The AI Task Frontier
Horizon, Forgiveness, and Directed Innovation

June 2026

Abstract

Why do cheap AI systems quickly commoditize some tasks while frontier labs continue to spend heavily on harder systems? This paper argues that the relevant state variable is not model quality on a scalar ladder but a task frontier. Tasks are hard along two economically distinct margins. Some require long chains of coordinated action; others leave little room to observe, correct, retry, or roll back mistakes. Innovation projects also have directions: they can move the horizon boundary, the low-forgiveness boundary, or both. The first result is a project-value rule: a project is valuable where it moves the scarce boundary. Search, tools, and test-time compute are most valuable where long but forgiving tasks are just out of reach; reliability, verification, monitoring, and rollback are most valuable where errors are costly and hard to reverse. A scalar quality ladder cannot generate these cross-region project-value reversals for a common project menu without adding the missing task heterogeneity back through project effects or costs. The paper then embeds the task frontier in a simple dynamic game with two front-running labs and a fringe of distillers. Distillation and cheap products erode rents in saturated task regions, which can increase foundational labs' incentives to race toward longer, higher-value tasks. But distillation also lowers the private prize from any frontier task that can be cheaply copied. The resulting investment direction is shaped by both task scarcity and appropriability: frontier labs move toward task regions where boundary value is high and cheap imitation is less effective. The model explains why low-cost products and frontier investment can rise together, and why frontier competition is often a race to define the next valuable task rather than a race to serve already saturated tasks.

JEL: O31, O33, L13, L40, D24.

1 Introduction

AI progress is often described as movement up a single quality ladder. That language is useful for ranking models, but it misses an economic margin. The same technical improvement can be decisive in one task market and almost irrelevant in another. Cheap models, distillation,

retrieval, and workflow integration can rapidly erode frontier premia in many deployed applications. At the same time, frontier labs continue to invest in reliability, verification, long-horizon control, tool use, and more demanding forms of agency.

This coexistence is hard to understand if the state of AI is one scalar quality index. If all tasks differ only in how much quality they require, then a better model should dominate a worse model wherever the quality gap matters, and cheap followers should simply compress markups where the quality gap is small. That view misses two related facts. First, tasks are hard for different reasons. Second, firms do not invest in "quality" in the abstract. They invest in projects that move some parts of the task frontier more than others.

The paper's primitive is the AI task frontier. A task is indexed by horizon and forgiveness. Horizon is the serial depth of the action path: the number of coordinated steps, dependencies, state contingencies, and context links needed for success. Forgiveness is the extent to which mistakes can be observed, corrected, retried, sandboxed, or absorbed before they create loss. Coding can be long but forgiving because tests, logs, and version control catch many errors. A medical dosage, driving maneuver, or payments-control decision can be shorter but less forgiving because a local error may be costly or irreversible. Autonomous scientific work can be both long and unforgiving.

This distinction changes the economics of innovation. A project that helps a system plan, search, use tools, or coordinate many steps mainly moves the horizon boundary. A project that improves base reliability, monitoring, verification, rollback, or institutional control mainly moves the low-forgiveness boundary. The same engineering advance can therefore be central in one task region and nearly redundant in another. The paper's first result makes this idea precise. The task-side value of a project equals the dot product of regional boundary scarcity and the project's effective movement per dollar. A project is valuable where it moves the scarce boundary.

This static logic gives the paper empirical content. A one-dimensional quality ladder with common project increments and costs cannot generate project-value reversals across task regions. A horizon-forgiveness frontier can. With many task regions and many projects, the model predicts sign and rank restrictions on the local project-value matrix: rows are task regions, columns are project directions, and entries are task-side frontier values per dollar. The restrictions are not labels attached after the fact. Task coordinates and project movements must be measured before the ranking exercise.

The static model also clarifies why cheap AI products can win in some markets while frontier systems remain valuable in others. A task region can have large current revenue and low marginal frontier value if many systems already serve it well. Cheap models and distilled products should win there when their cost advantage exceeds the remaining performance gap. A different region can have little current revenue and high frontier value if many valuable tasks sit just beyond a horizon or low-forgiveness boundary. Levels and slopes are different objects.

The second part of the paper adds a dynamic game. This is where the economics of frontier labs becomes sharper. Suppose two front-running labs compete while a fringe of distillers and cheap-product firms serves saturated tasks. If the existing task region is commoditized, then staying in place yields low rents. A lab can instead invest in defining a longer or less forgiving task that has higher value. The resulting game is not only a race to be better on today's tasks. It is a race to move the task frontier and thereby define tomorrow's scarce task market.

The dynamic game has two front-running labs and a low-cost fringe. Each frontier lab can stay in the saturated region or invest in a frontier project that opens a new task family. If one lab invests, it may obtain a lead prize from defining the new task. If both invest, the prize is dissipated by competition. The equilibrium depends on three objects: the prize from the new task, the cost of the frontier project, and the outside profit from staying in the saturated region. Distillers enter through that outside profit and through appropriability. They erode rents in already served tasks, which pushes frontier labs outward. But they also reduce the private prize from frontier tasks that can be cheaply copied, which pulls investment away from easily distilled task regions.

This gives a useful way to think about foundation labs and cheaper AI products in the same model. Distillers and cheap-product firms are not just passive imitators. They change the direction of frontier investment. They make already served, forgiving, codified tasks less attractive for frontier labs. They push frontier labs toward task regions where boundary value remains high and imitation is harder: longer tasks, less forgiving tasks, tasks requiring verification and monitoring, or tasks whose value comes from integration with hard-to-copy workflows and institutions. In this sense, distillation can accelerate frontier investment even as it reduces rents from any one frontier improvement.

The paper is theoretical. It contributes to task-based theories of technical change by putting the direction of AI progress on the task side rather than only on the worker-task substitution margin [Acemoglu and Autor(2011), Acemoglu and Restrepo(2018)]. It contributes to directed and step-by-step innovation by making the object of the race a moving task frontier rather than a scalar technology level [Aghion et al.(2001), Gilbert and Newbery(1982)]. The applied AI mechanisms in the paper–search, tools, verification, monitoring, distillation, and cheap inference–are not separate theories. They are ways to move different boundaries or change the private value of doing so.

The rest of the paper proceeds as follows. Section 2 defines the task frontier. Section 3 derives the project-value rule and the scalar benchmark restriction. Section 4 studies cheap followers and distillation in saturated regions. Section 5 introduces the two-period frontier race between two front-running labs. Section 6 shows how distillation changes the direction of frontier investment. Section 7 collects predictions and measurement implications. The appendix contains short proofs and derivations.

2 The Task Frontier

The task space is a measurable set $\mathcal{T} \subseteq \mathbb{R}^2$. A task is indexed by (h, f). The coordinate h is horizon: the serial depth of the required action path. The coordinate f is fragility, the inverse of forgiveness. Higher f means less slack for retry, correction, observation, rollback, or reversible error.

These are task-side requirements, not engineering primitives. A larger model, a better tool interface, a verifier, a monitoring system, a workflow redesign, or a cheaper inference stack can affect many engineering quantities. The theory asks how those changes translate into the task requirements an AI system can now satisfy.

Let an AI system have capability vector $q = (q_H, q_F)$. The first coordinate shifts the frontier over long tasks. The second shifts the frontier over fragile, low-forgiveness tasks. A transparent benchmark writes task success as

$$p(q, h, f) = G_H(q_H - h)G_F(q_F - f), \tag{1}$$

where G_H and G_F are increasing and differentiable. This product form is not the substantive contribution. It is a useful way to compute boundary values and to keep the two task-side margins visible.

Total task surplus is

$$W(q) = \iint_{\mathcal{T}} a(h, f)G_H(q_H - h)G_F(q_F - f)\, dh\, df, \tag{2}$$

where $a(h, f) \geq 0$ is a task-value density. For a region $R \subseteq \mathcal{T}$, write $W_R(q)$ for the same integral restricted to R.

The coordinates have operational content. Horizon can be disciplined by task families that vary chain length, dependency depth, context span, or the number of state-contingent actions. Fragility can be disciplined by variation in retryability, observability, rollback, verification coverage, tolerated failure probability, or the cost of a bad action. A raw increase in stages generally moves both coordinates. A clean horizon comparison holds the reduced fragility threshold fixed.

The sequential motivation is simple. Suppose a task has n essential stages. A local action succeeds with probability $G_F(q_F - \varphi)$. If the task permits ℓ retries at each stage, the stage success probability is

$$S_\ell(q_F, \varphi) = 1 - \{1 - G_F(q_F - \varphi)\}^{1+\ell}.$$

To keep task-level failure below ε, the system must satisfy

$$S_\ell(q_F, \varphi)^n \geq 1 - \varepsilon.$$

Equivalently, q_F must exceed a reduced fragility threshold

$$f(n,\ell,\varepsilon,\varphi) = \varphi + G_F^{-1}\Big(1 - \{1-(1-\varepsilon)^{1/n}\}^{1/(1+\ell)}\Big).$$

This threshold rises with serial depth n, rises with local burden φ, falls with retries ℓ, and rises when the tolerated failure probability ε falls. Horizon and forgiveness interact because serial depth compounds the residual local risk that remains after retries and safeguards.

2.1 Boundary Values

The marginal value of AI progress is the value mass close to the current frontier. Define the regional boundary values

$$B_H^R(q) = \frac{\partial W_R(q)}{\partial q_H}, \qquad B_F^R(q) = \frac{\partial W_R(q)}{\partial q_F}. \tag{3}$$

Under the benchmark in (1),

$$B_H^R(q) = \iint_R a(h,f) g_H(q_H - h) G_F(q_F - f)\, dh\, df,$$

$$B_F^R(q) = \iint_R a(h,f) G_H(q_H - h) g_F(q_F - f)\, dh\, df.$$

The first statistic is high when valuable tasks are just beyond the horizon boundary. The second is high when valuable tasks are just beyond the low-forgiveness boundary.

The cross-boundary statistic

$$C^R(q) = \iint_R a(h,f) g_H(q_H - h) g_F(q_F - f)\, dh\, df$$

measures value mass near the corner where both constraints bind. It is also the local complementarity term:

$$\frac{\partial B_H^R(q)}{\partial q_F} = \frac{\partial B_F^R(q)}{\partial q_H} = C^R(q) \geq 0.$$

For a finite capability step $\Delta = (\Delta_H, \Delta_F) \geq 0$,

$$W_R(q+\Delta) - W_R(q) = \int_0^{\Delta_H} B_H^R(q_H + u, q_F)\, du + \int_0^{\Delta_F} B_F^R(q_H + \Delta_H, q_F + v)\, dv. \tag{4}$$

Locally,

$$W_R(q+\Delta) - W_R(q) = \Delta_H B_H^R(q) + \Delta_F B_F^R(q) + o(\|\Delta\|).$$

Figure 1 summarizes the economics. Interior tasks can have large current value and low marginal frontier value because many systems already serve them. The right side is scarce in

horizon; the top side is scarce in forgiveness; the corner is where both bind.

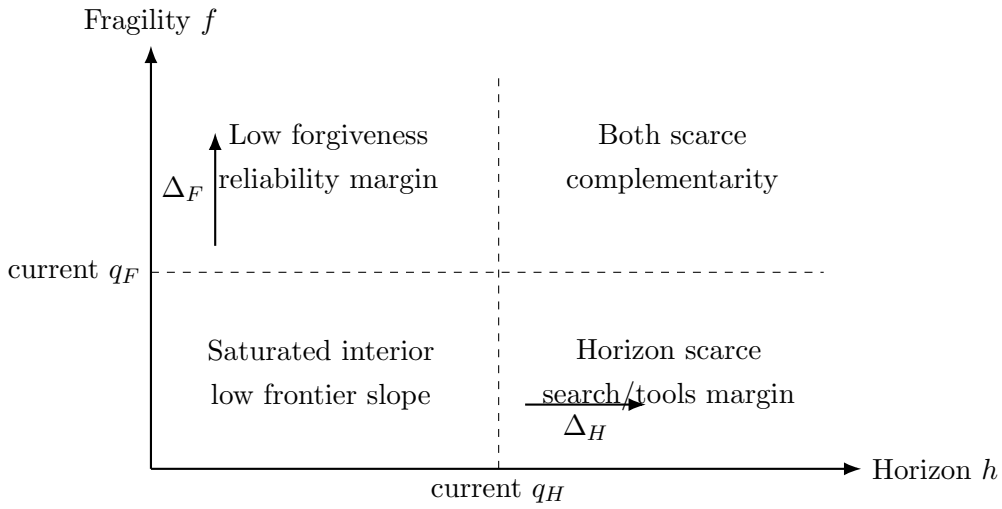

Figure 1: The AI task frontier. The value of an innovation project depends on which boundary it moves and which boundary is scarce in the task region.

2.2 Economic Task Regions

The frontier representation is useful because the same observed product market can sit in different parts of task space over time. Early in a market, tasks may be near a scarce boundary. Small improvements in planning, retrieval, tool use, reliability, or workflow integration can then create large value. Later, the same task family may move into the interior. Many systems can perform it; users learn how to adapt workflows; and evaluation becomes routine. The economic object changes from frontier movement to cost and distribution.

This distinction matters for interpreting AI adoption. A high-revenue task is not necessarily a high-frontier-value task. Customer support, routine copywriting, document summarization, and simple coding assistance can be economically large while having low marginal frontier slopes once the task is well served. By contrast, a frontier scientific assistant, a reliable autonomous operator, or a high-stakes monitoring system may have small current revenue because it is barely deployable, but high boundary value because a small movement of the frontier unlocks a valuable task family.

The framework therefore separates three questions that are often conflated. First, what is the current value of serving a task? Second, what is the marginal value of moving the frontier near that task? Third, how much of that value can a firm appropriate before cheap followers arrive? The first question governs current product revenue. The second governs the social and task-side value of innovation. The third governs private investment incentives.

This separation is also why the paper treats horizon and forgiveness as task-side coordi-

nates. A long task is not automatically valuable, and a fragile task is not automatically worth serving. Value comes from the density $a(h, f)$ near a boundary and from the private ability to appropriate the value created by moving that boundary. The frontier is therefore a way to organize economic margins, not a taxonomy of AI capabilities.

3 Projects and Static Directed Innovation

Firms choose projects, not coordinates. A project can raise capability, lower effective task thresholds, or change the cost of doing either. The project's engineering label is not the model object. Its model object is an effective frontier step

$$d_a = (d_{aH}, d_{aF})$$

and a cost $K_a > 0$. GPU supply, energy, interconnect, inference systems, financing, workflow design, and software efficiency enter by changing the feasible project menu, the cost K_a, or the effective step d_a. They are not additional task coordinates.

For a region R, the local project-value index is

$$\Lambda_{aR}(q) = B_H^R(q)\frac{d_{aH}}{K_a} + B_F^R(q)\frac{d_{aF}}{K_a}. \tag{5}$$

This is the paper's central object. It is a dot product. The region supplies scarcity prices for the two boundaries; the project supplies effective movements per dollar.

Keep two task regions in mind. In coding assistance, tests, logs, and version control make many mistakes corrigible, so search, tools, and longer inference can be valuable. In payments control, medical dosage, or driving maneuvers, a local error can be costly and hard to undo, so monitoring, verification, and rollback can dominate. The same project menu can therefore rank differently across task regions.

Proposition 1 (Project value and scalar minimality)**.** *Consider two projects y and z with costs $K_y, K_z > 0$ and effective steps d_y, d_z. In a region R with $B_H^R(q), B_F^R(q) > 0$, project y has higher task-side frontier value per dollar than project z if and only if*

$$\frac{B_H^R(q)d_{yH} + B_F^R(q)d_{yF}}{K_y} > \frac{B_H^R(q)d_{zH} + B_F^R(q)d_{zF}}{K_z}.$$

If y is more horizon-intensive in cost-normalized terms and z is more reliability-intensive,

$$\frac{d_{yH}}{K_y} > \frac{d_{zH}}{K_z}, \qquad \frac{d_{yF}}{K_y} < \frac{d_{zF}}{K_z},$$

then this comparison is equivalent to

$$\frac{B_H^R(q)}{B_F^R(q)} > \frac{d_{zF}/K_z - d_{yF}/K_y}{d_{yH}/K_y - d_{zH}/K_z}.$$

Hence a project-value reversal across two positive-boundary regions requires the two regional scarcity ratios to lie on opposite sides of the same project threshold.

In a scalar benchmark with one capability coordinate $\bar{q}$, region-specific local values $\tilde{B}^R(\bar{q}) > 0$, and region-invariant project increments and costs γ_a, K_a, the sign of the local project ranking is the sign of $\gamma_y/K_y - \gamma_z/K_z$ in every region. Therefore a scalar frontier cannot generate $y \succ_R z$ and $z \succ_{R'} y$ at the same state unless it adds project-region-specific productivities, costs, or adoption wedges.

The same logic has a many-region implication. Let M be the local project-value matrix with entries $M_{Ra} = \Lambda_{aR}(q)$. Let

$$b_R(q) = (B_H^R(q), B_F^R(q)), \qquad v_a = \left(\frac{d_{aH}}{K_a}, \frac{d_{aF}}{K_a}\right).$$

Then $M_{Ra} = b_R(q) \cdot v_a$. For two regions R, R' and projects a, a',

$$M_{Ra}M_{R'a'} - M_{Ra'}M_{R'a} = \{B_H^R B_F^{R'} - B_F^R B_H^{R'}\}\{v_{aH}v_{a'F} - v_{aF}v_{a'H}\}. \tag{6}$$

A scalar frontier representation implies rank at most one. The two-boundary model permits rank two only when regional scarcity vectors and project directions are both non-collinear. This is useful because it converts a qualitative theory into a restriction on held-out project values.

4 Saturation, Cheap Followers, and Distillation

The static frontier also explains why cheap products can dominate some tasks without eliminating frontier value elsewhere. Let L be a frontier system and M a lower-cost model or distilled product, with $q^L = q^M + \Delta$ and $\Delta \geq 0$. For a region R, let $c_i(R)$ be the cost of serving the region with system i, and let

$$\kappa_R = c_L(R) - c_M(R)$$

be the follower's cost advantage.

Proposition 2 (Cost-performance sorting)**.** *The lower-cost product is preferred on delivered*

net value in region R if and only if

$$\kappa_R \geq W_R(q^L) - W_R(q^M).$$

Moreover,

$$W_R(q^L) - W_R(q^M) = \int_0^{\Delta_H} B_H^R(q_H^M + u, q_F^M)\, du + \int_0^{\Delta_F} B_F^R(q_H^L, q_F^M + v)\, dv.$$

If $\bar{B}_R(q^M, q^L)$ bounds $B_H^R + B_F^R$ on the rectangle between the two systems, then

$$W_R(q^L) - W_R(q^M) \leq \bar{B}_R(q^M, q^L)\|\Delta\|_1.$$

The proposition is a levels-versus-slopes result. A region can have large current value $W_R(q)$ and still favor a lower-cost product if the remaining frontier gap has small boundary value. A different region can have small current revenue and high frontier value if valuable tasks sit near a horizon or low-forgiveness boundary.

Distillation is one way to create the low-cost product. It can turn expensive runtime procedures, tool-use routines, or frontier-model behavior into a cheaper future system [Hinton et al.(2015), Hsieh et al.(2023)]. But distillation is not equally effective everywhere. It is more powerful when the target task is forgiving, codified, observable, and easy to evaluate. It is weaker when value depends on rare failures, costly mistakes, monitoring, institutional workflow, or hard-to-copy verification. This means cheap followers are most likely to commoditize the interior of the task frontier and less likely to eliminate premia at scarce boundaries.

This distinction is important for interpreting market evidence. If cheap models win in coding autocomplete, customer support, routine document work, or search assistance, that does not imply frontier investment has low value everywhere. It implies those regions have become closer to the saturated interior. Frontier labs may then move toward tasks that are longer, less forgiving, or harder to distill.

5 A Two-Period Frontier Race

The static model explains which projects are valuable on the task side. It does not by itself explain why frontier labs keep investing when many deployed tasks appear saturated. A simple dynamic game adds the missing strategic margin.

There are two front-running labs, $i = 1, 2$, and a competitive fringe of distillers and cheap-product firms. The current served region is S. Because S is saturated, frontier improvements in S have low boundary value and cheap products put pressure on markups. Each front-running lab can either stay in S or invest in a frontier project that opens a new task family

N. The new family is longer, less forgiving, or both. It has private lead prize $P_N > 0$ before imitation, project cost $K_N > 0$, and distillability $\alpha_N \in [0, 1]$. A higher α_N means that cheap followers can copy or approximate the new task more quickly, reducing the appropriable prize to $(1 - \alpha_N)P_N$.

Let $\pi_S(\delta_S)$ be the profit from staying in the saturated region, where δ_S measures distiller pressure in S. Assume

$$\frac{\partial \pi_S(\delta_S)}{\partial \delta_S} < 0.$$

Distillers lower the outside value of staying in the old market.

The timing is:

1. In period 0, each frontier lab chooses I (invest in N) or S (stay).

2. If exactly one lab invests, it receives the lead payoff

$$(1 - \alpha_N)P_N - K_N.$$

3. If both labs invest, each receives expected payoff

$$\sigma(1 - \alpha_N)P_N - K_N,$$

where $\sigma \in (0, 1)$ captures prize splitting, duplicated effort, and rent dissipation.

4. A lab that stays receives $\pi_S(\delta_S)$.

The saturated-market payoff can be interpreted as the continuation value of competing in already served tasks. The frontier payoff is the value of defining the next task region.

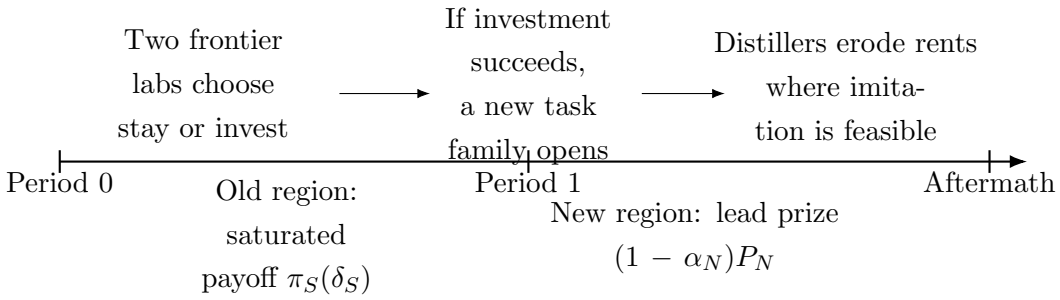

Figure 2: The two-period frontier race. Cheap followers lower the value of staying in saturated tasks and reduce the private prize from frontier tasks that are easy to copy.

The prize P_N should be read as the private value of defining a task family, not only as a product-market markup. It can come from being first to serve a high-value workflow, from setting the interface and evaluation standard, from accumulating deployment experience, or

from controlling complementary assets needed for reliable deployment. In task-frontier terms, P_N is high when a project opens a region with high boundary value and enough appropriability to reward the leader.

Proposition 3 (Frontier-race incentives)**.** *In the two-period game:*

1. *If*
$$\sigma(1-\alpha_N)P_N - K_N \geq \pi_S(\delta_S),$$
then investing is a dominant strategy and both frontier labs invest.

2. *If*
$$(1-\alpha_N)P_N - K_N \geq \pi_S(\delta_S) > \sigma(1-\alpha_N)P_N - K_N,$$
then the pure-strategy equilibria are the two preemption equilibria in which exactly one frontier lab invests.

3. *If*
$$(1-\alpha_N)P_N - K_N < \pi_S(\delta_S),$$
then neither lab invests.

An increase in distiller pressure in the saturated region expands the parameter set in which frontier investment occurs. An increase in the distillability of the new frontier task contracts it.

This proposition captures the economic reason frontier labs can keep moving even when many current tasks are saturated. Saturation lowers the value of incremental progress in the old region, but it also lowers the value of staying there. If a new task family has a large lead prize, the strategic response is to move outward. The race is not only to be better at existing tasks; it is to define a longer or less forgiving task that other systems cannot yet serve.

This mechanism is different from a pure scale race. In a scale race, firms spend because more quality wins the same market. Here, firms spend because the relevant market changes. The frontier lab wants to make a longer, harder, or less forgiving task economically serviceable before its rival does. Once that task is made routine, cheap products may again enter and push the frontier lab outward. The model therefore generates a moving frontier: saturation in one region can be the reason investment shifts to another.

The result also explains why frontier racing can coexist with commoditization. Cheap followers lower π_S, which makes the old market less attractive to frontier labs. That raises the incentive to invest in N. At the same time, if N itself is easy to distill, α_N is high and the lead prize falls. Distillers therefore have two opposing effects on frontier investment: they create escape-competition pressure, and they reduce appropriability where imitation is easy.

6 Distillation and the Direction of Frontier Investment

The dynamic game can be extended from a single frontier task N to a menu of candidate task regions. This is the setting closest to the observed AI industry. Frontier labs are not choosing only how much to invest. They are choosing what kind of frontier to push: long-horizon agents, reliable systems for fragile domains, tool-using workflows, verification and monitoring layers, or lower-cost inference.

Let $\mathcal{N}$ be a set of candidate frontier task regions. A frontier lab choosing project a for region $R \in \mathcal{N}$ gets private lead prize

$$P_{aR}(q) = \Omega_R \Lambda_{aR}(q) K_a,$$

where Λ_{aR} is the task-side project-value index and $\Omega_R > 0$ converts task-side value into an appropriable lead prize. The region has distillability $\alpha_R \in [0, 1]$. The private value of the project is

$$\Psi_{aR}(q) = (1 - \alpha_R) P_{aR}(q) - K_a. \tag{7}$$

The term Λ_{aR} captures task scarcity. The term $1 - \alpha_R$ captures appropriability in the presence of cheap followers. Both matter. A task region can have high social frontier value and low private frontier value if the improvement is immediately distilled. A different region can have moderate task-side value and high private value if cheap imitation is difficult.

Proposition 4 (Distillation and the direction of frontier investment)**.** *Consider two candidate project-region pairs (a, R) and (b, R'). A frontier lab strictly prefers (a, R) to (b, R') if and only if*

$$(1 - \alpha_R) P_{aR}(q) - K_a > (1 - \alpha_{R'}) P_{bR'}(q) - K_b.$$

Holding task-side project values fixed, an increase in α_R lowers the private value of investing in R. Holding frontier-task distillability fixed, an increase in distiller pressure in already saturated tasks lowers $\pi_S(\delta_S)$ and raises the incentive to invest in some frontier region rather than remain in S.

The proposition gives a clean interpretation of distillation. Cheap followers can make frontier labs invest more, but not necessarily in the same tasks. They reduce the value of remaining in saturated, forgiving, easily evaluated task markets. They also reduce the value of opening a new frontier market if that market is itself easy to copy. The net effect is directional: foundational labs are pushed toward tasks with high boundary value and low distillability.

This is where task reliability and task structure enter the industrial organization of AI. Distillation is most powerful when outputs are easy to evaluate and mistakes are forgiving. It is less powerful when the task requires persistent monitoring, formal verification, recovery from rare failures, institutional trust, or integration with real-world workflows. Reliability-intensive

tasks may therefore remain frontier-lab markets even when cheaper models are excellent on many routine tasks.

6.1 Three Channels

Distillation changes foundational investment through three channels.

The first channel is rent erosion in saturated tasks. When a task can be served by a cheaper distilled model, the profit from staying in that task falls. This does not merely hurt frontier labs. It can make frontier investment more attractive by lowering the outside option. In the dynamic game, this force appears as a reduction in $\pi_S(\delta_S)$.

The second channel is appropriability at the new frontier. If a new task can be quickly distilled after the frontier lab demonstrates it, the private prize from opening that task falls. This force appears as an increase in α_R, the distillability of the frontier task. It discourages investment in frontier regions whose outputs are easy to copy, benchmark, or reproduce with a smaller model.

The third channel is task redesign. Cheap models can change the task itself. They can make routine substeps inexpensive, move parts of a workflow into the saturated interior, and leave the frontier lab competing on the residual hard part: verification, monitoring, long-horizon coordination, or low-forgiveness control. This is why distillation affects not only the level of investment but also its direction.

The model does not imply that frontier labs always prefer fragile tasks. A low-forgiveness task may have high boundary value but also high development cost, regulatory cost, or deployment risk. The point is comparative. Distillation changes both the outside option and the private prize. It makes the direction of frontier investment depend not only on where the task frontier is socially scarce, but also on where cheap imitation can and cannot dissipate the lead.

7 Implications and Measurement

The model has implications at three levels: task values, product-market sorting, and dynamic investment. Table 1 summarizes the main restrictions.

Margin	Prediction	Why a scalar ladder misses it
Task-level project values	Search/tools can beat reliability in long forgiving regions, while reliability beats search/tools in low-forgiveness regions	Common scalar project increments imply the same project ranking across positive-slope regions
Rank restrictions	The local project-value matrix has rank two only when measured scarcity and project directions are both two-dimensional	A scalar frontier implies rank at most one after common project increments and costs are fixed
Cheap-product sorting	Distilled or cheaper systems win where boundary slopes are small relative to cost savings	Current market size is confused with the marginal value of frontier progress
Frontier racing	Foundational labs keep investing when saturated-market profits are low and the next task family has a high lead prize	The race is misread as wasteful overinvestment on existing tasks rather than competition to define new tasks
Direction under distillation	Cheap followers push frontier labs away from easily copied task regions and toward high boundary value, low-distillability tasks	Imitation is treated only as rent erosion, not as a force changing innovation direction

Table 1: Implications of the task-frontier model.

The measurement discipline follows from the same objects. First, task families must anchor the coordinates. Horizon-admissible families vary serial depth while holding the reduced fragility threshold fixed. Fragility-admissible families vary retryability, observability, reversibility, monitoring, or tolerated failure at fixed coordination depth. Second, project experiments must measure effective steps d_a and costs K_a before the ranking exercise. Third, held-out regions and projects should test whether signs and minors follow measured scarcity and measured project directions.

The dynamic implications require additional objects. One needs measures of saturated region profits, distiller pressure, frontier-task lead prizes, and task-specific distillability. The theory predicts that frontier labs should be more willing to move outward when cheap products erode current rents, but less willing to invest in frontier tasks whose output can be quickly

distilled. Observed investment should tilt toward tasks that combine high boundary value with low cheap-imitation pressure.

This also clarifies the paper's prediction power. A one-scalar quality model can say that better models are better and cheaper models matter when quality gaps are small. It cannot by itself predict which project direction should be valuable in which task region, when cheap products should increase frontier racing, or why distillation should redirect foundational investment toward harder-to-copy task structures. Those are the roles of horizon, forgiveness, boundary values, and appropriability.

7.1 What Would Move the Theory

The most direct evidence would not be a single benchmark score. It would be a region-by-project panel. Rows would be task families with independently measured horizon and forgiveness. Columns would be project directions such as inference-time search, tool-use integration, verifier training, monitoring, rollback, model compression, or workflow redesign. The entries would estimate task-side value per dollar, before using the same data to explain the ranking.

The dynamic part requires observing how cheap-product pressure changes frontier labs' choices. The model predicts that when distillation becomes stronger in a saturated task family, frontier labs should reduce investment aimed at that family and increase investment aimed at less saturated boundaries, provided those boundaries carry an appropriable lead prize. It also predicts that if a frontier capability is easy to distill immediately, investment should tilt away from that capability unless its task-side boundary value is very large.

This empirical discipline is demanding, but it is sharper than asking whether AI quality went up. The theory is about the direction of improvement, the region in which the improvement creates value, and the product-market structure that determines whether frontier labs can appropriate that value.

8 Conclusion

The economically relevant state of AI is not a scalar quality index. It is a frontier in task space. Tasks differ in horizon and forgiveness. Boundary values price movement of the horizon boundary, the low-forgiveness boundary, and the corner where both bind. A project is valuable where it moves the scarce boundary.

That static logic already explains why project rankings reverse across task regions and why cheap followers can win in saturated tasks while frontier premia persist elsewhere. The dynamic logic explains why frontier labs keep moving. When cheap products erode rents in already served tasks, foundational labs have stronger incentives to define the next valuable task. But if that next task is easily distilled, the private prize falls. The direction of frontier investment is therefore shaped by both task scarcity and appropriability.

This is a useful way to interpret current AI competition. Low-cost products and frontier investment are not opposites. Cheap products fill in the interior of the task frontier and reduce rents there. Frontier labs then race over the boundaries: longer tasks, less forgiving tasks, and task structures where reliability, verification, monitoring, and workflow integration remain scarce. Compute, tools, inference, distillation, and verification should therefore be evaluated not by their labels, but by which task boundary they move and how much of the resulting value can be appropriated before cheap followers arrive.

A Proofs and Derivations

Sequential fragility threshold

Let $\pi = G_F(q_F - \varphi)$. With ℓ retries, a stage fails only if all $1+\ell$ attempts fail, so $S_\ell(q_F, \varphi) = 1 - (1-\pi)^{1+\ell}$. The requirement $S_\ell(q_F, \varphi)^n \geq 1 - \varepsilon$ is equivalent to

$$1 - (1-\pi)^{1+\ell} \geq (1-\varepsilon)^{1/n},$$

and hence to

$$\pi \geq 1 - \{1 - (1-\varepsilon)^{1/n}\}^{1/(1+\ell)}.$$

Since G_F is increasing, this is equivalent to $q_F \geq f(n, \ell, \varepsilon, \varphi)$. The monotonicities follow from the threshold: more stages and a lower tolerated failure rate raise the required local success probability; more retries lower it; φ enters additively.

Boundary value identities

Differentiating (2) under the integral sign gives

$$\frac{\partial W_R(q)}{\partial q_H} = B_H^R(q), \qquad \frac{\partial W_R(q)}{\partial q_F} = B_F^R(q).$$

Differentiating B_H^R in q_F, and B_F^R in q_H, gives the same expression $C^R(q)$. For (4), move first in the horizon direction and then in the fragility direction:

$$W_R(q_H + \Delta_H, q_F) - W_R(q_H, q_F) = \int_0^{\Delta_H} B_H^R(q_H + u, q_F)\, du,$$

$$W_R(q_H + \Delta_H, q_F + \Delta_F) - W_R(q_H + \Delta_H, q_F) = \int_0^{\Delta_F} B_F^R(q_H + \Delta_H, q_F + v)\, dv.$$

Adding yields the finite-step expression.

Proof of Proposition 1. By differentiability, the first-order task-side frontier value of project a

in region R is $B_H^R(q)d_{aH} + B_F^R(q)d_{aF}$. Dividing by costs gives the first comparison. Under the cost-normalized intensity ordering, move the F-terms to the right side and divide by the positive term $B_F^R(q)(d_{yH}/K_y - d_{zH}/K_z)$. This yields the threshold condition. A reversal across two positive-boundary regions therefore requires the two regional scarcity ratios to straddle the common threshold. In the scalar benchmark, the local value difference is $\tilde{B}^R(\bar{q})(\gamma_y/K_y - \gamma_z/K_z)$, whose sign is common across positive-slope regions. Hence scalar quality cannot generate the reversal without project-region-specific wedges. □

Derivation of (6). Let B be the $|\mathcal{R}| \times 2$ matrix whose row is $b_R(q)$, and let V be the $|\mathcal{P}| \times 2$ matrix whose row is v_a. Then $M = BV^\top$. Expanding a two-by-two determinant gives (6). A scalar representation has $M_{Ra} = \lambda_R \kappa_a$, an outer product, and therefore rank at most one. □

Proof of Proposition 2. The lower-cost product is preferred if and only if

$$W_R(q^M) - c_M(R) \geq W_R(q^L) - c_L(R),$$

which rearranges to $\kappa_R \geq W_R(q^L) - W_R(q^M)$. Since $q^L = q^M + \Delta$, applying (4) to the regional surplus gives the integral expression. Bounding $B_H^R + B_F^R$ on the rectangle between the two systems gives the final inequality. □

Proof of Proposition 3. If the other lab invests, investing yields $\sigma(1 - \alpha_N)P_N - K_N$, while staying yields $\pi_S(\delta_S)$. If the displayed dominance condition holds, investing is a best response to investment. Because $\sigma < 1$, the payoff from being the sole investor is weakly larger, so investing is also a best response to staying. Hence both invest. If the middle inequalities hold, investing is a best response to staying, but staying is a best response to investment. The two asymmetric preemption outcomes are therefore the pure-strategy equilibria. If the no-investment inequality holds, investing is not a best response even when the other lab stays, so neither lab invests. Since π_S is decreasing in δ_S, higher distiller pressure relaxes the investment inequalities. Higher α_N lowers the appropriable prize and tightens them. □

Proof of Proposition 4. The first statement is the direct comparison of the private values in (7). Holding P_{aR} and K_a fixed, the derivative of Ψ_{aR} with respect to α_R is $-P_{aR} < 0$. Holding frontier-task distillability fixed, higher distiller pressure in the saturated region lowers $\pi_S(\delta_S)$, so it raises the payoff gain from choosing any frontier project over remaining in S. □

References

[Acemoglu and Autor(2011)] Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics*, Vol. 4B, 1043–1171. Elsevier.

[Acemoglu and Restrepo(2018)] Acemoglu, Daron, and Pascual Restrepo. 2018. "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment." *American Economic Review* 108(6): 1488–1542.

[Aghion et al.(2001)] Aghion, Philippe, Christopher Harris, Peter Howitt, and John Vickers. 2001. "Competition, Imitation and Growth with Step-by-Step Innovation." *Review of Economic Studies* 68(3): 467–492.

[Gilbert and Newbery(1982)] Gilbert, Richard J., and David M. G. Newbery. 1982. "Preemptive Patenting and the Persistence of Monopoly." *American Economic Review* 72(3): 514–526.

[Hinton et al.(2015)] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. "Distilling the Knowledge in a Neural Network." arXiv:1503.02531.

[Hsieh et al.(2023)] Hsieh, Cheng-Yu, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. "Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes." arXiv:2305.02301.

[Jimenez et al.(2023)] Jimenez, Carlos E., John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. "SWE-bench: Can Language Models Resolve Real-World GitHub Issues?" arXiv:2310.06770.

[Schick et al.(2023)] Schick, Timo, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. "Toolformer: Language Models Can Teach Themselves to Use Tools." arXiv:2302.04761.

[Snell et al.(2024)] Snell, Charlie, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. "Scaling LLM Test-Time Compute Optimally Can Be More Effective than Scaling Model Parameters." arXiv:2408.03314.

[Wang et al.(2022)] Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." arXiv:2203.11171.

[Wei et al.(2022)] Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." arXiv:2201.11903.

[Yang et al.(2024)] Yang, John, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. "SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering." arXiv:2405.15793.

[Yao et al.(2022)] Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. "ReAct: Synergizing Reasoning and Acting in Language Models." arXiv:2210.03629.