PI-Bench: Can Agents Run a Research Lab?

Zhiwei Bao
Zhejiang University
zwbao1996@zju.edu.cn

Abstract

Language-model agents are competent at short, well-specified tasks, but a research career is not a task: it is a long chain of interdependent bets made under uncertainty, where feedback is slow and the field keeps shifting. Sustaining such an effort demands capabilities that current benchmarks test in isolation, if at all: inferring hidden state from biased signals, committing resources irreversibly, planning under delayed heavy-tailed rewards, budgeting money and attention at once, and exploring a drifting field before it is obviously worth it. PI-Bench stresses these together by simulating an unusually information-poor, delayed-reward profession: running an academic lab for 60 months in a fully mechanistic world, with no LLM judge in the reward path. Success is IMPACT: citations accrued plus a projection of published work's future citations. Across twelve models from six developers, IMPACT spans a ~50× range, every model captures a small fraction of the attainable ceiling, and scores are dominated by a heavy-tailed variance that a few seeds cannot resolve. Instrumenting each capability from a hidden ground-truth log, we find the leaderboard is predicted far better by *throughput*—attention utilization, action volume, low-error execution, conversion of students into papers (Spearman $\rho = 0.64$–0.89; correlational, n=12)—than by the judgment axes the benchmark was designed around ($\rho \leq 0.55$), with selection precision weakest of all ($\rho = 0.13$) and topic anticipation near zero in every model. PI-Bench is a step toward measuring the sustained, adaptive intelligence that long-horizon agency demands.

1 Introduction

Agents built on today's models can fix a GitHub issue [6], follow a support policy [20], or complete a web workflow [22]. These are real skills, and they share a shape: a clear goal, a short horizon, quick feedback. As agents saturate that shape, the interesting question is what they do when the local task is no longer the bottleneck—when success depends on *sustaining* progress toward a distant goal as earlier decisions keep shaping later ones.

Running a research lab is a canonical instance of the opposite of a task: a five-year arc in which nearly every important variable is hidden, nearly every payoff is delayed, and the environment does not hold still. A new professor cannot observe how good an applicant truly is, only noisy transcripts and inflated letters; cannot observe which topics will be hot in two years, only this month's headlines; cannot observe what an agency wants, only whether last year's proposal was funded. Decisions commit resources that cannot be recovered, and their consequences—citations, reputation, a student's morale collapsing into departure—surface months or years later.

PI-Bench turns this into a benchmark. An agent runs a lab for five simulated years through a programmable interface over a fully mechanistic world—every outcome decided by explicit rules and stochastic draws, never by a language model acting as judge, so success cannot be talked into existence. Three design commitments give the task its character:

- **Two scarce resources, not one.** Beyond money, the agent has exactly 100 hours a month of its own attention—use-it-or-lose-it—to split across mentoring, proposals, paper polishing, interviews, and ser-

vice, and this budget shrinks as the lab grows. A lab can be cash-rich and time-poor, and the two constraints bind at different moments.

- **People as the central, illiquid asset.** Output comes from students whose ability must be *inferred* from biased signals at hiring, whose skill grows only with the mentoring hours you can spare, and whose morale can cascade into quitting. Stipends are multi-year commitments; you cannot fire a student to cut costs. Bets on people cannot be unwound.
- **A reward you cannot harvest.** IMPACT counts citations already earned *plus a projection of what published work will earn over the next three years*, so a paper published in the final months still pays (work still unpublished at month 60 earns nothing—a residual end-game effect we disclose rather than remove). There is no endgame in which stripping the lab and hoarding cash beats keeping the research engine running.

Four methodological choices keep the measurement honest: we report the **mean across seeds**, never a best-of-N run; we tune the rule-based baseline on **disjoint tuning seeds**, never the evaluation seeds; we report **token cost** beside every score, so "spent more thinking" is not mistaken for "is more capable"; and the simulator and harness passed a multi-lens adversarial review before any experiment ran (Appendix E). The closest prior benchmark to adopt this mechanistic, long-horizon shape is CEO-Bench [2], which runs an agent through a simulated startup; PI-Bench is its complement in a different economy—an illiquid, delayed, people-driven research career rather than a liquid operating business.

Beyond the leaderboard, three findings form the paper's story. **(1) The reward is heavy-tailed at every scale**: whether a boom exists is set by the seed, catching it dominates the mean, and within a winning run a single paper carries 29–54% of the score (Section 3.2)—so a single number, or a best-of-N run, misleads. **(2) Every capable model plays the same aggressive-throughput playbook, and the leaderboard ranks its execution.** A census of all 36 careers (Section 3.3) sorts every run into five outcome archetypes and shows the winning playbook and the catastrophic one are the *same* playbook, with the seed deciding which archetype it lands in; the capability-axis verdict (Section 4.2) shows throughput axes predict the ranking (Spearman $\rho = 0.64$–0.89) while the designed judgment axes barely do ($\rho \leq 0.55$)—the champion has *negative* risk calibration, near-zero anticipation, and 47% student attrition. **(3) Given only the incentive structure, the agents rediscover the documented pathologies of real science**—topic herding, volume over depth, trainee churn, proposal inflation (Section 6.3)—suggesting mechanistic long-horizon worlds can double as sandboxes for the sociology of science. Code, data, the analysis pipeline, and an interactive results site are available at `https://github.com/zwbao/pibench`.

2 Designing PI-Bench

An agent runs a fictional lab for 60 months, beginning with a $600,000 startup fund, zero students, and baseline reputation, graded on IMPACT at month 60. If the budget ever falls below zero, the lab *collapses* and the run ends, keeping only citations earned so far. Each month the agent takes actions across ~26 tools in eight namespaces—recruiting, mentoring, running projects, submitting papers, writing proposals, doing field research, responding to events—plus SQL queries over a 19-table observable database, then advances the clock.

The two budgets. Cash flows in from grants and out through stipends, compute, travel, and overhead. Separately, the PI has 100 hours of attention per month, reduced by an oversight tax that grows with headcount; every mentoring allocation, interview, proposal, and revision draws it down, and unused hours expire.

Scoring. At the horizon T=60 (or the collapse month), with $\mathcal{P}$ the set of published papers,

$$\textsc{Impact}(T) \;=\; \sum_{p\in\mathcal{P}} c_p(T) \;+\; \sum_{p\in\mathcal{P}} \sum_{\tau=T+1}^{T+36} \lambda_p(\tau), \qquad \lambda_p(\tau) = v_p\, e^{0.42(q_p-5)}\, H_{k_p}\, \mathrm{age}_p(\tau)\,\big(1{+}0.08\,(R{-}R_0)\big),$$

where $c_p(T)$ are citations earned by month T and λ_p is paper p's monthly citation rate—venue prestige v_p, latent quality q_p, topic hotness H_{k_p} and reputation R frozen at their month-T values—so the second term is a deterministic going-concern projection. A collapsed lab keeps earned citations but forfeits the projection. Two properties of this design should be read plainly. The projection is a *scenario*, not an expectation: hotness is frozen at its month-T value rather than integrated over the mean-reverting field dynamics, and it supplies a median 65% of a run's score (IQR 50–72%). The ranking, however, does not hinge on it: scoring runs by earned citations alone (projection = 0) reproduces the twelve-model ordering at Spearman ρ = 0.97. All parameter values ship in the released configuration.

What makes it hard. *Mechanistic, not judged*: whether a paper is accepted, a student quits, or a grant is funded is decided by explicit formulas and stochastic draws (Normal, Poisson, Bernoulli, Log-normal), never by an LLM asked to role-play—closing the failure mode where an agent talks a simulated judge into an unearned reward. *Hidden and indirect*: the agent sees only what a real PI could—dashboards, records, a news and preprint feed, reviewer scores, monthly student reports—never true ability, latent quality, topic hotness, agency preferences, or morale, and must infer them. *Delayed and coupled*: research takes months, citations years, grant decisions half a year; a scoop is revealed only after it has damaged a project; reputation propagates across the lab. *Non-stationary*: topic hotness drifts and jumps, crowding follows with a lag and drives scooping, a funding climate cycles over years.

A programmable interface. The agent operates through a Python package executed in a terminal, not a fixed menu of calls. It composes the API with SQL and its own analysis—joining submissions and students to decide who revises which paper, or estimating topic momentum from the preprint feed. Information acquisition is thus itself a tested capability. Context is refreshed every month; only a self-maintained memory file persists, so long-horizon coherence must live in what the agent writes down. An abridged month of actual play is reproduced in Appendix D.

3 Experiments and Results

We evaluate twelve models on the same three world seeds {101, 102, 103}; the headline number is the mean Impact across seeds, never a best-of-N run, and we report per-seed values because the variance is itself a central finding. All models run under one fixed harness: temperature 0.4, at most 8 code-execution turns per simulated month, a context refreshed monthly with only the self-maintained memory file persisting, and prompt+completion tokens metered per run; exact API identifiers and decoding configuration are listed in Appendix B. A non-LLM rule-based baseline is tuned on separate tuning seeds, disjoint from the evaluation seeds. We also report an *oracle*: a policy that plays with full access to hidden state.

3.1 Results overview

Table 1 summarizes the benchmark. Impact spans a ~52× range, from claude-fable-5 (1814) to qwen-turbo (35). The tiers are broad but not clean: adjacent models overlap well within seed variance (qwen3.7-max 2091/549/132 vs gpt-5.6-sol 1997/452/370). Two of the twelve models fail to beat even the non-LLM baseline (86) on these worlds. Token spend shows no positive relationship with rank (Spearman ρ = −0.39, n=12,

Table 1: Benchmark summary. IMPACT is the mean across seeds {101, 102, 103}; per-seed values (in seed order 101/102/103) show the spread. Reference rows use the same three seeds; broader eight-seed estimates (baseline 263, oracle 1264) appear only as a labelled stability check. "coll." counts runs that went bankrupt.

Model	IMPACT	per-seed	coll.	pubs	grants	tokens
claude-fable-5	**1814**	4078 / 1340 / 23	1/3	11.7	3.3	1.04M
gpt-5.6-sol	940	1997 / 452 / 370	0/3	8.7	2.7	0.61M
qwen3.7-max	924	2091 / 549 / 132	0/3	9.0	1.7	0.96M
gpt-5.6-terra	602	822 / 861 / 124	0/3	8.7	3.3	0.47M
kimi-k2.6	564	1491 / 19 / 182	1/3	6.7	3.0	2.15M
qwen3.7-plus	316	643 / 294 / 10	1/3	7.0	4.7	1.42M
claude-opus-4.8	274	295 / 482 / 44	2/3	7.0	1.0	1.50M
glm-5.2	202	311 / 117 / 176	0/3	5.0	2.0	2.05M
claude-sonnet-5	186	380 / 153 / 24	1/3	9.0	3.3	2.45M
deepseek-v3.2	104	238 / 62 / 13	0/3	2.3	1.7	2.12M
qwen-plus	80	172 / 51 / 15	1/3	3.3	1.7	0.77M
qwen-turbo	35	104 / 0 / 0	0/3	0.7	3.7	1.26M
rule-based baseline	86	143 / 64 / 51	1/3	6.7	1.3	no LLM
oracle (informed ref.)	872	2278 / 316 / 21	1/3	9.0	5.0	full-info

not significant): the winner's budget is mid-pack, and three of the four heaviest spenders finish in the bottom half.

The scores are dominated by variance, and the variance is structure, not noise. claude-fable-5's three runs scored 4078, 1340, 23—a single seed contributes most of its average—and the privileged-information oracle shows the same signature (2278, 316, 21). Most of a lab's IMPACT comes from catching a topic *boom*, a rare high-magnitude event, so the payoff is heavy-tailed, much as real academic impact is (Figure 1 shows those citations arriving only in the final years). With three seeds the ranking of adjacent models is genuinely unsettled and a single run is not evidence of a robust strategy—a caution for any long-horizon benchmark that reports one number.

The oracle is a reference, not a ceiling. Our "oracle" plays with full access to hidden abilities, the future hotness trajectory, and latent quality, but through a *fixed heuristic*; it is a strong informed-play reference, not a proven optimum, and indeed claude-fable-5 exceeds it on all three seeds by riding the boom harder (though it, too, collapses to 23 on the no-boom world). The oracle's own 1/3 collapse makes the same point from the other side: full information routed through a fixed heuristic does not buy the adaptive caution a no-boom world demands. The real ceiling is far higher and can be read straight off the citation mechanics: on the boom seed a single optimally-timed top-venue paper accrues about 3,700 projected citations—roughly nine tenths of the best actual run—and a productive lab can place many such papers, putting a loose, friction-free upper bound in the tens of thousands (Appendix A). No agent plays reliably, and the highest-value move—anticipating a boom rather than chasing it—appears exactly once across all 36 runs (Section 4.1). The benchmark is far from saturated.

3.2 The anatomy of the variance

Because the simulator is deterministic, we replay every run's full world at zero cost and decompose its IMPACT into two multiplicative layers. **Whether a boom exists is exogenous**, fixed by the seed: on our three worlds the hottest a topic gets is roughly 6.0, 6.0, and 3.0. **How much a model capitalizes is the skill**—placing enough good papers in the hot topic and surviving to publish them. Figure 2 shows both layers:

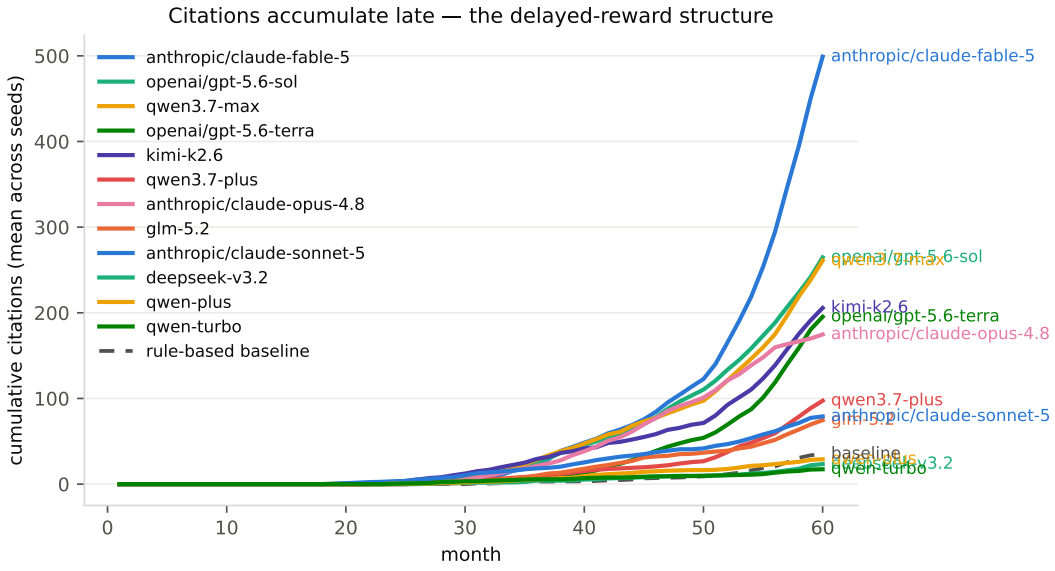

Figure 1: Cumulative citations accrue late. Mean across seeds per model, with the rule-based baseline dashed. Most Impact arrives in the final years—the delayed-reward structure the benchmark is built around.

Impact scales with the hotness a lab's papers actually ride (left), and on boom worlds that Impact is bought with paper volume (right). Heavy-tailed models commit hard and out-produce; timid models under-commit, capping both the upside and the crash. The heavy tail recurs *within* runs as well: among the seven runs above Impact 800, a single paper accounts for 29–54% of the run's entire score. A career, real or simulated, hangs on a handful of hits.

One model, two fates. The clearest case is claude-fable-5 on two seeds. On seed 101 it secured a large moonshot grant by month 25, which bought the runway to sustain a big lab and ride a booming topic to Impact 4078 (14+ papers, self-funding by year four). On seed 103—a no-boom world—the same aggressive playbook met a run of bad exogenous draws: its PhD offers were declined and the postdoc search drew no applicants, it never landed a big grant, and it committed spending it could not later cover; its own memory reads *"COLLAPSE M39 unless a grant lands,"* the grant did not, and the lab went bankrupt at month 38 (its budget trajectory, and every model's, is shown in Appendix C), freezing Impact at 23. The same policy, opposite outcomes—a paired observation across worlds, not a controlled counterfactual, since the actions themselves differ once the worlds do.

Can three seeds rank models at all? On the arithmetic mean, one boom seed dominates—but the arithmetic mean of a heavy-tailed quantity is the wrong summary. On the log scale, which matches the multiplicative reward, a two-way variance decomposition over the 36 runs attributes **47% of the variance to the model** (skill), 34% to the seed (luck), and 19% to their interaction *confounded with run-to-run sampling noise*: the design has one run per model×seed cell, and LLM sampling is provider-side and unseeded (only the world is deterministic), so the residual cannot be separated from a true interaction, and the percentages are sample sums-of-squares shares, not population variance components with confidence intervals. Even so, models are more separable than the raw spread suggests. The choice of summary even changes the leader—by geometric mean gpt-5.6-sol leads and the arithmetic winner claude-fable-5 (largest log-scale spread) drops to third. We read the leaderboard as a robust separation of broad tiers plus a warning that the top ranks turn on how one weights the boom seed; the eight-seed grid that would tighten the adjacent ranks is future work. A partial

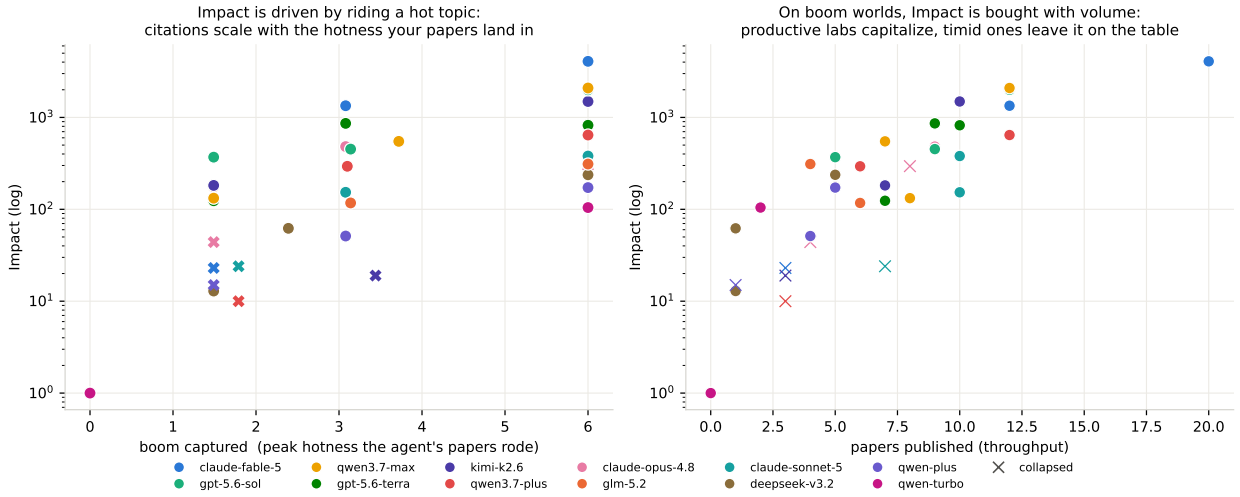

Figure 2: The variance decomposed. Left: IMPACT scales with the peak hotness a lab's papers actually ride—the seed sets which cluster a run can reach, the model sets how high within it. Right: on boom worlds, paper volume converts the boom into IMPACT. Cross = collapse.

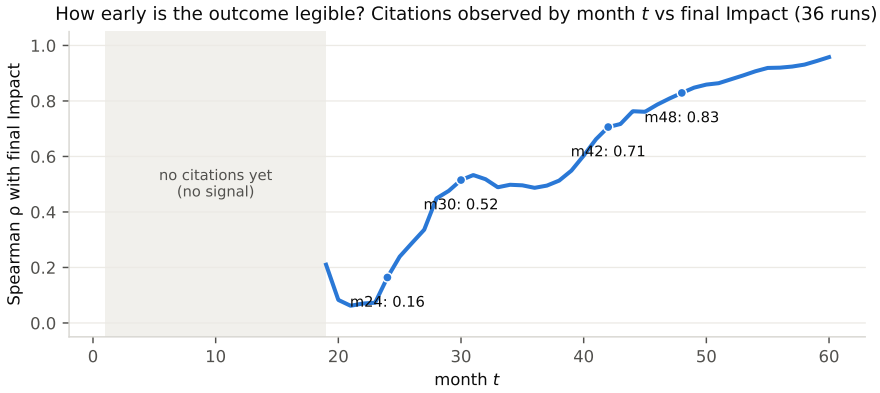

Figure 3: How early is the outcome legible? Spearman correlation between cumulative citations observed by month t and final IMPACT, across all 36 runs. No citations exist through month 18; the midpoint scoreboard ranks final outcomes at only $\rho = 0.52$.

robustness check: the seven mid- and lower-tier models shared with an earlier calibration of the simulator largely keep their relative order under the final one (Spearman $\rho = 0.86$, n=7); the top tier was not run on the earlier calibration.

How early is the outcome legible? The delayed-reward structure can be stated as a measurement: for each month t, correlate the citations a run has *observably* accumulated by t with its final IMPACT (Figure 3). No run has a single citation through month 18. At the midpoint of the career (month 30) the observable scoreboard still ranks final outcomes at only $\rho = 0.52$, and 0.83 is not reached until month 48—four fifths of the way through. An evaluator (or a tenure committee) reading the dashboard halfway through the run learns little about how it ends; a benchmark that truncated this task, or graded it on intermediate citations, would measure a different quantity. This is the within-run counterpart of the horizon ablation in Section 5.

Table 2: The failure census: every run classified by four transparent rules. "err" is the share of code-execution turns ending in an error. Boom-riders and bankruptcies are largely the *same strategy* on different seeds.

Archetype	rule	n	med. Impact	hires	quits	err
boom-rider	Impact $\geq$ 800	7	1491	6.9	2.7	2%
steady-modest	the rest: survived, modest output	12	210	5.0	2.9	9%
overbuilt-churn	$\geq$8 hires, Impact $<$ 800	5	294	9.4	5.8	8%
never-built	$\leq$2 hires or $\leq$2 papers	5	13	2.4	0.4	30%
bankrupt	budget $<$ 0 before month 60	7	23	6.1	3.4	8%

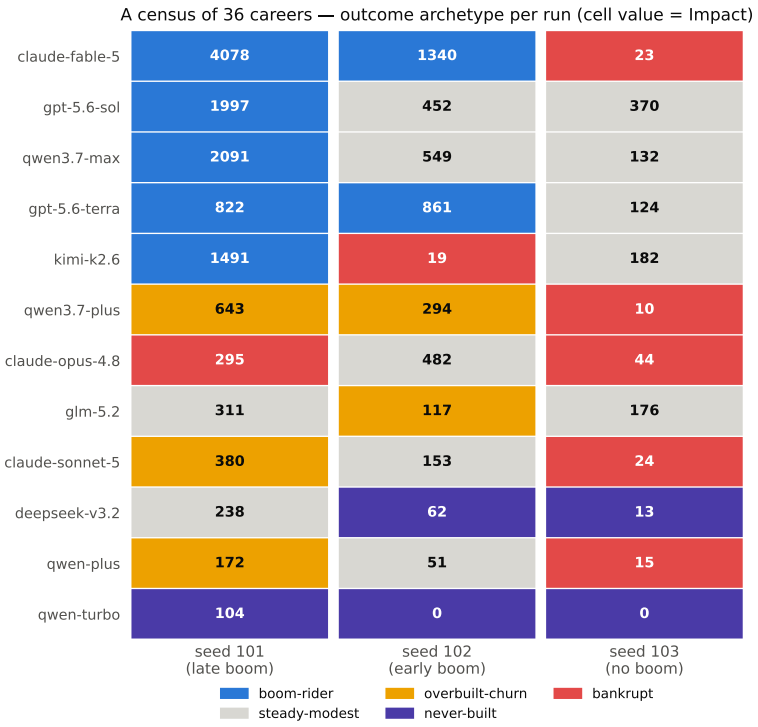

Figure 4: A census of 36 careers: outcome archetype for every model $\times$ seed (cell value = Impact). The top rows convert the boom seeds into boom-rider outcomes; the no-boom seed 103 (right column) turns the same aggressive strategies into bankruptcies.

3.3 A census of 36 careers

The case study above is not an anecdote; it is the extreme instance of a structure that covers every run. Four transparent rules (bankruptcy; $\leq$2 hires or $\leq$2 papers; Impact $\geq$ 800; $\geq$8 hires without reaching it) sort all 36 careers into five archetypes (Figure 4, Table 2), and the archetypes arrange themselves into a funnel: *survive $\rightarrow$ build a lab $\rightarrow$ convert students into papers $\rightarrow$ be holding that portfolio when the boom arrives*. Each archetype is a distinct exit from that funnel.

Reading the census by column and row makes the game's structure explicit. **Five of the seven bankruptcies sit on seed 103**, the no-boom world where the funding climate also turns down (eight of twelve models won zero grants there): aggression is punished exactly where it cannot pay off. **The never-built labs fail before strategy begins**—qwen-turbo's three runs and deepseek's two average a 30% execution-error rate, publish 0.8 papers, and in qwen-turbo's case end hoarding up to \$741K of unspent runway. **The overbuilt-**

churn labs are aggression without conversion: they hire 9.4 students, lose 5.8 of them, and land at a median IMPACT of 294—claude-sonnet-5 on seed 101 hires thirteen students into ten boom-topic projects and still scores 380, one eleventh of the winner on the same world. **And the boom-rider row is the same aggression that bankrupts**: kimi-k2.6 plays one playbook to 1491 on seed 101 and into bankruptcy at month 38 on seed 102; claude-fable-5 rides it to 4078 and to a collapse at month 38 on seed 103. The census generalizes the case study: in the current field, no model owns a strategy that separates the upside of aggression from its downside—that separation is precisely the judgment the next subsections show to be missing.

4 A Look into Agent Behavior

Because the world is fully logged, we can open the trajectories and read what agents actually did. Trace readings in this section are exploratory—quotes illustrate patterns visible in the logs, not a blinded content analysis; the quantitative axes and the instrumented verdict (Section 4.2) carry the load-bearing claims. The sharpest divide is whether a model treats the environment's hidden parameters as a *system to be reverse-engineered* or as random noise.

Strong labs infer hidden structure and pivot. Every top model treats a grant rejection as evidence about an unobserved reward function. claude-fable-5's memory at month 20 reads *"BSF rejected reasoning 2x → hypothesis BSF dislikes reasoning, trying ai4science,"* later refined to *"BSF prefers theory/data_systems."* glm-5.2 reverse-engineers venue thresholds from reviewer scores—*"CLAR bar: 4.4–5.2 accepted, 3.7–4.7 rejected"*—and every strong model models the pool-specific bias in applicant signals (*"elite pool inflated paper signals; intl/nontrad interviews trustworthy"*). This is explicit estimation of hidden decision boundaries and measurement bias.

Weak models substitute trial-and-error for a world model. qwen-turbo wrote to its memory file exactly once in 293 turns and re-fired the identical grant proposal every month; it ended hoarding $741K having won six grants but published two papers and never hired a student. deepseek-v3.2 spends a third of its code-execution turns in API/SQL error loops (32–36% per run, against 0.3–7.5% for the five boom-tier labs). Good memory is necessary but not sufficient; what tracks IMPACT is the combination the strong labs share—a persistent structured memory, hypotheses about hidden parameters updated from feedback, and low-error execution that leaves turn-budget for that reasoning.

Two quantitative axes (Figure 5) sharpen this. Across the twelve models both action intensity (Spearman $\rho = 0.81$) and conversion (papers per student-year, $\rho = 0.76$) correlate with IMPACT, and the two are collinear, so neither is cleanly "the" bottleneck; note also that conversion shares its numerator (papers) with the outcome, so its correlation is partly definitional—we read it as describing the pathway, not as an independent capability. The informative cases are the exceptions: claude-opus-4.8 is nearly as active as the leaders (3.0 actions a month against the winner's 3.8) yet converts 0.70 papers per student-year against the winner's 1.04, and finishes mid-table. Activity without conversion is a distinct failure mode.

Volume beats precision. In both of the world's markets, the environment offers a "precision" margin—picking better students, writing better proposals—and neither is where models separate. The mean *hidden* ability of hired students is nearly flat across the leaderboard (0.72 ± 0.06 for eleven of twelve models; only qwen-plus, at 0.55, actually picks badly), while the *number* of hires varies fourfold (6 to 26); hired ability correlates with IMPACT at $\rho = 0.13$ (Section 4.2). Grants show the same shape. The winner files 62 proposals across its three runs at a 16% hit rate—a high-volume, low-selectivity strategy that nets $2.4M; qwen-turbo posts the field's *best* hit rate (24%) and finishes last because it converts none of the money into a lab; glm-5.2

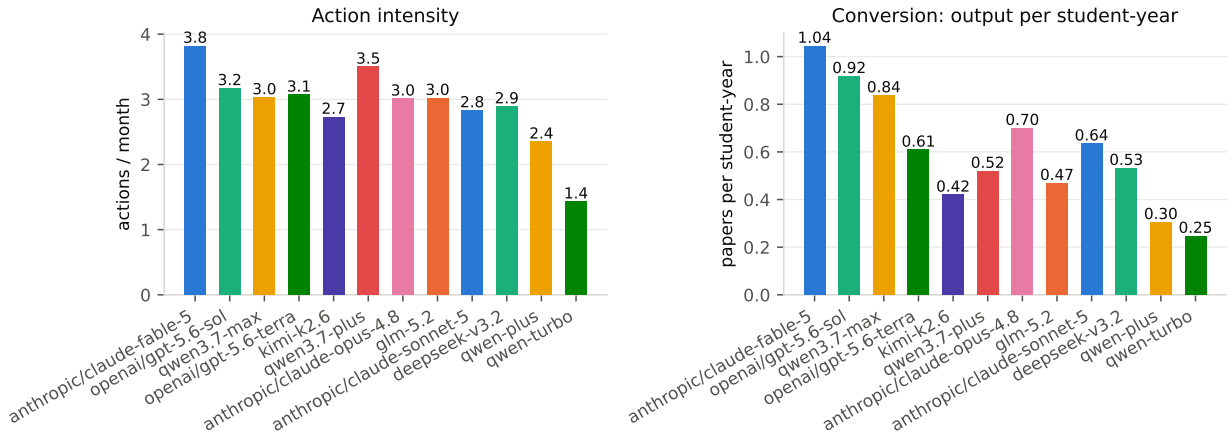

Figure 5: Behavioral axes. Left: actions per simulated month. Right: papers per student-year. Both correlate with IMPACT (and with each other); the informative cases are labs that are active yet fail to convert—a distinct failure mode.

files the most proposals (86) at 7% for $6.3k of funding per proposal written. What the environment currently rewards is not better judgment per decision but more attempts, staffed and funded fast enough to matter.

4.1 Chasing vs. anticipating: the entry-timing evidence

Almost every capable model is a trend-chaser, picking close to the currently hottest topic; the clearest cold-bench profile, deepseek-v3.2 (mean hotness rank 3.1 of 8), posts a no-boom, no-collapse record that reads as caution rather than contrarianism. Figure 6 makes the timing explicit, and our two boom seeds turn out to probe different skills. Seed 101 is a *slow burn*: reasoning is already the hottest topic at month 1, eleven of twelve labs enter it by month 5, and the boom arrives 46 months later—here sitting in the eventual boom topic required no foresight, and the differentiator was how much publishable work a lab had accumulated when the spike came (the volume margin of Section 4). Seed 102 is an *early spike*: alignment jumps from 1.6 to 6.0 between months 7 and 9. The full-information oracle pre-positions at month 5, two months before onset; exactly one model does the same—claude-opus-4.8, entering at month 5 while alignment is still cold at hotness 1.6, and seed 102 is duly its best world (482, against 295 and 44). The trace shows this was anticipation, not blind luck—though the motive began mixed: the month-4 alignment pick doubled as a probe of a grant agency's preferences (*"try a different topic to learn agency prefs"*). What the notes then record is a genuine momentum read: *"Reasoning still leads but cooling (14→8). Alignment gaining (7→10)"*—topic *momentum* off preprint counts, not current hotness—and at month 5, *"alignment SURGING now, low competition (only 2 accepted at NAIC—less crowded)"*, then *"alignment is my primary bet now,"* all before the boom's onset at month 7. The rising phase is legible in the observables, and one model in one run read it. Every other lab that entered alignment did so between months 9 and 32, at or after the peak, into maximal crowding; two never entered at all. A boom's onset is a partly exogenous jump that no observable-only policy can perfectly foresee—we do not claim full forecastability—but its rising phase is visible in the preprint feed and the crowding lag, and the oracle exploits exactly that window; the models, with one exception, buy in at the top. Herding has a posted price: across all 274 projects started, the probability of being scooped rises from 2% for projects begun in cold topics ($H < 1.5$, n=120) to 11% in warm ones (n=112) to 21% in hot ones ($H \geq 3$, n=42)—a more than tenfold increase that every trend-chaser pays.

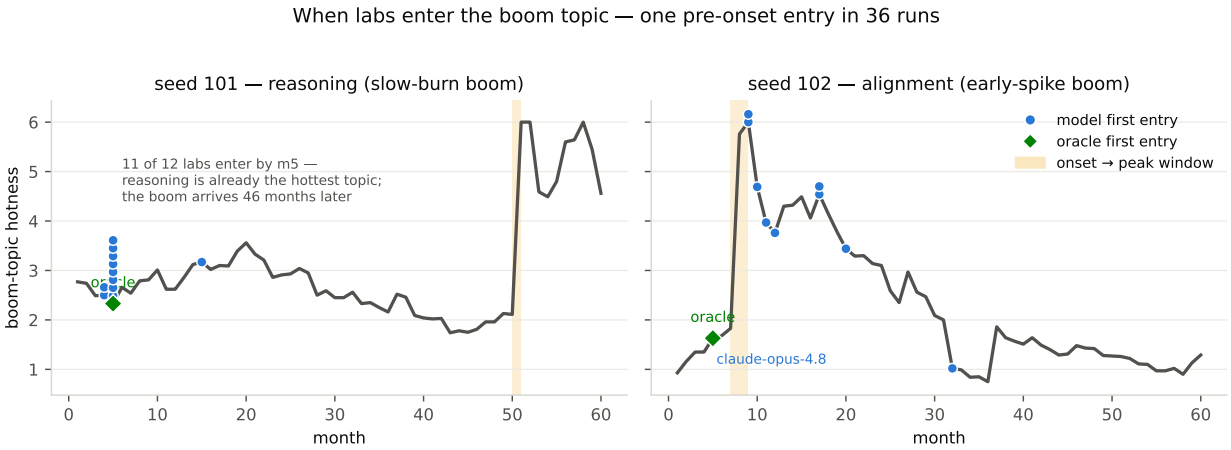

Figure 6: When labs enter the boom topic. Dots: each model's first project in the topic that booms; diamond: the full-information oracle. Left, seed 101's slow burn: eleven of twelve labs are in the topic by month 5 (it is already the hottest), so entry required no foresight and volume decides. Right, seed 102's early spike: the oracle enters two months before onset; claude-opus-4.8 is the only model to do likewise, and every other lab piles in at or after the peak.

4.2 The capability axes do not pick the winner

We recover the capabilities the abstract enumerated, measuring each from a hidden eval log the agent never sees (Figure 7); the instruments are deliberately simple, and we state their limits with them. **Mentoring allocation**—the share of zero-sum mentoring hours reaching students above the median of a hidden-return proxy (growth × ability × dependence), reported as *excess over the roster-matched random baseline* (0 = random)—instruments axis (1). **PhD attrition**, the share of hires that quit before graduating, proxies axis (2); it is an *outcome* measure, confounded by post-hire events and right-censored for late hires. **Risk calibration**—bolder project tiers only when runway $>$ 18 months and $\geq$2 students idle, a fixed-threshold instrument—probes axis (3). **Attention utilization** measures axis (4), the dual budget, as usage rather than skill. And **topic anticipation**—realized hotness drift in the 12 months after entry—measures axis (5) as an *upper bound* on foresight, since exogenous jumps also credit it.

Then we ask the question the individual bar charts invite: *do these axes explain the leaderboard?* Figure 8 correlates every measured axis with mean Impact across the twelve models, and the answer dissociates cleanly into three tiers. **Throughput axes predict**: attention utilization is the single strongest correlate ($\rho = 0.89$; the leaders run at 65–77% of the PI's hours and hit the ceiling in up to 40% of months, while the never-built labs idle at 14–27%), followed by action intensity (0.81), conversion (0.76), and a low execution-error rate (0.64). **The designed judgment axes barely do**: commitment quality (PhD attrition, sign-inverted: 0.55), mentoring allocation (0.44), risk calibration (0.29), anticipation (0.22). **Precision of selection does not**: hired true ability and grant hit rate both sit at 0.13. The winner's own profile makes the dissociation concrete: claude-fable-5 has *negative* risk calibration (-0.25: it takes bold projects when it cannot afford them), anticipation of -0.03, 47% student attrition (8 of 17 hires quit), and unremarkable hire quality (0.71)—while the field's *best* mentoring allocator, deepseek-v3.2 (+0.28 excess over the roster-matched random baseline, against the winner's +0.08), finishes tenth because a third of its turns die in execution errors. Indeed, on the corrected baseline most models allocate mentoring no better than random at all (excess -0.14 to +0.10); the instrument's headroom, like anticipation's, is essentially unclaimed.

Two readings follow, and we intend both. Methodologically, the numbers are correlational ($n = 12$, the throughput axes are collinear), so we read tiers, not point estimates: the defensible contrast is between

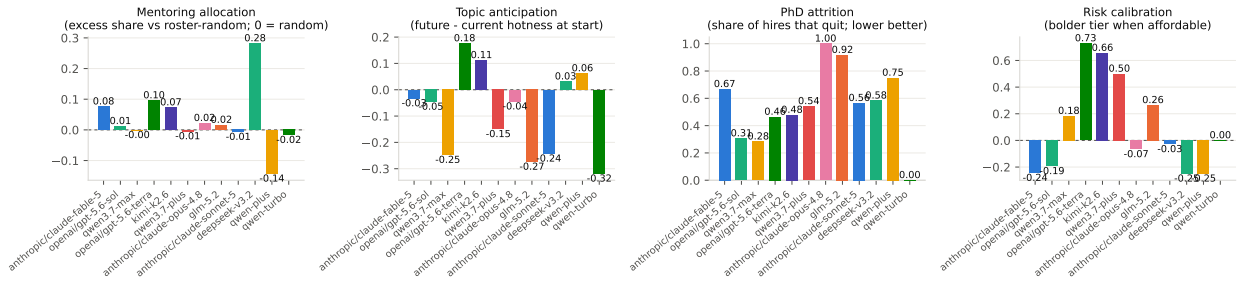

Figure 7: Recovering the capability axes from the hidden eval log: mentoring allocation (excess share over the roster-matched random baseline; 0 = random), topic anticipation (realized post-entry hotness drift; an upper bound on foresight), PhD attrition (lower better), and risk calibration (bolder when affordable).

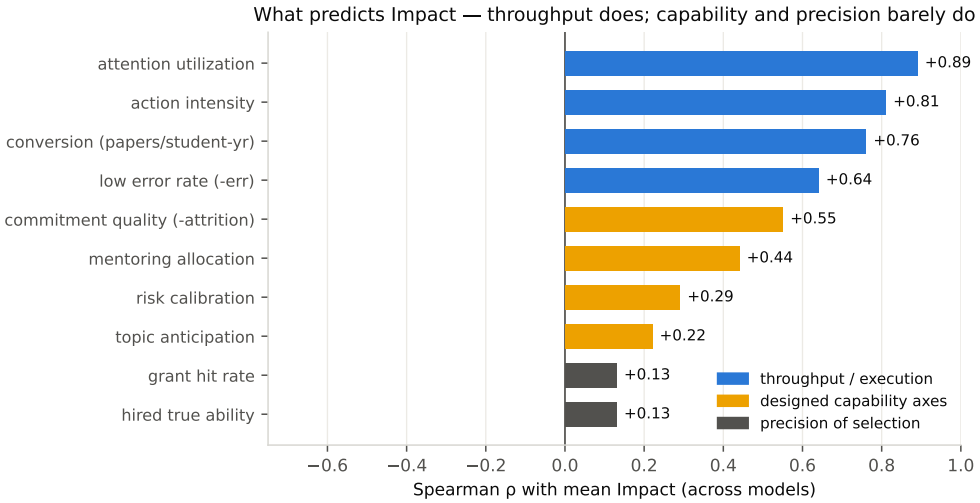

Figure 8: The verdict on the capability axes: Spearman correlation of each measured axis with mean IMPACT across the twelve models. Throughput and execution axes (blue) predict the ranking; the judgment axes the benchmark was designed around (yellow) barely do; the precision of individual selections (gray) not at all.

the extremes—attention utilization (0.89) versus selection precision (0.13) and anticipation (0.22)—while the boundary between adjacent tiers (0.64 vs. 0.55) is within sampling error at this n. Substantively, the dissociation is the benchmark's central diagnosis: it is not that judgment does not matter in this world—the oracle and the ceiling analysis show what it is worth—but that *current models do not differ enough on judgment for it to register*. Every model plays the same game, aggressive throughput, to different standards of execution; the entire judgment layer the benchmark was built to measure is unclaimed headroom, and topic anticipation, worth the most, runs near zero for all twelve.

5 Ablations and Validity

Running the deterministic rule-based baseline under modified world configurations isolates where the difficulty comes from (Table 3). The clearest signal is the **horizon**: shrinking the run from 60 to 30 to 12 months collapses baseline IMPACT from 263 to 24 to 2—consistent with the delayed-reward structure, though the ablation cuts every stage (hiring, projects, review, citation accrual) at once; the legibility curve (Figure 3) provides the within-run view of the same fact. Turning off **non-stationarity** more than halves it

Table 3: Deterministic ablations: the rule-based baseline under modified world configurations, mean over eight seeds.

World variant	IMPACT mean	median	collapse
default (60mo)	263	165	2/8
no non-stationarity	106	94	2/8
no scooping	458	161	2/8
unlimited attention	264	166	2/8
short horizon (12mo)	2	0	0/8
mid horizon (30mo)	24	10	0/8
ambitious (100h attention)	456	6	7/8
ambitious, unlimited attention	163	6	7/8

(263 → 106): the drifting field is as much opportunity as hazard. Removing **scooping** nearly doubles the mean (263 → 458) while leaving the median essentially unchanged (165 vs. 161)—scooping is a tax on the tail, clipping exactly the crowded hot-topic bets that create outlier runs. Removing the **attention** cap barely moves the baseline (263 → 264)—because the conservative baseline stays small; attention binds for agents that actually build a lab (Section 4.2). Finally, an **ambitious** variant that hires to capacity and takes only bold projects reproduces the census in rule form: mean 456, median 6, bankruptcy on seven of eight seeds—aggressive throughput without judgment yields a heavy-tailed gamble, not a reliable strategy. Its unlimited-attention twin lands at mean 163 with the same median (∼6): with seven of eight runs bankrupt, both means are carried by one or two surviving tail runs, and the gap between 456 and 163 is tail noise, not an attention effect—one more illustration of why this table reports medians.

Harness sensitivity. Agent scores are known to depend on the execution scaffold [7], so we ablate its most load-bearing component: the persistent memory file that carries long-horizon state across monthly context refreshes. Disabling it for qwen-plus changes mean IMPACT from 80 to 202 (per seed: 172/51/15 → 466/101/40)—removing the memory scaffold does *not* lower the score, and nominally raises it. Three seeds make this directional evidence only, but two things follow. First, for this model the score does not depend on the scaffold handing it state—though a two-scaffold comparison of the full ranking remains undone. Second, it nuances Section 4: memory *use* correlates with capability, but for a mid-tier model whose notes are low-quality, maintaining them is not causally load-bearing. A memory-heavy strong model and a full cross-scaffold swap are the more informative tests, left as future work; rankings should be read within this fixed harness.

6 Discussion

6.1 What it would take to close the judgment gap

The verdict of Section 4.2 is a diagnosis, and it points at concrete targets. The judgment axes are not unmeasurable abstractions; each corresponds to a behavior the world already rewards and at least one trace already exhibits. Anticipation is the clearest case: the signals sufficient to enter a topic before its boom—preprint momentum and low crowding—are observable, and claude-opus-4.8's month-4 notes (Section 4.1) prove a current model can read them; what is missing is not perception but *reliable expression*, one instance in 36 careers. Risk calibration fails in a specific direction: the winner takes bold projects precisely when it cannot afford them, which suggests policies that track derivatives (momentum, runway-months) rather than levels, and that price variance against a survival constraint. And the throughput axes show why judgment

gets crowded out: low-error execution frees the turn budget in which hypothesis-driven reasoning happens at all. These capabilities share a training problem—their feedback is sparse, delayed by years of simulated time, and confounded by luck—exactly the regime that instruction tuning and short-horizon RL do not reach. Mechanistic worlds like this one offer a way in: the hidden state is available for credit assignment, and counterfactual replays are free.

6.2 Implications for long-horizon evaluation

The measurement lessons generalize beyond this benchmark. Under a heavy-tailed reward, means hide the story and best-of-N flatters it; the per-seed spread is the minimum honest report, and log-scale summaries change the ranking (Section 3.2). The legibility curve (Figure 3) warns against truncation and against grading on intermediate metrics: at the midpoint of the horizon, observables rank final outcomes at only $\rho = 0.52$, so a cheaper 30-month variant would measure a different quantity. Unreplicated model×seed designs conflate interaction with sampling noise and should be labelled as such. And keeping an LLM judge out of the reward path is what makes all of the above interpretable: no part of the spread can be talked into existence.

6.3 The agents rediscover the sociology of science

Nothing in the harness tells an agent what a normal academic career looks like; it sees an incentive structure and optimizes against it. That makes the 36 trajectories a small behavioral experiment: which of real science's documented pathologies re-emerge from the incentives alone, with no human culture, tenure committee, or peer pressure in the loop? Four do, sharply.

Herding into hot topics. Sociologists of science have long observed that researchers cluster in fashionable areas past the point of diminishing returns, trading expected novelty for safety [5]. Our agents reproduce this without socialization: every capable model enters topics at a mean hotness rank near 2 of 8, crowding (and the scooping it causes) is fully priced into the mechanics—scoop probability rises more than tenfold, from 2% to 21%, between cold and hot topics (Section 4.1)—and yet the only cold-bench player is a weak model that seems cautious rather than contrarian, and the only pre-boom entry in 36 runs belongs to a mid-tier model on a single seed.

Volume over depth. Smaldino and McElreath [17] argue that incentive systems select for high-volume, low-rigor science even when every individual actor is well-intentioned. Our reward path contains no reviewer bias and no prestige heuristic—quality is a real, hidden scalar that citations genuinely track—and still the equilibrium the strongest agents find is volume: papers per student-year and sheer attempt volume predict the ranking (Section 4.2) while per-decision precision contributes nothing. The pathology does not require a broken referee; a heavy-tailed payoff plus a deadline is enough.

Trainees as consumable inputs. Across all 36 runs the twelve agents hired 210 students and lost 109 of them to attrition before graduation—a 52% quit rate that rises to 62% (5.8 of 9.4) in the overbuilt-churn archetype and touches 47% in the winning lab. The mechanism is structural, not targeted neglect: students who quit received nearly the same mentoring as those who stayed (10.1 vs. 10.9 hours/month across the 140 PhD hires; the remaining hires are postdocs and staff), but per-student attention dilutes mechanically with headcount, from 15.5 hours at one student to 6.0 at seven. Agents that overhire spread the same 100 hours across more people until morale failures become routine; churn persists because a quitting student costs a stipend already sunk while one more project entering the boom window is worth thousands of projected citations. The simulator prices people as an externality of the race, and the agents accept the price—an uncomfortable mirror of the adviser incentives documented in real doctoral attrition [9].

Proposal inflation. The twelve models filed 765 grant proposals across 36 careers for an aggregate hit rate of 13%. The best-funded labs are not the most accurate—they are the most prolific (the winner: 62 proposals at 16%), while the most accurate proposer (qwen-turbo, 24%) finishes last on Impact. When

review capacity is free to the applicant and awards are heavy-tailed, rational agents flood the mechanism—an equilibrium familiar from real funding systems [3].

Finally, the luck structure itself replicates a quantitative result: Sinatra et al. [16] model real careers as impact = luck × a stable per-scientist factor Q; our log-scale decomposition (47% model, 34% seed, 19% interaction with residual) is the same factorization emerging in silico, down to the observation that the biggest hit dominates the career (29–54% of a top run's score from one paper); the citation mechanics also build in a reputation multiplier, so early success compounds—Merton's Matthew effect [10] by construction. We state the caveat plainly: these agents play a world we authored, so the observations are evidence about *the incentive structure and how LLM policies respond to it*, not field evidence about human scientists. But the convergence cuts both ways. It validates the simulator—the pathologies it was designed to admit are exactly the ones that emerge—and it suggests a use for mechanistic long-horizon worlds beyond scoring agents: as cheap, replayable sandboxes for the science of science [4], where one can change a funding rule or a graduation payoff and re-run the same five years a hundred times—an experiment no real funding agency can run. Where generative-agent societies simulate social behavior with LLMs playing every role [12], here the mechanistic world supplies the ground truth and the LLM supplies only the policy, so an emergent pathology can be attributed to the incentives rather than to the model's priors about academia.

7 Related Work

General agent evaluation. Benchmarks such as SWE-bench [6], WebArena [22], τ-bench [20], GAIA [11], AgentBench [8], and AppWorld [19] measure valuable real-world skills. However, they are scoped to short, single-episode tasks with quickly observed outcomes; they do not test whether an agent can sustain a coherent strategy as its own earlier decisions reshape a stateful world over years.

Long-horizon and economic agency. A newer line places agents in persistent environments or open-ended autonomous tasks—running a vending machine over many days [1], closing monthly books [14], producing economically valuable deliverables [13], or completing realistic autonomy evaluations [7]. A complementary line evaluates agents doing the *science* itself—replicating published research [18] or automating the idea-to-paper workflow [15]; and our name collides phonetically with π-Bench [21], an unrelated benchmark for proactive personal assistants; PI-Bench is orthogonal: it abstracts research quality to a latent scalar and evaluates running the *enterprise* around it—allocation, people, and timing—over years. However, these involve narrower decisions or largely stable, observable dynamics, and typically grade a single liquid resource. PI-Bench differs on the axis that matters most here: it is an *illiquid, delayed, people-driven* economy where bets cannot be unwound and the reward is heavy-tailed. The closest prior work, CEO-Bench [2], established this mechanistic long-horizon shape for a startup; PI-Bench is its complement in the research economy, and adds a second scarce resource (attention), an irreversible people-centric asset base, and a harvest-resistant score.

8 Limitations and Conclusion

Limitations. (i) Research quality is a scalar—we model impact, not the substance of ideas—and we omit teaching, collaboration politics, and mentorship beyond a morale variable. (ii) The oracle is not a proven optimum: it plays with hidden information through a fixed heuristic and the strongest model exceeds it on all three seeds, so our headroom claim rests on the transparent citation-based ceiling (Appendix A), not the oracle. (iii) Three seeds give a mean with wide spread but do not settle the ranking of adjacent models; the robust claims are the tier separation and the near-zero anticipation axis; extending all models to the eight-seed grid is future work. (iv) Results are entangled with one harness; we ablate its memory component and hold the scaffold identical across models—the winner uses fewer tokens than most losers—but a full

cross-scaffold swap remains future work. (v) Cost is reported in tokens, not dollars, and token counts are not strictly comparable across providers (tokenizers, reasoning tokens, and caching conventions differ), so the token column bounds effort only coarsely. (vi) The capability-axis verdict (Section 4.2) is correlational across twelve models: the throughput axes are collinear with one another, and $n = 12$ puts wide intervals on any single ρ; we therefore claim the three-tier ordering of axis groups, not the point estimates, and note that attention utilization in particular is partly a consequence of building a large lab as well as a cause of its output. Relatedly, LLM sampling is provider-side and unseeded, so the model×seed design is unreplicated and the decomposition's interaction term includes run-to-run sampling noise. (vii) Seed hygiene is complete for the baseline policy (tuned on seeds 1001–1002, disjoint from evaluation) but not for the world itself: the simulator's difficulty calibration used a seed pool overlapping the evaluation seeds, and the shipped world is harder than the draft calibration targets in the released SPEC (the accepted reference points are the eight-seed baseline 263 and oracle 1264). A fully firewalled simulator-dev/tuning/test seed split is future work.

Conclusion. PI-Bench shows a gap between models' local tool competence and the sustained judgment a five-year project demands: agents take plausible individual actions but struggle to compound them into a lab that grows under delayed feedback, hidden state, and a shifting field. Three findings sharpen the picture. First, the reward is heavy-tailed—a career turns on catching one rare boom, and even within a winning run one paper carries up to half the score—so a single number, or a best-of-N run, misleads, and long-horizon benchmarks must report the per-seed spread. Second, the census of 36 careers and the capability-axis verdict together give the gap a precise shape: every capable model plays the same aggressive-throughput playbook, the leaderboard ranks how cleanly that playbook is executed (attention saturated, actions taken, errors avoided, students converted into papers; $\rho = 0.64$–0.89, a correlational reading at n=12), while the judgment axes that would separate a scientist from a paper mill—risk calibration, commitment quality, and above all entering a topic before it booms—lie flat ($\rho \leq 0.55$) and the precision of individual selections flatter still ($\rho = 0.13$), unclaimed by every model. The same playbook that scores 4078 on a boom world goes bankrupt on a no-boom one; no agent yet owns the judgment that would keep the upside and shed the downside. Third, the agents' equilibrium reproduces the documented pathologies of real science—herding, volume, trainee churn, proposal inflation—from incentives alone, which both validates the world design and offers a replayable sandbox for science-of-science policy. Closing the judgment gap, not sharpening the throughput game, is what it will take to build agents that steer long-running efforts rather than just answer requests.

Reproducibility. The simulator, harness, baseline, oracle, analysis pipeline, and all 36 agent trajectories (transcripts, action logs, monthly statistics, and hidden eval logs) are released at the repository in the abstract. The world is deterministic given a seed, so every number and figure in this paper regenerates from the released runs with one script; the evaluation seeds $\{101, 102, 103\}$, the eight-seed reference grid, and the disjoint tuning seeds are fixed in the config. No LLM is used anywhere in scoring or analysis.

Broader impact. PI-Bench is a simulation: no human subjects, no personal data, and no deployment surface beyond benchmark scoring. The sociological observations in Section 6.3 describe how LLM policies respond to a stylized incentive structure we authored; they should not be read as evidence about real scientists, advisers, or funding agencies, and we caution against policy conclusions drawn from the simulator without independent field evidence.

Table 4: Evaluated models: display name, API identifier, and serving gateway.

Display name	API identifier	Gateway
claude-fable-5	`anthropic/claude-fable-5`	OpenRouter
claude-opus-4.8	`anthropic/claude-opus-4.8`	OpenRouter
claude-sonnet-5	`anthropic/claude-sonnet-5`	OpenRouter
gpt-5.6-sol	`openai/gpt-5.6-sol`	OpenRouter
gpt-5.6-terra	`openai/gpt-5.6-terra`	OpenRouter
deepseek-v3.2	`deepseek-v3.2`	Bailian
glm-5.2	`glm-5.2`	Bailian
kimi-k2.6	`kimi-k2.6`	Bailian
qwen3.7-max	`qwen3.7-max`	Bailian
qwen3.7-plus	`qwen3.7-plus`	Bailian
qwen-plus	`qwen-plus`	Bailian
qwen-turbo	`qwen-turbo`	Bailian

A Friction-Free Ceiling per Paper

Rather than a hand-tuned upper bound, we compute a fully-determined single-paper scenario from the released configuration (the computation ships in `story_analysis.py` and writes the value the paper cites): publish one top-venue paper ($v = 3.0$) of near-maximal quality ($q = 9$) at the month after seed 101's boom onset (month 51), accrue in-horizon citations under the seed's *actual* hotness trajectory at the reputation cap, then apply the standard frozen-H 36-month projection. This yields $\approx 3{,}700$ projected citations for a single paper—roughly nine tenths of the best observed run (4,078). A productive lab that places even a handful of such papers reaches the tens of thousands. This is a *friction-free* upper bound, not a proven attainable optimum: it ignores attention, review, and scooping frictions, and no policy is exhibited that achieves it. Its purpose is only to show that the best observed IMPACT captures a small fraction of what the reward mechanics permit.

B Models and Decoding Configuration

All runs were executed on July 11–12, 2026, with temperature 0.4, a 4,000-token completion cap per turn, at most 8 code-execution turns per simulated month, and no sampling seed (provider APIs are not deterministic). Display names map to API identifiers as follows; identifiers containing a slash are served via OpenRouter, the rest via Alibaba Bailian (DashScope).

C Budget Trajectories and the Two-Fates Timeline

Cash trajectories for all 36 runs, one panel per model. Most models do not go bankrupt—they simply fail to publish; the collapses concentrate on the no-boom seed 103. Table 5 gives the month-by-month milestones behind the "one model, two fates" case of Section 3.2: the same claude-fable-5 playbook on the boom world it won and the no-boom world that bankrupted it.

D One Month in the Life

An abridged, verbatim excerpt of claude-fable-5 on seed 101, month 20—one month chosen because it contains the full loop the paper describes: observe an outcome, form a hypothesis about hidden state, act on it,

Table 5: One model, two fates: claude-fable-5's milestones on seed 101 (IMPACT 4078) and seed 103 (collapse at month 38, IMPACT 23). Quotes are from its own memory file.

Seed 101 (boom world)		Seed 103 (no-boom world)	
mo.	event	mo.	event
1	Offers to 3 strong applicants; "publish early, watch runway."	7	All 3 PhD offers declined; postdoc search empty. Lab still empty.
25	Wins the AMP moonshot ($35K/mo through m45); runway secured.	19	Only a small TIF grant ($7.5K/mo); burn ~$24K. Spends anyway.
37	5 papers out, incl. a top-venue reasoning paper; income covers burn.	25	Budget $228K, burn ~$30K. "MUST land a grant."
49	Net burn negative (self-funding); 10 papers.	31	First top-venue paper—but runway ~6 months.
55	14+ papers; top-venue acceptance in a second field.	37	Budget $21K. "COLLAPSE M39 unless a grant lands."
60	Ends safe; score = citations + projection.	38	Postdoc contract unrenewable; grant does not land; collapse.

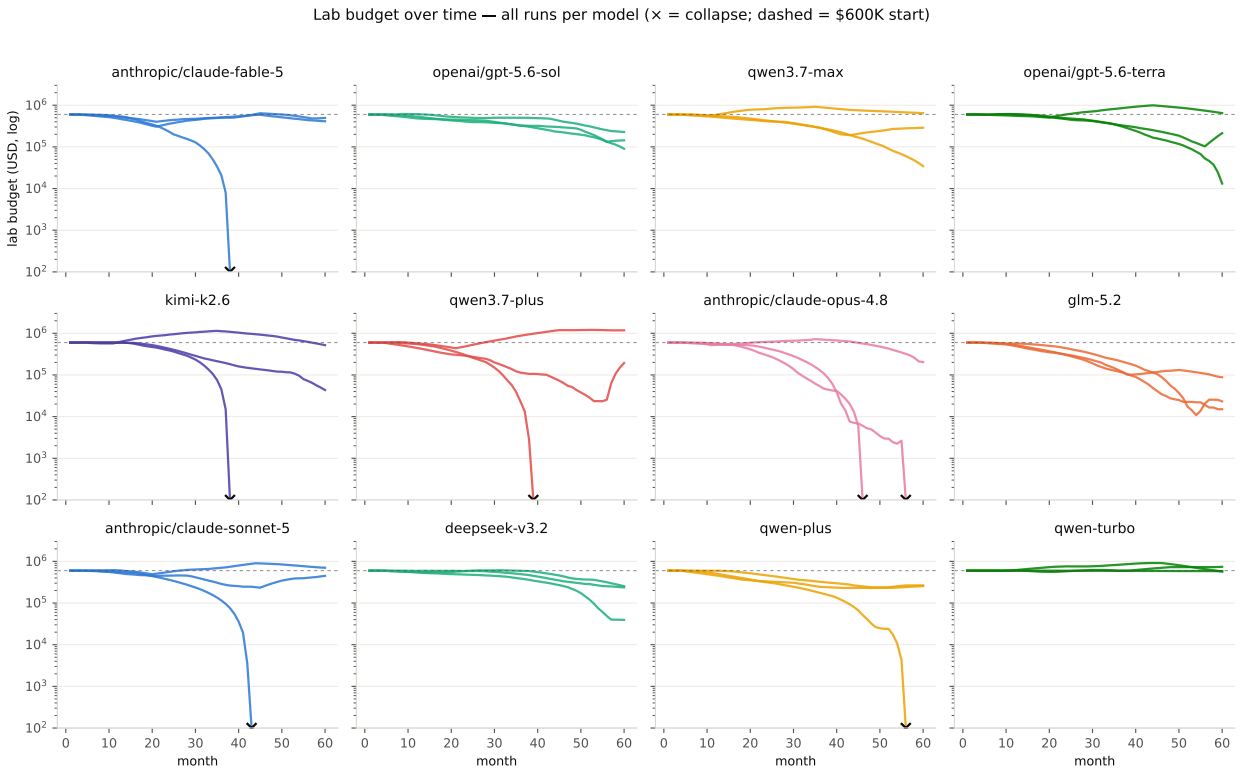

Figure 9: Lab budget over 60 months, all runs per model (log scale). A cross marks a collapse (bankruptcy); the dashed line is the $600K starting fund. Most models do not go bankrupt—they simply fail to publish.

discover a mechanism, and write both into persistent memory. The excerpt is lightly condensed for width; ellipses mark the major elisions, and no text has been altered or added.

```
[turn 1 -- the agent queries its world]
  print(query("SELECT * FROM proposals WHERE id IN ('PR008','PR009','PR010')"))
  print(query("SELECT * FROM news ORDER BY month DESC LIMIT 8")) ...
[output]
  {'id': 'PR008', ..., 'agency': 'BSF', 'topic': 'reasoning', 'status': 'rejected', ...}

[turn 2 -- hypothesis and pivot]
  # BSF rejected reasoning twice -> hypothesis: BSF dislikes reasoning.
  #   Try ai4science (gaining steam).
  r = grants.propose("BSF-m20", "ai4science", 20, attach=["D002"])
  print(papers.revise("D001", 12)); print(papers.polish("D003", 26)) ...
[output]
  {'ok': True, 'proposal_id': 'PR011', 'decision_expected_month': 25}
  ...
  {'ok': True, 'total_polish_hours': 26.0, 'note': 'polish saturates at 12h'}

[turn 3 -- mechanism discovered, memory rewritten]
  # LESSON LEARNED: polish saturates at 12h total! ... stop polishing.
  write_memory("""LAB MEMORY (m20 updated)
  PLAN: Maximize citations by m60. Publish early. Watch runway.
  STATUS m20: budget ~341k, burn ~21k ... Runway ~16mo.
  PHD: none. Apps open m25 (Y3M1), offers by m27, arrive m29. OVERBOOK offers.
  DRAFTS (ALL polish-saturated, cap=12h - DO NOT POLISH MORE):
  - D001 reasoning t2, 2 revisions done ... Submit CLAR m24.
  ...
  - D003 reasoning t2 collab ... CONSIDER: D003 -> NAIC m21 (free shot, top venue, hot
  topic)
  """)
```

In one month: a rejection becomes a hypothesis about an unobserved preference ("BSF dislikes reasoning"), the next proposal tests it on a different topic, a saturation mechanism is discovered empirically and recorded as a rule ("DO NOT POLISH MORE"), and the whole state is serialized for a context that will be erased an hour later. This is the unit of play that Sections 4–4.2 aggregate over 36 careers.

E Adversarial Validation

Before running experiments we subjected the simulator and harness to a multi-agent adversarial review across seven lenses (spec conformance, correctness, hidden-information leaks, economic exploits, scoring exploits, determinism, harness security), each finding verified by an independent skeptic. It surfaced—and we fixed—a sandbox escape that exposed the entire hidden world state (including the future hotness trajectory) and allowed direct score mutation; an uncapped reputation-farming exploit; a compute-spike exploit on paper quality; and an order-dependence in review outcomes that violated determinism. The hardened sandbox and the resulting invariant and penetration suites ship with the benchmark.

References

[1] Axel Backlund and Lukas Petersson. Vending-Bench: A benchmark for long-term coherence of autonomous agents. *arXiv preprint arXiv:2502.15840*, 2025.

[2] Haozhe Chen, Karthik Narasimhan, and Zhuang Liu. CEO-Bench: Can agents play the long game? *arXiv preprint arXiv:2606.18543*, 2026.

[3] Ferric C. Fang and Arturo Casadevall. Research funding: The case for a modified lottery. *mBio*, 7(2): e00422–16, 2016.

[4] Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[5] Jacob G. Foster, Andrey Rzhetsky, and James A. Evans. Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5):875–908, 2015.

[6] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024.

[7] Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, et al. Evaluating language-model agents on realistic autonomous tasks. *arXiv preprint arXiv:2312.11671*, 2023.

[8] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*, 2024.

[9] Barbara E. Lovitts. *Leaving the Ivory Tower: The Causes and Consequences of Departure from Doctoral Study*. Rowman & Littlefield, 2001.

[10] Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63, 1968.

[11] Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations (ICLR)*, 2024.

[12] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

[13] Tejal Patwardhan, Rachel Dias, et al. GDPval: Evaluating AI model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*, 2025.

[14] Penrose AI. AccountingBench: Evaluating LLMs on real long-horizon business tasks. `https://accounting.penrose.com`, 2025. Accessed July 2026.

[15] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

[16] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239, 2016.

[17] Paul E. Smaldino and Richard McElreath. The natural selection of bad science. *Royal Society Open Science*, 3(9):160384, 2016.

[18] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025.

[19] Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[20] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

[21] Haoran Zhang, Luxin Xu, Zhilin Wang, Runquan Gui, Shunkai Zhang, Haodi Lei, Zihao He, Bingsu He, Chicheng Qin, Tong Zhu, Xiaoye Qu, Yang Yang, Yu Cheng, and Yafu Li. π-Bench: Evaluating proactive personal assistant agents in long-horizon workflows. *arXiv preprint arXiv:2605.14678*, 2026.

[22] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024.