as "labor-augmenting" in the aggregate.

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.

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.

Baqaee and Farhi (2019) shows that in production networks, micro shocks can have macro consequences and nonlinearities/higher-order terms matter.

Baqaee and Burstein (2021) and Comin, Lashkari, and Mestieri (2021) take into account income effects.

Amdahl's law as the perfect-complements benchmark Amdahl's law in computer science says the speedup from improving one component is bounded by the unimproved fraction. This corresponds to the perfect-complements case.

Illustrations

Indifference Curve

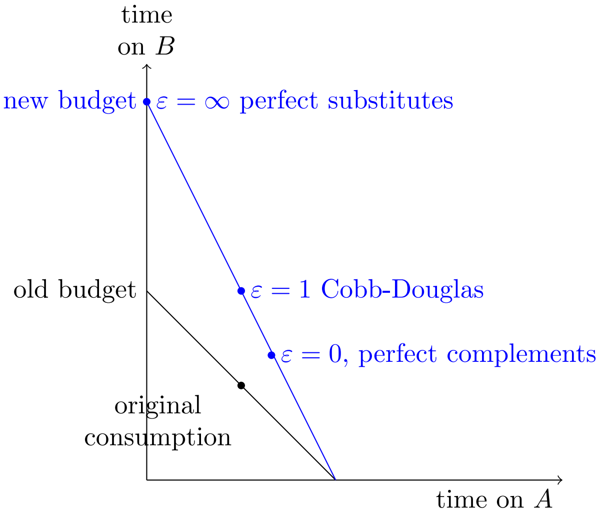
time
on B
new budget
$\varepsilon = \infty$ perfect substitutes
old budget
$\varepsilon = 1$ Cobb-Douglas
$\varepsilon = 0$, perfect complements
original
consumption
time on A

Figure 1: Budget constraint and optimal allocations under different elasticities

Demand Curve

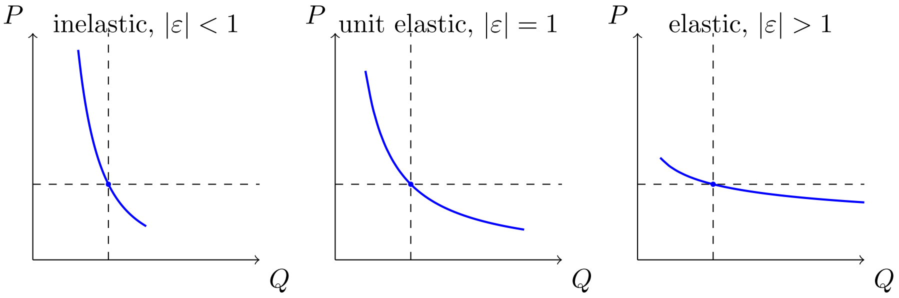

Figure 2: Demand curves with different price elasticities

2026-01-17 | AI & substitution

Puzzle:

1. **Task speedups bigger than aggregate speedups.** LLMs seem to be having large speedups for tasks that make up a large share of peoples' time, yet this doesn't seem to reflect actual productivity.
2. **Resolution:** LLMs are mostly making new tasks affordable, rather than lowering the cost of existing tasks. It's like making a Cadillac the same cost as a Toyota, instead of lowering the cost of all cars uniformly.
3. **How to test this.** We want to know the effects of LLMs on both (A) task speed; (B) task selection.
   - **Ask people which tasks they would select with & without AI.**
   - **Add a time cost to tasks.** See how task selection will change if you add artificial time costs, e.g. you have to do CAPTCHA for 1 minute.
   - **Observe effect of latency.**
4. **Note: elasticity of substitution not useful.**

Model: Two Quality Levels

Suppose you can spend your time on just two tasks, t_1 and t_2, they are perfect substitutes, each has a quality (q_i) and a time-price (p_i), we normalize the first price to 1.

$$\begin{aligned} U &= t_1 + q_2 t_2 \\ 1 &= t_1 + p_2 t_2 \quad \text{(budget constraint)} \end{aligned}$$

We can embed this into a model with multiple different activities, CES across activities, perfect substitution within activities:

$$\begin{aligned} U &= \sum_{i=1}^{n} \left((t_{i,L} + t_{i,H})^{\eta}\right)^{1/\eta} \\ 1 &= \sum_{i=1} (t_{i,L} + p_{i,H} t_{i,H}) \end{aligned}$$

Model: Independent Tasks

There are n different tasks, each has payoff u_i and time-cost p_i.

$$\begin{aligned} U &= \sum_{i=1}^{n} u_i \times \mathbb{1}\{t_i \geq p_i\} \\ \sum_{i}^{n} t_i &= 1 \end{aligned}$$

If the time-cost of *existing* tasks falls by $1/\beta$ then utility increases by β. But if the time-cost of *new* tasks (that you're not already doing) falls then the utility increase is much less than $1/\beta$. This is especially true if there's a minimum time-cost to doing a task, e.g. every task takes at least 15 minutes.

claims:

1. The relative time-cost of the old tasks at the new prices will be a lower-bound on the effective time increase.
2. The relative time-cost of the new tasks at the old prices will be an upper-bound on the effective time-increase.
3. **Two quality levels.** There n different activities, and two qualities, L and H. The two quality-levels are perfect substitutes:

$$U = \sum_{i=1}^{n} \left((x_{i,L} + x_{i,H})^{\eta}\right)^{1/\eta}$$

How to test:

1. **SWE tasks.**
   - Write out all tasks, check which ones you would do with Human vs Augmented.
   - Add a time cost to each task – you have to fill out CAPTCHA for 5 minutes.
2. **Chatbot.**
   - Add a latency cost. Note that this utility function does not have constant returns to scale, so needs some adjustment to apply to work tasks.

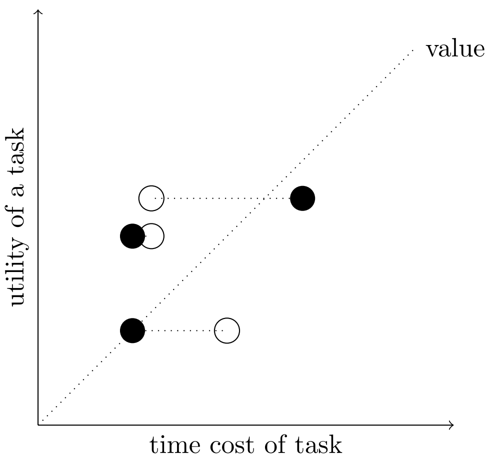

Figure 3: Quality and Time-cost of Different Tasks

2026-01-19 | The Cadillac Theory of LLMs

The Cadillac Theory of LLMs

TLDR: Estimates of LLM productivity which are based on time-savings often overestimate the true impact because using an LLM fundamentally changes the type of task you do. LLMs make it cheaper to drive a Cadillac, but they don't change the price of a Ford.

explanation

People are using LLMs to do work that would otherwise take them many hours. But there's reason to believe that LLMs are mostly speeding people up in tasks that they wouldn't otherwise be doing, and for that reason the aggregate efficiency improvement is much smaller than implied by the naive calculation.

An analogy: suppose a Cadillac is 10% better than regular car but costs 10x as much. Now you move to a country where everything is identical except that Cadillacs are the same price as a regular car, and so everyone drives a Cadillac. A simple calculation would say that living standards are much higher: the price of cars has fallen by 10X, and expenditure on cars is the same (meaning elasticity of substitution is 1), so welfare must be substantially higher. Though in fact.

In the same way, we see people sending queries to LLMs that would've taken them hours to do themselves, but a large share are Cadillac queries, things they would never have considered doing themselves:

- "identify this species of finch"
- "summarize this long PDF"
- "write me a literature review on this topic I'm marginally interested in"

The critical modelling difference is how you divide up tasks, whether *coarse* or *fine*. If coarse then you fail to pick up differences in quality. If you use fine tasks then the assumption of constant-elasticity-of-substitution and Hulten's thoerem are not applicable.

relation to theory

- hulten's theorem doesn't apply if all cost reductions are on tasks you don't currently do –>
- if you use coarse tasks then miss the quality effects ; if you use fine tasks then CES and hulten aren't applicable

applications

- Anthropic (2025) look at Claude queries and estimate (1) the typical speedup is 5X; (2) they account for 10% of tasks. Assuming elasticity of substitution is 1 (Hulten's thm) they get an aggregate 20% speedup (=5^0.1). However I'm skeptical that the tasks they see being performed by Claude are a representative set of tasks that people would do themselves.

related literature

Anthropic/Saffron (2025) "How AI is transforming work at Anthropic"

> "27% of Claude-assisted work consists of tasks that wouldn't have been done otherwise,

> "Claude fixes a lot of "papercuts". 8.6% of Claude Code tasks involve fixing minor issues that improve quality of life, like refactoring code for maintainability (that is, "fixing papercuts") that people say would typically be deprioritized. These small fixes could add up to larger productivity and efficiency gains.

Microsoft

> "Different productivity measurement strategies will likely also be needed if people are bypassing existing tasks and doing new ones instead, rather than simply doing existing tasks faster or better.

Anthropic (2026):

> "We measure whether users could have completed tasks without Claude, and the years of education needed to understand both user prompts and Claude's responses.

theory

- TLDR: the Cadillac theory doesn't naturally fit with a CES model of time allocation, it fits better with discrete tasks.
- The most common way of estimating AI-based productivity improvements is to combine (1) the size of speedup (β); (2) the share of tasks sped up (s); (3) the elasticity of substitution between exposed and non-exposed tasks (ε), assuming that the elasticity is constant (CES).
- This setup implies that the productivity effects of LLMs must be large, independent of the elasticity of substitution. If people are using LLMs for a significant share of tasks ($s \gg 0$), and the speed-up is significant ($\beta \gg 1$), then the aggregate productivity increase will be large for *any* value of ε.

notes

- Speeding you up on tasks outside your domain. Doing a literature review.
- Examples.

misc

Chen, Jeon, and Kim (2014)

They find that searching for answer to questions online is 4X faster than searching offline in a library:

> "Using a random sample of queries from a major search engine, we conduct an experiment to compare online and offline search experiences and outcomes. We find that participants are significantly more likely to find an answer on the Web. Restricting to the set of queries which participants find answers in both treatments, the average search time is 22 minutes offline, and 7 minute online.

Note they are using a distribution of queries from web search, not from library search.

1. **Tom doing literature reviews.** Suppose I spend 5 minutes each day using ChatGPT to do literature reviews, & the reviews would've taken me 8 hours to do without ChatGPT's help, meaning this is a roughly 100X speedup. On the surface this seems to be doubling my aggregate productivity, i.e. I'm now doing 16 hours work in an 8 hour day.

   To get the true time-saving we need to know how substitutable these tasks (literature reviews) are for other tasks. Suppose for argument that ChatGPT hasn't changed the share of my time I'm spending on literature reviews, it's just made me far more productive. This implies an elasticity of substitution of 1. It also implies that

Cadillac theory

tweets

Cadillac tasks: I believe many estimates of LLM productivity boosts are over-estimates because people are using them for cadillac tasks: things that would take you a long time unaided, but have only marginal additional value.

- "identify this species of finch"
- "write me a literature review on this topic I'm marginally interested in"
- "write tests for this whole codebase"
- "proof read and spell check this long document"

LLms let me do 80 hours an 8-hour workday, but this doesn't represent a true 10X of my productivity.

Two basic ways of estimating productivity impacts: (1) estimate time-savings at a task level; (2) self-reports of time-savings.

If we estimate task-level time savings it's important whether you're selecting tasks from the pre-LLM world or post-LLM world: they will give lower and upper bounds on the true productivity improvement. (additionally, if you just choose tasks from chatbot logs then you're disproportinately getting things where the difference is large).

Alternatively we can ask people "how much time are you saving?" It's a hard question to answer partly because it has three different interpretations with very different answers: new time with old tasks, old time with new tasks, or new time with new tasks (technically Laspeyres, Paasche, and Compensating Variation). Ideally we want to say "how much more overall work are you getting done in a workday?"

Cadillac tasks: I think many estimates of LLM productivity boosts from task-level time-savings typically overestimate the true impact because LLMs cause substitution towards "cadillac" tasks: things that would take a long time unaided, but have only marginal additional value.

A common way of estimating LLM productivity effects is to multiply (1) the share of tasks LLMs are used for; (2) the task-level speedup. This typically implies a large aggregate impact. You can also adjust for elasticity of substitution but this doesn't make a huge difference (Hulten's theorem says it's 2nd-order).

However this can be misleading if using coarse task categories, e.g. "programming", "research". This is a problem because LLMs can change the specific mix within these categories, and people will substitute towards tasks that are relatively cheaper with LLMs.

An analogy: suppose a Cadillac is 10% better than a regular car, but costs 10X as much. You move to a country where everything is the same except here Cadillacs are the same price as Chevys. Superficially it looks like your standard of living has rocketed: when evaluated at the old prices your car expenditure has gone up by 10X. But in reality your welfare is only marginally higher. The welfare calculation can be fixed by dividing expenditure shares more finely: i.e. looking at specific products, rather than just cars in general.

In the same way, we see people sending queries to LLMs that would've taken them hours to do themselves, but a large share are arguably Cadillac queries, things they would not have thought worth doing without the LLM:

- "identify this species of finch"
- "proof read and spell check this long document"
- "write me a literature review on this topic I'm marginally interested in"

This will be a problem when LLMs disproportionately lower the time-cost of tasks that you don't otherwise do. This will be a problem if your pool of tasks comes from logs of a chatbot, because they are *selected* on tasks that people expect LLMs to help with. Additionally a general theory of LLMs is that they share knowledge, and so they'll be used far more for *new* tasks, than for existing tasks.

2026-01-23 | learning from randomization

1. If AI causes people to change their selection of tasks then we can only get lower and upper bounds on the time-saving, even if we observe the time required for every task, both with AI and without AI.
2. However if we observe choices when you randomize AI-allowed vs AI-disallowed, that allows you to better approximate the true value.
3. A worked example:
   - Suppose task A takes 1 hour, and task B takes 2 hours, and you choose to do task A.
   - Suppose LLMs let you do do task B in 1 hour, and now you switch to doing task B.
   - How much time have you saved?
     - Upper bound 1 hour (if B is twice as valuable as A)
     - Lower bound zero hours (if B has same value as A)
   - Now we say you're randomized, so you first have to choose task A or B, and you only later find out whether B will take 1 or 2 hours.
   - You will choose task B if-and-only-if the expected benefit is greater than the expected cost (0.5) hours. You've therefore tightened your bounds on the efficiency benefit.
4. Intuition: suppose people in the uplift don't choose any AI-heavy tasks, compared to people outside the uplift study. Then the value of AI must be less than half the nominal time-savings from switching to AI-heavy tasks.

2026-01-25 | substitution

OK can you integrate these motivating questions & answers at the top. Don't need to use my exact wording.

1. Suppose AI makes you more productive by a factor A on tasks which are a share s of your time, e.g. 50% more productive in 10% of your time. How much more productive will you be overall?
2. How does this depend on:
   - whether you reallocate your time to maximize output
   - the substitutability between sped-up tasks and other tasks
   - if we observe the *post-AI* task-shares instead of pre-AI task shares
3. Suppose you observe (1) someone's pre-LLM and post-LLM task choice across a set of tasks (suppose you either do a task or don't, unit demand); (2) the per-task time-required for all tasks, both pre-AI and post-AI.
   - We want to show that the productivity increase will be between Laspeyres and Paasche.

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.

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. https://doi.org/10.3386/w8337.

Baqaee, David Rezza, and Ariel Burstein. 2021. "Welfare and Output with Income Effects and Demand Instability." Working Paper. https://www.semanticscholar.org/search?q=Welfare%20and%20Output%20with%20Income%20Effects%20and%20Demand%20Instability.

Baqaee, David Rezza, and Emmanuel Farhi. 2019. “The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten’s Theorem.” *Econometrica* 87 (4): 1155–1206. https://doi.org/10.3982/ecta15202.

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/pdf/2507.09089.pdf.

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.

Chen, Yan, Grace YoungJoo Jeon, and Yong-Mi Kim. 2014. “A Day Without a Search Engine: An Experimental Study of Online and Offline Searches.” *Experimental Economics* 17 (4): 512–36. https://doi.org/10.1007/s10683-013-9381-9.

Comin, Diego, Danial Lashkari, and Martín Mestieri. 2021. “Structural Change with Long-Run Income and Price Effects.” Working Paper. https://doi.org/10.3982/ecta16317.

Deaton, Angus, and John Muellbauer. 1980. “An Almost Ideal Demand System.” *American Economic Review* 70 (3): 312–26. https://www.semanticscholar.org/search?q=An%20Almost%20Ideal%20Demand%20System.

DeSerpa, Allan C. 1971. “A Theory of the Economics of Time.” *The Economic Journal* 81 (324): 828–46. https://doi.org/10.2307/2230320.

Hausman, Jerry A. 1981. “Exact Consumer’s Surplus and Deadweight Loss.” *American Economic Review* 71 (4): 662–76. https://www.semanticscholar.org/search?q=Exact%20Consumer%27s%20Surplus%20and%20Deadweight%20Loss.

Hulten, Charles R. 1978. “Growth Accounting with Intermediate Inputs.” *The Review of Economic Studies* 45 (3): 511–18. https://doi.org/10.2307/2297252.

Willig, Robert D. 1976. “Consumer’s Surplus Without Apology.” *American Economic Review* 66 (4): 589–97. https://www.semanticscholar.org/paper/745fa39279d59c6f6b14dce4a38bcf098774c2ad.