

The Catch-Up Gradient

Why the AI frontier is getting harder to extend and easier to approach, and what that is worth to the players behind it

WORKING PAPER · CHAPTERS 1 AND 2

SEPTEMBER 2026

SUMMARY

Two costs in AI are moving in opposite directions. The cost of extending the frontier has grown by roughly $2.4\times$ a year for most of a decade, and the largest laboratories now spend at the scale of national infrastructure to stay in front. The cost of reaching a capability level the frontier demonstrated a year or two earlier is falling by roughly $3\times$ a year, faster than the frontier is getting dearer. Because both curves move, every year of lag buys a larger discount. On public estimates, matching a two-year-old frontier now costs one to two percent of what leading costs.

The mechanism is unusual. In earlier technology races the leaders' products diffused to followers. In this race the product is also a means of production: released weights write code, generate training data, grade experiments, and teach smaller models. A follower inherits not only the leaders' recipes but a share of their labor.

This paper argues three things. First, the gradient is real, measurable, and steep (Chapter 1). Second, it is a window rather than a law: it depends on continued open-weight releases, on a liquid market for rented compute, and on followers choosing directions the leaders are not optimizing. Third, because everything that diffuses is by construction cheap, durable value in this regime sits in what does not diffuse: an organization's experimental record, its environments and evaluations, its research velocity, and its model lineage (Chapter 2). Those assets can only be built by participating, and they are priced in years rather than dollars.

The practical conclusion is a sized bet. A research program at the trailing edge of the frontier now costs on the order of ten to thirty million dollars a year, roughly one percent of the cost of leading, and buys ownership of an improvement process rather than a subscription to someone else's.

Thesis. The frontier gets harder to extend and easier to approach, and the gap between those two costs widens every year. Followers inherit the models, methods, tools, data-generation techniques, and increasingly the research labor that the leaders create. The scarce asset therefore shifts from possessing intelligence to owning a process that improves it.

The mechanism

1.1 Two curves, not one race

Conventional discussion treats frontier AI as a single race. The laboratory with the most compute, data, and talent advances the state of the art, and everyone else falls progressively farther behind. That description is accurate for one curve and silent about the second.

Table 1. Two different problems

DIMENSION	LEADER	FOLLOWER
Objective	Move the capability frontier forward	Reach a capability level already demonstrated
Economics	Pay for discovery: the failed runs, the dead ends, the scale needed to find out what works next	Exploit accumulated discoveries, published recipes, and trained teachers
What drives cost	Compute at the next scale, multiplied by exploration overhead	Compute at the old scale, divided by everything learned since
Trend	Rising about $2.4\times$ a year	Falling about $3\times$ a year

The asymmetry is simple to state. A leader must discover the path from capability 100 to 110, and discovery is paid for in failed runs at frontier scale. A follower can later reproduce most of capability 100 without replaying the cost of invention, and each year the reproduction gets cheaper. The frontier can therefore become more expensive at the same time that yesterday's frontier becomes dramatically cheaper. Figure 1 shows the shape of the claim.

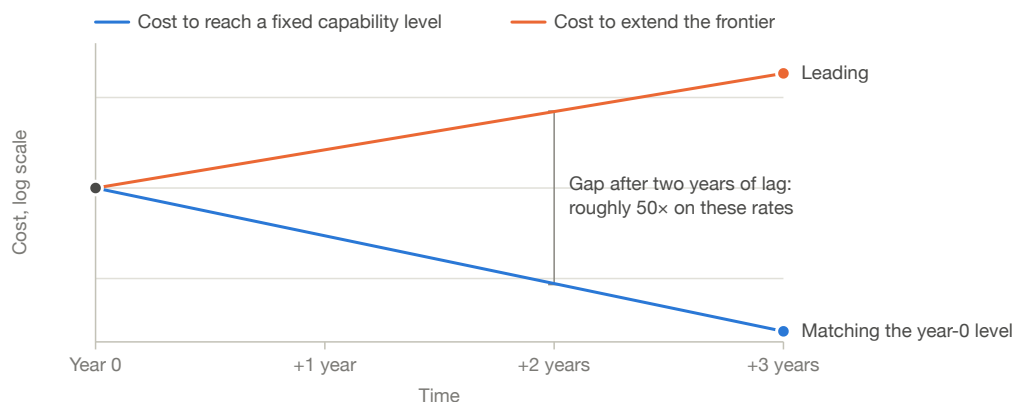


Figure 1. The catch-up gradient. Both curves begin at the same point, because reaching today's frontier today costs what the frontier costs. Leading then costs about $2.4\times$ more each year (Epoch AI's estimate for the largest runs) while matching the year-0 level costs about $3\times$ less (Table 2). The vertical gap is the gradient, roughly $7\times$ wider for each year of lag. Illustrative rates, not measurements.

Table 2. What the two curves look like in public estimates

QUANTITY	ESTIMATE	SOURCE
Cost of the largest frontier training runs	Growing about $2.4\times$ a year since 2016; GPT-4 near \$78M of compute, Gemini Ultra near \$191M; billion-dollar runs projected by 2027	Epoch AI 2024; Stanford AI Index 2024
Compute needed to reach a fixed level of language-model performance	Halving roughly every eight months from algorithmic progress alone (about $2.8\times$ a year)	Ho et al., Epoch AI 2024
Inference price for GPT-3.5-level capability	Down more than $280\times$ between November 2022 and October 2024	Stanford AI Index 2025
Training cost of a roughly GPT-4-class model	Near \$78M (GPT-4, 2022–23) to \$5.6M of rented compute for DeepSeek-V3’s final run (December 2024), which matched closed models released six to nine months earlier	AI Index 2024; DeepSeek-V3 technical report
Rental price of an H100 GPU-hour	From roughly \$4–8 in 2023 to roughly \$2–3 by 2025, with no capital outlay or allocation queue	Public rental-market pricing
Hyperscaler AI capital spending	More than \$300B in 2025 across the four largest US cloud companies	Company guidance, 2025

The DeepSeek-V3 figure covers the final training run only; the report excludes prior research and ablations, which for a first-time recipe are probably several times larger. That exclusion is the point: a follower who inherits the published recipe and the released weights pays something closer to the final-run figure. Algorithmic progress (about $2.8\times$ a year) combined with hardware price-performance (about $1.3\times$ a year) implies the cost of reaching a fixed capability falls about $3.5\times$ a year; this paper uses $3\times$ as a round, conservative figure. The observed decline for GPT-4-class training, roughly $10\times$ over about two years, is consistent with it.

Illustrative arithmetic. If leading costs $2.4\times$ more each year and matching a fixed level costs $3\times$ less, then matching the frontier of one year ago costs about one seventh of leading; the frontier of two years ago, about one fiftieth; the frontier of three years ago, about a quarter of one percent. The discount compounds because both curves move. A follower does not have to be clever to benefit from it. It only has to be late.

1.2 Why the follower's curve falls faster

Someone creating a given generation of model for the first time must discover thousands of interacting choices: architecture, data mixture, curriculum, optimization, post-training, tool use, evaluation, serving, and failure recovery. A later entrant inherits most of those answers. The cost of reaching a fixed capability level at a given date decomposes into four terms, and all four fall with time:

1. **Compute required**, given the best recipes of the day. Algorithmic progress alone halves it roughly every eight months, and a teacher model collapses it further: distilling a demonstrated capability from released weights is far cheaper than discovering it.
2. **Price per unit of compute**. Hardware price-performance improves each year, and the rental market has turned a capital expenditure with a waiting list into an hourly operating expense.
3. **Labor to reproduce the recipe**. Recipes are published, reference implementations are open, and the models themselves now write and debug much of the code.
4. **Cost of failed attempts**. A follower knows which paths worked, so its exploration overhead is a fraction of the leader's.

The second term is shared with the leaders. The other three are specific to followers, and the first is where the subsidy is strongest, because the output of the race is a productive input to the next round. Open-weight models generate training data, write code, design experiments, interpret papers, construct evaluations, diagnose training failures, and act as distillation teachers. Closed frontier systems can also serve as teachers where their terms permit it, which for competing models they generally do not. Followers are therefore not merely copying old papers. They recruit older frontier intelligence as research labor.

The evidence for the teacher effect is now direct. DeepSeek's R1 release in January 2025 shipped distilled variants from 1.5 to 70 billion parameters that inherited much of the reasoning ability of a far larger model. The same month, a Berkeley team trained Sky-T1, a 32-billion-parameter reasoning model, for under \$450 of compute. Weeks later the s1 project matched a leading closed reasoning model on competition mathematics by fine-tuning an open 32-billion-parameter model on 1,000 curated examples for 26 minutes on 16 H100s. These are narrow capabilities, and distillation transfers narrow capabilities most cheaply; broad capability still costs real compute. But the direction and magnitude are unmistakable: once a strong open base exists, the marginal cost of transferring a capability that took a frontier-scale run to demonstrate can be the price of a good dinner.

1.3 The unusual part: the product is the means of production

Technology races have always leaked. Patents expire, engineers move, reverse engineering works. What is different here is what leaks. In earlier races the leaders' products diffused to followers, but the products were not themselves productive inputs to the next round: a competitor's aircraft did

not design your next aircraft, and a rival's chip did not lay out your next chip. A rocket company does not periodically hand outsiders a slightly older team of its best engineers. An AI laboratory effectively does. Through released weights, distillation targets, published methods, and model access, older frontier intelligence becomes part of the follower's production function. The diffusion of the product is the diffusion of labor.

The subsidy is not charity, and it is worth being precise about why it exists, because that determines how long it lasts. Every major supplier of open weights has a strategic motive. Meta commoditizes the model layer that sits beside its advertising business. Chinese laboratories such as DeepSeek, Alibaba's Qwen team, Moonshot, and Zhipu use open releases to win adoption, talent, and standards while insulating themselves from hardware sanctions. OpenAI's 2025 open-weight release under an Apache license courts developers. NVIDIA's open Nemotron models sell GPUs. Mistral and the sovereign programs behind it sell independence from American platforms. Several independent motives make the supply more durable than any one actor's choice. It remains a choice, and Section 1.6 returns to what happens if it changes.

1.4 The research threshold and what actually recurses

Let C^* denote the capability at which a model becomes a competent contributor to AI research. Below C^* , humans supply nearly all of the research judgment. Above it, a model can increasingly propose experiments, implement them, run evaluations, diagnose failures, generate better data, modify training recipes, and produce stronger research agents. For illustration, suppose the strongest open-weight release of a given year sits at or above C^* . The question that matters for a small laboratory then changes from whether it can reproduce the largest frontier training run to whether it can assemble a system whose research capability crosses the threshold needed to improve its own research process.

Three measurements show how close that threshold is. METR finds that the length of software and research tasks frontier models can complete at 50% reliability, measured by how long the same tasks take human professionals, has doubled about every seven months since 2019 and faster in the most recent years. On METR's RE-Bench, a suite of realistic machine-learning research problems, the best agents outscored human experts when both had a two-hour budget, while the humans pulled ahead when given eight hours or more. And in May 2025 Google DeepMind reported that AlphaEvolve, an evolutionary coding agent, had sped up a kernel in Gemini's training stack by 23%, cutting Gemini's training time by about 1%, and had recovered 0.7% of Google's global compute through scheduling improvements. Models are already competent research labor for bounded tasks and increasingly competent agents for tasks lasting hours. Autonomous multi-week research has not been demonstrated as of this writing. The economics change well before it is, because the flywheel does not need a model that runs the laboratory. It needs one that contributes to it.

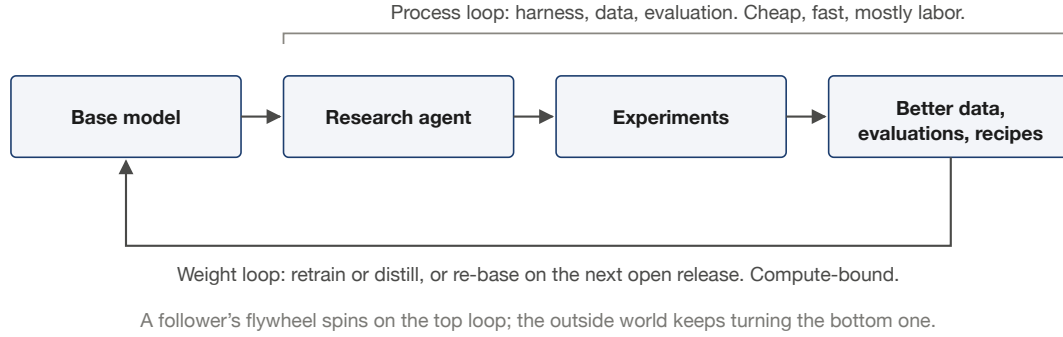


Figure 2. Two loops, priced differently. The top loop improves the research process and is mostly labor, increasingly the model's own. The bottom loop improves the weights and is bound by compute; a follower gets a new turn of it with each open-weight release.

The distinction in Figure 2 matters because the two loops are priced differently. Improving the process is cheap: better harnesses, better synthetic data, better environments, and better evaluations are mostly labor, and the labor is increasingly the model's own. Improving the weights is expensive, and the leaders have the same loop with a thousand times the compute. A follower's flywheel therefore spins on the process loop, and it gets a new turn of the weight loop each time a stronger open model is released. A first iteration that is only 20 to 30% better at research still compounds, because what it produces is data, evaluations, and recipes that make the next iteration better and the next base model more useful. The central economic variable becomes research progress per GPU-hour, not benchmark prestige or consumer-chat quality.

1.5 Where the gradient flattens

The catch-up advantage is not infinite. It is strongest when the target capability has already been demonstrated and weakens as a follower approaches the true frontier. Once the teacher no longer knows the answer and the next step requires novel, expensive experiments, compute and infrastructure are decisive again. Far behind the frontier the advantage is also small, for a different reason: the models there fall below C^* , the capability is already a commodity, and there is nothing left to win. Between those two regions lies the position this paper is about.

Table 3. The follower's advantage by position

POSITION	FOLLOWER ADVANTAGE	WHAT BINDS
At the frontier	Low	Novel discovery, compute at the next scale, infrastructure; the teacher does not know the answer
One or two generations behind	High	Research quality and experiment velocity; teachers are competent, recipes and tools have diffused, directions are still open
Far behind	High but worth little	Integration and reproduction of capabilities that are already commodities; models below C^*

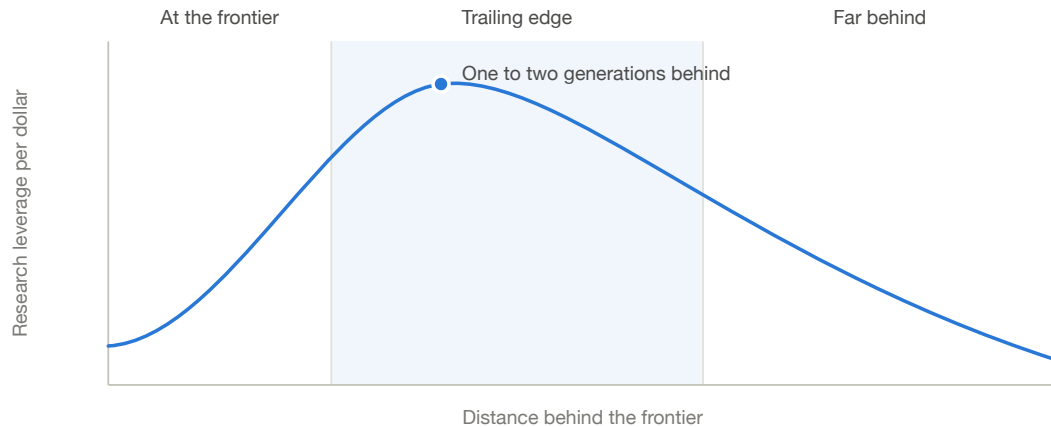


Figure 3. Why the optimum is interior. Leverage is the product of two factors that move in opposite directions: how competent the available teacher is, which falls with distance behind the frontier, and how much of the path someone else has already paid for, which rises with it. At the frontier, discovery cost dominates and the teacher does not know the answer. Far behind, the models sit below C^* and the capability is already a commodity. Between them, teachers are competent, recipes have diffused, and directions are still open. The curve is stylized, and its peak moves forward every time the open-weight frontier does.

1.6 Objections, and what survives them

The leaders get the same flywheel with a thousand times the compute, so the gap widens. True, and irrelevant to the follower's objective. A follower does not need to win the general frontier; it needs to reach a threshold and then choose a direction. The leaders' compute is committed to frontier runs and to serving hundreds of millions of users, and their research agenda is constrained by product roadmaps, inference economics, safety and policy obligations, and the need for broad capability. The follower's advantage is direction, not speed. Chapter 2 makes this precise.

The subsidy is a policy choice and can be withdrawn. Also true. The most prominent Western supplier of open weights signaled in 2025 that its most capable future models might not be released; closed laboratories forbid distillation in their terms; the EU already attaches obligations to models above a compute threshold and its open-source exemptions stop there; export controls could reach models as well as chips. The right reading is that the gradient is structural in mechanism and contingent in supply. It is a window. Two consequences follow: fork what you depend on while the door is open, and note that even a closed door leaves published methods, a rental compute market, and hardware price-performance, which together still push the follower's curve down, just more slowly.

Followers' own discoveries diffuse too, so no lead lasts. Correct, and this is the central design constraint of Chapter 2. FlashAttention came from a Stanford student, rotary position embeddings from a small Chinese company, direct preference optimization from Stanford, Mamba from CMU and Princeton, and the efficiency recipe that repriced the industry in early 2025 (multi-head latent attention, GRPO, FP8 training at scale) from DeepSeek. Most of those inventors captured little of the value, because they published. A participant that wants to capture value must hold assets that do not diffuse. Section 2.2 says what those are.

The followers who caught up were not small. DeepSeek trained V3 on a cluster of about 2,000 H800s while the leaders were building sites of 100,000 GPUs, a compute disadvantage near $50\times$ at the run level, but it was backed by a quantitative hedge fund and years of accumulated infrastructure. Moonshot and Mistral raised well over a billion dollars each. Contesting the general frontier from behind is still a game for the hundreds of millions. The game this paper is about is different: a fixed open base, a research loop, and a chosen direction, at roughly a hundredth of that price. Section 2.8 sizes it.

Core claim. The AI race has a built-in catch-up mechanic: the frontier gets harder to extend and easier to approach, and the gap widens every year. Once trailing models become competent research labor, the mechanic feeds itself, because the thing being diffused is the thing that does the catching up. The mechanism is structural; the supply of open weights that carries it is a strategic choice, so the gradient is a window rather than a law.

CHAPTER 2

What it is worth

If the catch-up gradient is real, the practical choice for most organizations is not whether to use frontier AI. It is whether to remain a consumer of intelligence or become a participant in its improvement. A consumer benefits from falling capability costs. A participant benefits from the same falling costs plus the compounding returns of ownership, learning, experimentation, and search.

A smaller actor does not need to reproduce the absolute frontier to matter. It needs enough capability to enter the research loop, enough compute to run meaningful experiments, and enough independence to pursue directions the frontier laboratories are not optimizing for.

A spectator inherits better models. A participant inherits better models and compounds a research process of its own.

2.1 Three positions

There are three rational positions, and the gradient improves all three. Surfing gets cheaper every year, which is exactly why the bar for participating is set by what surfing and customizing cannot buy.

Table 4. Surf, customize, or participate

POSITION	WHAT IT BUYS	WHAT COMPOUNDS	RATIONAL FOR
Surf	Cheap access to improving commodity intelligence through APIs and open models	Application and distribution knowledge; nothing about the models themselves	Most companies. If the goal is a product built on the strongest available intelligence, letting the leaders absorb the cost of model development is an extraordinary subsidy.
Customize	Proprietary capability in a domain, through fine-tuning, retrieval, and tooling on open or closed models	Domain data, workflows, and evaluation sets	Companies whose data or workflows are distinctive but whose edge does not depend on how models are made.
Participate	Ownership of a research trajectory: weights, experiments, infrastructure, lineage, and research velocity	Everything in Section 2.2, including the possibility of a local frontier	Actors with a proprietary reward signal or environment, a sovereignty requirement, very large inference volume, or a research direction the leaders have set aside.

2.2 What diffuses, and what does not

The gradient carries codified knowledge: weights, papers, code, recipes, benchmarks. It carries them in months. It does not carry the other half of what a laboratory knows: which examples matter, which reward signals create pathologies, which benchmarks predict genuine ability, which mixtures, curricula, tools, and architectural changes help, and which experiments reliably waste compute. That knowledge lives in ablation records, failure databases, environments, evaluation judgment, and the craft of a team that has run the loop many times. It is the frontier laboratories’ actual moat: their internal scaling laws and ablation databases are worth more than any checkpoint, and none of it ships with the weights.

Table 5. Two kinds of knowledge

DIFFUSES IN MONTHS	DIFFUSES SLOWLY OR NEVER
Weights and distilled descendants	The record of interventions and outcomes: what actually caused a model to improve
Papers, recipes, reference implementations	Failure databases and the negative results that steer the next thousand experiments
Benchmarks and public evaluation suites	Environments and reward signals built from proprietary work, and evaluations that predict real ability
Training and serving infrastructure	Research velocity: how fast the organization can design, run, diagnose, and iterate
Yesterday’s frontier model	A lineage of descendant checkpoints and the team that produced them

Three consequences follow. First, the durable asset is the research organization, not the weights. A particular checkpoint ages in months. The pipelines, harnesses, synthetic-data systems, reinforcement environments, experiment automation, failure databases, and thousands of negative results that improve checkpoints do not. Each new open-weight release can be dropped into an increasingly mature research organism: the model improves exogenously because the outside world advances, and the organization improves endogenously because it has learned how to turn one model generation into the next.

Second, every serious experiment produces proprietary information, and over time the laboratory accumulates something more valuable than a dataset: a growing record of interventions and outcomes, evidence about what actually causes its systems to improve. Frontier laboratories hold enormous internal stores of this kind. A participant begins building its own on the first day, and cannot begin any other way.

Third, research velocity compounds. The first training run, distributed reinforcement-learning system, or autonomous experiment loop is difficult; repetition converts each into infrastructure. The organization gets better not only because its models improve but because it can design, execute, diagnose, and iterate experiments faster and more cheaply. A better strategic metric than parameter count or benchmark score is research velocity: the number of high-quality model-improvement experiments completed per dollar per week. A spectator does not accumulate it. A participant does.

The economic point is that everything in the right-hand column of Table 5 is denominated in time. It cannot be bought later at any price, because the purchase is the years of running the loop. A late entrant starts with a better base model and an empty right-hand column.

2.3 The local frontier: where the leaders' gradient does not point

The large laboratories search only a fraction of the space of possible AI systems. Their research is constrained by commercial usefulness, inference economics, product roadmaps, safety requirements, regulation, reliability, latency, and the need to produce broadly capable systems. A smaller independent laboratory can optimize for a very different objective: the best AI researcher, the highest rate of scientific discovery per GPU-hour, long-horizon software engineering, continual learning, or recursive model improvement itself. That creates the possibility of a local frontier: a participant may remain behind on general capability while becoming the world leader in a narrower direction.

The risk is absorption. General models have repeatedly overtaken specialized ones, and a direction that is merely under-explored will be explored by a leader as soon as it looks valuable. A direction is defensible when it passes at least one of four tests:

1. It sits outside the leaders' objective. Broad capability served at scale to consumers is their objective; anything that trades generality for depth, or serving economics for accuracy, is not.

2. It depends on a reward signal, environment, or corpus the leaders cannot access: proprietary simulators, instrumented industrial or scientific processes, regulated data, or expert feedback that exists in one organization.
3. It exploits a deployment constraint they do not serve: air-gapped, on-device, sovereign, latency-bound, or below a cost floor that would not pay for a frontier model.
4. Its moat is the environment, the evaluation, and the lineage rather than the checkpoint, so that even a leaked recipe does not transfer the position.

The base rate for outsiders is better than the compute disparity suggests. Beyond the algorithmic contributions listed in Section 1.6, Midjourney led image generation for a period with a team of about a dozen and no outside capital, and DeepSeek forced the industry to reprice the cost of frontier-class reasoning with a fiftieth of the leaders' run-level compute. What those cases share is a chosen direction and an owned loop, not a bigger cluster.

2.4 A portfolio of asymmetric options

Participation creates repeated shots at nonlinear discoveries. Most experiments fail or yield increments, but an improvement in memory, optimization, reinforcement learning, routing, synthetic-data generation, continual learning, or agent architecture can return far more than it cost. A small laboratory cannot usually neutralize a $1,000\times$ compute disadvantage by doing the same thing more cheaply. It can sometimes neutralize part of it by finding a better algorithmic path, and the history in Section 1.6 shows that such paths are found by small teams with some regularity.

The right frame is a portfolio of options, and the lottery metaphor understates the agency involved. Each experiment is an option whose premium is compute plus calendar time and whose payoff is heavy-tailed: many null results, some useful improvements, and a small chance of a step change that creates leadership in one direction. Automation lowers the premium and multiplies the draws. An automated research system running hundreds or thousands of well-instrumented experiments is repeatedly sampling the design space of possible systems, and the sampling is adaptive: each result updates the prior for the next. The value of the portfolio is set by its tail. A single $2\times$ improvement in the cost-efficiency of the chosen direction is worth, to the laboratory that finds it, half of its own compute budget every year thereafter, and to the market it enters, potentially far more. That is why the relevant unit of account is draws per dollar rather than dollars.

2.5 Capability sovereignty, priced as insurance

Closed frontier models provide access to intelligence, but the user does not own the substrate. Pricing, rate limits, product policy, safety rules, model behavior, deployment regions, tool permissions, and model availability all change on the supplier's schedule. Model retirement is now an annual event; in 2025 a leading laboratory withdrew its most widely used model on the day it

launched the successor and restored it only after public protest. With open weights and a working training and inference stack, a participant can preserve, fork, specialize, retrain, benchmark, deploy offline, and continue a model lineage independently.

This is not absolute independence. Hardware supply, power, regulation, software ecosystems, and licenses still matter. But it is a substantially stronger position, and it has a price that can be compared with alternatives: the expected cost of a forced migration when a model is withdrawn or repriced, the value of the regulated, air-gapped, or sovereign markets that a hosted API cannot enter, and the inference margin from serving a distilled model of one's own at a fraction of API prices for narrow, high-volume tasks. The strategic asset is a capability that remains available and modifiable even if the organizations that created the surrounding ecosystem change direction.

2.6 One or two generations behind is the highest-leverage position

Figure 3 gives the mechanism. At the frontier, experiments are extraordinarily expensive and the answers are unknown. Far behind it, the models are not capable enough to be strong research collaborators and the capability is already a commodity. One or two generations behind, the models cross the threshold needed to contribute to research while architectures, papers, tools, and implementation knowledge have diffused. This is a form of temporal arbitrage: the frontier laboratories pay the discovery cost, followers inherit most of the resulting knowledge after a delay, and followers then redirect their capital toward specialization and experimentation.

The position is not static. Its peak moves forward with every open release, so a participant must re-base: the organization is the constant, and the base model is a replaceable part. A laboratory that has built its loop around one checkpoint is behind; one that has built it to accept the next checkpoint is riding the gradient.

2.7 Intellectual property becomes a lineage

A participant that begins with an open model does not have to remain a fine-tuner. Continued pre-training, proprietary datasets, reward environments, architecture changes, evaluation systems, and repeated descendant checkpoints gradually create an independent technical lineage, and the valuable property comes to reside across the whole system rather than in any single weight file.

Open base model → proprietary research variant → improved descendant → novel training recipe → specialized architecture → **independent model family**

The important transition occurs when the organization owns enough of the improvement process that the next model is meaningfully the product of its own research trajectory rather than of the last release it downloaded.

2.8 Sizing the bet

“Small” needs a number. Table 6 gives three tiers at 2025 rental prices of roughly \$2.50 per H100-hour, with people costed at fully loaded rates. The figures are illustrative and round; the ratios between tiers are the point.

Table 6. Three tiers of participation (illustrative, 2025 prices)

TIER	SCALE	ANNUAL COST	WHAT IT BUYS
Directional research on a fixed base	64–256 rented GPUs; 6–15 people	\$5–15M	Post-training, reinforcement environments, distillation, evaluation, continued pretraining of models up to roughly 30B parameters; several thousand instrumented experiments a year; a lineage of specialized descendants. The tier this paper is mainly about.
Own the recipe	1,000–4,000 GPUs; 30–80 people	\$40–150M	Continued pretraining at scale, architecture research, mid-size pretraining from scratch, a near-frontier position in one direction.
Contest the general frontier	20,000+ GPUs; hundreds of people	\$500M–1B+	The DeepSeek, Moonshot, Mistral, and xAI game, with a price that rises $2.4\times$ a year. Out of scope here.

Three kinds of actor should read the first two rows as an opportunity. Investors, because the trail-edge laboratory with a chosen direction is a venture-scale bet whose entry price fell roughly tenfold in two years and whose payoff is an option portfolio plus a set of non-diffusing assets. Corporations, because the boundary between customizing and participating has moved: an organization with a proprietary reward signal or environment, a sovereignty requirement, or inference volume large enough that a distilled model of its own pays for itself now has a reason to cross it. And national programs, because the first two tiers put a sovereign research capability within the budget of a mid-sized country, where before only the third tier existed.

2.9 The cost of waiting

Waiting has real value, and an honest version of this argument concedes it. Later entrants pay less for compute, inherit better models, and avoid the dead ends the early participants paid to find. Under uncertainty, delaying an irreversible investment is often correct. Three things, however, cannot be bought later.

1. **The right-hand column of Table 5.** Experimental record, velocity, and lineage are accumulated by running the loop, and the clock does not run faster for a late entrant with more money.
2. **The window.** The supply of open weights is a strategic choice by a handful of actors. An organization that has forked, verified, and continued a lineage keeps its position if the supply narrows; one that planned to start next year does not.

3. **The direction.** A local frontier claimed by another participant, with an environment and an evaluation moat behind it, is expensive to displace, and the number of directions that pass the tests in Section 2.3 is finite.

The observer who waits for better open models will receive them. The participant receives the same models while accumulating years of infrastructure, experimental memory, automation, and institutional knowledge. Today's model will not remain valuable. The reason to enter is that it is already capable enough to help build the machinery that improves tomorrow's. The objective is a sovereign research organism rather than a smaller copy of a frontier laboratory: a capable open base model that is replaced with each release, rented compute behind an experiment scheduler, persistent experimental memory that records every run and every failure, evaluations tied to a chosen direction, automated training and reinforcement loops, an environment or reward signal the leaders do not have, and the freedom to pursue unusual directions.

If the catch-up gradient democratizes access to capable models, its deeper consequence is that it also democratizes entry into the search for the next kind of intelligence.

Chapter 2 thesis. If capable models are becoming cheaper faster than the absolute frontier is becoming accessible, the scarce asset shifts from possession of intelligence to possession of an improvement process. The participant's durable advantage is not being one generation less behind. It is owning the machinery, the record, and the optionality required to choose where to go next.

What would change this conclusion

The argument rests on measurable trends, so it can be wrong in measurable ways. Any of the following would weaken or reverse it:

- The lag between the best open-weight and the best closed models, currently something like six to twelve months on standard evaluations, grows past two years, or major suppliers stop releasing.
- The cost of reaching a fixed capability level stops falling at roughly $3\times$ a year, for instance because algorithmic progress stalls or rental compute prices rise.
- Automated research turns out to need frontier-scale compute to be useful, so that C^* is only reachable at the frontier.
- General models keep absorbing specialized directions faster than participants can build environment and evaluation moats.
- Regulation restricts open weights above the capability levels that matter, or attaches obligations that only the largest actors can meet.

The indicators to watch are correspondingly plain: the open-versus-closed lag on fixed evaluations, the price of a unit of capability over time, the share of significant algorithmic advances originating outside the three or four largest laboratories, and results on automated-research benchmarks such as RE-Bench.

SOURCES AND NOTES

1. Epoch AI, *The rising costs of training frontier AI models* (Cottier et al., 2024): amortized hardware and energy cost of the largest runs growing about $2.4\times$ a year since 2016, with billion-dollar runs projected by 2027.
2. Stanford HAI, *AI Index Report 2024*, citing Epoch AI: estimated training compute cost of about \$78M for GPT-4 and \$191M for Gemini Ultra.
3. Epoch AI, *Algorithmic progress in language models* (Ho et al., 2024): compute needed for a fixed level of performance halving roughly every eight months, with a wide confidence interval.
4. Stanford HAI, *AI Index Report 2025*: inference cost for GPT-3.5-level performance down more than 280-fold from November 2022 to October 2024.
5. DeepSeek-AI, *DeepSeek-V3 Technical Report* (December 2024): 2.788M H800 GPU-hours, about \$5.6M at \$2 per GPU-hour, final run only; the model was trained on 2,048 H800s. DeepSeek-AI, *DeepSeek-R1* (January 2025): distilled variants from 1.5B to 70B parameters.
6. NovaSky (UC Berkeley), *Sky-T1-32B-Preview* (January 2025): trained for under \$450. Muennighoff et al., *s1: Simple test-time scaling* (2025): 1,000 examples, 26 minutes on 16 H100s.
7. METR, *Measuring AI ability to complete long tasks* (2025) and *RE-Bench* (2024).
8. Google DeepMind, *AlphaEvolve* (May 2025): 23% kernel speedup and about 1% reduction in Gemini training time; 0.7% of global compute recovered.
9. Outsider algorithmic advances: FlashAttention (Dao et al., Stanford, 2022); rotary position embeddings (Su et al., Zhuiyi, 2021); direct preference optimization (Rafailov et al., Stanford, 2023); Mamba (Gu and Dao, CMU and Princeton, 2023); multi-head latent attention, GRPO, and FP8 training at scale (DeepSeek, 2024).
10. Capital spending, rental prices, and the open-weight lag are drawn from 2025 company guidance, public rental-market pricing, and public evaluations; they are order-of-magnitude figures, and the argument does not depend on any of them to better than a factor of two.