The Training Bottleneck and Model Collapse

Jon Smirl

2026-03-01

The String Quartet Problem

In 1967, William Baumol made a simple but profound observation about the performing arts. A string quartet that took 40 minutes to perform in 1800 still takes 40 minutes in 2025. You cannot speed it up by adding more musicians, better concert halls, or faster violins. Meanwhile, manufacturing productivity has increased a hundredfold. The result is that live performance becomes relatively more expensive every year — not because musicians got worse, but because everything else got cheaper (Baumol1967).

This is Baumol's cost disease: when a "stagnant sector" is yoked to a "progressive sector," the stagnant sector's relative cost rises without bound. Baumol used it to explain why healthcare and education keep getting more expensive. But the same logic applies, with striking precision, to the economics of artificial intelligence.

Training Is the Stagnant Sector

Modern AI has two distinct activities: **training** and **inference**. Training is the process of building a model — feeding it data, adjusting billions of parameters, teaching it to recognize patterns. Inference is the process of using that trained model to answer questions, generate text, or classify images.

These two activities scale very differently. Inference is the "progressive sector." Once a model is trained, you can run copies of it on thousands of machines simultaneously. Better hardware makes each query cheaper. Distributing inference across a mesh of devices is straightforward engineering. The cost per query falls reliably with scale.

Training is the string quartet. To push the frontier of AI capability — to make models genuinely smarter, not just faster — you need all of the following simultaneously: massive datasets, enormous compute clusters, and algorithmic innovations. These inputs are **strongly complementary**. Having twice as much data does not help if you lack the compute to process it. Having a brilliant new algorithm does not help if you lack the data to train it on.

In the CES framework, this complementarity is captured by a low value of the substitution parameter ρ. Training has near-Leontief complementarity: the inputs must be used together in roughly fixed proportions. The CES aggregate for training output looks something like:

$$F_{\text{train}} = \left(\frac{1}{3}\left[x_{\text{data}}^{\rho} + x_{\text{compute}}^{\rho} + x_{\text{algorithm}}^{\rho}\right]\right)^{1/\rho}, \quad \rho \ll 0$$

When ρ is strongly negative, the aggregate is dominated by the weakest input. No amount of compute compensates for missing data. No dataset compensates for inadequate algorithms. Training capability grows only as fast as the slowest-improving input.

Inference, by contrast, has much higher ρ — closer to zero or even positive. You can trade off compute for latency. You can use different hardware architectures. You can quantize models to run on cheaper chips. The inputs are substitutable.

This **training-inference bifurcation** *training-inference-bifurcation* explains why training stays centralized while inference distributes. Centralized facilities are necessary for training precisely because the inputs are so complementary — you need everything in one place. But once the model exists, inference can scatter to wherever hardware is cheapest.

The Hierarchical Ceiling

The training bottleneck is not merely an inconvenience. It is a fundamental constraint on how fast AI capability can grow. In the hierarchical framework described in *economy-has-layers*, each level of the economy sets a ceiling on the levels above it. The slowest level determines the pace.

For AI, the hierarchy looks like this: semiconductor learning curves improve hardware over decades. Training exploits that hardware over months to years. Inference uses trained models over days. Applications built on inference change over hours.

The *baumol_limit* theorem makes the constraint precise: the capability growth rate of the mesh (distributed AI) is bounded above by the frontier training rate g_Z, no matter how efficiently inference scales. You can distribute inference across a million devices, each running at negligible cost, and the models those devices run will still only be as capable as centralized training has made them.

This ceiling is not a prediction about what will happen — it is a structural feature of complementary production. As long as training requires simultaneous access to data, compute, and algorithms in fixed proportions, no amount of progress in inference changes the rate at which models get smarter.

Model Collapse: When AI Eats Its Own Tail

There is a second, subtler threat to AI capability, and it comes not from a bottleneck but from a feedback loop.

As AI-generated text, images, and code proliferate across the internet, an increasing fraction of the data used to train new models is itself machine-generated. (Shumailov2023) demonstrated what happens next: models trained on the output of previous models progressively lose the diversity and richness of the original human-generated data. They called this phenomenon **model collapse**.

The CES framework provides a precise accounting of why this matters. The diversity premium — the bonus that heterogeneous inputs provide over homogeneous ones — is proportional to curvature K times the squared coefficient of variation:

$$\text{Superadditivity bonus} \propto K \times \mathrm{CV}^2$$

where $K = (1 - \rho)(J - 1)/J$ is the curvature of the CES aggregate. When inputs are diverse (high CV), the bonus is large. When inputs are homogeneous (low CV), the bonus vanishes. *diversity-premium*

Model collapse is the process by which CV shrinks. The mechanism involves effective diversity M — the number of independent information sources contributing to training data. With M truly independent sources, each generation of self-referential training reduces variance at the rate:

$$\text{Variance retained} = \left(1 - \frac{1}{M}\right)^n$$

where n is the number of generations. At $M = 10$ independent sources, each generation retains 90 percent of the original variance — a tolerable 10 percent loss. At $M = 3$, each generation retains only 67 percent. After five generations, barely 13 percent of the original variance survives. The diversity that made the model valuable has been consumed.

This connects directly to the correlation robustness result from CES theory. One of the roles of CES curvature K is extracting value from idiosyncratic variation — the uncorrelated, independent signals that each input contributes. The bonus from this extraction is proportional to K^2. But self-referential training systematically destroys idiosyncratic variation. Models trained on model output produce correlated outputs. The independent signals converge toward a shared mean.

This is the same mechanism that causes diversification to fail in financial crises. In normal times, assets have idiosyncratic returns, and a portfolio benefits from combining them. In a crisis, correlations spike toward one, idiosyncratic variation disappears, and the diversification benefit evaporates. Model collapse is the data equivalent: training sources that were once independent become correlated, and the K^2 bonus vanishes.

The Grossman-Stiglitz Connection

There is an elegant parallel with a classic result in financial economics. (GrossmanStiglitz1980) proved that perfectly efficient markets are impossible because if all information were already in prices, no one would have an incentive to gather information. Markets need a diversity of informed participants to function.

AI faces the same paradox. If all training data comes from AI models that have already processed the available information, there are no independent signals left to learn from. The system needs fresh human-generated data — the equivalent of informed traders — to maintain the diversity that keeps models useful. The effective diversity M is not the number of data sources but the number of *independent* information signals.

Two Constraints, One Lesson

The training bottleneck and model collapse are different problems, but they point to the same structural reality. AI capability is not limited by inference speed, hardware availability, or deployment scale. It is limited by two deeper constraints: the complementarity of training inputs (which bounds capability growth to the slowest-improving factor) and the diversity of training data (which erodes when the system feeds on itself).

Both constraints are instances of the CES framework in action. The bottleneck arises from low ρ in the training production function — strong complementarity means the weakest link governs. Model collapse arises from shrinking CV in the training data — reduced diversity means the superadditivity bonus erodes.

For anyone building on AI, the implication is practical: scaling inference is necessary but not sufficient. The binding constraints are upstream — in the slow, expensive, irreducibly complementary process of frontier training, and in the diminishing supply of genuinely independent data to train on. [@Baumol1967; @Shumailov2023]

References