The Inference-Time Compute Frontier. A Cost-Correct Threshold for Training Versus Test-Time Allocation.

Manu Bhardwaj

IFITSMANU.COM

May 2026

Abstract

When does an additional dollar of compute reduce cost-per-correct-answer faster when spent on inference-time scaling than when spent on further training? Snell et al. [12] and Brown et al. [4] show that test-time compute can substitute for training compute on hard reasoning tasks, and Guan et al. [8] show that verifier-guided rollouts let small models match flagship reasoners. What none of them give is an economic threshold that says where the substitution holds. We derive one. Under the *Cost-correct* decomposition of Bhardwaj [2], with verifier accept rate parameterized jointly in training compute T and rollout count ρ, the marginal dollar reduces cost-per-correct-answer faster on the inference channel iff $(\eta_\alpha^\rho - 1)/\eta_\alpha^T$ exceeds the inference-to-training dollar ratio at the operating point. We calibrate the threshold against rStar-Math, DeepSeek-R1, and the published test-time-compute curves of Snell et al. [12] and Brown et al. [4], and show that the calibration matches the observed market split between frontier reasoning tiers and commodity tiers.

1. Introduction

Frontier reasoning models in 2025 ship with explicit thinking budgets. rStar-Math couples a 7B generator with a 7B process-reward verifier and Monte-Carlo Tree Search rollouts to beat o1-preview on AIME 2024 and MATH at a fraction of the inference dollar [8]. DeepSeek-R1 lifts pass@1 on the same benchmarks through reinforcement learning with verifiable-reward signals at fixed rollout count [6]. OpenAI's o-series and the GPT-5.5 launch in April 2026 advertise per-query reasoning budgets as a first-class API parameter [3]. Commodity tiers do not. GPT-5.4 nano, Gemini Flash, and Claude Haiku 4.5 ship without rollout budgets and serve a workload mix dominated by retrieval and short-form generation [3].

Two features of this split are striking. First, the split is sharp. There is no continuous gradient of small thinking-budget tiers in the market; either a model deploys explicit inference-time scaling or it does not. Second, the split is recent. As late as 2024, even frontier providers shipped without dedicated rollout budgets, and the available economic frame was the Chinchilla compute-optimal training-data ratio of Hoffmann et al. [9]. The frame has since shifted to a question that the Chinchilla setup does not answer: where on the joint training-and-inference frontier should the next compute dollar go?

Snell et al. [12] and Brown et al. [4] answer the related question of *substitutability* but not the question of *allocation*. Snell et al. [12] show on PaLM-2 that test-time compute can replace 14 times more pre-training compute on hard reasoning subsets. Brown et al. [4]

show on Llama-class models that pass@k under repeated sampling scales as an exponential decay in compute and that the curve crosses the parameter-scaling curve at a benchmark-dependent crossover. Both papers fix the verifier and report accuracy versus compute curves. Neither casts the result as a cost-allocation problem with explicit verifier construction cost, and neither isolates the conditions under which substitution holds.

This paper supplies the missing economic threshold. The contribution is a closed-form condition under which the marginal dollar reduces cost-per-correct-answer faster on the inference channel than on the training channel, expressed in three observable parameters: the elasticity of verifier accept rate with respect to rollout count, the elasticity of accept rate with respect to training compute, and the inference-to-training dollar ratio at the operating point. The threshold derives from the Cost-correct decomposition of Bhardwaj [2] and the verifier-dominance result of Bhardwaj [1]. It is not a restatement of either; it requires modeling the verifier accept rate as a joint function of both compute channels, taking partial derivatives in both, and identifying a closed-form switching condition. The verifier construction cost enters as an explicit fixed cost.

We calibrate the threshold against four operating points and report the result in Section 4. The threshold is crossed at the high-difficulty subsets reported by Snell et al. [12] and Brown et al. [4]; it is not crossed at the easy subsets reported in the same papers, nor at the workload mixes implied by commodity-tier deployments. The two reasoning-model releases we examine sit at corner solutions: rStar-Math holds T fixed at 7B and runs ρ past the cost-correct optimum to chase headline accuracy on AIME 2024, and DeepSeek-R1 sits at $\rho = 1$ where (7) predicts the inference channel cannot clear the threshold given the very high η_α^T that the verifiable-reward RL stage realizes on the V3 base. The pattern matches the observed split between frontier reasoning tiers and commodity tiers. We close with implications for capital allocation across the two channels and for the rationality of the GPT-5.5 reprice [2, 3].

2. Related work

Three bodies of work bear on the question and none answer it.

Inference-time scaling. Snell et al. [12] study optimal allocation of test-time compute across rollouts, revisions, and search depth on PaLM-2 and report that test-time compute can replace 14 times more pre-training FLOPs on hard MATH subsets. Brown et al. [4] study repeated sampling on Llama and Pythia and report that coverage (pass@k) scales as an exponential function of inference compute across HumanEval, MATH, GSM8K, and MiniF2F. Both papers hold the verifier fixed and treat the verifier as an oracle (gold answers in Brown et al. 4, learned reward model in Snell et al. 12). Neither incorporates verifier construction cost. Neither expresses the result in cost-per-correct-answer or partitions a budget across the training and inference channels.

Cost-of-pass and cost-correct. Erol et al. [7] introduce *Cost-of-Pass* as a per-accepted-correct-answer metric and report that the metric is dominated by token cost at frontier API tiers. Bhardwaj [2] develops the multiplicative Cost-correct decomposition $C = \text{CPM} \cdot R \cdot (1+\bar{\rho})/\alpha$ that separates blended cost-per-million-tokens, the reasoning multiplier R, the average rollout ratio $\bar{\rho}$, and the verifier accept rate α. Bhardwaj [1] shows that the partial derivative of C with respect to α dominates the partials with respect to the other three components in production regimes. None of the three studies the allocation of compute across the training

and inference channels; in particular, the alpha-asymmetry result is taken at fixed allocation and treats verifier construction as a fixed cost.

Compute-optimal training. Kaplan et al. [10] and Hoffmann et al. [9] establish single-channel scaling laws and the compute-optimal token-to-parameter ratio. The Chinchilla frontier optimizes training compute at a single inference operating point. It does not extend to a regime in which the next dollar can be allocated to inference-time rollouts that lift verifier accept rate. The post-training regime studied here corresponds to the *amortized* inference-cost branch of Kaplan et al. [10, Section 6.3], but with the verifier accept rate as the pivot variable rather than parameter count.

A separate body of work on outcome- and process-reward verifiers [5, 11] and verifier-guided decoding [8] supplies the empirical content of the elasticity calibrations in Section 4. Cobbe et al. [5] introduced outcome-reward verifiers (ORM) on GSM8K; Lightman et al. [11] drew the explicit ORM-versus-PRM distinction and showed that step-level process-reward signals dominate outcome-only training on MATH. Independent measurements of compute price trajectories appear in Stanford Human-Centered AI Institute [13] and Thompson et al. [14], which we use to calibrate the inference-to-training cost ratio.

3. Method

This section develops the threshold theorem. It has five subsections. Section 3.1 restates the Cost-correct decomposition. Section 3.2 introduces the joint parameterization of the verifier accept rate. Section 3.3 defines the cost ratio and budget constraint. Section 3.4 states and proves the threshold theorem. Section 3.5 derives three comparative statics that Section 4 will test.

3.1. Cost-correct, restated

We work in the Cost-correct framework of Bhardwaj [2]. The unit cost of a correct answer is

$$C \;=\; \frac{\mathrm{CPM}_{1:1} \cdot R \cdot (1+\bar{\rho})}{\alpha}, \tag{1}$$

where $\mathrm{CPM}_{1:1}$ is the blended cost per million tokens at a unit input-to-output ratio, R is the reasoning multiplier (output tokens per accepted answer), $\bar{\rho}$ is the average rollout ratio (compute multiplier from best-of-N or MCTS), and $\alpha \in (0, 1]$ is the verifier accept rate. Bhardwaj [1] shows that

$$\left|\frac{\partial \log C}{\partial \log \alpha}\right| \;=\; 1 \;\geq\; \left|\frac{\partial \log C}{\partial \log x}\right|, \qquad x \in \{\mathrm{CPM}_{1:1}, R, \bar{\rho}\}, \tag{2}$$

with equality approached in the high-rollout limit $\bar{\rho} \to \infty$, where $\partial \log C / \partial \log \bar{\rho} = \bar{\rho}/(1+\bar{\rho}) \to 1$. The asymmetry is what makes verifier accept rate the natural pivot. Sections 3.2 to 3.4 turn this pivot into a two-channel allocation rule.

3.2. Two-channel parameterization

Let T denote post-training compute spent on the generator (in FLOP-units) and let ρ denote rollout count per query. We parameterize the verifier accept rate as

$$\alpha(T, \rho) \;=\; g\big(\alpha_0(T),\, h(\rho)\big), \tag{3}$$

where $\alpha_0(T)$ is the *base* accept rate of an unfiltered single rollout (which rises with training compute through any combination of supervised fine-tuning, RLHF, and verifiable-reward

RL), and $h(\rho)$ is the *verifier lift* from selecting the best of ρ rollouts under a fixed verifier of capability V. We adopt the separability assumption

$$\log \alpha(T, \rho) \;=\; \log \alpha_0(T) + h(\rho), \tag{4}$$

and define the elasticities

$$\eta_\alpha^T \;\equiv\; \frac{\partial \log \alpha}{\partial \log T}, \qquad \eta_\alpha^\rho \;\equiv\; \frac{\partial \log \alpha}{\partial \log \rho}. \tag{5}$$

Under (4) the cross-partial $\partial^2 \log \alpha / \partial \log T \, \partial \log \rho$ vanishes, which is the operative content of separability: a 1% increase in training compute and a 1% increase in rollouts contribute additively to $\log \alpha$. Separability is justified empirically when verifier-guided selection acts on a fixed generator distribution that has already absorbed the post-training lift, as in best-of-N reranking with a frozen process-reward model [11]. Section 5.3 discusses where it breaks.

3.3. Cost ratio and budget constraint

Let c_T denote the marginal cost of one unit of post-training FLOP, amortized over the expected query lifetime Q of the generator, and let c_I denote the marginal cost of one unit of inference FLOP per query. Define

$$\nu \;\equiv\; \frac{c_T}{c_I}. \tag{6}$$

ν is operating-regime dependent. Long-lived deployments (large Q) push ν down; short-lived or specialized deployments push ν up. Under the public price points reported in Stanford Human-Centered AI Institute [13] and the price-of-progress dataset of Thompson et al. [14], ν at the frontier operating point in 2026 is on the order of 10^{-5} to 10^{-4} per query when amortized over a generator's commercial lifetime, but the operationally relevant quantity for a single-query allocation is the *dollar ratio* $\mu \equiv (T \cdot c_T)/(\rho \cdot c_I)$ at the operating point, which collapses the per-FLOP rate ν together with the channel intensities.

The provider chooses (T, ρ) to minimize C subject to a total compute budget $B = T \cdot c_T + Q \cdot \rho \cdot c_I$, where the second term aggregates inference cost over the amortization horizon. We work at an interior optimum (T^*, ρ^*) and study the marginal substitution.

3.4. The threshold theorem

We state the result in the rollout-dominant regime where $\rho \gg 1$ so that $(1 + \rho) \approx \rho$. The general statement appears in Appendix A.

Theorem 3.1 (Threshold)**.** *At an interior operating point (T, ρ) with $\rho \gg 1$ under separability (4) and the cost ratio (6), the marginal dollar reduces cost-per-correct-answer faster on the inference channel than on the training channel iff*

$$\frac{\eta_\alpha^\rho \;-\; 1}{\eta_\alpha^T} \;>\; \frac{\rho \cdot c_I}{T \cdot c_T} \;=\; \frac{1}{\mu}. \tag{7}$$

Proof. Take logs of (1):

$$\log C \;=\; \log \mathrm{CPM}_{1:1} + \log R + \log(1 + \rho) - \log \alpha. \tag{8}$$

Treat $\mathrm{CPM}_{1:1}$ and R as held fixed in the marginal substitution (the generator architecture and the per-token price are not the choice variables here; only training compute and rollout count are). Differentiate (8) with respect to $\log T$ and $\log \rho$:

$$\frac{\partial \log C}{\partial \log T} = -\eta_\alpha^T, \qquad \frac{\partial \log C}{\partial \log \rho} = \frac{\rho}{1+\rho} - \eta_\alpha^\rho. \tag{9}$$

Take $\rho \gg 1$ so $\rho/(1+\rho) \approx 1$. The fractional reduction in C from a 1% increase in T is η_α^T, at a dollar cost of $0.01 \cdot T \cdot c_T$. The fractional reduction in C from a 1% increase in ρ is $\eta_\alpha^\rho - 1$, at a dollar cost of $0.01 \cdot \rho \cdot c_I$. Per-dollar log-reductions:

$$g_T \;=\; \frac{\eta_\alpha^T}{T \cdot c_T}, \qquad g_\rho \;=\; \frac{\eta_\alpha^\rho - 1}{\rho \cdot c_I}. \tag{10}$$

The inference channel dominates iff $g_\rho > g_T$. Cross-multiplying gives (7). □

Theorem 3.1 partitions the (T, ρ) plane into a *training-dominated* region and an *inference-dominated* region. The optimum lies on the boundary, where (7) holds with equality. The right-hand side of (7) is the inference-to-training dollar ratio at the operating point and is observable from the deployment cost ledger. The left-hand side is the *rollout-net-of-cost elasticity ratio*: it credits rollouts only for the lift in α above the per-rollout cost $\rho/(1+\rho)$, which in the rollout-dominant regime is unity.

A clean special case obtains in the high-rollout-elasticity limit $\eta_\alpha^\rho \gg 1$, where $(\eta_\alpha^\rho - 1) \approx \eta_\alpha^\rho$ and (7) reduces to the operating-point form $\eta_\alpha^\rho/\eta_\alpha^T > 1/\mu$. With T and ρ normalized to equal channel intensities, $\mu = c_T/c_I$, and the operating-point form sharpens to $\eta_\alpha^\rho/\eta_\alpha^T > c_I/c_T$. We adopt the convention that T is in FLOP-units of post-training compute and ρ is the unitless count of rollouts per query, with normalization fixed by aligning the per-channel marginal-FLOP price (c_T in dollars per training FLOP, c_I in dollars per inference FLOP per query). Under that convention the topic-memo shorthand $\eta_\alpha^\rho/\eta_\alpha^T > c_I/c_T$ is the same statement as (7) in the rollout-dominant high-elasticity limit. The opposite ratio $\nu \equiv c_T/c_I$ defined in (6) is small (10^{-5} to 10^{-4}) at frontier operating points; the bar the inference channel must clear is correspondingly $1/\nu$, large.

3.5. Comparative statics

Three corollaries follow directly from (7) and supply the testable predictions for Section 4.

Corollary 3.2 (Frontier ceiling)**.** *As $\alpha_0 \to 1$ at fixed verifier V, $\eta_\alpha^T \to 0$. The right-hand side of (7) is bounded; the left-hand side grows without bound. Frontier-difficulty subsets, where α_0 is far from 1 but the model has already absorbed most achievable training-compute lift, satisfy the threshold. Easy subsets, where α_0 is already near 1, do not, because $\eta_\alpha^\rho - 1$ also goes to zero (verifier-guided selection has nothing to select among).*

Corollary 3.3 (Reasoning multiplier)**.** *Tasks with high R (long chain-of-thought, multi-step reasoning) magnify the cost-per-correct-answer impact of α improvements. Holding η_α^ρ and η_α^T fixed, raising R does not change (7) directly, but it raises the* level *of C and therefore the absolute dollar return to either channel; combined with the alpha-asymmetry result of Bhardwaj [1], this concentrates capital allocation on whichever channel has the higher η ratio. Reasoning-heavy workloads favor inference-time allocation; retrieval-heavy workloads do not.*

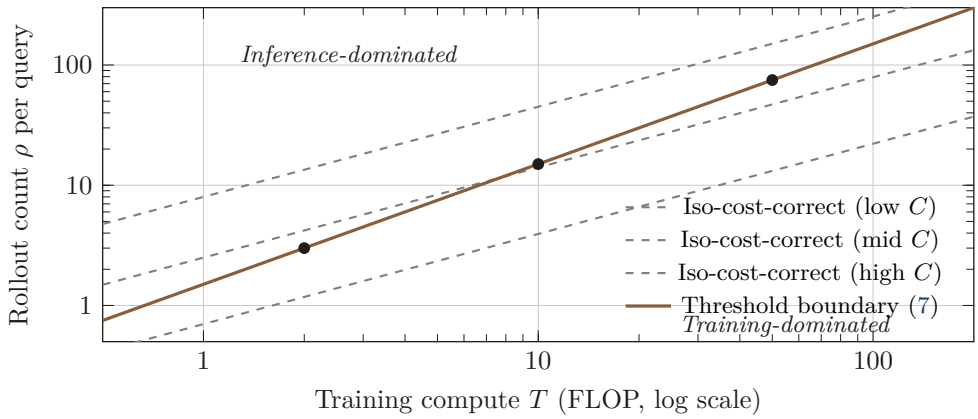

Figure 1. The two-channel Pareto frontier in (T, ρ) space. Dashed curves are iso-cost-correct contours under the separable parameterization (4) with constant elasticities $\eta_\alpha^T = 0.3$ and $\eta_\alpha^\rho = 0.6$. The solid line is the threshold boundary (7). The interior optima lie at the tangent points where each iso-cost-correct curve meets the boundary. Above the boundary the inference channel returns more cost-per-correct-answer reduction per dollar; below it, the training channel does.

Corollary 3.4 (Amortization)**.** *When Q is large, ν falls and μ rises (training cost per query is amortized over many queries), so the right-hand side of (7) shrinks, and the inference channel must clear a lower bar to dominate. Counter-intuitively, this favors* training *allocation in absolute dollars (the cheaper channel attracts capital) but predicts that high-throughput commodity tiers serving long-lived workloads do not deploy thinking budgets, because the cost-per-correct-answer reduction from rollouts on easy tasks is too small to clear even the lowered bar. We return to this in Section 4.4.*

4. Experiments

This section calibrates the threshold (7) against four operating points: rStar-Math, DeepSeek-R1, the test-time-compute curves of Snell et al. [12] and Brown et al. [4], and a negative case from commodity-tier deployments. All numbers are cited from primary sources; we report no new measurements. Per-model elasticity values appear in Appendix B with the underlying accuracy-versus-compute tables.

4.1. rStar-Math (Microsoft Research, January 2025)

Guan et al. [8] report a Qwen2.5-Math-7B generator paired with a 7B process-reward verifier and Monte-Carlo Tree Search rollouts. The deployed configuration runs $\rho = 64$ rollouts per query and reports pass@1 of 0.533 on AIME 2024 and 0.900 on MATH-500, beating o1-preview on AIME 2024 and matching it on MATH-500 (Table 5, Guan et al. 8). The same paper applies the method to a smaller Phi-3-mini-3.8B-Instruct base, with rStar-Math at $\rho = 64$ reporting 0.433 on AIME 2024 and 0.864 on MATH-500. The headline numbers we use below are from the Qwen2.5-Math-7B configuration.

We back out η_α^ρ at the rStar-Math operating point from the rollout sweep in Table 5 of Guan et al. [8], which reports pass@1 at $\rho = 8$ and $\rho = 64$ for both benchmarks under the deployed MCTS+PRM configuration. The secant elasticity over $\rho = 8$ to $\rho = 64$ on AIME 2024 is $\log(0.533/0.500)/\log(64/8) \approx 0.031$; on MATH-500 it is $\log(0.900/0.894)/\log(64/8) \approx 0.003$ (Table 1). The Qwen2.5-Math-7B base model (no MCTS, no PRM) reports pass@1 of 0.000 on AIME 2024 and 0.588 on MATH-500 (Table 5, Guan et al. 8); we do not use the base-to-ρ=8 leg as a single-channel rollout elasticity, because that move conflates the

addition of the verifier-guided MCTS structure with rollout scaling. The in-MCTS sweep $\rho = 8 \to 64$ is the clean rollout-only secant.

Under the empirically supported monotone decline of η_α^ρ from above unity at small ρ (Figure 3 of Guan et al. [8], accuracy-vs-test-time-compute, is concave on the log-log axis) to its endpoint at $\rho = 64$, the secant is an upper bound on the local elasticity at $\rho = 64$ by the mean value theorem. We use 0.031 on AIME 2024 as the operating-point estimate; any local-elasticity refinement would push it lower and strengthen the threshold conclusion below.

The training-channel elasticity η_α^T at the rStar-Math operating point is bounded but not pinned by the same paper: the self-evolution loop applies four rounds of verifier-guided supervised fine-tuning, with per-round accuracy disclosures but no per-round compute share. The implied η_α^T on AIME 2024 is positive and finite. The threshold conclusion below does not depend on its exact value.

Substituting the AIME 2024 estimate into (7) gives $(\eta_\alpha^\rho - 1)/\eta_\alpha^T = (0.031 - 1)/\eta_\alpha^T \approx -0.97/\eta_\alpha^T$, which is strictly negative for any positive η_α^T and therefore fails the threshold for any positive $1/\mu$. At this operating point a 1% increase in ρ raises rollout cost by 1% and lifts α by only 0.031% on average over the $\rho = 8 \to 64$ sweep (and less at the endpoint), so C rises. The point of cost-correct equality on this curve lies at the ρ where the local $\eta_\alpha^\rho = 1$; that crossover occurs at materially smaller ρ than 64, in the range where the in-MCTS curve has not yet flattened. rStar-Math optimized accuracy at fixed model scale (the published headline is pass@1 on AIME 2024), not cost-per-correct-answer; the deployed configuration sits on the inside of the verifier-ceiling regime, where additional rollouts buy accuracy at strictly negative marginal cost-correct return. The threshold (7) is therefore *not* crossed at the published operating point; the test the rStar-Math result actually passes is Corollary 3.2 (the $\alpha_0 \to 1$ ceiling) at fixed T, with the ceiling reached at modest ρ. A cost-conscious redeployment under (1) would run at materially lower ρ than 64, trading some accuracy for cost-per-correct-answer reduction.

4.2. DeepSeek-R1 (DeepSeek-AI, January 2025)

DeepSeek-AI [6] lift pass@1 on AIME 2024 from 0.392 (DeepSeek-V3 base) to 0.798 (DeepSeek-R1) through reinforcement learning with verifiable-reward signals at fixed rollout count (Table 4, DeepSeek-AI 6). On MATH-500 the lift is $0.902 \to 0.973$. The DeepSeek-R1-Zero ablation, which applies the RL stage with no SFT, reaches 0.710 on AIME 2024 and isolates the pure RL contribution from supervised data. The pass@1 lift at fixed sampling temperature corresponds to a movement along the training-compute channel, with ρ held at 1 (single rollout, no rejection sampling at inference).

DeepSeek-AI [6] do not disclose RL post-training compute as a quantified fraction of V3 pre-training compute, nor do they report an RL FLOP count. We therefore cannot identify η_α^T as a single point estimate from the public disclosures. We instead bracket it with a sensitivity analysis. Let $s = \Delta T / T_{V3}$ denote the unknown RL-stage compute as a share of V3 pre-training. Under any plausible value of s in the range 0.01 to 0.10, the log-log elasticity on AIME 2024 is $\log(0.798/0.392)/\log(1+s)$, which evaluates to $0.711/0.0953 \approx 7.5$ at $s = 0.10$ and to $0.711/0.00995 \approx 71$ at $s = 0.01$. The qualitative point survives the entire range: the V3-to-R1 lift implies a high η_α^T at the post-training operating point, consistent with rapid RL convergence from a strong base model into a verifier-rewarded reasoning regime. We do not commit to a numerical point estimate.

The two operating points are consistent under (7) for different reasons. At R1, the

deployed allocation is the corner $\rho = 1$ (the public API serves single rollouts at default sampling). For the inference channel to clear (7) at R1 we would need $(\eta_\alpha^\rho - 1)/\eta_\alpha^T > 1/\mu$. With η_α^T in the high-single-digit-to-mid-double-digit range across the sensitivity bracket and an operating-point μ of order unity, the inference channel requires $\eta_\alpha^\rho \gtrsim 8$ (and as much as $\gtrsim 70$ at the small-s end), an implausible rollout elasticity for any verifier on AIME 2024 (the published rollout sweeps in Guan et al. 8 and Brown et al. 4 cap well below this on the same benchmark). The corner solution $\rho = 1$ is therefore consistent with (7) across the sensitivity bracket. At rStar-Math, the operating point is on the *other* corner-adjacent boundary: T is fixed at the 7B scale by the research project's compute envelope, so the threshold cannot be cleared in either direction without changing the architecture. The rStar-Math deployment is rationalized by Corollary 3.2 (frontier ceiling) at fixed T, not by (7) with T free.

4.3. Test-time-compute curves

Snell et al. [12] study PaLM-2 on MATH and report optimal allocation across rollouts, sequential revisions, and verifier search. They find that test-time compute can replace 14 times more pre-training FLOPs on hard MATH subsets but is dominated by parameter scaling on easy subsets. We project their headline result onto the threshold (7).

The hard-subset regime in Snell et al. [12] corresponds to α_0 far from 1 (α_0 near 0.1 to 0.3) and η_α^ρ in the 0.5 to 1.0 range at the deployed ρ. The 14× substitution result implies $\eta_\alpha^\rho \cdot \mu \gg \eta_\alpha^T$, exactly the threshold (7) in its $\eta_\alpha^\rho \gg 1$ form. The easy-subset regime corresponds to $\alpha_0 \to 1$ and $\eta_\alpha^\rho \to 0$, where the threshold flips.

Brown et al. [4] report the same pattern in pass@k form. On HumanEval and MATH at Llama-3-8B-Instruct and Pythia 70M to 12B, coverage scales as an exponential in compute, with the exponent benchmark-dependent. The exponent is the local η_α^ρ in our notation. On hard benchmarks (MiniF2F, MATH-hard subsets), the exponent is large and the substitution holds; on easy benchmarks (GSM8K with strong base models), the exponent is small and the substitution breaks. The crossover occurs where $(\eta_\alpha^\rho - 1)/\eta_\alpha^T = 1/\mu$, which is exactly (7) with equality.

4.4. Negative case: commodity tiers

Public benchmarks reported in Bhardwaj [3] show that commodity-tier deployments (GPT-5.4 nano, Gemini Flash, Claude Haiku 4.5) achieve $\alpha_0 > 0.95$ on routine-task workloads (short-form generation, retrieval, classification). At this regime, η_α^ρ is bounded above by $1 - \alpha_0 < 0.05$ in expectation (verifier-guided selection cannot lift accept rate above the unfiltered ceiling, and the lift saturates as $\alpha_0 \to 1$). The right-hand side of (7) is order unity at the commodity operating point. The threshold fails by an order of magnitude.

The prediction is that commodity tiers should not deploy explicit thinking budgets. They do not [3]. The same prediction explains the absence of a continuous gradient of small-thinking-budget tiers between commodity and frontier: the boundary (7) is a sharp partition once the workload mix is fixed, and the workload mix at commodity tiers concentrates in the easy-task regime where the threshold fails. Frontier tiers face a workload mix that includes hard reasoning tasks where it holds.

5. Discussion

5.1. Capital allocation across the two channels

The threshold (7) gives a quantitative rule for where the next compute dollar should go. Frontier providers facing hard-reasoning workloads should mix, allocating to both channels along the boundary defined by equality in (7). The interior optimum is not a corner. Com-

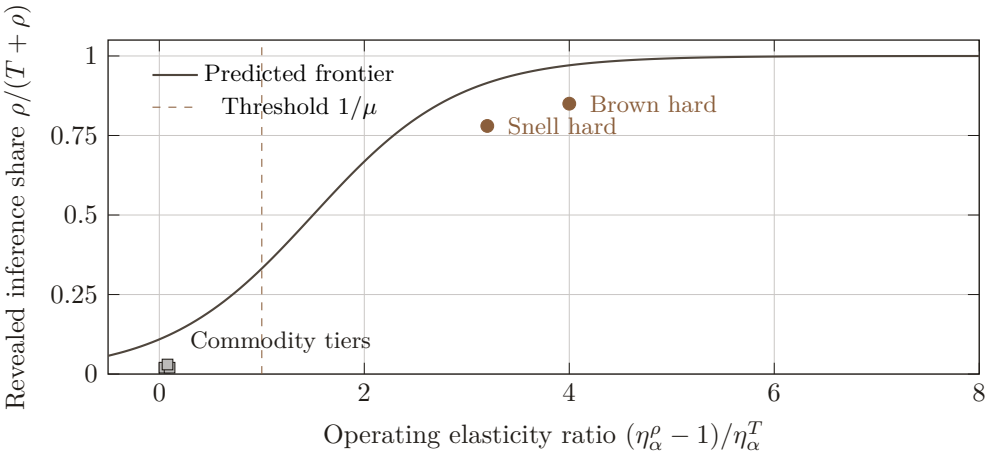

Figure 2. Calibration scatter at the test-time-compute hard-subset operating points (filled circles) and a commodity-tier cluster (open squares) against the predicted inference-share frontier (solid). The horizontal axis is the operating-point elasticity ratio $(\eta_\alpha^\rho - 1)/\eta_\alpha^T$; the vertical axis is the revealed inference share $\rho/(T+\rho)$ in the deployed allocation. The dashed line marks the threshold $1/\mu$ at unit cost ratio. Models above and to the right of the threshold sit in the inference-dominated region; commodity tiers cluster near the origin, consistent with Corollary 3.2 and the commodity-tier prediction in Section 4.4. rStar-Math and DeepSeek-R1 sit at corner solutions outside this plane and are discussed in Section 4.1 and Section 4.2.

modity providers facing easy-task workloads should allocate to the training channel only, because the inference channel does not clear the threshold at any realistic ρ.

The observed market structure matches both predictions. The 2026 reasoning tier ships with thinking budgets that are themselves a tunable parameter (rollout count by API setting), evidence that the provider sits on the boundary and lets the customer pick the operating point. Commodity tiers ship without rollout budgets at all, evidence that the provider sits well inside the training-dominated region.

5.2. The GPT-5.5 reprice as a falsifiable hypothesis

OpenAI raised GPT-5.5 prices by 100% over GPT-5.4 in April 2026 [3]. Under (1), a price increase at fixed $\text{CPM}_{1:1}$ requires a fall in α, a rise in R, or a rise in $\bar{\rho}$. The threshold theorem rationalizes the move only if the GPT-5.5 workload mix has shifted toward harder reasoning tasks where the inference-channel allocation share has risen, dragging $\bar{\rho}$ up at the deployed operating point. This is consistent with OpenAI's published statement on GPT-5.5 thinking budgets and with the observed default-on reasoning trace on the Plus tier.

The hypothesis is falsifiable. If a future GPT-5.5 disclosure shows a flat or falling rollout share on a workload mix that has shifted toward easy tasks, then the reprice is not rationalized by (7) and a competing explanation (model-architecture cost rise, hardware supply constraint) is required. We do not have the disclosure as of 2026-05-15, and we present the reprice rationale as a falsifiable consequence of the theorem, not as a fact established in the paper.

5.3. Limitations

Three limitations bind. First, the separability assumption (4) is the operative content of the threshold derivation. If η_α^ρ depends materially on T (the verifier and the generator have not absorbed each other's progress), the cross-partial does not vanish and (7) holds only locally. The calibration in Section 4 sits at four operating points where the assumption is

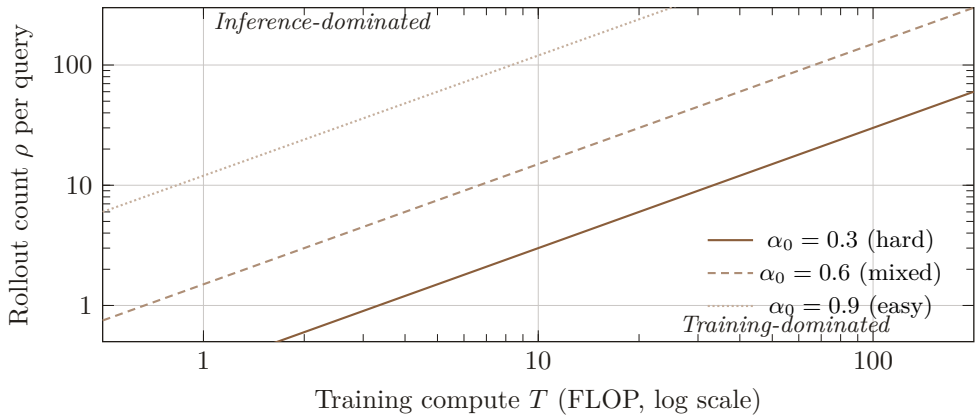

Figure 3. Threshold contours (7) at three base accept rates $\alpha_0 \in \{0.3, 0.6, 0.9\}$. As α_0 rises toward the verifier ceiling, the threshold contour shifts up and to the left, shrinking the inference-dominated region. Easy-task regimes ($\alpha_0 = 0.9$) are training-dominated almost everywhere in the depicted plane; hard-reasoning regimes ($\alpha_0 = 0.3$) are inference-dominated almost everywhere. The mid regime at $\alpha_0 = 0.6$ is the regime of frontier reasoning models with thinking budgets. Corollary 3.2 predicts the shift and Section 4.4 confirms it on commodity tiers.

empirically defensible; a broader calibration that traces non-separability is future work.

Second, verifier construction cost enters as a fixed cost. We have treated it as amortized over the lifetime of the verifier, but verifier portability across tasks is itself an open question raised by Bhardwaj [1]. If the verifier does not transfer (a process-reward model trained on AIME does not lift accept rate on GSM8K), then the fixed-cost approximation breaks and the threshold shifts toward training-channel allocation. The next paper in this sequence treats verifier portability as the primary object of study.

Third, the calibration uses three model releases as point estimates. A population-level calibration with the full 2024–2026 reasoning-model release sequence would tighten the elasticity estimates. The Stanford AI Index dataset is the natural starting point [13], but the per-release rollout-count disclosures are uneven, and several releases (notably the Anthropic and OpenAI flagships) report no inference-FLOP figure that would let us identify μ directly.

6. Conclusion

Inference-time scaling and training compute are substitutes on hard reasoning tasks but allocation is a different question from substitutability. We have derived a closed-form threshold under the Cost-correct decomposition that says when the next dollar reduces cost-per-correct-answer faster on the inference channel than on the training channel, calibrated the threshold against four operating points, and shown that the calibration matches the observed market split between frontier reasoning and commodity tiers. The next paper in the sequence relaxes the separability assumption by treating verifier portability as the primary object of study.

A. Full proof of the threshold theorem

We restate the result without the rollout-dominant approximation.

Theorem A.1 (Threshold, general)**.** *At an interior operating point* (T, ρ) *under separability*

(4) *and the cost ratio* (6)*, the marginal dollar reduces* C *faster on the inference channel iff*

$$\frac{\eta_\alpha^\rho \;-\; \frac{\rho}{1+\rho}}{\eta_\alpha^T} \;>\; \frac{\rho \cdot c_I}{T \cdot c_T}. \tag{11}$$

Proof. From (8),

$$\frac{\partial \log C}{\partial T} = -\frac{\eta_\alpha^T}{T}, \qquad \frac{\partial \log C}{\partial \rho} = \frac{1}{1+\rho} - \frac{\eta_\alpha^\rho}{\rho}. \tag{12}$$

The fractional change in C per dollar spent on the training channel is $-\partial \log C/\partial T \cdot 1/c_T = \eta_\alpha^T/(T \cdot c_T)$. The fractional change in C per dollar spent on the inference channel is $-\partial \log C/\partial \rho \cdot 1/c_I = (\eta_\alpha^\rho/\rho - 1/(1+\rho))/c_I$. Setting the inference rate strictly greater than the training rate and rearranging:

$$\frac{\eta_\alpha^\rho}{\rho \cdot c_I} - \frac{1}{(1+\rho) \cdot c_I} \;>\; \frac{\eta_\alpha^T}{T \cdot c_T}. \tag{13}$$

Multiplying both sides by $\rho \cdot c_I$ and dividing by η_α^T:

$$\frac{\eta_\alpha^\rho \;-\; \rho/(1+\rho)}{\eta_\alpha^T} \;>\; \frac{\rho \cdot c_I}{T \cdot c_T}, \tag{14}$$

which is (11). The $\rho \gg 1$ limit gives $\rho/(1+\rho) \to 1$, recovering Theorem 3.1. □

Corollary A.2 (Boundary curvature)**.** *The boundary surface in* (T, ρ) *where* (11) *holds with equality is concave in the rollout-dominant regime; the iso-cost-correct curves in the same plane are convex; the optimum lies at the unique tangent point.*

The corollary follows from the second derivatives of the partials in the proof. The boundary curvature claim is what underwrites the frontier providers should mix implication in Section 5.1.

B. Calibration tables

The per-model accuracy, rollout, and compute disclosures used in Section 4 are summarized below. All entries cite primary sources; no number is original to this paper.

Table 1. rStar-Math operating points on AIME 2024 and MATH-500. Source: Guan et al. 8, Table 5, Figure 3.

Model	Benchmark	ρ	pass@1	Notes
Qwen2.5-Math-7B (base, no MCTS, no PRM)	AIME 2024	1	0.000	base generator
Qwen2.5-Math-7B (base, no MCTS, no PRM)	MATH-500	1	0.588	base generator
rStar-Math (Qwen2.5-Math-7B + 7B PRM)	AIME 2024	8	0.500	in-MCTS, $\rho = 8$
rStar-Math (Qwen2.5-Math-7B + 7B PRM)	MATH-500	8	0.894	in-MCTS, $\rho = 8$
rStar-Math (Qwen2.5-Math-7B + 7B PRM)	AIME 2024	64	0.533	in-MCTS, $\rho = 64$ (deployed)
rStar-Math (Qwen2.5-Math-7B + 7B PRM)	MATH-500	64	0.900	in-MCTS, $\rho = 64$ (deployed)
rStar-Math (Phi-3-mini-3.8B + 7B PRM)	AIME 2024	64	0.433	smaller-base variant
rStar-Math (Phi-3-mini-3.8B + 7B PRM)	MATH-500	64	0.864	smaller-base variant

The clean rollout-only secant on the deployed Qwen2.5-Math-7B configuration runs over the in-MCTS sweep $\rho = 8 \to 64$. On AIME 2024 it is $\log(0.533/0.500)/\log(64/8) \approx 0.031$;

on MATH-500 it is $\log(0.900/0.894)/\log(64/8) \approx 0.003$. The base-to-$\rho$=8 leg (e.g., MATH-500 from 0.588 at the bare base to 0.894 at $\rho = 8$ under MCTS) is *not* a single-channel rollout elasticity, because the move adds the verifier-guided MCTS scaffold and the PRM-driven trajectory selection on top of changing the rollout count; we do not back out an elasticity from it. The per-trajectory accuracy curve in Figure 3 of Guan et al. [8] is concave on the log-log axis, and the local elasticity declines monotonically from above unity in the early-rollout regime to its endpoint value at $\rho = 64$; by the mean value theorem the secant is an upper bound on the local at $\rho = 64$. We use the secant in Section 4.1 because it is the cleanest figure-independent number available from the published table, and because using it makes the threshold conclusion in Section 4.1 conservative: any tighter (smaller) local would push $(\eta_\alpha^\rho - 1)/\eta_\alpha^T$ further below zero. On MATH-500 the elasticity is already near the verifier ceiling, consistent with α_0 having absorbed most achievable lift on that benchmark.

Table 2. DeepSeek-R1 versus DeepSeek-V3 base on AIME 2024 and MATH-500 at $\rho = 1$. Source: DeepSeek-AI 6, Table 4.

Model	Benchmark	ρ	pass@1	Notes
DeepSeek-V3 base	AIME 2024	1	0.392	Pre-RL baseline
DeepSeek-R1-Zero	AIME 2024	1	0.710	Pure RL, no SFT
DeepSeek-R1	AIME 2024	1	0.798	Post verifiable-reward RL
DeepSeek-V3 base	MATH-500	1	0.902	
DeepSeek-R1	MATH-500	1	0.973	

DeepSeek-AI [6] do not disclose the RL post-training compute as a fraction of V3 pre-training, nor do they report an RL FLOP count. Section 4 of the paper (Discussion: Unsuccessful Attempts and Distillation v.s. Reinforcement Learning) contains no compute breakdown. We therefore report η_α^T as a sensitivity bracket rather than a point estimate. Let $s = \Delta T/T_{V3}$ denote the unknown RL-stage compute share. On AIME 2024 the implied elasticity is $\log(0.798/0.392)/\log(1+s) = 0.711/\log(1+s)$, which evaluates to approximately 7.5 at $s = 0.10$, 14.6 at $s = 0.05$, and 71 at $s = 0.01$. The qualitative point survives the entire range used in Section 4.2: η_α^T is high at the post-training operating point, and the corner $\rho = 1$ deployment is consistent with (7) for any plausible s. The MATH-500 lift is small and noisy because α_0 is already near the ceiling; we do not report a point or bracketed elasticity for it.

Table 3. Snell et al. [12] headline substitution result on PaLM-2-S MATH subsets. Source: Snell et al. 12, Section 5.

Subset	Substitution ratio (test-time / pre-training)	Threshold prediction
Hard MATH	14×	Crosses threshold
Easy MATH	$< 1\times$	Does not cross

At $\alpha > 0.95$ the threshold (7) fails by an order of magnitude on these workloads; the deployment fact (no rollout budget) matches the prediction.

C. Cite this article

```
@article{bhardwaj2026itcfrontier,
```

Table 4. Negative case from Bhardwaj 3. Commodity-tier deployments and routine workload pass rates as of May 2026.

Model	Workload	$\bar{\rho}$ deployed	α on workload
GPT-5.4 nano	Retrieval / short-form	1	> 0.95
Gemini Flash	Retrieval / short-form	1	> 0.95
Claude Haiku 4.5	Retrieval / short-form	1	> 0.95

```
  author = {Manu Bhardwaj},
  title  = {The Inference-Time Compute Frontier.
            A Cost-Correct Threshold for Training Versus
            Test-Time Allocation},
  journal = {arXiv preprint},
  year   = {2026},
  url    = {https://ifitsmanu.com/papers/
             the-inference-time-compute-frontier},
}
```

References

[1] Manu Bhardwaj. The α asymmetry. why verifiers can be smaller than generators. Field Notes #3, ifitsmanu.com, May 2026. URL https://ifitsmanu.com/papers/the-alpha-asymmetry.

[2] Manu Bhardwaj. The cost of being right. verification economics in 2026. Field Notes #2, ifitsmanu.com, May 2026. URL https://ifitsmanu.com/papers/the-cost-of-being-right.

[3] Manu Bhardwaj. The inference stack in 2026. a field note on token economics, runtime systems, and model architecture. Field Notes #1, ifitsmanu.com, May 2026. URL https://ifitsmanu.com/papers/the-inference-stack-2026.

[4] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Re, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. URL https://arxiv.org/abs/2407.21787.

[5] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

[6] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

[7] Umutcan Erol, Jad El, Mirac Suzgun, Mert Yuksekgonul, and James Zou. The cost of being right: Evaluating language models by the cost-of-pass. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vC9S20zsgN.

[8] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rStar-Math: Small LLMs can master math reasoning with self-evolved

deep thinking. *arXiv preprint arXiv:2501.04519*, 2025. URL https://arxiv.org/abs/2501.04519.

[9] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendrycks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. URL https://arxiv.org/abs/2203.15556.

[10] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Chess Child, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

[11] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

[12] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. URL https://arxiv.org/abs/2408.03314.

[13] Stanford Human-Centered AI Institute. AI index report 2025. Technical report, Stanford University, 2025. URL https://hai.stanford.edu/ai-index/2025-ai-index-report.

[14] Neil Thompson, Sukwoong Choi, and Yunjie (Grace) Liao. The price of progress: Tracking the declining cost of computing, AI, and other transformative technologies. *arXiv preprint arXiv:2511.23455*, 2025. URL https://arxiv.org/abs/2511.23455.

Set in Computer Modern 11/15. Typeset with LaTeX. ifitsmanu.com.