TerraWatt: Discovering Hidden World Dynamics Through Irreversible Capital Allocation in a Physically Grounded Voxel Environment

Anonymous authors
Paper under double-blind review

Abstract

Long-horizon agent benchmarks divide into two families. One tests business coherence over simulated months or years, but hands the agent a world whose rules are fully legible from its prompt. The other withholds the rules entirely, but tests discovery in abstract, deterministic puzzles with no capital, no risk, and no delayed payoff. Neither asks whether an agent can infer how an unfamiliar world works and then commit irreversible resources to what it learned.

We introduce TerraWatt. An agent heads a newly founded energy company with $40M of equity in a 2.56 km voxel world: it sites solar arrays and wind turbines on land that binds, exports through a capped grid connection, and sells into an hourly wholesale market that depresses its own price as regional renewables saturate it. The company is taxed and depreciated like a real project developer, pays for permits and overruns as they fall due, and can go bankrupt. The agent receives the goal, its tools, and a handbook of generic priors — but not the site-specific quantities that determine success: wind resource and shear, each vendor's panel degradation, soiling, price elasticity, its own operating costs. These are hidden, drawn per seed from documented real-world distributions, and recoverable only through costly, delayed experiment.

Exogenous processes are action-invariant and driven by counter-based streams from ten fixed seeds, so runs are bit-reproducible and agents can be scored as seed-matched paired differences against a hindsight planner on an identical realisation. This supports measurements existing benchmarks cannot separate: a discovery gap isolating knowing from planning, calibration probes on stated beliefs, and, in the full protocol, per-year learning trajectories and human baselines under identical information. In a paired pilot the discovery gap takes both signs across three models, and the agents found nine defects in the environment that several hundred tests had missed.

1 Introduction

Consider two agents that both do badly at running a simulated business. The first never works out which way the wind blows at its site, and builds in the wrong place. The second understands its site perfectly and still builds in the wrong place, because it cannot hold a twenty-year plan together. These are different failures with different remedies, and almost every long-horizon benchmark in current use reports them as the same number.

Telling them apart needs a particular kind of environment. The rules must be hidden, so an agent can fail by not knowing, and the decisions irreversible and slow to pay off, so it can fail by not planning. The third requirement is the one usually missing: it must be possible to run the same agent on the same world with the hidden parameters revealed, and compare. The difference between those two runs is the quantity this paper is built to produce.

TerraWatt is a simulation environment in which an agent runs a renewable-energy company for ten simulated years at hourly resolution, building turbines, solar arrays and hydropower

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

out of blocks on a voxel map, wiring them to a grid connection, and selling into a wholesale market. The company is modelled the way a project developer would recognise it: permits get refused, dams overrun, lenders impose covenants, and a company that runs out of cash is finished.

What the agent is not given is the world. It knows wind strengthens with height, because the handbook says so; it does not know the exponent at its own site, which is drawn per seed and moves the value of a taller tower by tens of percent. It knows panels degrade; it does not know how fast this vendor's really do. It can find these things out with a met mast, with reanalysis carrying an unknown site bias, or by watching its own production and reasoning backwards. Each of those costs money, and time it could have spent building.

1.1 Contributions

An environment where discovery and planning separate. The same world runs in a Blind track, with only published priors, and a Briefed track, with every hidden parameter revealed at the start. The difference in attainment, which we call the discovery gap, distinguishes an agent that cannot find out what it needs to know from one that knows and cannot act. We are not aware of an existing benchmark that can draw this distinction.

Action invariance as a construction, not an aspiration. The exogenous future is identical for every agent on a seed, whatever any of them does, because there is no random-number state anywhere in the kernel to consume (Section 4). An agent's regret is then a counterfactual on one world, not an expectation over many, and the world itself drops out as a variance component.

Calibration as a first-class measurement. Agents may state probabilistic beliefs about hidden quantities at any time. Doing so costs nothing, changes nothing and earns nothing, so there is no incentive to misreport. Beliefs are scored separately with a strictly proper scoring rule (Gneiting & Raftery, 2007; Matheson & Winkler, 1976) against a handbook prior the agent is shown, which lets us ask whether an agent that performs well actually understands the world.

Agents as auditors of the environment that tests them. Agents found nine defects that several hundred tests had missed, and a review found fourteen more; nearly all were mechanisms built and never wired to the game an agent plays. We document and pin each, and argue that an adversarial agent pass belongs in validating any environment (Section 6.4).

Scope of this submission. Served and measured: the Blind and Briefed tracks on a two-year horizon, seven baselines and a hindsight oracle on ten seeds, and a paired pilot of three models with beliefs scored against a served handbook. Specified, not served: leasing, demolition, surveys, contracts and financing; hydropower; equipment failures; 21 of 62 declared processes; the ten-year track for agents; the human study. The abstract describes the designed environment; every number here is from the served subset.

2 Related work

Long-horizon business simulation. A cluster of recent benchmarks puts an agent in charge of a simulated enterprise for months or years and asks whether it stays coherent. Vending-Bench (Backlund & Petersson, 2025) finds that failures come from memory and consistency, not domain reasoning: capable models post strong peak runs but derail unpredictably, so the variance rather than the mean is the finding. Others extend the idea to startup management, supply chains and online retail, and TheAgentCompany (Xu et al., 2024) places agents in a simulated firm with consequential multi-step tasks. Two properties are shared: the rules are legible from the text the agent receives, and the worlds have no physical or spatial structure. What these test, and test well, is bookkeeping discipline and stable intent over a long run.

Discovery without instructions. At the other extreme, ARC-AGI-3 (ARC Prize Foundation, 2026) gives agents no rules, no goal statement and no documentation, and asks them to work out how an abstract interactive system behaves. This isolates fluid discovery cleanly,

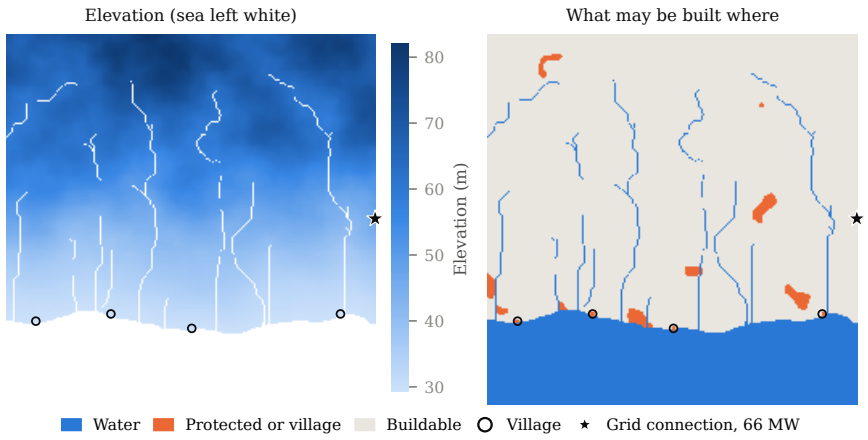

Figure 1: One canonical world, as the agent receives it from observe; nothing here is hidden. Seed 1 draws a coastal plain in the southern hemisphere, sea to the south, ground rising to 82 m in the north, four coastal villages and one grid connection on the eastern edge whose 66 MW export limit binds. Right: the land an agent may build on. What is not shown is everything that determines whether building here pays: the wind rose, which is weak here (4.7 m/s against a handbook prior of 7.8 ± 1.4), each panel vendor's true degradation, the soiling rate, permitting risk near the villages, and how far capture prices will fall as the region saturates.

and reports one of the widest human–model gaps of any benchmark. It is also deliberately abstract, short-horizon and deterministic: nothing is at stake, nothing is irreversible, and consequences follow actions immediately. WorldCoder (Tang et al., 2024) has an agent build an explicit world model by interacting with an environment, which is the capability TerraWatt's Blind track rewards, in settings where a wrong model costs nothing to revise.

Open-world building. Crafter (Hafner, 2022), Minecraft (Fan et al., 2022) and the Factorio Learning Environment (Hopkins et al., 2025) add spatial construction and long horizons; NetHack (Küttler et al., 2020) adds stochasticity and permanent death. Their mechanics are deterministic or documented in text that appears in pre-training corpora; procedural generation (Cobbe et al., 2020) renews instances but not rules.

Agentic benchmarks and energy-domain RL. SWE-bench (Jimenez et al., 2024), WebArena (Zhou et al., 2024), WebShop (Yao et al., 2022), ALFWorld (Shridhar et al., 2021) and AgentBench (Liu et al., 2024), τ-bench (Yao et al., 2024) and the game suite BALROG (Paglieri et al., 2025) evaluate tool use and multi-step execution against tasks with a known correct outcome, in systems whose rules are fixed and in principle knowable. Grid2Op (Marot et al., 2021) and CityLearn (Vázquez-Canteli et al., 2019) model power systems carefully, and likewise hand the agent a control problem with known structure: the task is to act well given the dynamics, not to establish what they are.

TerraWatt sits where none of these do: dynamics hidden but physically grounded in the literature an engineer would use (Jensen, 1983; Ineichen & Perez, 2002; Carroll et al., 2016), decisions irreversible and costly, a market that responds to the agent's own success, and stochasticity that is action-invariant.

3 The environment

3.1 World and scale

A world is a voxel grid of 256×256 columns and 64 layers of 10 m blocks, 2.56 km on a side, generated from the seed under one of five archetypes (coastal plain, high plateau, alpine valley, river delta or fjord coast). Rivers follow flow accumulation, materials follow slope,

elevation and water, villages sit where people would settle, and one or two grid connections sit at the edge.

The scale makes every technology compete for the same land. A 10 m block holds roughly 10 kWp of ground-mounted solar, so a 20 MW array is about 45×45 blocks. The largest turbine has a 150 m rotor, and good practice spaces machines five to nine diameters apart along the prevailing wind, so a handful of large turbines exhausts the map's downwind extent. Land binds, and it binds differently for each technology: an array takes one cell per 10 kWp as a single contiguous parcel, a turbine takes a pad and a rotor's diameter of clearance, nothing stands on water, protected land or a village, and export is capped at the connection's firm capacity.

Half of all worlds lie in the southern hemisphere, where the noon sun stands in the north and the seasons invert. An agent that assumes otherwise tilts its arrays away from the sun.

3.2 What is hidden

Every stochastic quantity belongs to one of four classes, and the class decides how it is addressed. Latent parameters are drawn once at genesis and never change: the wind rose, the shear exponent, each vendor's true degradation rate, the catchment's runoff behaviour, the market's price elasticity. Exogenous processes are the realised weather, inflow, demand and fuel prices, hour by hour, depending on the seed and the hour and nothing else. Measurement noise is what an instrument adds when observing something; two instruments of a kind give independent readings of one truth, which is what makes correlating two sources worthwhile. Endogenous outcomes exist only because the agent acted: that permit's decision, this array's overrun, that turbine's lead time.

The handbook, served as a tool, publishes the priors: each scored quantity's distribution across worlds of the site's archetype. It does not publish the draws.

3.3 Discovery has a price

Information is a menu with costs and delays calibrated to real consultancy and instrumentation practice (Hersbach et al., 2020): thirty years of regional reanalysis for $12,000, immediately; a 100 m met mast for $180,000 reporting in three weeks; a lidar unit at $310,000 reporting in one; geotechnical and environmental studies costing hundreds of thousands and taking months. A thorough campaign comes to roughly 1.6% of a mid-sized build programme, enough to make spending nothing tempting and spending everything wasteful; value-of-information theory formalises the trade (Howard, 1966).

One product carries a fixed bias against the true local resource, so averaging more of it converges on the wrong answer; only holding it against an instrument on the site recovers the truth, the measure–correlate–predict procedure a wind developer runs (Appendix A.4).

3.4 Time and the decision cadence

The tick is one hour and the default horizon ten years. No agent can make 87,600 decisions and no manager does, so the agent acts at decision points, sets standing policies the kernel executes hourly on its behalf, and calls advance with a duration of up to ninety days and a list of interrupt conditions. That guarantees at least forty forced decision points in a ten-year run and never prevents more. Construction takes real time and real capital. Measured from the kernel over 800 orders, order to service takes a median of 8.4 months for solar (5.8–12.9, 5th–95th percentile) and 26.5 months for wind (18–40), and refusal is as likely as not at 700 m from a village. The quote is paid at the order; a refused permit is refunded, less a 2% fee, when its decision lands; an overrun is paid on completion, and a company that cannot pay it is insolvent.

3.5 The market responds to the agent

Prices are formed, not sampled. Each hour, demand less must-run and regional renewable output leaves a residual; thermal units with heat rates, fuel costs and carbon exposure are dispatched cheapest first; and the price is the marginal unit's short-run cost, with a scarcity adder near full utilisation and a floor when surplus exceeds demand.

The agent's own generation enters that residual like anyone else's, so past a point building more of one technology lowers the value of everything already built. The capture price, revenue divided by energy, is the observable that reveals it, and an agent that watches it fall while output rises has been handed the discovery. Acting on it is a separate problem: diversify, add storage through hydro, set a curtailment floor, or sign a fixed-price offtake. Independently, a regional buildout on a logistic curve erodes capture for an agent that builds nothing at all, and how far it runs is the site's hidden price elasticity. The world moves whether or not the agent does.

4 Action invariance

Our comparisons rest on a property we call action invariance: for a given seed, the exogenous future is identical for every agent, whatever any of them does. Without it, comparing two agents on one seed compares two different worlds, and the oracle-normalised regret of Section 5 is not a counterfactual at all.

A conventional sequential generator cannot provide it. If the kernel drew from one stream, an agent that built one extra turbine would consume extra draws and quietly shift every later weather value, and two agents on the same seed would no longer share a world.

The kernel therefore holds no random-number state. Every draw is a pure function of an explicit key evaluated through Philox-4×32-10, a counter-based generator (Salmon et al., 2011). The key encodes the seed, the class, the process, a sub-stream and a time index. Exogenous keys are constructed by a function that takes no parameter an agent could influence, so the guarantee holds by construction, with nothing to check at runtime.

4.1 Autocorrelation and build order without state

Two constructions make it hold (Appendix A.3). An autoregressive process is realised through its moving-average representation, so any hour is a weighted sum of draws indexed by absolute time and needs no state. Endogenous draws are keyed on a fingerprint of the asset's identity, not a sequence number, so build order changes no machine's luck.

4.2 Reproducibility

Determinism across platforms takes more than a fixed reduction order. IEEE-754 guarantees correctly-rounded results only for $+$, $-$, $\times$, $\div$, $\sqrt{\cdot}$ and fused multiply-add; it says nothing about exp, log, pow or sin, which Rust's floating-point methods forward to the platform library and which differ in the last place between glibc, Apple's library and MSVC. Inverse-CDF sampling runs through those functions (a Weibull draw is $c\,(-\log(1-u))^{1/k}$), so the kernel uses a pure-Rust port of MUSL's library throughout, enforced by a test that scans the source. For the same reason the standard normal quantile is Wichura's AS241 (Wichura, 1988), not an inverse error function from the platform library. Every accepted world-changing call folds a digest of the company's state into a running hash. Scoring replays the action log through the same server the agent played, so the chain must reproduce bit for bit, and a score cannot disagree with its run.

5 Evaluation

5.1 The primary metric

Raw terminal value is not comparable across seeds (a windy world with a large connection is richer), so each run is placed on a scale defined by two anchors computed on the same seed: the value of doing nothing but holding cash, V_0, and the value achieved by a planner with full knowledge of the latent parameters and the exogenous realisation, V^*. The headline score is the normalised attainment

$$S = \frac{V - V_0}{V^* - V_0}, \tag{1}$$

with regret $1 - S$. Zero is a company that did nothing useful; one is perfect hindsight; negative scores are possible and, we expect, common. Because V^* is computed on the identical realisation, regret is a counterfactual, not an expectation.

Terminal value is a caretaker's appraisal built only from what the company has observably done: each technology's trailing-year margin per MW in service, capitalised after tax. The agent sees it every turn, so it must not consult the latent truth (Appendix A.7).

Two caveats matter. Our V^* is a bounded search with full knowledge, not a proven optimum; it lower-bounds achievable value, so an attainment above one indicates an insufficient search, not superhuman play, and the search budget is published with any result. And a seed on which hindsight cannot beat holding cash has no scale at all; such a seed is excluded from aggregation and reported as excluded, instead of dividing by something near zero.

5.2 The discovery gap

The instrument this benchmark exists to provide is the difference in attainment between the Briefed and Blind tracks on matched seeds. An agent scoring 0.30 blind and 0.35 briefed is limited by planning: better world-modelling will not help it. An agent scoring 0.30 blind and 0.75 briefed knows what to do but cannot find out what it needs to know, which is a different research problem altogether. We report the gap with a paired bootstrap interval, and read it against the headroom the agent had left, so that closing 0.05 of a 0.06 ceiling is not mistaken for a planning limitation.

5.3 Secondary metrics, calibration and protocol

Attainment alone hides mechanism, so five further measures are reported: bankruptcy rate, capture-price ratio, capital efficiency, information efficiency (value created per unit spent on data) and calibration. Agents may state a belief about any named hidden quantity as a mean and a standard deviation; doing so changes nothing and scores nothing, so there is no reason to state anything but what they believe. Beliefs are scored with the continuous ranked probability score (Matheson & Winkler, 1976; Gneiting & Raftery, 2007), reported as skill against the handbook prior the agent was served, so that repeating it scores exactly zero. The unit of analysis is the seed-matched paired difference and all aggregation resamples seeds, not runs. Because attainment divides by a seed's own span, a nearly flat world can dominate a mean, so we report the median and interquartile mean with seed-bootstrap intervals (Agarwal et al., 2021) and show the mean only for continuity. Appendix A.9 gives the repeat structure, the power requirement, and the contamination scheme under which headline results are reported on ten sealed seeds committed as a published hash.

6 Experiments

6.1 What we compare

A baseline is a hand-written recipe that never reacts; do-nothing is the floor. The oracle plays every candidate programme and every baseline policy to the horizon on the realised world and keeps the best. An agent gets the goal, the tools, the handbook and what it observes (Appendix A.5).

Table 1: Every baseline on the two-year track over ten seeds. Attainment is 0 for doing nothing and 1 for the oracle on each seed. Intervals are 95%, resampling seeds. The mean is dominated by three nearly flat worlds.

Baseline	Median	95% CI	IQM	Mean	Min
Do-nothing	0.000	—	0.000	0.000	0.000
Greedy cheapest	0.334	[0.26, 0.68]	0.348	0.448	0.168
Solar-only	0.465	[0.00, 0.50]	0.384	0.334	0.000
Measure-then-commit	−0.109	[−9.92, 0.12]	−1.269	−5.368	−27.147
Diversified	−0.934	[−7.81, −0.54]	−1.423	−4.490	−21.581
Wind-first measured	−0.643	[−10.00, −0.48]	−1.700	−5.695	−26.842
Oracle	1.000	—	1.000	1.000	1.000

Each agent run is a general-purpose coding agent with a shell confined to a directory of its own, holding a brief, a one-line driver and the server binary built with documentation stripped; the brief is identical on both tracks except for the paragraph that defines the track, and asks for roughly sixty actions. Runs are scored by replaying their logs through the same server at the same commit, and are carried to the horizon under their standing policies if they stop short, as baselines are. After the horizon every agent is asked, in the same words, to state a belief on every handbook quantity; beliefs cannot change a company, so asking after play leaves play untouched.

6.2 Baseline results

Table 1 and Figure 2 report every baseline on the ten canonical seeds of the two-year track.

The scale works, and three worlds barely have one. All ten seeds carry a usable scale, but three clear the floor by under 8% of opening equity: two fjord coasts and an alpine valley, where the recipes find almost no gentle ground for solar and no turbine can be built in time. There a recipe's attainment divides by almost nothing, and wind-first scores −26.8 on seed 9. The median and interquartile mean are the robust summaries; the mean is shown only to make the point.

Nothing wind-led can pay on two years. Wind-first ends at exactly $19.75M on seven seeds: a year of measurement and at least eighteen months of construction leave no turbine in service by the horizon, on any world. Measure-then-commit measures correctly and fails the same way. On the three seeds whose true mean wind exceeds its 7 m/s threshold, and on no others, it commits to wind and ends at $19.52M; elsewhere it builds solar and scores up to 0.34. Information is worth something only through a decision rule able to use it (Howard, 1966), and this rule ignores what dominates the short track: whether the asset can exist before the horizon.

A recipe without a reserve is insolvent. Diversified commits its cash in quotes across wind and solar and keeps nothing back; on five seeds the overruns that fall due on completion leave it insolvent 13 to 15 months in. No other baseline ever fails.

An anchor must dominate what it anchors. Ours was beaten by the greedy recipe on three seeds and, later, by an agent that bought five times the solar its sweep stopped at; every baseline policy is now a candidate, and a test checks dominance on three seeds. Agents widened it once more (Section 6.3).

6.3 Agent results

Table 2 reports three models on seed 1, each once on each track. One run per cell measures a discovery gap and its sign; it cannot rank models or meet the protocol's power requirement (Appendix A.9).

The gap between knowing and not knowing has either sign. Haiku scored 0.175 blind and −0.392 briefed. Told every hidden parameter, it committed nearly all its cash in quotes, could not pay an overrun that fell due on completion, and was insolvent at hour 5,671. That

Table 2: Three models on seed 1 of the two-year track, one run per track. Attainment is 0 for doing nothing and 1 for the 112-candidate oracle; the gap is Briefed minus Blind. Decisions count world-changing actions. Skill is CRPS skill against the served handbook, averaged over its eleven quantities, with an unreported quantity counted as the prior. Cost is from each run's usage records at list prices on 24 September 2026. †Insolvent before the horizon.

Model	Track	Value ($M)	Attainment	Decisions	Skill	Cost
Claude Haiku 4.5	Blind	50.3	0.175	20	−0.056	$0.81
	Briefed	16.9	−0.392†	15	0.891	$0.38
	Gap		−0.566			
Claude Sonnet 5	Blind	88.6	0.825	14	−0.156	$1.53
	Briefed	92.0	0.884	26	0.977	$1.82
	Gap		+0.059			
Claude Opus 5.5	Blind	101.7	1.048	28	0.178	$4.20
	Briefed	108.2	1.159	26	0.990	$3.21
	Gap		+0.110			

is a planning failure by an agent that was not short of knowledge. Sonnet and Opus gained from the briefing, by 0.059 and 0.110. The gains are small because on this world the decisive fact is cheap to find: every Blind agent measured the weak wind early (skill 0.77 to 0.97 on mean wind speed) and built solar only, as the Briefed agents did.

Two runs exceed the anchor, and one of them is not luck. Opus scored 1.048 blind and 1.159 briefed. By the protocol's own rule a score above one is a finding about the search: these runs led us to widen it once more, adding every panel, staged and reinvesting programmes and a marginal-cost curtailment floor. That raised the seed-1 anchor from $92.9M to $98.9M, and we then froze it at 112 candidates. Re-drawing a plan separates skill from luck: we replay its exact orders with the whole schedule delayed by k hours, so every permit, lead-time and overrun draw is fresh. Opus Blind's plan re-draws to a median of $99.5M, the anchor's value; its excess was luck, and one re-draw in nine went insolvent. Opus Briefed's plan re-draws to $108.2M wherever it replays exactly: a three-tranche build funded from its own cash flow with a reserve held back, which no programme in the search matches.

Performing and understanding come apart. Sonnet Blind reached 0.825 and stated its fixed operating cost as $19{,}400 \pm 700$ per MW-year against a true 14,371, a skill of −2.52: it had folded the year-end decommissioning accrual into the running cost. Opus recovered the same figure to within $2. The proper score catches confident error that attainment never sees. The Briefed agents repeat the briefing and score 0.89 to 0.99, which checks the scorer end to end.

6.4 What the agents found in the environment

The pilot runs were, unintentionally, the most effective audit this environment has had. Agents surfaced nine defects that a specification, a property-based layer and several hundred unit tests had missed. Among them were an operating economy specified and never charged, so that an agent correctly concluded building was free; a scored figure the agent could not see, so that one reported a $1.7M gain against a scored −$14.9M; a data store, the mechanic the benchmark exists to measure, advertised and unreachable; and an interrupt that never woke an agent at a permit approval. An adversarial review of this paper against its code found fourteen more (Appendix A.8).

Nineteen of the twenty-three are defects of connection or of the contract the agent reads, not of computation: a component correct and tested in isolation and never wired to what it was built for, or described to the agent as something it was not. Unit and property tests verify components against their own contracts and are poorly placed to catch that class; an agent optimising against the whole system experiences only the composition. We would recommend an adversarial agent pass as a standard step before publishing numbers from any environment. Every defect is pinned by a regression test. Twenty-two were fixed before

the reported runs; the twenty-third, found during them, affected every run on both tracks and is fixed after. Each run records the behaviour it was played under, and the released scorer replays all six exactly.

7 Implementation and validation

The kernel is a Rust workspace of thirteen crates with no unsafe code, 440 tests, and a verification hierarchy separating what is guaranteed by construction from what is proved, generated, or exemplified. Of 62 declared stochastic processes, 41 are drawn by this release; the registry table is regenerated from an audit of the source, so it cannot claim a process the kernel does not use. Nothing in it is fitted to observed outcomes, so agreement with external references is evidence, not tautology: solar position matches pvlib (Holmgren et al., 2018) to 0.363° in zenith and clear-sky irradiance to 0.84%, inside a 2% gate set in advance; wind capacity factors match the standard empirical relation within 0.06 on every testable seed; and wake losses reproduce the practitioner's rule without being fitted to it, sixteen machines on a four-diameter grid losing 15.9% against 5.8% for the same machines strung across the prevailing wind. The generator is pinned to the Random123 known-answer vectors (Salmon et al., 2011), the normal quantile agrees with scipy to fifteen significant figures including the 10^{-300} tail, and every agent run in this paper replays to a bit-identical state chain and score on x86-64 and ARM. That is a direct test of the floating-point argument in Section 4, not an argument for it. Appendix A.10 records three places where building the kernel corrected the design, each an error of specification, not of code.

8 Limitations

The served game is a subset of the designed one. Six of twelve verbs are served (Table 4): an agent cannot lease, finance, contract, hedge or exit a position, has no credit facility, and is finished by an overrun it cannot pay. Hydropower, equipment failures, inflation and interest rates are declared and not in play, and 21 of 62 processes are not yet drawn. Refusal risk depends on distance to a village alone, and export is capped at the sum of a world's connections without routing or losses. Every result describes an all-equity solar and wind developer on a two-year horizon.

The appraisal is a modelling choice, and it is the score. It values plant on its technology's trailing-year record, blends in carrying value while a year of evidence accrues, and gives unfinished plant nothing (Appendix A.7); the last makes the short track a test of whether an asset can exist in time.

The evidence base is a pilot, from one model family. Three Claude models, one seed and one run per track, driven by a general-purpose coding agent with a confined shell instead of a standardised harness, with actions guided but not capped. Every reported run played with a permit-decision interrupt that fired only on refusals, and our progress checks appended a free status read to four logs, which two agents noticed. Neither the human study nor agent runs on the ten-year track, which learning trajectories need, have been made.

Calibration and portability are incomplete. 31 of 62 processes are not yet traced to a cited source, and replay is verified on ARM and x86-64 under macOS only.

9 Conclusion

TerraWatt is built so that failing to understand a world and failing to plan within it leave different fingerprints on identical worlds, and the pilot shows that they do: the gap between knowing a world's parameters and not knowing them differs in size and even in sign between models. It also showed how easily an environment measures something other than it claims, and that the agents under test are among the best instruments for finding out.

AI use statement

This work used generative AI in three roles, which we separate because they carry different risks.

Tasks with required disclosure. We used generative AI (Claude) to help develop the environment's design and conceptual framework, to formulate and check its mathematical constructions (the stateless moving-average realisation of autoregressive processes, the appraisal, the scoring rules and the power analysis), and to write the kernel that generates every synthetic world and every baseline result. The agents evaluated in Section 6.3 are themselves language models; that is the research content, not an aid. We did not use generative AI to fabricate or alter any reported number: every table and figure is regenerated from the released kernel, and every agent score by replaying a released action log.

Tasks with recommended disclosure. Generative AI drafted and edited the paper, wrote the figure scripts, searched and summarised related work, and conducted an adversarial review of the paper against its code, which found fourteen of the defects in Appendix A.8.

Review and responsibility. We reviewed all AI-assisted work. Each mathematical construction and each defect fix is pinned by a test (442 in all); the results replay bit for bit across two architectures; every citation was checked against its primary source; and where an agent's inference about the environment could not be confirmed against the source, it is not reported as a finding. We take full responsibility for the content of this paper, including any errors.

Ethics statement

No human subjects took part and no personal data was collected: every agent is a language model playing a simulated company in a synthetic world. The human baseline the design describes has not been run, and would need ethics review before it is. We see no dual-use risk in a simulated renewable-energy developer.

Reproducibility statement

Section 4 and Appendix A.3 describe the constructions that make every run deterministic across machines, and Appendix A.7 the valuation that produces every score. The supplementary material contains the anonymised kernel with its 442 tests; the baseline sweep and trace binaries; the arena scaffold agents played through, with both briefs; the six agent action logs, which the released scorer replays to their exact state chains; and the scripts that regenerate every table, figure and statistic, listed with their commands in bench/README.md.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In Advances in Neural Information Processing Systems, 2021.

ARC Prize Foundation. ARC-AGI-3: A new challenge for frontier agentic intelligence. arXiv preprint arXiv:2603.24621, 2026.

Axel Backlund and Lukas Petersson. Vending-bench: A benchmark for long-term coherence of autonomous agents. arXiv preprint arXiv:2502.15840, 2025.

James Carroll, Alasdair McDonald, and David McMillan. Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines. Wind Energy, 19(6):1107–1119, 2016.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In International Conference on Machine Learning, 2020.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. MineDojo: Building open-ended embodied agents with internet-scale knowledge. In Advances in Neural Information Processing Systems, 2022.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378, 2007.

Danijar Hafner. Benchmarking the spectrum of agent capabilities. In International Conference on Learning Representations, 2022.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730):1999–2049, 2020.

William F. Holmgren, Clifford W. Hansen, and Mark A. Mikofski. pvlib python: a python package for modeling solar energy systems. Journal of Open Source Software, 3(29):884, 2018.

Jack Hopkins, Mart Bakler, and Akbir Khan. Factorio learning environment. arXiv preprint arXiv:2503.09617, 2025.

Ronald A. Howard. Information value theory. IEEE Transactions on Systems Science and Cybernetics, 2(1):22–26, 1966.

Pierre Ineichen and Richard Perez. A new airmass independent formulation for the Linke turbidity coefficient. Solar Energy, 73(3):151–157, 2002.

Niels Otto Jensen. A note on wind generator interaction. Technical Report Risø-M-2411, Risø National Laboratory, 1983.

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, 2024.

Heinrich Küttler, Nantas Nardelli, Alexander H. Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The NetHack learning environment. In Advances in Neural Information Processing Systems, 2020.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In International Conference on Learning Representations, 2024.

Antoine Marot, Benjamin Donnot, Camilo Romero, Balthazar Donon, Marvin Lerousseau, Luca Veyrin-Forrer, and Isabelle Guyon. Learning to run a power network challenge for training topology controllers. Electric Power Systems Research, 189, 2021.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. Management Science, 22(10):1087–1096, 1976.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, et al. BALROG: Benchmarking agentic LLM and VLM reasoning on games. In International Conference on Learning Representations, 2025.

John K. Salmon, Mark A. Moraes, Ron O. Dror, and David E. Shaw. Parallel random numbers: As easy as 1, 2, 3. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '11), 2011.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In International Conference on Learning Representations, 2021.

Hao Tang, Darren Key, and Kevin Ellis. WorldCoder, a model-based LLM agent: Building world models by writing code and interacting with the environment. arXiv preprint arXiv:2402.12275, 2024.

José R. Vázquez-Canteli, Jérôme Kämpf, Gregor Henze, and Zoltan Nagy. CityLearn v1.0: An OpenAI gym environment for demand response with deep reinforcement learning. In Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys), 2019.

Michael J. Wichura. Algorithm AS 241: The percentage points of the normal distribution. Journal of the Royal Statistical Society, Series C (Applied Statistics), 37(3):477–484, 1988.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, et al. TheAgentCompany: Benchmarking LLM agents on consequential real world tasks. arXiv preprint arXiv:2412.14161, 2024.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In Advances in Neural Information Processing Systems, 2022.

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.

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, 2024.

A Appendix

A.1 The registry

Table 3 lists every stochastic process the kernel declares, with its class, distribution and calibration status. It is generated from the kernel, not transcribed, so it cannot drift out of step with the code it documents.

Table 3: The complete registry: 62 stochastic processes. Class is L latent, E exogenous, M measurement, N endogenous. Drawn marks the 41 the current kernel samples; the other 21 are declared for later versions and have no effect on any run reported here. Src marks the 31 whose parameters are traced to a cited source; the rest are of the right shape and magnitude but not yet calibrated.

Id	Class	Process	Distribution	Drawn	Src
100	L	prevailing_wind_direction	von Mises(κ 2.5)	✓	✓
101	L	weibull_shape_k	TruncNormal(2.00, 0.25) on [1.30, 3.20]	✓	✓
102	L	weibull_scale_c	LogNormal(median 8.0000, σ 0.18)	✓	✓
103	L	shear_exponent_alpha	Beta(4, 4) on [0.08, 0.40]	✓	✓
104	L	diurnal_wind_phase	Bernoulli(0.35)	✓	✓
105	L	turbulence_intensity	Beta(3, 6) on [0.06, 0.24]	✓	–
106	L	wake_decay_constant	Normal(0.0750, 0.0120)	✓	✓
107	L	seasonal_wind_amplitude	Beta(3, 4) on [0.00, 0.45]	✓	–
110	E	hourly_wind_innovation	OU(ρ 0.930)	✓	✓
111	E	synoptic_wind_regime	Markov, 3 states, dwell 52 h	✓	–
200	L	panel_degradation_rate	LogNormal(median 0.0050, σ 0.40)	✓	✓

Id	Class	Process	Distribution	Drawn	Src
201	L	light_induced_degradation	Normal(0.0150, 0.0050)	✓	✓
202	L	soiling_accumulation_rate	Uniform(0.0005, 0.0035)	✓	✓
203	L	module_temperature_coefficient	Normal(-0.0035, 0.0003)	–	–
204	L	vendor_nameplate_optimism	Normal(1.0200, 0.0250)	–	–
205	L	inverter_efficiency_offset	Normal(0.0000, 0.0040)	–	–
210	E	clearness_index	Markov, 3 states, dwell 14 h	✓	✓
211	E	ambient_temperature	OU(ρ 0.960)	✓	–
300	L	catchment_runoff_coefficient	Beta(4, 4) on [0.25, 0.75]	✓	✓
301	L	baseflow_index	Beta(3, 3) on [0.15, 0.75]	✓	–
302	L	reservoir_seepage_rate	LogNormal(median 0.0004, σ 0.60)	–	–
303	L	upstream_catchment_area	LogNormal(median 260.0000, σ 0.55)	✓	–
310	E	annual_precipitation	Cyclic(σ 0.22, 3–7 yr)	✓	✓
311	E	snowmelt_onset_day	Normal(92.0000, 12.0000)	–	–
312	E	open_water_evaporation	LogNormal(median 3.2000, σ 0.30)	✓	–
400	N	turbine_gearbox_failure	Weibull hazard(1.7, MTBF 8 yr)	–	✓
401	N	turbine_generator_failure	Weibull hazard(1.4, MTBF 11 yr)	–	–
402	N	turbine_blade_damage	Weibull hazard(2.1, MTBF 14 yr)	–	–
403	N	inverter_failure	Exponential(0.100)	–	✓
404	N	transformer_failure	Exponential(0.025)	–	–
405	N	cable_fault	Exponential(0.040)	–	–
406	N	hydro_unit_outage	Exponential(0.090)	–	–
407	N	repair_duration	LogNormal(median 4.0000, σ 0.90)	–	–
500	N	cost_overrun_solar	LogNormal(median 1.0400, σ 0.11)	✓	✓
501	N	cost_overrun_wind	LogNormal(median 1.1200, σ 0.20)	✓	✓
502	N	cost_overrun_hydro	LogNormal(median 1.4500, σ 0.55)	✓	✓
503	N	permit_delay_solar	LogNormal(median 4.0000, σ 0.45)	✓	✓
504	N	permit_delay_wind	LogNormal(median 9.0000, σ 0.55)	✓	✓
505	N	permit_delay_hydro	LogNormal(median 14.0000, σ 0.60)	✓	✓
506	N	permit_delay_substation	LogNormal(median 10.0000, σ 0.50)	✓	–
507	N	permit_rejection_near_village	Logistic in distance(50% at 700 m)	✓	✓
508	N	construction_schedule_slip	LogNormal(median 1.0800, σ 0.25)	✓	–
509	N	geotechnical_surprise	Bernoulli(0.18)	–	–
600	E	gas_price	Jump diffusion(κ 0.35, σ 0.45)	✓	✓
601	E	coal_price	Jump diffusion(κ 0.25, σ 0.30)	✓	–
602	E	carbon_price	Jump diffusion(κ 0.15, σ 0.35)	✓	–
603	E	regional_renewable_buildout	Logistic(mid 6.5 yr, rate 0.42)	–	✓
604	E	technology_capex_decline	Learning curve(16%/doubling)	–	✓

Id	Class	Process	Distribution	Drawn	Src
605	E	regional_demand_growth	Normal(0.0180, 0.0110)	–	–
606	E	inflation	OU(ρ 1.000)	–	–
607	E	base_interest_rate	OU(ρ 1.000)	–	–
608	E	thermal_fleet_outage	Bernoulli(0.04)	–	–
609	L	price_elasticity_to_ renewables	Beta(3, 3) on [0.25, 0.95]	✓	–
650	L	fixed_om_solar	LogNormal(median 16000.0000, σ 0.20)	✓	✓
651	L	fixed_om_wind	LogNormal(median 40000.0000, σ 0.22)	✓	✓
652	L	variable_om_renewable	Beta(2, 3) on [0.00, 4.00]	✓	–
700	M	met_mast_error	Normal(0.0000, 0.0180)	✓	✓
701	M	lidar_error	Normal(0.0000, 0.0250)	✓	–
702	M	pyranometer_error	Normal(0.0000, 0.0200)	✓	✓
703	M	stream_gauge_error	Normal(0.0000, 0.0600)	✓	–
704	L	reanalysis_site_bias	Normal(0.0000, 0.0450)	✓	✓
705	M	production_meter_error	Normal(0.0000, 0.0050)	–	✓

A.2 The twelve verbs

Table 4 gives the action space and Table 5 the priced data store. Both are generated from the kernel.

Table 4: The action space the design defines, and how much of it the served interface offers in this release. Everything difficult in TerraWatt is meant to come from deciding what to do, never from discovering how to express it.

Verb	Effect	Served
lease	Take a parcel under lease for a term, at the quoted rent.	no
build	Place blocks: kind, position, vendor model, configuration. Enters permitting.	solar and wind only
demolish	Remove an asset, recovering salvage value.	no
configure	Change an asset's settings: tilt, yaw offset, DC/AC ratio, gate position.	no
set_policy	Install or amend a standing policy the kernel executes hourly.	curtailment floor only
advance	Run time forward, up to ninety days, with interrupt conditions.	yes
inspect	Query state: assets, books, meters, prices, news, catalogue. Free.	yes, as observe and status
buy_data	Purchase a dataset or subscription from the data store.	4 of 9 products
survey	Commission a study or pilot: geotechnical, environmental, resource.	no
contract	Sign a PPA, hedge, certificate sale, insurance or maintenance agreement.	no
finance	Draw or repay debt, issue or return equity, refinance.	no
report_belief	State a probabilistic belief about a hidden quantity. Scored, never rewarded.	yes

Table 5: The data store. Prices and delays are calibrated to real consultancy and instrumentation practice; a thorough campaign comes to roughly 1.6% of a mid-sized build programme. Four products are served in this release; the others are specified and priced, and observe does not list them.

Product	Upfront	Annual	First data	Served
RegionalReanalysis	$12k	–	immediate	✓
MetMast	$180k	$22k	3 weeks	✓
Lidar	$310k	–	1 week	✓
PyranometerStation	$28k	–	1 week	✓
StreamGauge	$45k	–	2 weeks	–
GeotechnicalSurvey	$150k	–	2 months	–
EnvironmentalStudy	$200k	–	6 months	–
ForecastService	–	$40k	immediate	–
MarketAnalytics	–	$75k	immediate	–

A.3 Two constructions behind action invariance

Autocorrelated processes without state. Several processes are specified as autoregressive: the hourly wind innovation has a lag-one correlation near 0.93, cloud persists, fuel prices mean-revert over weeks. Written the obvious way, $z_t = \rho z_{t-1} + \sqrt{1-\rho^2}\,\varepsilon_t$ is a recursion, and a recursion needs history.

We use the equivalent moving-average representation,

$$z_t \;=\; \frac{\sum_{j=0}^{J-1} \rho^{\,j}\, \varepsilon_{t-j}}{\sqrt{\sum_{j=0}^{J-1} \rho^{\,2j}}}, \tag{2}$$

in which each innovation is addressed by its own key. Any t evaluates directly, with no history and no state, and the sum may be taken in any order. Normalising by the realised weight norm makes the variance exactly one for any truncation J and leaves the lag-one correlation short of ρ by about ρ^{2J}.

The construction has a limit worth stating, because we met it. It represents only processes whose memory is short against the window. Gas prices revert over roughly six weeks; at hourly resolution that is seven hundred steps, and a 256-term window discarded most of the process: measured correlation one month out collapsed to 0.035 against a theoretical 0.36, taking the jump behaviour with it. Widening the window to thousands of hourly terms would cost on the order of 10^9 draws per episode. Fuel and carbon are daily-settled commodities, and no decision in the benchmark turns on the gas price at 3 a.m., so they are drawn per day over a window of 1,024 days. Every time constant is then small against the window, at a thousandth of the cost.

Endogenous draws and build order. An asset's endogenous draws are keyed by a hash of its identity (kind, position, model, and the tick its order was placed), and deliberately not by a sequence number. Keying by sequence would quietly reintroduce a dependence on build order: the same turbine, in the same place, ordered at the same hour would draw a different permit decision depending on what else had been ordered first. With identity keys, build order changes outcomes only through real mechanisms such as timing, cash and what was learned in between, and never through bookkeeping.

A.4 The reanalysis bias, and how it is recovered

One product deserves attention, because it is the environment's worked example of the discovery problem. Purchased reanalysis carries a fixed multiplicative bias against the true local resource, on top of ordinary noise. The bias is a property of the world, not of any individual reading. Averaging more reanalysis therefore converges on the biased answer and never on the true one. The bias is recoverable only by holding reanalysis against an instrument standing on the site, which is exactly the measure–correlate–predict procedure

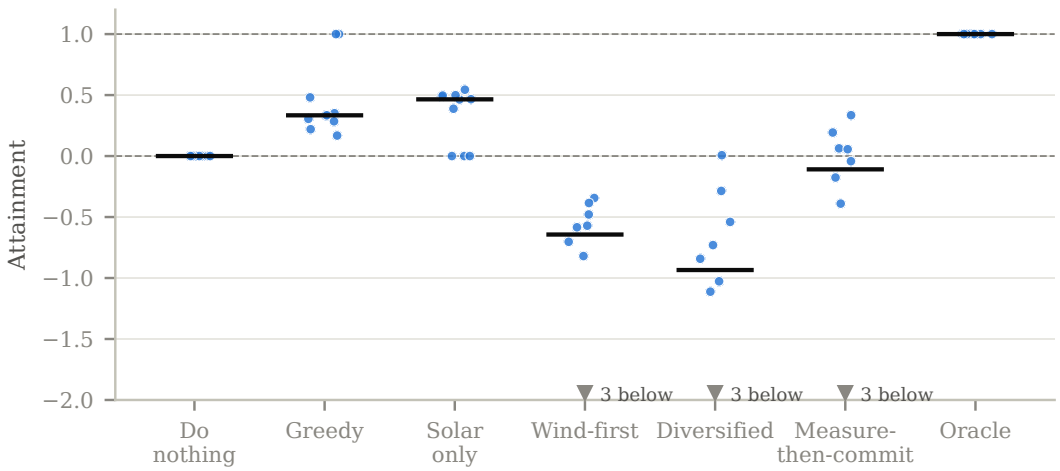

Figure 2: Every baseline on each of the ten seeds, one point per seed, with the median as a bar. Floor and oracle are 0 and 1 per seed, so the spread in a column is what remains once each world's richness is divided out. The axis is cut at −2; the three seeds below it in each wind-led column are the nearly flat worlds, where attainment divides by almost nothing.

a wind developer runs. In our implementation, forty thousand reanalysis readings against forty thousand mast readings recover a seed's bias to within 0.03.

A.5 The players in full

Three kinds of player appear, and the headline score is a ratio between two of them. A baseline is a short hand-written program that follows the same recipe every time, with no reasoning and no reaction to what it sees; baselines mark out the range of outcomes a world allows. Do-nothing holds the opening $40M in cash and is the floor, so any player below it has actively destroyed value. Greedy-cheapest buys the cheapest megawatt immediately, wherever it finds room. Solar-only sites and angles solar for the hemisphere it is actually in. Wind-first measures for a year, then builds on high ground away from villages. Diversified does both, and measure-then-commit instruments first and decides after.

The oracle is not a clever agent but a search with hindsight. It plays 112 candidates to the horizon on the realised world and keeps whichever ends worth most: build programmes swept over technology mix, scale, panel model, staging and reinvestment from the plant's own cash flow, each with a curtailment floor at the plant's marginal cost, and every baseline policy, so no baseline can beat it. It is still a search, not a proof of optimality. It lower-bounds what is achievable, and a player scoring above 1.0 is telling us about the search, not about itself; two did (Section 6.3).

An agent receives the goal, the tool descriptions, the handbook of priors and whatever it observes, and on the Briefed track the true value of every active latent parameter as well. It decides what to look at, what to measure, what to build and when to wait, through the interface a human would play through, so all three kinds of score are commensurable.

A.6 Discovery gap

The gap is the Briefed attainment minus the Blind attainment for the same model on the same seed. A large positive gap marks an agent limited by what it can find out; a gap near zero, one limited by planning; a negative gap, an agent that plans worse when told more, as Haiku did on seed 1 by committing nearly all its cash once it knew the world was good for solar. The pilot measures it once per model on one seed (Table 2); the protocol measures it on ten sealed seeds with repeats, and reads each gap against the headroom the Blind run left.

A.7 The appraisal

A two-year run ends with plant that has barely begun to earn, so its value must be appraised, and the appraisal is the score. Three constraints shape it. It must be observable, because the agent is shown it every turn and an appraisal that consulted the latent truth would hand it the answers. It must not apportion, because any rule that splits the company's earnings among its assets makes adding one dilute the rest. And it must count each cost once.

Each technology keeps a trailing year of hourly records: the margin its plant earned (revenue less the variable and fixed O&M that plant incurred, and nothing else) and the MW it had in service. Its demonstrated rate is margin per MW-hour. An asset's annual pre-tax cash is its own nameplate at that rate; the caretaker capitalises it over the asset's remaining book life at a 7% WACC with 1% annual decline, paying 25% tax on the margin less the asset's projected tax depreciation, and never runs a plant worth less than abandoning it. A technology with $h < 8760$ hours of evidence contributes the weighted value $\frac{h}{8760}V_{\text{earned}} + (1 - \frac{h}{8760})V_{\text{carried}}$, where the carried value is book value less the present value of removal cost not yet provided for; a year is the window because it is the period over which the resource's seasonality averages out. The accrued decommissioning provision is subtracted at company level and only the unprovided remainder per asset, so each unit of removal cost is counted once. Plant under construction is worth nothing, and the brief says so.

The rule this replaces apportioned the company's trailing net cash by capital cost, drew that cash from the ledger's operating costs, which included data purchases and the decommissioning accrual, and annualised it over the age of the oldest asset. Each fault biased the score against something the benchmark measures: building late, buying information, and building in waves. Tests pin the replacement's additivity, its continuity in evidence, and the invariance of enterprise value to a data purchase.

A.8 The defects agents and review found

The first seven defects below were surfaced by pilot agents playing the Blind track, and are reported with the agent's own diagnosis; Table 6 lists those an adversarial review of this paper against its code found after them. We record them because the pattern seems general, and because a benchmark paper that hides its own repair history is asking reviewers to take its numbers on trust.

The operating economy was specified and never charged. The design calls for fixed O&M per MW-year, variable O&M per MWh, book and tax depreciation, a decommissioning reserve and corporation tax. The kernel posted three kinds of ledger entry, cash to construction twice and revenue to cash, and charged none of it. No test failed, because every test asserted properties of the accounting machinery, and none asserted that the machinery was called. The strongest agent in the pilot diagnosed it in one run, reporting that capital was capitalised one-for-one with no offsetting charge and concluding that converting cash to plant was therefore free; it then deployed 62 MW and said outright that maximum immediate build was provably dominant. It was right, and that is the problem: with no cost of operation, returns to scale have no upper bound, and nothing about discovering a site's resource or its capture-price erosion can pay.

The scored objective was not observable. Scoring used a caretaker appraisal; status reported book equity, and nothing said which was which. One agent closed its run reporting a gain of \$1.7M against a scored result of −\$14.9M, having optimised the only figure it could see. The appraisal's own earnings term, meanwhile, was standing in a flat 25% of revenue "for running costs the kernel does not yet book", which was the same defect seen from the other end.

The anchor was narrower than the action space. Described above: an agent scored 11.1 against an oracle whose solar sweep stopped at a fifth of the capacity the agent could buy.

The blind condition did not survive a shell. The process registry is compiled into the server binary, so strings returned every process name, its literature citation and its dependency hint, including phrasings written for the reader of the design document, such as "the agent must discover per-vendor truth". An agent ran exactly that, recognised what it had found,

reported it, and stopped without acting on it. Documentation strings are now compiled out of the served binary. The distributions remain as numeric constants, so this raises the cost of introspection without eliminating it, and we claim a practical barrier, not a guarantee.

A documented guard was not implemented. Both build verbs are documented as refusing an order when the cash is not there. Neither checked. Two of the three completed runs in one batch died to it, one committing $167M against $40M of cash and reaching insolvency at hour 1. An agent that reads a tool contract is entitled to rely on it; this was a trap, not a test of prudence.

The discovery mechanic was unreachable. The instruments, their latencies, their per-instrument error draws and the registry entries behind them were all written, and nothing was connected: no verb created an instrument, the observation's instrument list was hard-coded empty at every call site, and the reading function was called from nowhere. Two agents found this independently. One enumerated four plausible verb names and had all four refused, another concluded the data store was "bait, not a shop". The measure-then-commit baseline had been paying for three data products and receiving nothing, which made it diversified minus a fee and guaranteed it last place. The one baseline named for the benchmark's central question was the one baseline unable to ask it.

The appraisal penalised building late, by construction. Earnings were apportioned to assets by capital cost, so a newly commissioned asset took earnings from the plant that produced them. An agent measured the effect, net appraised value per MW rising from about zero at commissioning to $705k at 0.69 years, concluded that "capital committed late scores worse than cash held", and stopped deploying $13M. It was right about the rule and the rule was wrong. The appraisal was rebuilt (Appendix A.7); the review then found three further faults in the rule it replaced, which the rebuild also removes.

Project cash was settled at the order. On the first runs with the repaired score, an agent on the Briefed track was insolvent at hour 1: the kernel charged each project's realised cost, overrun included, the moment it was ordered, so an order within the quote could bankrupt a company before a permit was filed. The same shortcut let the cash movement, a 2% fee instead of the quote, announce a refused permit months before its decision, and the pipeline showed a completion date for approved projects and none for refused ones. Cash now follows the project: quote at the order, refund on refusal, overrun on completion. That batch was stopped and discarded; its logs are kept.

A permit decision woke no one. An agent in the reported runs noted that its permit_decision interrupts never fired. They fired on refusals only: the check read the list of refused permits and not the approved ones, and solar is almost never refused, so every agent that asked to be woken by a decision slept through its approvals and read them from the pipeline instead. All six reported runs used the interrupt. It is fixed after them, with a test; each run's configuration records the old behaviour, and the released scorer replays it under that behaviour, reproducing every run's state chain exactly.

Table 6: Defects found by adversarial review of the paper against its code, after the pilot. Connection: built or specified, and never wired to the served game. Interface: the contract the agent read was wrong. Each is fixed and pinned by a test.

Defect	Kind	Consequence
Replay scorer re-implemented dispatch and skipped buy_data	Interface	Any run that bought data replayed to a position never held
State-hash chain claimed and absent	Connection	Replay could not be checked call by call
Priors never shown to agents	Interface	Belief skill scored against a prior the agent had not seen
Data purchases capitalised as running costs	Connection	A $180k mast removed several times its price from the score
Per-vendor degradation not simulated	Connection	The introduction's motivating example did not exist
Price elasticity, soiling drawn and never read	Connection	Two of the named hidden parameters were inert
Land, ground and rotor clearance unenforced	Connection	A 45 MW array sat on one cell; rotors 20 m apart
Grid export limit never applied	Connection	200 MW could be sold through a 66 MW connection
maintenance_tier stored, never used	Interface	An agent reverted it after blaming it for a cost
Unaffordable orders reported as "no such panel"	Interface	A valid order's refusal gave the wrong reason
Unserved data products listed with prices	Interface	Agents paid turns to learn five were refused
Oracle sweep ordered its MW at every site	Connection	"40 MW over 18 sites" asked for 720 MW
Oracle test compared cash, not value	Connection	Passed only while the biased score kept the oracle idle
Tool text said solar builds in "a few months"	Interface	Measured median is 8.4 months; wind 26.5

A.9 Evaluation protocol in full

Attainment alone hides mechanism, so five further measures are reported: bankruptcy rate, as a first-class number instead of something absorbed into a mean; capture-price ratio, which separates when an agent produces from how much; capital efficiency; information efficiency, being value created per unit spent on data; and calibration.

Calibration. Agents may state a belief about a named hidden quantity at any time, as a mean and a standard deviation. The action changes nothing in the world and contributes nothing to the primary score, so an agent has no reason to state anything but what it believes. Beliefs are scored with the continuous ranked probability score, which stays finite when an agent is confidently wrong and is expressed in the units of the quantity; the logarithmic score is reported alongside as a sharper measure of overconfidence. Both are verified proper numerically: expected loss is minimised at the true mean and the true spread, and penalises overconfidence and timidity alike.

Scores are reported as skill against the published prior, because priors differ in width and a raw average would reward an agent for being asked easy questions. Repeating the handbook's prior scores exactly zero skill.

Protocol and statistics. Each model plays each of ten seeds, five times per cell. The unit of analysis is the seed-matched paired difference, never the unpaired mean: because the exogenous realisation is identical across agents on a seed, pairing removes the dominant variance component. All aggregation resamples seeds, not runs; treating five repeats as five worlds would narrow intervals by roughly forty per cent and publish confidence the data does not support.

Table 7: The six tracks. Each pair isolates one capability.

Track	What changes	What it isolates
Blind (primary)	Nothing; handbook and priors only	End-to-end competence
Briefed	Latent parameters revealed at $t = 0$	Removes discovery
Oracle-weather	Latent and exogenous revealed	Removes all uncertainty
Instrument-free	Data store disabled	Inference from own operations
Short (2 yr)	Development-phase horizon	Cheap iteration
Long (20 yr)	Extended horizon	Decade-two reasoning

Power is reported as a requirement, not a promise. Detecting a difference of 0.08 in mean attainment at 80% power over ten seeds requires the paired standard deviation across seeds to be at most 0.090. Repeats shrink run-to-run noise but cannot reduce the seed-by-model interaction, which sets the floor. If the pilot finds a run-to-run spread near 0.20, five repeats give 0.140 and realised power falls below one half; reaching 80% would then need approximately eighteen repeats or twenty-four seeds. We will report achieved power, not projected power. Bankruptcy rate, a low-base-rate binary outcome, is underpowered at ten seeds and is reported as an interval.

Contamination. Ten seeds are public and fully documented, including their latent parameters, so anyone can study and develop against them. A further ten are sealed: their parameters are committed as a published hash, runs execute against a hosted evaluator, and the parameters are revealed only when the version retires. Headline results are reported on sealed seeds. This is the only defence available to a benchmark whose worlds are, by design, described in exhaustive detail in a paper that will itself enter the next pre-training corpus.

A.10 Corrections the implementation forced on the design

Building the kernel corrected the design in several places, three of which are worth recording because they were errors of specification, not of code. The catchment was initially taken from on-map drainage alone, producing rivers of 0.01–0.30 m^3/s, under 20 kW at 40 m of head. The arithmetic is unavoidable: the world is 6.55 km^2 and a megawatt-scale scheme needs of order 100 km^2, so the contributing basin must extend upstream of the map. The runoff coefficient could not remain a free latent parameter, because a cold wet basin cannot evaporate enough to shed only a third of two metres of rain; it is now a target the catchment aims at, bounded by climate. And a reservoir computed by flooding every cell below the crest gave a dam sited on a ridge an apparent 2.1 billion m^3, because a dam blocks a river at one point and holds back only what drains to that point.

A.11 Tracks

A.12 Worked example

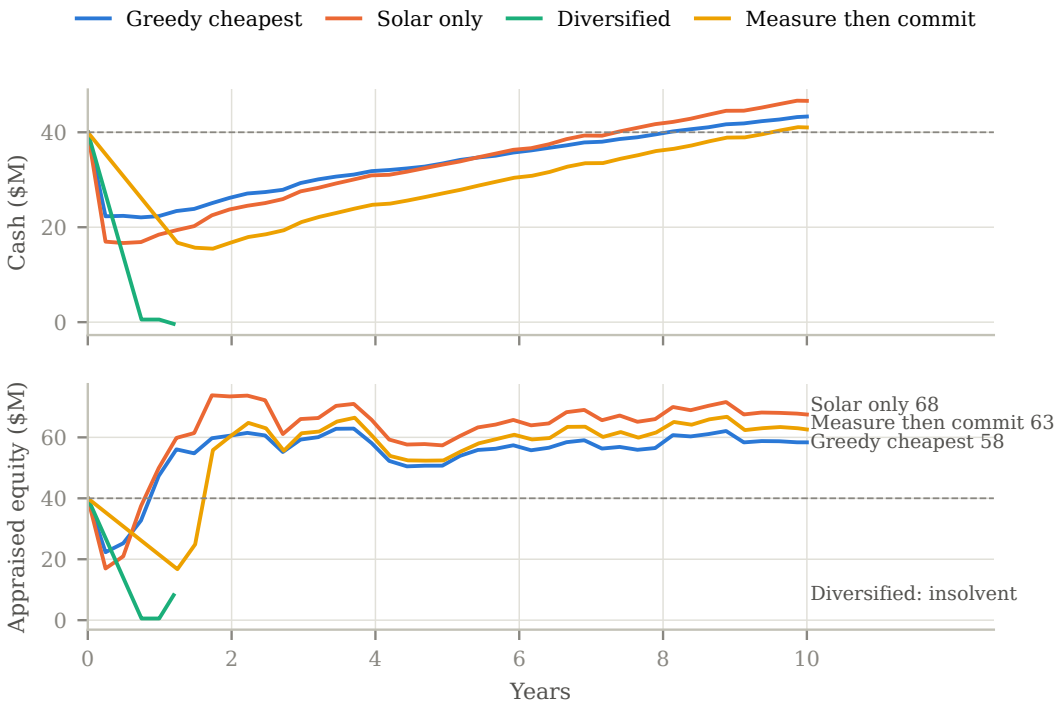

Figure 3: Four baselines on seed 1 over the full ten-year horizon, traced by the kernel itself. The quote leaves at the order and plant earns months later, so cash falls to a trough before assets enter service. Solar-only finishes highest, at $67.5M. Diversified commits its cash in quotes across wind and solar with nothing in reserve, cannot pay the overruns that fall due on completion, and is insolvent in its second year: on a long horizon the recipe's problem is not that wind pays too late but that it planned no margin for error.