Agentic Racing
Task Horizons, Temporary Leads, and Compute Rents

June 2026

Abstract

This paper develops a theory of frontier AI racing in which the innovation prize is not an exogenous rank prize. It is the value of newly feasible task horizons. A task has log horizon h, capability is q, and completion probability is a shifted reliability curve $G(q-h)$. In a task economy with value density $a_t(h)$, the marginal value of capability is the boundary statistic

$$B_t(q) = \int a_t(h) g(q-h)\, dh, \qquad g = G'.$$

Thus frontier value is local in task space: it depends on value density near the current reliability boundary, not on average benchmark performance. Bounded task menus saturate, while moving agentic task frontiers need not. I then embed this statistic in an innovation-race model. A successful frontier step creates a temporary lead that decays at rate μ. The winner obtains appropriable flow value and the loser suffers defensive displacement. Both are proportional to the same task-horizon step value, so the private race prize is $(\pi+\beta)S_t(q,\Delta)/(r+\mu)$. Labs choose race effort in a contest for the lead, while compute is a common input with price $R(X)$ increasing in aggregate frontier demand X. In symmetric equilibrium, race intensity is governed by $B_t(q)$, scaling progress $e_t q'(C_t)$, lead persistence, business stealing, and compute-rent elasticity. Relative to a joint-profit benchmark, equilibrium over-races when the contest share-shifting value of the lead exceeds average compute cost at the joint optimum. Relative to a social planner, the sign also depends on consumer surplus, spillovers, serving crowd-out, and safety externalities. The contribution is a sufficient-statistic theory: AI race pressure is high when valuable tasks are dense near the agentic-time boundary and temporary leads make that boundary privately appropriable or defensively important.

JEL: L13, O31, O33, L40, D74.

1 Introduction

Frontier AI competition is often described as a race over model quality. That description hides the central economic margin. Frontier systems are becoming able to complete longer tasks: a short answer, a memo, a code change, a multi-hour debugging assignment, a multi-day software project, or a longer chain of tool-using work. The relevant frontier is an agentic-time frontier.

This paper builds a theory around that frontier. A task has duration or effective step length n, and I write its log horizon as $h = \log n$. A model has capability q. Capability is the location of a reliability frontier over task horizons. Tasks far behind the frontier are already feasible. Tasks far beyond the frontier are still too unreliable. The marginal economic value of improvement comes from tasks close to the boundary.

The first contribution is a sufficient-statistic result. If $a_t(h)$ is the economic value density of tasks at horizon h, and completion probability is a shifted reliability curve $G(q-h)$, then

$$\frac{\partial}{\partial q}\int a_t(h)G(q-h)\,dh = \int a_t(h)g(q-h)\,dh, \qquad g = G'.$$

This boundary statistic is the value density of tasks near the current agentic-time frontier, smoothed by the reliability kernel. It is not the average value of all tasks and not a generic benchmark score. If the task menu has bounded support, marginal capability value eventually vanishes. If valuable task horizons keep appearing near the moving frontier, marginal value can remain positive.

The second contribution is to derive the race prize from this boundary statistic. The current frontier step has value

$$S_t(q,\Delta) = W_t(q+\Delta) - W_t(q).$$

If a lab reaches the step first, it has a temporary lead. During the lead, the winner earns appropriable flow value from newly feasible tasks and the loser suffers defensive displacement through customer adoption, integration, data feedback, standards, or follow-on option value. If the lead decays at rate μ, the productive lead value and defensive loss are

$$L_t = \frac{\pi S_t(q,\Delta)}{r+\mu}, \qquad D_t = \frac{\beta S_t(q,\Delta)}{r+\mu}.$$

Thus the private race prize is not an arbitrary rank prize. It is

$$P_t = L_t + D_t = \frac{\pi+\beta}{r+\mu}S_t(q,\Delta),$$

and locally it is proportional to the boundary statistic $B_t(q)$.

The third contribution is an equilibrium race model with compute as a common input. Two labs choose race effort x_i. Aggregate frontier demand is $X = x_1 + x_2$. The probability that the frontier step is completed is $\Lambda(X)$, and conditional on completion lab i wins the lead with contest share x_i/X. Compute has rental price $R(X)$, increasing in aggregate frontier demand.

Lab i's cost is $x_i R(X)$. In a symmetric interior equilibrium with $X = 2x$,

$$\frac{1}{2}\Lambda'(X)(L_t - D_t) + \frac{\Lambda(X)}{2X}(L_t + D_t) = R(X) + \frac{X}{2}R'(X).$$

The first term is the private speed effect. The second is the contest share-shifting effect. The right side is the private marginal compute cost. Substituting the lead values shows that all task and scaling primitives enter through

$$S_t(q, \Delta) \simeq e_t q'(C_t)\varepsilon B_t(q).$$

This is the theory's main race statistic.

The final contribution is a welfare comparison. A joint-profit planner for the two labs chooses X to maximize

$$\Lambda(X)(L_t - D_t) - XR(X).$$

At an interior joint optimum X^J, the decentralized symmetric race has locally excessive aggregate effort if and only if

$$\frac{\Lambda(X^J)}{X^J}(L_t + D_t) > R(X^J).$$

This condition is interpretable. Equilibrium racing is pushed above the joint-profit benchmark when the private contest value of shifting the lead is large relative to average compute cost. Defensive displacement β raises this share-shifting motive even when it does not raise joint lab surplus. Compute scarcity also matters because each lab internalizes only its own marginal effect on the common input price.

The paper therefore differs from a generic innovation-race model. The prize is not free-standing. The same boundary task value creates productive lead rents, defensive racing motives, and the measurable object that should be compared to compute rents. The model also avoids the claim that racing is always socially wasteful. Relative to social welfare, racing can be excessive or insufficient depending on consumer surplus, spillovers, serving crowd-out, and safety externalities. The theory identifies the sufficient statistics that determine the sign.

2 Task Horizons and Boundary Value

A task consists of $n > 0$ effective steps. Let $h = \log n$ denote the log task horizon. A model with capability q completes the task with probability

$$p(q, h) = G(q - h),$$

where $G : \mathbb{R} \to [0, 1]$ is continuously differentiable, increasing, and has derivative $g = G'$. The shifted form says that capability moves the reliability frontier in log task-horizon space.

The Poisson reliability case used in time-horizon discussions is

$$p(q,h) = \exp[-\exp(h-q)]. \tag{1}$$

It is the shifted curve $G(s) = \exp[-\exp(-s)]$, with density

$$g(s) = \exp[-s - \exp(-s)].$$

The 50 percent completion horizon solves $G(q-h) = 1/2$, so $h = q - \log(\log 2)$. Capability is therefore a log task-horizon shifter, up to a constant.

Let $a_t(h) \geq 0$ be the economic value density of tasks with log horizon h at date t. This is the task-based primitive of the paper. In an empirical implementation it should be built from economically valuable work, not from exam items alone. For an interval H, define task surplus

$$W_t(q;H) = \int_H a_t(h)G(q-h)\,dh. \tag{2}$$

Theorem 1 (Boundary sufficient statistic)**.** *Let H be an interval. Suppose that g is bounded, $\int_{-\infty}^{\infty} g(s)ds = 1$, and for every compact set $Q \subset \mathbb{R}$,*

$$\int_H a_t(h) \sup_{q\in Q}\{G(q-h) + g(q-h)\}\,dh < \infty.$$

Then $W_t(q;H)$ is differentiable and

$$W_t'(q;H) = B_t(q;H) \equiv \int_H a_t(h)g(q-h)\,dh. \tag{3}$$

If $H = (-\infty, \bar{h}]$, $A_H = \int_H a_t(h)dh < \infty$, and $g(s) \to 0$ as $s \to \infty$, then

$$B_t(q;H) \to 0 \quad \text{as } q \to \infty.$$

If, for a given q, there exist $w > 0$ and $a_0 > 0$ such that $[q-w, q+w] \subset H$, $a_t(h) \geq a_0$ on that interval, and $g(s) > 0$ on $[-w, w]$, then

$$B_t(q;H) \geq a_0 \int_{-w}^{w} g(s)ds > 0.$$

The theorem is the paper's first substantive object. It says that the marginal value of capability is the value density of tasks near the current reliability boundary. Bounded menus of limited-step tasks saturate. Expanding menus of valuable agentic tasks need not.

Remark 1 (Why the boundary statistic is AI-specific)**.** *In standard innovation-race models, the prize is usually exogenous or summarized by a scalar market size. Here the prize is en-*

dogenous to task horizons. A capability step is valuable when it moves the completion boundary through dense economic task mass. This makes the location of task value in horizon space a state variable for competition.

3 Scaling and Frontier Steps

The task-horizon theorem treats q as the capability index. Scaling laws map training compute into local movements in q. Let cumulative effective compute be C, and suppose loss follows

$$L(C) = L_\infty + \chi(C+\kappa)^{-\eta}, \qquad L_\infty, \chi, \kappa, \eta > 0, \tag{4}$$

with

$$q(C) = -\log L(C). \tag{5}$$

Let $y = C + \kappa$ and $b = L_\infty/\chi$. Then

$$q'(C) = \frac{\eta}{y(1+by^\eta)} > 0, \qquad q''(C) = -\frac{\eta[1+(1+\eta)by^\eta]}{y^2(1+by^\eta)^2} < 0. \tag{6}$$

A physical training project of size ε creates $e_t\varepsilon$ units of effective compute. The exact capability step is

$$\Delta q_t(C,\varepsilon) = q(C + e_t\varepsilon) - q(C),$$

and locally

$$\Delta q_t(C,\varepsilon) = e_t q'(C)\varepsilon + O(\varepsilon^2). \tag{7}$$

The exact task-surplus value of a frontier step is

$$S_t(q,\Delta;H) = W_t(q+\Delta;H) - W_t(q;H). \tag{8}$$

For a small step,

$$S_t(q,\Delta;H) = \Delta B_t(q;H) + O(\Delta^2). \tag{9}$$

Combining (7) and (9), the local value of a physical training project is

$$S_t(C,\varepsilon;H) \simeq e_t q'(C)\varepsilon B_t(q(C);H). \tag{10}$$

This is the race numerator. It is high when scaling progress is steep and when there is valuable task density at the current agentic-time boundary.

4 Temporary Leads and Defensive Value

The model now derives the private race prize. A successful frontier step moves capability from q to $q + \Delta$, opening task-surplus value $S_t(q, \Delta)$. The first lab to reach the step receives a temporary lead. The lead ends at an exponentially distributed time τ with rate $\mu > 0$, representing catch-up, diffusion, imitation, open-source replication, worker mobility, or independent rediscovery. The discount rate is $r > 0$.

During the lead, the winner can appropriate flow share $\pi \geq 0$ of the newly feasible task surplus. The loser suffers flow defensive displacement $\beta \geq 0$ times the same task-surplus band. The parameter β captures customer lock-in, data feedback, integration loss, standards displacement, and follow-on option value. It is not a free-standing harm parameter; it loads on the same boundary task value.

Lemma 1 (Temporary lead value)**.** *Let $S = S_t(q, \Delta)$. If the lead duration τ is exponential with rate μ, then the winner's productive lead value and the loser's defensive loss are*

$$L(q, \Delta) = \frac{\pi S}{r + \mu}, \qquad D(q, \Delta) = \frac{\beta S}{r + \mu}. \tag{11}$$

The private prize from shifting the lead from the rival to the lab is

$$P(q, \Delta) = L(q, \Delta) + D(q, \Delta) = \frac{\pi + \beta}{r + \mu} S_t(q, \Delta). \tag{12}$$

Locally,

$$P(C, \varepsilon) \simeq \frac{\pi + \beta}{r + \mu} e_t q'(C) \varepsilon B_t(q(C)). \tag{13}$$

This lemma is the main conceptual difference from a generic race model. Productive innovation value and defensive race pressure are generated by the same task-horizon boundary. If $\beta = 0$, racing is only about appropriating productive flow. If $\beta > 0$, falling behind makes the same task band strategically important even when the step is not very profitable in joint terms.

5 Equilibrium Race Intensity

Two symmetric labs choose race effort $x_i \geq 0$. Let $X = x_1 + x_2$. The probability that the frontier step is completed in the relevant race window is $\Lambda(X)$, where $\Lambda(0) = 0$, $\Lambda' > 0$, and $\Lambda'' \leq 0$. Conditional on completion and $X > 0$, lab i wins the temporary lead with contest share

$$\sigma_i(x_i, x_j) = \frac{x_i}{X}.$$

Compute is a common input. The rental price of frontier compute is $R(X)$, with $R(X) > 0$ and $R'(X) \geq 0$. Lab i's compute cost is $x_i R(X)$. Given $L = L(q, \Delta)$ and $D = D(q, \Delta)$, lab i's expected payoff is

$$U_i(x_i, x_j) = \Lambda(X)\{\sigma_i L - (1-\sigma_i)D\} - x_i R(X). \tag{14}$$

Proposition 1 (Symmetric race condition)**.** *Any interior symmetric equilibrium $x_i = x_j = x > 0$, with $X = 2x$, satisfies*

$$\frac{1}{2}\Lambda'(2x)(L-D) + \frac{\Lambda(2x)}{4x}(L+D) = R(2x) + xR'(2x). \tag{15}$$

Equivalently, in aggregate effort X,

$$\frac{1}{2}\Lambda'(X)(L-D) + \frac{\Lambda(X)}{2X}(L+D) = R(X) + \frac{X}{2}R'(X). \tag{16}$$

Using (11), the private marginal race value is

$$\frac{S_t(q,\Delta)}{r+\mu}\left[\frac{\pi-\beta}{2}\Lambda'(X) + \frac{\pi+\beta}{2X}\Lambda(X)\right]. \tag{17}$$

Locally, $S_t(q,\Delta) \simeq e_t q'(C_t)\varepsilon B_t(q_t)$. Hence task values and scaling laws affect equilibrium race intensity only through the boundary statistic $B_t(q_t)$ and the local scaling step $e_t q'(C_t)\varepsilon$.

The proposition separates two forces. The speed term $\Lambda'(X)(L-D)/2$ is valuable when completing the step sooner has positive net joint value for the two labs. The share-shifting term $\Lambda(X)(L+D)/(2X)$ is valuable because unilateral effort increases the probability that the lab, rather than its rival, receives the lead. Defensive displacement β raises the share-shifting term even though it lowers the net joint value $L-D$.

Corollary 1 (Boundary-density comparative statics)**.** *Suppose the interior symmetric equilibrium is unique, the private marginal value in (16) is decreasing in X, and the private marginal compute cost is increasing in X. If the bracketed term in (17) is positive, equilibrium aggregate race effort is increasing in the boundary task statistic $B_t(q_t)$, scaling efficiency e_t, the local scaling slope $q'(C_t)$, and lead persistence $1/(r+\mu)$. It is decreasing under an upward shift in the compute-rent schedule R.*

The corollary gives the paper's main prediction. Frontier racing should be intense when valuable tasks are dense near the current agentic-time frontier, when scaling still moves that frontier, when temporary leads are persistent, and when compute rents are not too high.

6 Joint-Profit and Social Benchmarks

The decentralized equilibrium is not automatically inefficient. It must be compared to a benchmark. Start with the joint-profit benchmark for the two labs. Since one lab's defensive loss is the other lab's competitive advantage, the total lab payoff from completing the step is $L - D$. A joint-profit planner choosing aggregate effort X solves

$$\max_{X \geq 0} \Lambda(X)(L - D) - XR(X). \tag{18}$$

An interior joint optimum X^J satisfies

$$\Lambda'(X^J)(L - D) = R(X^J) + X^J R'(X^J). \tag{19}$$

Proposition 2 (Private over-racing relative to joint profit)**.** *Assume the symmetric private equilibrium and the joint-profit solution are interior and locally unique. At an interior joint optimum X^J, decentralized symmetric effort is locally higher than the joint-profit benchmark if and only if*

$$\frac{\Lambda(X^J)}{X^J}(L + D) > R(X^J). \tag{20}$$

Using temporary lead values, this is

$$\frac{\Lambda(X^J)}{X^J} \frac{\pi + \beta}{r + \mu} S_t(q, \Delta) > R(X^J). \tag{21}$$

Locally, $S_t(q, \Delta) \simeq e_t q'(C_t) \varepsilon B_t(q_t)$.

The over-racing condition is the sharper version of the simple dissipative race region. It says that the private contest value of reallocating the lead must be high relative to average compute cost at the joint optimum. The condition becomes easier to satisfy when the boundary task band is valuable, when defensive displacement is high, or when leads last longer. Compute scarcity enters through $R(X)$ and through the joint optimum X^J. In the private first-order condition each lab internalizes only $XR'(X)/2$, while the joint planner internalizes $XR'(X)$.

A social planner may disagree with the joint-profit planner. Let $S_t^W(q, \Delta)$ denote the total social surplus from the frontier step, including consumer surplus and nonappropriated productivity gains. Let $\Psi(X)$ denote external costs of aggregate frontier compute, including serving crowd-out, grid constraints, safety externalities, or other costs not priced into $R(X)$. The social planner solves

$$\max_{X \geq 0} \Lambda(X) S_t^W(q, \Delta) - XR(X) - \Psi(X), \tag{22}$$

with interior condition

$$\Lambda'(X^S) S_t^W(q, \Delta) = R(X^S) + X^S R'(X^S) + \Psi'(X^S). \tag{23}$$

Thus a private AI race can be socially excessive or socially insufficient. Defensive displacement and unpriced compute externalities push toward excessive racing. Consumer surplus, knowledge spillovers, and low appropriability push toward insufficient racing. In all cases, the task-horizon boundary statistic remains the object that maps capability into frontier value.

7 Measurement Implications

The theory is a theory paper, but it is disciplined by measurable sufficient statistics. The empirical object is not average benchmark performance. It is

$$B_t(q_t) = \int a_t(h) g(q_t - h) dh.$$

Time-horizon evaluations discipline q_t: they estimate the human-time duration of tasks an agent can complete at a fixed success probability. GDP-style task evaluations discipline $a_t(h)$: they identify economically valuable deliverables and attach labor time, cost, and quality comparisons. Scaling studies discipline $q'(C_t)$ and e_t. Compute markets discipline $R(X)$ and $R'(X)$. Lead half-lives discipline μ. Product-market evidence disciplines π and β.

The model implies a break-even exercise. For a candidate frontier step, estimate or bound

$$e_t q'(C_t) \varepsilon B_t(q_t), \qquad \frac{\pi + \beta}{r + \mu}, \qquad R(X),\ R'(X),\ \Lambda(X).$$

These objects determine decentralized race intensity through (16), joint over-racing through (21), and social racing through (23). Public data alone usually cannot sign all wedges. But the theory says exactly which private primitives are missing and how they enter.

8 Conclusion

The paper's central claim is that frontier AI racing is a task-horizon race. Capability moves a reliability frontier over economically valuable task horizons. The marginal value of a capability step is therefore local to the boundary. Temporary leads convert the boundary task band into productive rents and defensive losses. Compute scarcity converts aggregate race intensity into a common-input externality.

This structure gives a theory contribution beyond a reduced-form race matrix. The innovation prize is derived from a task economy; defensive racing is derived from temporary lead dynamics; and equilibrium effort is characterized by a contest with common compute rents. The same sufficient statistic, $B_t(q_t)$, governs productive value, defensive pressure, and the empirical break-even exercise.

A Proofs

Proof of Theorem 1. Differentiate (2) with respect to q:

$$\frac{\partial}{\partial q}G(q-h) = g(q-h).$$

The domination assumption permits differentiation under the integral sign and gives (3). If $H = (-\infty, \bar{h}]$, then for every fixed $h \le \bar{h}$, $q - h \to \infty$ as $q \to \infty$, so $g(q-h) \to 0$. Since g is bounded and a_t is integrable on H, dominated convergence implies $B_t(q; H) \to 0$. If $a_t(h) \ge a_0$ on $[q-w, q+w]$, then

$$B_t(q; H) \ge a_0 \int_{q-w}^{q+w} g(q-h)dh.$$

With $s = q - h$, this equals

$$a_0 \int_{-w}^{w} g(s)ds > 0.$$

□

Derivation of (6). Write $y = C + \kappa$ and $b = L_\infty/\chi$. Then

$$L(C) = \chi y^{-\eta}(1 + by^{\eta})$$

and

$$q(C) = -\log\chi + \eta\log y - \log(1 + by^{\eta}).$$

Differentiating gives

$$q'(C) = \frac{\eta}{y} - \frac{b\eta y^{\eta-1}}{1+by^{\eta}} = \frac{\eta}{y(1+by^{\eta})}.$$

Differentiating again gives

$$q''(C) = -\frac{\eta[1 + (1+\eta)by^{\eta}]}{y^2(1+by^{\eta})^2} < 0.$$

□

Proof of Lemma 1. For an exponential lead duration τ,

$$\mathbb{E}\left[\int_0^{\tau} e^{-rt}dt\right] = \int_0^{\infty} e^{-rt}\Pr(\tau > t)dt = \int_0^{\infty} e^{-(r+\mu)t}dt = \frac{1}{r+\mu}.$$

Multiplying this duration factor by the winner's flow payoff πS and the loser's flow loss βS gives (11). The private prize from increasing the probability that the lab wins rather than loses is $L + D$, which gives (12). The local approximation follows from (10). □

Proof of Proposition 1. For $X > 0$,

$$U_i = \Lambda(X)\{\sigma_i(L+D) - D\} - x_i R(X).$$

The derivative of the contest share is

$$\frac{\partial \sigma_i}{\partial x_i} = \frac{x_j}{(x_i + x_j)^2}.$$

At a symmetric profile $x_i = x_j = x$, $X = 2x$, $\sigma_i = 1/2$, and $\partial\sigma_i/\partial x_i = 1/(4x) = 1/(2X)$. Therefore

$$\frac{\partial U_i}{\partial x_i} = \Lambda'(X)\frac{L-D}{2} + \Lambda(X)\frac{L+D}{2X} - R(X) - xR'(X).$$

Setting this equal to zero gives (15) and (16). Substituting (11) gives (17). Substituting the local frontier value (10) gives the final statement. □

Proof of Corollary 1. Under the stated monotonicity and uniqueness assumptions, the equilibrium aggregate effort X^* is defined by equality between a decreasing private marginal value schedule and an increasing private marginal cost schedule. If the bracketed term in (17) is positive, increasing $B_t(q_t)$, e_t, $q'(C_t)$, or $1/(r+\mu)$ shifts the private marginal value schedule upward and therefore raises X^*. An upward shift in R shifts the private marginal cost schedule upward and lowers X^*. □

Proof of Proposition 2. The joint-profit first-order condition is (19). Evaluate the decentralized private marginal payoff (16) at X^J. Using $\Lambda'(X^J)(L-D) = R(X^J) + X^J R'(X^J)$,

$$\begin{aligned}
M^P(X^J) &= \frac{1}{2}\Lambda'(X^J)(L-D) + \frac{\Lambda(X^J)}{2X^J}(L+D) - R(X^J) - \frac{X^J}{2}R'(X^J) \\
&= \frac{\Lambda(X^J)}{2X^J}(L+D) - \frac{1}{2}R(X^J).
\end{aligned}$$

The decentralized equilibrium lies locally above the joint-profit benchmark exactly when this private marginal payoff is positive at X^J. This is equivalent to (20). Substituting (11) gives (21), and the local boundary approximation follows from (9) and (7). □

References

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

[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.

[Acemoglu and Restrepo(2019)] Acemoglu, Daron, and Pascual Restrepo. 2019. "Automation and New Tasks: How Technology Displaces and Reinstates Labor." *Journal of Economic Perspectives* 33(2): 3–30.

[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.

[Bresnahan and Trajtenberg(1995)] Bresnahan, Timothy F., and Manuel Trajtenberg. 1995. "General Purpose Technologies: Engines of Growth?" *Journal of Econometrics* 65(1): 83–108.

[Hoffmann et al.(2022)] Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, John Rae, Oriol Vinyals, and Laurent Sifre. 2022. "Training Compute-Optimal Large Language Models." arXiv:2203.15556.

[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.

[Kaplan et al.(2020)] Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." arXiv:2001.08361.

[Kwa et al.(2025)] Kwa, Thomas, Ben West, Joel Becker, Ann Deng, Kiara Garcia, Megan Kinniment, Nick Rush, Sevon Von Arx, Ryan Bloom, Tobias Broadley, Huy Du, Ben Goodrich, Nikola Jurkovic, Lawrence Miles, Sawyer Nix, Tony Lin, Nikhil Parikh, David Rein, Lucas Sato, Hanna Wijk, Daniel Ziegler, Elizabeth Barnes, and Lawrence Chan. 2025. "Measuring AI Ability to Complete Long Software Tasks." arXiv:2503.14499.

[OpenAI(2025)] OpenAI. 2025. "GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks." arXiv:2510.04374.

[Rein et al.(2025)] Rein, David, Joel Becker, Amy Deng, Seraphina Nix, Chris Canal, Daniel O'Connel, Pip Arnott, Ryan Bloom, Thomas Broadley, Katharyn Garcia, Brian Goodrich, Max Hasin, Sami Jawhar, Megan Kinniment, Thomas Kwa, Aron Lajko, Nate Rush, Lucas Sato, Sydney Von Arx, Ben West, Lawrence Chan, and Elizabeth Barnes. 2025. "HCAST: Human-Calibrated Autonomy Software Tasks." arXiv:2503.17354.

[Wijk et al.(2024)] Wijk, Hanna, Tony Lin, Joel Becker, Kiara Garcia, Megan Kinniment, Lucas Sato, and coauthors. 2024. "RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents." arXiv:2411.15114.