GCAM-GYM: A BENCHMARK FOR AGENT-BASED CLIMATE POLICY SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Comparing climate-policy search methods requires accounting for differences in policy control, available observations, feedback timing, and simulator budgets. We introduce GCAM-Gym, a Gymnasium-compatible environment that supports full-policy evaluation and stepwise control through a common interface and exposes the components of the policy objective. We compare seven search and control systems in three settings: target attainment and cost, cost-temperature trade-offs, and one-step delayed feedback. Each system is evaluated with ten seeds per setting, for a total of 3,210 attempted GCAM evaluations. Four systems share a budget of 24 attempts per seed; among these, CMA-ES achieves the highest interquartile mean (IQM) of the best returns from completed policy evaluations in all three settings. We release the evaluation archive and analysis tools to reconstruct the reported comparisons without running GCAM. Code available at https://anonymous.4open.science/r/ICLR-GCAM-GYM.

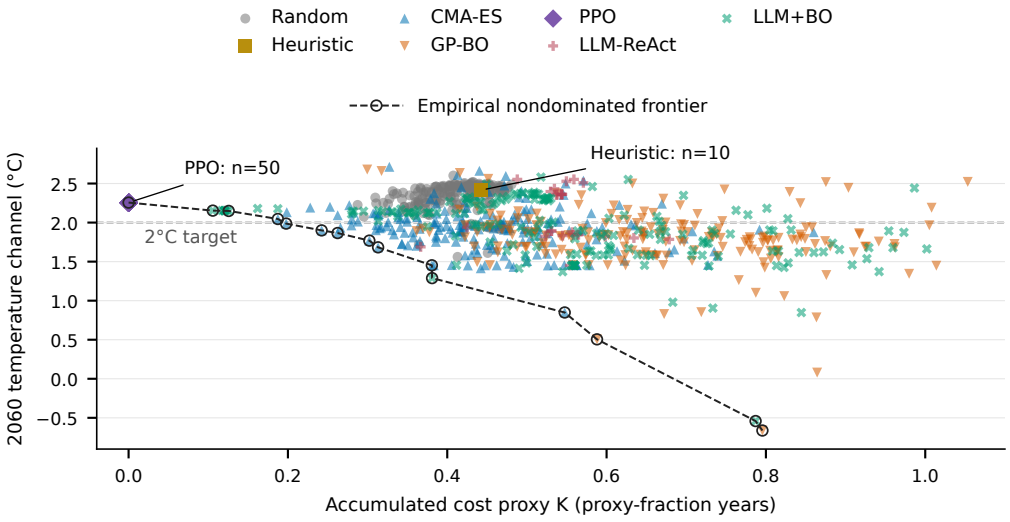

Figure 1: **Different policy-search methods explore different cost–temperature outcomes.** Each point is a completed policy evaluation in GCAM-Gym, colored by method. The horizontal axis shows an accumulated policy-cost proxy; the vertical axis shows the simulator's temperature readout for 2060. Lower cost and temperature lie toward the bottom left; the horizontal dashed line marks the study's 2°C target. Connected open circles mark outcomes for which no other recorded point has both no greater cost and no higher temperature, with at least one strictly lower. This empirical frontier is not a validated policy optimum: measurement-affected negative temperatures are retained (Section 5.4). The 834 completed evaluations use method-specific budgets (Table 2); repeated outcomes overlap.

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.

1 INTRODUCTION

Climate-policy design involves decisions whose consequences unfold over decades and interact across sectors. A carbon-price trajectory changes incentives in energy and land systems, with climate consequences mediated by emissions and the carbon cycle. Integrated assessment models provide a computational framework for studying these interactions. GCAM couples energy, water, land, climate, and economic systems (Calvin et al., 2019).

Searching for policies with a simulator can follow different interaction patterns. A schedule optimizer proposes a complete policy for evaluation, while a stepwise controller chooses actions as observations arrive. Comparing their performance requires specifying what each method controls, which observations it receives and when, and how many evaluation attempts it uses, including failed attempts. A benchmark for GCAM-based policy search therefore needs to support both interaction patterns and link proposed policies to executed trajectories, feedback histories, and evaluation outcomes.

We introduce **GCAM-Gym**, a Gymnasium-compatible benchmark environment for climate-policy search. Methods specify a global carbon-price trajectory, either as a complete policy or through stepwise actions, using a shared action decoder and GCAM backend. The environment returns structured observations of climate and energy-system responses, together with cost and distributional proxies. A feedback wrapper supports a one-step delay in the observations and rewards delivered to controllers, and the evaluation protocol counts both completed and failed attempts toward each system's budget. Individual reward components are returned alongside the scalar objective so that users can inspect their contributions to a policy's score. The accompanying archive links proposed policies, executed trajectories, feedback, and execution status; analysis tools reconstruct the reported comparisons without running GCAM.

We use GCAM-Gym to compare seven systems spanning fixed policies, schedule search, and observation-driven control: random search, fixed heuristics, CMA-ES, Gaussian-process Bayesian optimization, PPO pretrained on a mock backend, LLM-ReAct, and LLM-initialized Bayesian optimization. We examine whether these systems find target-reaching policies and at what cost, which cost–temperature trade-offs they attain, and how they perform under one-step delayed feedback. Each system is evaluated with ten seeds in each setting, using the configurations and budgets specified in Section 4.

Among the four systems evaluated with a common budget of 24 attempts per seed, CMA-ES achieves the highest interquartile mean of the best returns from completed policy evaluations in all three settings.

2 RELATED WORK

Climate-policy environments. Integrated assessment models connect economic decisions to climate outcomes (Nordhaus, 1992; 2017), providing one setting for machine learning in climate mitigation (Rolnick et al., 2022). GCAM provides the coupled simulator underlying this resource (Calvin et al., 2019). RICE-N studies climate negotiations, agreements, and cooperation among regional decision makers (Zhang et al., 2025). GCAM-Gym focuses on single-agent global carbon-price control. Its evaluation unit is a policy trajectory produced through a shared simulator interface, with time-indexed responses retained for inspection. Shared Socioeconomic Pathways describe alternative futures and their energy, land-use, and emissions implications (O'Neill et al., 2014; Riahi et al., 2017); robust decision making examines policies across uncertain futures (Lempert et al., 2006). These differ from our three evaluation settings. The AI Economist studies taxation with coadapting agents and a planner (Zheng et al., 2022), rather than single-agent carbon-price search.

Reusable optimization evaluation. Gym standardizes environment interfaces, while the DeepMind Control Suite and bsuite provide continuous-control and diagnostic tasks (Brockman et al., 2016; Tassa et al., 2018; Osband et al., 2020). COCO and the Black-Box Optimization Challenge provide complementary optimizer-evaluation settings (Hansen et al., 2021; Turner et al., 2021). HPOBench demonstrates how common interfaces and isolated configurations support reproducible optimizer comparisons (Eggensperger et al., 2021). GCAM-Gym applies this resource-oriented approach to policy trajectories in a coupled climate simulator. In addition to an objective value, an attempted

evaluation produces a sequence of actions, observations, reward components, and execution statuses. This record supports both endpoint comparisons and inspection of when feedback reaches a controller.

Search objectives and feasibility. Random search provides a nonadaptive reference (Bergstra & Bengio, 2012); surrogate-based global optimization and GP-based tuning motivate sample-efficient search (Jones et al., 1998; Snoek et al., 2012). TuRBO uses local models, and SCBO extends trust-region search to constraints (Eriksson et al., 2019; Eriksson & Poloczek, 2021). Explicit constraint modeling and safe exploration differ from assigning penalty scores to failures (Gardner et al., 2014; Sui et al., 2015); our adapters do not inherit their feasibility or safety guarantees. Displaying attained cost–temperature pairs likewise differs from running a Pareto-search algorithm such as NSGA-II (Deb et al., 2002).

Language models and delayed interaction. Chain-of-thought prompting studies intermediate reasoning, while Reflexion retains verbal feedback across trials (Wei et al., 2022; Shinn et al., 2023). OPRO generates candidates from evaluated solutions, and LLAMBO integrates language models into BO, including warm-starting (Yang et al., 2024; Liu et al., 2024). Our LLM+BO uses language-generated initialization followed by conventional GP search, rather than reproducing LLAMBO. Research on randomly delayed interaction explicitly models action and observation delays (Ramstedt et al., 2021); our comparisons remain descriptive, not evaluations of a delay-correcting algorithm.

Baseline methods and statistics. The suite uses established approaches: CMA-ES (Hansen, 2016), Gaussian-process optimization through BoTorch (Balandat et al., 2020), PPO (Schulman et al., 2017), and an adaptation of the ReAct interaction pattern (Yao et al., 2023). Their actual configurations are part of the resource, including initialization, pretraining, and failure feedback. For aggregate reporting, we use IQMs and seed-bootstrap intervals motivated by robust empirical evaluation (Agarwal et al., 2021); the independent-unit and aggregation conventions of this study are specified below. Prior work highlights implementation and run-level variability (Henderson et al., 2018; Engstrom et al., 2020); bootstrap resampling provides the underlying statistical foundation (Efron, 1979).

3 The GCAM-Gym Interface

3.1 State, observation, and action spaces

The environment implements the Gymnasium `reset`/`step` interface (Towers et al., 2025). The backend initializes the simulator and executes each period; the environment decodes actions, assembles observations, and returns rewards and episode-end flags (Figure 2A). We evaluate policies with GCAM 9.1 from an initial state in 2025 to an endpoint in 2060, using H decision periods of Δt years each. Table 1 lists the configuration. A stylized mock backend supports interface development and PPO pretraining; all reported policy evaluations use the native GCAM backend.

State and observation spaces. The internal simulator state $x_t \in \mathcal{X}$ comprises the coupled model variables and execution state. Agents receive an observation $o_t \in \mathcal{O}$: a dictionary of twelve climate, energy, cost, and distributional fields that flattens to 22 scalars (Table 4). These observations summarize the simulator response; their Markov sufficiency is not assumed. The solver-iteration field is a constant-zero placeholder, and execution status is returned separately. Controllers can use the dictionary, a flattened vector, or a fixed-key-order serialization for an LLM. Appendix A gives the declared bounds and field ordering.

Action spaces. The API action space is $\mathcal{A}_{\text{API}} = [0, 1]^4$, with coordinates for carbon price, low-carbon subsidy, coal phaseout, and carbon removal. Only carbon price affects the native simulator in this study, giving an effective normalized control space $\mathcal{A}_{\text{native}} = [0, 1]$. Nonzero inactive coordinates, which can occur in PPO outputs, are ignored by the native backend. The active coordinate a_t maps to an integer price $p_t \in \{0, \ldots, p_{\max}\}$ through

$$p_t = \text{round}\left[p_{\max} \frac{\exp(\beta a_t) - 1}{\exp(\beta) - 1}\right]. \tag{1}$$

Here $p_{\max}$ sets the price ceiling and β controls the curvature of the map. The backend treats p_t as a price per tonne of CO_2 and writes κp_t to the native carbon market, using the carbon-mass conversion

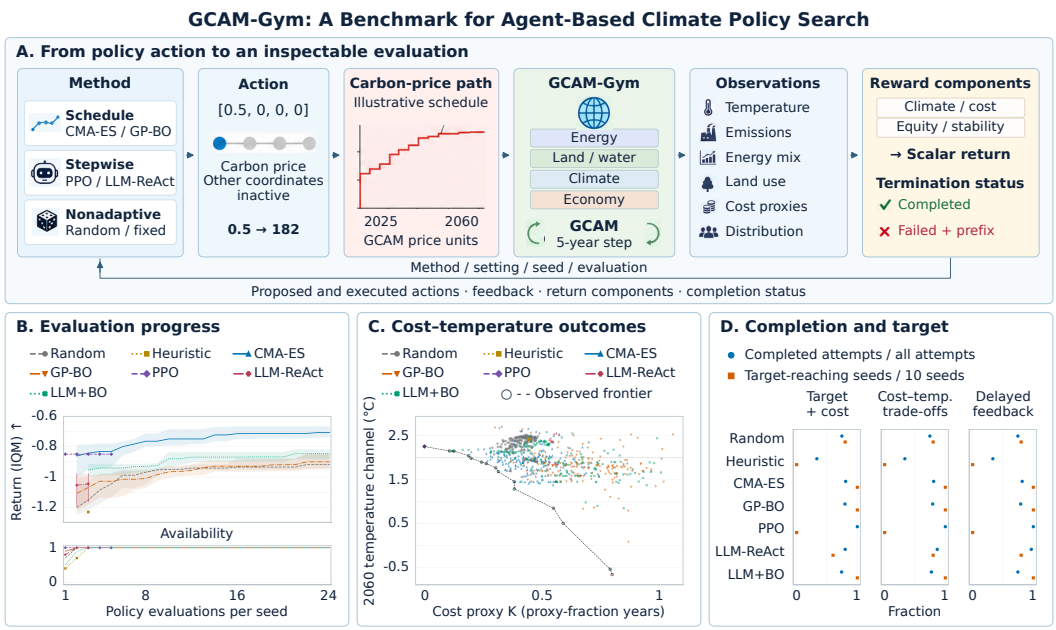

Figure 2: **GCAM-Gym links policy actions to inspectable outcomes. A** Shared evaluation interface; the price path is illustrative. **B** Target-and-cost progress: best-return IQM with pointwise 95% seed-bootstrap intervals, and the fraction of ten seeds with a completed policy. Failed evaluations count toward the budget; PPO is evaluated, not trained. **C** All 834 completed cost–temperature evaluations; lower-left means lower cost and temperature. The horizontal dashed line marks 2°C; connected circles show the observed nondominated frontier, not a validated optimum. Measurement-affected negative temperatures remain (Section 5.4). **D** Completion and target attainment across three settings. Budgets differ (Table 2); total computation is not matched.

factor κ in Table 1. The native market is documented in 1990 dollars per tonne of carbon. The interface performs no currency-year adjustment, and the monetary units of the input datasets have not been independently validated (Appendix A).

A schedule optimizer selects $(a_1, \ldots, a_H) \in [0,1]^H$ before execution; a stepwise controller selects each action from the available feedback. Both use the same decoder and period execution. Search-domain restrictions are system-specific: random search and LLM initialization use monotone schedules, while CMA-ES and subsequent BO proposals use the full normalized schedule domain (Appendix B).

Interaction and episode boundaries. Reset supplies the initial observation o_0. At decision $t = 1, \ldots, H$, the environment decodes the action, advances the simulator by Δt years, and returns o_t, reward r_t, execution status, and episode-end flags. An episode that reaches the configured endpoint has `truncated=True`. A reported solver failure sets `terminated=True` and returns a failure penalty; other execution failures are also recorded as failed attempts. Appendix A specifies the penalty. Section 3.3 describes the delayed-feedback variant of this interaction.

3.2 Objectives and Recorded Outcomes

For successfully executed periods, the scalar reward combines climate, cost, equity, and electricity-price stability components:

$$r_t = w_{\text{clim}} r_t^{\text{climate}} + w_{\text{cost}} r_t^{\text{cost}} + w_{\text{eq}} r_t^{\text{equity}} + w_{\text{stab}} r_t^{\text{stability}}, \qquad R = \sum_{t=1}^{H} r_t. \tag{2}$$

Table 1: Fixed interface and objective parameters used in the reported evaluations. Price values follow the native monetary convention described in the text; temperature thresholds apply to the recorded Hector channel.

Parameter	Meaning	Value
H	Decision periods per complete episode	7
Δt	Duration of each period	5 years
N	Regions in the emissions aggregation	5
$p_{\max}$	API carbon-price ceiling	1000 per tCO_2
β	Curvature of the action decoder	3
κ	Carbon-mass price conversion factor	$44/12$
$(w_{\text{clim}}, w_{\text{cost}}, w_{\text{eq}}, w_{\text{stab}})$	Reward weights	$(0.40, 0.30, 0.15, 0.15)$
T_{ref}	Period-wise climate penalty threshold	1.5°C
T_{goal}	Endpoint target threshold	2°C
α_C, α_E	Cost and stability penalty scales	$10,\ 5$
δ_E	Relative price-increase threshold	0.05

The complete-episode return R is undiscounted. The components are

$$\begin{aligned} r_t^{\text{climate}} &= -\max(0, T_t - T_{\text{ref}})^2, \\ r_t^{\text{cost}} &= -\alpha_C C_t, \qquad r_t^{\text{equity}} = -G_t, \\ r_t^{\text{stability}} &= -\alpha_E \max\left(0, \frac{E_t}{E_{t-1}} - 1 - \delta_E\right). \end{aligned} \tag{3}$$

Here T_t is the recorded Hector temperature, C_t and G_t are the proxies defined below, and E_t is the electricity-price index. The climate term penalizes temperatures above T_{ref}; the stability term penalizes relative price increases above δ_E. The coefficients α_C and α_E scale the component penalties before weighting. All three evaluation settings use the fixed parameters in Table 1.

The reward threshold T_{ref} applies at each period; the distinct endpoint threshold T_{goal} defines target attainment. Both operate directly on Hector's emitted `global_tas` channel, without an additional anomaly-baseline adjustment. The climate reference period for this archived configuration has not been independently established, so these thresholds define tests on the recorded channel. Appendix A describes the temperature reader and missing-data behavior.

Cost proxy. Let Q_t denote summed native CO_2 query values and Y_t summed `GDP_Trial` demand. We use a clipped revenue-to-output proxy and its time accumulation:

$$C_t = \text{clip}_{[0,1]}\left(\frac{p_t \kappa \max(Q_t, 0)}{Y_t}\right), \qquad K = \Delta t \sum_{t=1}^{H} C_t. \tag{4}$$

The numerator approximates carbon-tax revenue by multiplying the prescribed price, expressed per tonne of carbon, by nonnegative aggregate emissions. The denominator is the queried output measure. The implementation substitutes zero for the ratio when Y_t is zero or NaN. We report K in proxy-fraction years under this unit convention; it measures accumulated revenue relative to output, not discounted consumption loss or calibrated welfare. Native query conversions are specified in Appendix A.

Distributional proxy. For each of the N R5 regions, we clip aggregate emissions below at zero and divide by population after unit conversion to obtain per-capita emissions $z_{t,r}$. The regional mapping and conversion are specified in Appendix A. Sorting these values in increasing order as $z_{t,(r)}$, we calculate the equal-region Gini coefficient

$$G_t = \frac{\sum_{r=1}^{N} (2r - N - 1) z_{t,(r)}}{N \sum_{r=1}^{N} z_{t,(r)}}. \tag{5}$$

The coefficient is set to zero when the sum is nonpositive. It describes variation among regional per-capita emissions values with equal weight per region; it does not measure within-region inequality or the distribution of policy costs.

Table 2: System configurations and attempted GCAM evaluations per seed and setting. Initial candidate evaluations and failed attempts count toward these budgets; mock pretraining and LLM inference are separate. The schedule length H is defined in Table 1.

System	Search or control configuration	Attempts
Random	Seeded monotone random walk; no return-based adaptation	24
Heuristic	Three fixed H-period schedules; order rotated by seed	3
CMA-ES	H normalized coordinates; population 6, initial $\sigma = 0.2$	24
GP-BO	Three Sobol initial points; single-task GP, standardized outputs, and log expected improvement	24
PPO	One mock-trained model per seed; 30,000 requested training steps; deterministic native control	5
LLM-ReAct	Stepwise observations and recent action/reward/status history; constrained JSON action	3
LLM+BO	Ten LLM-generated monotone initial schedules, then the same GP/log-EI search as GP-BO	24

Attempt records. Each evaluation record is indexed by method, setting, seed, and attempt, and links the proposed policy to executed actions, feedback, and execution status. Failed attempts consume budget and retain their available trajectory prefixes. The reported incumbent is the best complete-episode return. For subsequent proposals, the available CMA-ES and BO adapter implementations assign penalized scores to failures. Appendix B specifies these scores and their source provenance.

3.3 Three evaluation settings

The three settings share the simulator configuration, horizon, and scalar objective. Each has a separately executed set of method–seed evaluations. The first two emphasize different outcome summaries under immediate feedback; the third changes feedback timing.

Target and cost. For each seed, we record the best complete return, whether any complete policy satisfies $T_H \leq T_{\text{goal}}$, and the minimum K among its target-reaching policies. Here T_H is the endpoint temperature. The minimum-cost target-reaching policy may differ from the policy with the best return. Target-reaching cost is undefined for a seed with no qualifying policy.

Cost–temperature trade-offs. We retain (K, T_H) for every complete evaluation to describe the outcomes visited by each scalar-reward system. Figure 1 displays the full recorded point sets.

Delayed feedback. A wrapper delays observations and rewards delivered to the agent by one decision step while preserving the simulator's period sequence. It returns the reset observation and zero reward during warmup. At episode end, all pending rewards are summed into the final feedback reward and the queues are cleared. Evaluation uses the underlying rewards to compute R in Equation 2. Schedule-search systems precommit their actions within each episode; PPO and LLM-ReAct choose actions as observations arrive.

4 Experimental Protocol

4.1 Seven configured systems

Table 2 summarizes the seven systems and their evaluation budgets. Random search, the fixed heuristic, CMA-ES (Hansen, 2016), GP-BO (Balandat et al., 2020), and LLM+BO select complete schedules; PPO (Schulman et al., 2017) and LLM-ReAct (Yao et al., 2023) select actions from stepwise observations. The comparison includes each system's policy representation, initialization, pretraining, and feedback rules. Appendix B gives the fixed schedules, additional optimizer settings, failure scores, and configuration sources.

Search and initialization. For random search, the seed and attempt index determine the sampled price path. For the heuristic, the seed changes the order of the fixed schedules. CMA-ES searches the full normalized schedule domain (Hansen, 2016). The two BO systems use the same GP/log-expected-improvement implementation after initialization (Balandat et al., 2020; Ament et al., 2023),

also over the full domain. Their initial designs differ in both source and size (Table 2), so the GP-BO–LLM+BO comparison combines these two choices. Failed evaluations remain in the budget; their use in subsequent proposal generation is specified in Appendix B.

Pretrained and feedback-driven control. PPO uses one mock-trained model for each of the ten seeds (Schulman et al., 2017). Each model receives 22 observation scalars and emits four normalized controls, clipped to $[0, 1]$ for native evaluation. Its parameters remain fixed across repeated evaluations, with inactive controls handled as specified in Section 3. Training-step allocations in Table 2 are requested counts; Appendix B reports checkpoint evidence for actual training steps. LLM-ReAct receives the agent-facing observation, task identifier, step index, and latest three action/reward/status entries (Yao et al., 2023). It returns a four-element JSON action with inactive coordinates fixed to zero. The experimental LLM configuration is `gpt-5.6-sol` with reasoning effort `high`; retry limits and the disabled fallback policy are specified in Appendix B.

4.2 Allocation and Statistical Readouts

Native allocation. An evaluation cell is one system–setting–seed combination. Each system has ten cells in each of the three separately executed settings defined in Section 3.3, giving 210 cells and 3,210 attempted evaluations. The first two settings use immediate feedback with the same simulator and reward configuration; the third introduces the one-step delay. Random search, CMA-ES, GP-BO, and LLM+BO form the common-budget group with 24 attempts per cell. Table 3 additionally reports the heuristic, PPO, and LLM-ReAct at their respective allocations in Table 2. Budgets count evaluation attempts, including failures, rather than total computation. Episode timing omits some preempted work and optimizer overhead; mock pretraining and LLM inference are excluded from the attempt counts.

Endpoints and uncertainty. For a fixed system and setting, let $\mathcal{C}_s(b)$ contain the complete evaluations among seed s's first b attempts. With $R_{s,j}$ denoting the complete episode return of attempt j, the seed-level incumbent is

$$I_s(b) = \max_{j \in \mathcal{C}_s(b)} R_{s,j}. \tag{6}$$

It is undefined when $\mathcal{C}_s(b)$ is empty. We report the IQM across the ten seed-level incumbents only when all ten are defined, and show the available-seed count separately. The IQM retains the middle 50% of empirical mass with fractional weights at quartile boundaries (Agarwal et al., 2021). Pointwise 95% percentile intervals use 10,000 bootstrap resamples of the ten seed-level incumbents, recomputing the IQM in each resample. Repeated evaluations of a PPO model contribute to its seed-level incumbent, not additional bootstrap units. Appendix C gives the resampling seed and pairwise comparison procedures.

Target attainment and cost. A seed attains the target if at least one complete policy satisfies $T_H \leq T_{\text{goal}}$. We report the number of such seeds out of ten. For each target-reaching seed, the cost endpoint is the minimum accumulated proxy K among its qualifying policies. Its all-seed IQM is reported only when every seed has at least one qualifying policy; otherwise the cost summary is undefined, even if all seeds have complete returns. The temperature threshold and cost proxy are defined in Section 3.2.

Delayed-feedback contrasts. At each system's allocated budget b, we match seed labels across the target-and-cost and delayed-feedback settings and compute

$$\Delta_s = I_s^{\text{delayed}}(b) - I_s^{\text{target}}(b). \tag{7}$$

These post-hoc comparisons summarize the ten differences by their median, range, and sign counts (Appendix C). They describe differences between separately executed runs. Matching labels does not control all variation in policy generation, simulator execution, or interruptions, so the contrasts do not identify a causal effect of delay.

Table 3: Evaluation results at the attempt budgets in Table 2. Best complete return is summarized by its IQM across ten seeds and a 95% seed-bootstrap interval; higher is better. Completion is reported as completed/all attempts. Target counts are seeds with at least one complete policy satisfying the recorded-channel criterion $T_H \leq T_{\text{goal}}$. Measurement limitations are discussed in Section 5.4.

Method	Best return: IQM [95% CI]	Completed / attempted	Target seeds
Target and cost			
Random	-0.918 [-0.945, -0.848]	179/240	8/10
Heuristic	-1.229 [-1.229, -1.229]	10/30	0/10
CMA-ES	-0.708 [-0.741, -0.667]	194/240	10/10
GP-BO	-0.900 [-0.948, -0.849]	191/240	10/10
PPO	-0.850 [-0.850, -0.850]	50/50	0/10
LLM-ReAct	-1.045 [-1.162, -0.978]	24/30	6/10
LLM+BO	-0.847 [-0.881, -0.818]	178/240	10/10
Cost–temperature trade-offs			
Random	-0.918 [-0.944, -0.848]	179/240	8/10
Heuristic	-1.229 [-1.229, -1.229]	10/30	0/10
CMA-ES	-0.724 [-0.759, -0.686]	194/240	10/10
GP-BO	-0.885 [-0.925, -0.830]	190/240	10/10
PPO	-0.850 [-0.850, -0.850]	50/50	0/10
LLM-ReAct	-1.032 [-1.171, -0.934]	26/30	8/10
LLM+BO	-0.821 [-0.828, -0.820]	185/240	10/10
Delayed feedback			
Random	-0.918 [-0.944, -0.848]	179/240	8/10
Heuristic	-1.229 [-1.229, -1.229]	10/30	0/10
CMA-ES	-0.708 [-0.741, -0.668]	196/240	10/10
GP-BO	-0.885 [-0.924, -0.830]	190/240	10/10
PPO	-0.850 [-0.850, -0.850]	50/50	0/10
LLM-ReAct	-1.021 [-1.100, -0.989]	29/30	8/10
LLM+BO	-0.839 [-0.873, -0.798]	179/240	10/10

5 Evaluation Results

5.1 Recorded returns and search progress

CMA-ES has the highest best-return IQM among the four 24-attempt systems: -0.708 in target and cost, -0.724 in cost–temperature trade-offs, and -0.708 under delayed feedback (Table 3). LLM+BO has a higher IQM than GP-BO in each setting. The archived rank-based tests do not detect a difference between their seed-level endpoints after Holm correction across 18 comparisons (adjusted $p = 0.7099$ in each setting; Appendix C). These tests address endpoint distributions rather than the IQM difference specifically.

Figure 3 in Appendix C tracks best complete return against attempt budget. PPO's flat trace reflects repeated evaluation of fixed policies; the heuristic's all-seed IQM is defined only at its final budget. Figure 5 shows all three settings and return availability.

5.2 Target cost and cost–temperature outcomes

In the target-and-cost setting, CMA-ES, GP-BO, and LLM+BO each find a policy satisfying $T_H \leq T_{\text{goal}}$ in all ten seeds. Their IQMs of the minimum qualifying cost proxy K are respectively 0.333, 0.458, and 0.434 proxy-fraction years, with 95% intervals $[0.276, 0.380]$, $[0.414, 0.507]$, and $[0.373, 0.479]$. Random search meets the criterion in eight seeds and LLM-ReAct in six; the heuristic and PPO meet it in none. Only the first three systems therefore have defined all-seed target-cost summaries under the protocol in Section 4.

The separately executed cost–temperature setting yields 834 complete coordinate pairs (K, T_H) (Figure 1). PPO's 50 pairs coincide at $(0, 2.256°\text{C})$, and the heuristic's ten complete pairs coincide at $(0.442, 2.415°\text{C})$; both temperatures exceed T_{goal}. The search systems and LLM-ReAct produce

more varied coordinates. Repeated outcomes remain in the record, with overlapping markers in the figure.

The pooled point set has 15 distinct nondominated coordinates: no other recorded point has both no greater cost and no higher temperature, with a strict improvement in at least one. PPO occupies the zero-cost end. Negative-temperature points from GP-BO and LLM+BO also enter this set. Runtime defects described in Section 5.4 leave the physical interpretation of these coordinates and their dominance relations unverified. The plotted set is retained to show the outcomes used by the archived search and analysis.

5.3 Completion and target attainment

Among the 3,210 attempts, 2,493 are recorded as complete. Post-collection diagnosis classifies the remaining 717 as 714 numerical failures and three infrastructure failures; Appendix D maps these categories to the original worker statuses. Figure 4 reports completion fractions per attempt and target-attainment fractions per seed at each system's allocated budget. PPO completes all evaluations without meeting the temperature criterion. In the target-and-cost setting, CMA-ES completes 194 of 240 attempts and finds a qualifying policy in every seed. Execution completion alone therefore does not guarantee target attainment.

Delayed feedback. The median seed-level contrast Δ_s is -0.014 for LLM-ReAct and $+0.016$ for LLM+BO; PPO retains zero carbon-price actions. These separately executed comparisons are descriptive. Appendix C gives the ranges, sign counts, and executed-sequence comparisons.

5.4 Measurement Scope

The runtime diagnostics identified state-mutating residual checking, a residual-scale error, and recovery/read paths that bypass required validation. Omitting the checking path in a matched diagnostic changed one policy's recorded endpoint temperature from $-0.6625^{\circ}\mathrm{C}$ to $1.964^{\circ}\mathrm{C}$; this was not a validated replacement runtime (Appendix D). Such changes can affect rewards, target qualification, and nondominated outcomes. Their effect on the relative performance of the evaluated methods remains unresolved. A physically validated comparison requires new evaluations and adaptive search histories under a qualified runtime; rescoring selected endpoints alone is insufficient. Appendix E describes the replay resource and native-reuse requirements.

6 Limitations

Runtime measurement defects leave the physical interpretation of the results unresolved (Section 5.4). The study covers only carbon-price control through 2060, using fixed proxy objectives and system-specific evaluation budgets.

7 Conclusion

GCAM-Gym combines a shared policy-evaluation interface, configured search and control systems, and attempt-level records with numerical replay. In the archived runs, CMA-ES has the highest best-return IQM among the four equal-attempt-budget systems, and execution completion does not guarantee target attainment. Runtime diagnostics leave the physical interpretation of these policy outcomes and the ranking unresolved. The resource supports inspection of the recorded comparisons and development of a validated benchmark; establishing policy performance under a qualified runtime requires new evaluations with clean adaptive search histories.

AI USE STATEMENT

We used generative AI tools to assist with coding, experiment management, manuscript editing, and draft figure illustrations. The authors reviewed the resulting code, experimental settings, text, and figures and take responsibility for the final manuscript and accompanying artifacts.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. *Advances in Neural Information Processing Systems*, 2021.

Sebastian Ament, Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected Improvements to Expected Improvement for Bayesian Optimization. *Advances in Neural Information Processing Systems*, 36, 2023.

Maximilian Balandat et al. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. *Advances in Neural Information Processing Systems*, 2020.

James Bergstra and Yoshua Bengio. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

Greg Brockman et al. OpenAI Gym. arXiv:1606.01540, 2016. Preprint.

Katherine Calvin et al. GCAM v5.1: representing the linkages between energy, water, land, climate, and economic systems. *Geoscientific Model Development*, 12:677–698, 2019. doi:10.5194/gmd-12-677-2019.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002. doi:10.1109/4235.996017.

Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi:10.1214/aos/1176344552.

Katharina Eggensperger et al. HPOBench: A Collection of Reproducible Multi-Fidelity Benchmark Problems for HPO. *NeurIPS Datasets and Benchmarks*, 1, 2021.

Logan Engstrom et al. Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO. *International Conference on Learning Representations*, 2020. arXiv:2005.12729.

David Eriksson and Matthias Poloczek. Scalable Constrained Bayesian Optimization. *International Conference on Artificial Intelligence and Statistics*, PMLR 130:730–738, 2021.

David Eriksson, Michael Pearce, Jacob Gardner, Ryan D. Turner, and Matthias Poloczek. Scalable Global Optimization via Local Bayesian Optimization. *Advances in Neural Information Processing Systems*, 32, 2019. arXiv:1910.01739.

Jacob Gardner, Matt Kusner, Zhixiang Xu, Kilian Weinberger, and John Cunningham. Bayesian Optimization with Inequality Constraints. *International Conference on Machine Learning*, PMLR 32(2):937–945, 2014.

Nikolaus Hansen, Anne Auger, Raymond Ros, Olaf Mersmann, Tea Tušar, and Dimo Brockhoff. COCO: A Platform for Comparing Continuous Optimizers in a Black-Box Setting. *Optimization Methods and Software*, 36(1):114–144, 2021. doi:10.1080/10556788.2020.1808977.

Nikolaus Hansen. The CMA Evolution Strategy: A Tutorial. arXiv:1604.00772, 2016. Preprint.

C. A. Hartin, P. Patel, A. Schwarber, R. P. Link, and B. P. Bond-Lamberty. A simple object-oriented and open-source model for scientific and policy analyses of the global climate system—Hector v1.0. *Geoscientific Model Development*, 8:939–955, 2015. doi:10.5194/gmd-8-939-2015.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep Reinforcement Learning that Matters. *AAAI Conference on Artificial Intelligence*, 32, 2018. arXiv:1709.06560.

S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13:455–492, 1998. doi:10.1023/A:1008306431147.

Robert J. Lempert, David G. Groves, Steven W. Popper, and Steve C. Bankes. A General, Analytic Method for Generating Robust Strategies and Narrative Scenarios. *Management Science*, 52(4):514–528, 2006. doi:10.1287/mnsc.1050.0472.

Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large Language Models to Enhance Bayesian Optimization. *International Conference on Learning Representations*, 2024. arXiv:2402.03921.

William D. Nordhaus. An Optimal Transition Path for Controlling Greenhouse Gases. *Science*, 258(5086):1315–1319, 1992. doi:10.1126/science.258.5086.1315.

William D. Nordhaus. Revisiting the social cost of carbon. *Proceedings of the National Academy of Sciences*, 114(7):1518–1523, 2017. doi:10.1073/pnas.1609244114.

Brian C. O'Neill et al. A new scenario framework for climate change research: the concept of shared socioeconomic pathways. *Climatic Change*, 122:387–400, 2014. doi:10.1007/s10584-013-0905-2.

Ian Osband et al. Behaviour Suite for Reinforcement Learning. *International Conference on Learning Representations*, 2020. arXiv:1908.03568.

Simon Ramstedt, Yann Bouteiller, Giovanni Beltrame, Christopher Pal, and Jonathan Binas. Reinforcement Learning with Random Delays. *International Conference on Learning Representations*, 2021. arXiv:2010.02966.

Keywan Riahi et al. The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, 42:153–168, 2017. doi:10.1016/j.gloenvcha.2016.05.009.

David Rolnick et al. Tackling Climate Change with Machine Learning. *ACM Computing Surveys*, 55(2), article 42, 2022. doi:10.1145/3485128.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. arXiv:1707.06347, 2017. Preprint.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. *Advances in Neural Information Processing Systems*, 36, 2023. arXiv:2303.11366.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*, 25, 2012. arXiv:1206.2944.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe Exploration for Optimization with Gaussian Processes. *International Conference on Machine Learning*, PMLR 37:997–1005, 2015.

Yuval Tassa et al. DeepMind Control Suite. arXiv:1801.00690, 2018. Preprint.

Mark Towers et al. Gymnasium: A Standard Interface for Reinforcement Learning Environments. *Advances in Neural Information Processing Systems*, 38, 2025. arXiv:2407.17032. API documentation: Farama Foundation. Env API. *Gymnasium Documentation*. Documentation snapshot accessed September 26, 2026. https://gymnasium.farama.org/api/env/.

Ryan Turner et al. Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020. *NeurIPS 2020 Competition and Demonstration Track*, PMLR 133:3–26, 2021.

Jason Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35, 2022. arXiv:2201.11903.

Chengrun Yang et al. Large Language Models as Optimizers. *International Conference on Learning Representations*, 2024. arXiv:2309.03409.

Shunyu Yao et al. ReAct: Synergizing Reasoning and Acting in Language Models. *International Conference on Learning Representations*, 2023.

Tianyu Zhang et al. AI for Global Climate Cooperation: Modeling Global Climate Negotiations, Agreements, and Long-Term Cooperation in RICE-N. *International Conference on Machine Learning*, PMLR 267:76332–76360, 2025.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*, 8(18):eabk2607, 2022. doi:10.1126/sciadv.abk2607.

Table 4: Observation schema. Sizes count scalar coordinates. Cost and distributional proxies are defined in Section 3.2; field ordering, unit conversions, and missing-data behavior are specified in Appendix A.

Field	Size	Units	Construction
`year`	1	calendar year	Simulator decision year
`temperature`	1	°C	Hector `global_tas`, without rebasing
`co2_concentration`	1	ppmv	Hector CO_2 channel
`emissions_global`	1	$GtCO_2$/yr	Sum of five regional emissions
`emissions_r5`	5	$GtCO_2$/yr	Mapped R5 aggregate emissions
`elec_mix`	7	fractions	Electricity-generation shares
`energy_price_index`	1	ratio	Electricity price relative to reset reference
`policy_cost_frac_gdp`	1	proxy fraction	Clipped revenue-to-output ratio C_t
`cdr_deployed`	1	$GtCO_2$/yr	Aggregate sequestration, converted to CO_2 mass
`inequality_gini`	1	dimensionless	Equal-region Gini G_t of per-capita emissions
`carbon_price_effective`	1	price/tCO_2	Native market price divided by κ
`solver_iterations`	1	placeholder	Constant zero

A Space Bounds, Unit Conventions, and Query Fallbacks

Table 4 gives the observation schema; Section 3.2 defines the objective construction. This appendix specifies field dimensions and units, declared space bounds, additional unit conventions, and query fallbacks.

Declared observation space. The environment declares a Gymnasium `Dict` of `Box` spaces with the componentwise bounds in Table 5. Year and solver iterations use integer coordinates; other fields use floating-point coordinates. These are declared API bounds; observations that violate them remain in the archive (Appendix D). Regional arrays follow the order `R5_OECD90+EU`, `R5_REF`, `R5_ASIA`, `R5_MAF`, and `R5_LAM`. The seven electricity-share coordinates are coal, gas, oil, nuclear, hydro, wind/solar, and bioenergy/CCS/other. The worker source bundled with the resource specifies the mapping from native regions and subsectors to these groups. Flattening follows the Gymnasium dictionary-space order; LLM serialization sorts the field keys.

Table 5: Declared observation bounds in the archived environment. Bounds apply separately to every coordinate of vector-valued fields; field dimensions and units are given in Table 4.

Field	Lower bound	Upper bound
`year`	2025	2100
`temperature`	0	10
`co2_concentration`	300	1500
`emissions_global`	−100	150
`emissions_r5`	−50	100
`elec_mix`	0	1
`energy_price_index`	0.1	10
`policy_cost_frac_gdp`	0	1
`cdr_deployed`	0	25
`inequality_gini`	0	1
`carbon_price_effective`	0	1000
`solver_iterations`	0	10^6

Action decoding and native units. Equation 1 defines the decoder. For example, the active coordinate $a_t = 0.5$ decodes to $p_t = 182$. The API field `carbon_price_usd` is written to the native carbon market after multiplication by κ; the reported effective carbon price divides the queried market price by the same factor. The monetary convention and its validation limit are stated in Section 3.

Query conversions. The worker converts native emissions and sequestration query sums to reported $GtCO_2$/yr by multiplying by κ and dividing by 10^3. For per-capita emissions, it multiplies the reported regional emissions by 10^9 and the native population values by 10^3 before division. These factors implement an MtC-to-$GtCO_2$ conversion and a thousands-of-people-to-people conversion. The cost proxy in Equation 4 instead uses the native emissions sum Q_t and native `GDP_Trial` demand sum Y_t directly. Its interpretation as a revenue-to-output fraction requires compatible monetary units for the price and output queries; independent dataset-level validation remains outstanding.

Temperature channel. Hector is a reduced-complexity global climate model (Hartin et al., 2015). The reader selects the last matching non-spinup `global_tas` entry for the requested year from Hector's output stream. Section 3.2 defines how this channel enters the reward and target criterion, including the unresolved reference-period convention. Re-execution requires the corresponding Hector configuration.

Electricity-price index and failure penalties. The electricity-price index divides the mean queried electricity-market price by the first available reference price, normally obtained at reset. A zero reference produces an index of one; a missing current price can produce NaN. The stability expression in Equation 3 has no additional guard for a zero previous index. The evaluation return is defined in Equation 2; PPO's training discount is specified separately below.

On a reported solver failure, the environment replaces the usual reward components with scaled running component minima. Each minimum is initialized at -10^{-3} and updated after successfully executed periods; the failure components are the running minima multiplied by $\lambda_{\text{fail}} = 1.5$. The scalar penalty uses the same reward weights as a successful period. These interaction penalties are distinct from the optimizer training scores described in Appendix B.

Missingness and execution status. The archived worker also contains query fallbacks: absent regional emissions default to zero, missing regional population entries default to 1.0 before the per-capita unit conversion, unavailable sequestration may become zero, and unavailable effective carbon price may become zero. These are implementation behaviors, not independent observations of zero. In particular, the historical temperature reader could use a previous-year value when a requested observation was unavailable; this behavior is one of the defects in Appendix D. Environment failure penalties influence interaction, while incomplete trajectories have no complete-policy incumbent value in the reported analysis.

B Baseline Configuration and Feedback

Table 2 summarizes the complete systems. The settings below supplement the main protocol with exact schedules, training parameters, and implementation provenance.

Schedule-search settings. The inspected adapter's random walk starts uniformly at an integer in $[0, 40]$ and adds independent integer increments from $[0, 45]$, capped at 1,000. The seeded generator depends on both the seed cell and evaluation index. CMA-ES (Hansen, 2016) has recorded bounds $[0, 1]^7$, initial standard deviation 0.2, and population size six; the inspected adapter sets the initial center to $(0.25, \ldots, 0.25)$. With 24 attempts, the allocation covers four populations. These coordinates are decoded through the shared nonlinear price map.

GP-BO has three recorded Sobol initial points. The inspected implementation uses BoTorch's `SingleTaskGP` with one-output standardization (Balandat et al., 2020) and log expected improvement (Ament et al., 2023), limits fitting to 25 optimizer iterations, and selects from a 128-point scrambled Sobol candidate pool. LLM+BO uses ten initial candidates from the recorded LLM warm-start manifest, followed by the same GP/log-EI implementation. Its initial candidates are constrained to nondecreasing integer price schedules; the later BO search uses the full normalized coordinate box. Exact fitting settings are source-level descriptions, not all independently locked historical metadata.

In the inspected adapter, a failed BO observation at zero-based attempt index j receives training utility $-10^6 - j$; a complete observation receives its recorded return. CMA-ES minimizes negative complete return, with failed fitness $10^6 + j$. These proposal-generation labels are not saved as checkpoint fields

and differ from the environmental step penalty. The GP delegates default covariance and likelihood choices to its installed dependency; native re-execution must bind the adapter and dependencies. The replay shipped here uses saved records and performs no refitting.

Fixed schedules. The heuristic evaluates the following seven-period integer price sequences, rotating their execution order by seed:

$$\begin{aligned} \text{mild} &= (30, 47, 68, 91, 116, 150, 190), \\ \text{moderate} &= (45, 60, 80, 105, 130, 165, 205), \\ \text{low} &= (0, 1, 3, 5, 7, 12, 19). \end{aligned}$$

PPO training and evaluation. The ten training manifests contain ten distinct model hashes. Each requests 30,000 mock-training steps with four environments, `MlpPolicy`, rollout length 256, batch size 64, ten epochs, learning rate 3×10^{-4}, and discount $\gamma = 0.99$ (Schulman et al., 2017). The inspected seed-zero checkpoint records 30,720 actual timesteps. The training environment has 22 observation scalars and four action coordinates. Native evaluation uses deterministic prediction and clips actions to $[0, 1]$; unsupported coordinates can remain nonzero in the records but do not produce native interventions. The five evaluations per model and view are not additional training runs.

LLM control. The stepwise controller adapts the ReAct interaction pattern (Yao et al., 2023). It receives the current agent-facing observation, step index, task identifier, and the most recent three action/reward/status entries. Its response schema contains only a four-element normalized `action`; unsupported coordinates must be zero. It does not emit a public interleaved reasoning transcript. The configured model is `gpt-5.6-sol`, reasoning effort `high`, with a 2,048-token response limit, 60-second timeout, at most two retries, and a six-second minimum interval. Fallback policies are disabled. Deployment endpoints and credentials are excluded from the resource files.

Configuration provenance. Per-attempt metadata and frozen training and warm-start manifests establish recorded settings. Other details, including the CMA initial center, GP fitting settings, and optimizer failure labels, describe the inspected adapter but lack a contemporaneous binding of every implementation byte. The archived package preserves environment and worker definitions. The selected export omits native binaries and optimizer dependencies; the README distinguishes these provenance levels from portable arithmetic replay.

C Statistical Details and Additional Views

Table 3 reports results for all evaluated methods with endpoint intervals, completion counts, and target attainment. Here we give the statistical procedures, three-setting evaluation curves, descriptive delayed-feedback contrasts, and action-record details.

All 21 endpoint rows are retained. At prefixes where any seed has no complete policy, the all-seed return IQM is undefined. Minimum target-cost IQMs similarly require all ten seeds to have a target-reaching policy. The native record's numerical failure taxonomy and the measurement boundary apply to these saved outcomes. Degenerate observed bootstrap intervals do not imply certainty about another training run or another simulator configuration.

For the heuristic, complete returns are available for four, seven, and ten seeds after the first, second, and third attempts, respectively, in each setting. Its all-seed IQM is therefore defined only at the final budget. The seed-dependent ordering of the fixed schedules produces this availability pattern.

The archived analysis uses 10,000 percentile seed-bootstrap resamples with RNG seed 20260916. Its IQM retains the middle 50% empirical mass with fractional boundary weights (Agarwal et al., 2021). The protocol was fixed after collection and before the first comparative analysis. Six pairwise search comparisons per view form an 18-comparison family with Holm adjustment (Holm, 1979) for exact two-sided pooled-average-rank label-permutation tests. A separate family uses paired seed-label sign flips as a dependence sensitivity. GP-BO versus LLM+BO has adjusted rank-test $p = 0.7099$ in each view. These saved intervals and tests are retained; the portable replay independently reconstructs their input summaries, not the resampling or permutation procedures.

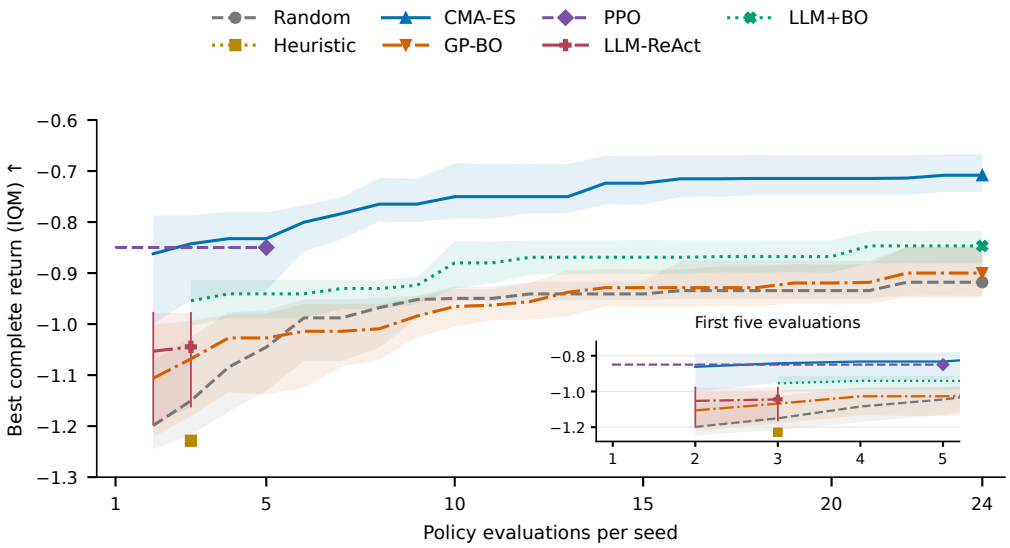

Figure 3: Best complete return versus attempted evaluations in the target-and-cost setting. Curves show ten-seed IQMs and pointwise 95% seed-bootstrap intervals; short traces retain vertical interval bars. Markers identify each system's final budget, and the inset enlarges the first five attempts. Failed attempts count toward the budget, and estimates remain unplotted until all ten seeds have a complete return. The axis excludes pretraining and LLM inference; PPO parameters remain fixed.

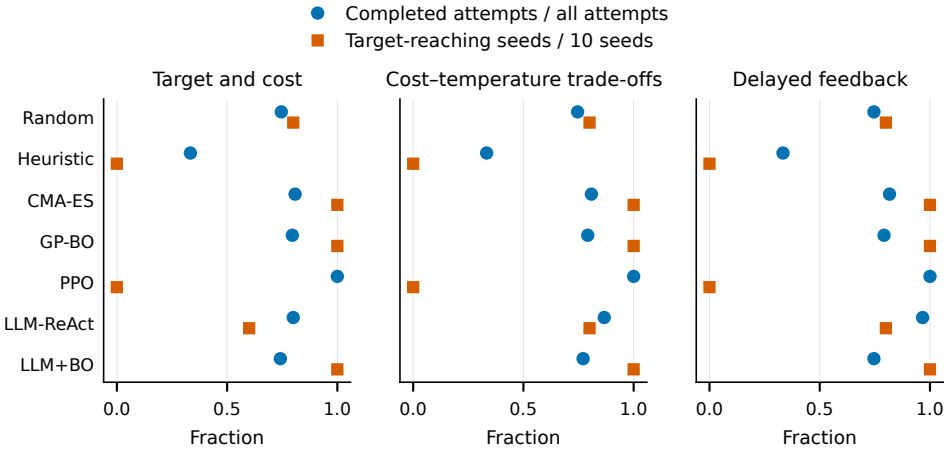

Figure 4: Completion and target attainment across the evaluated methods. Circles denote completed attempts divided by all attempts; squares denote seeds with any complete policy satisfying $T_{2060} \leq 2°\mathrm{C}$, divided by ten. Shapes are vertically offset to preserve coincident observations. The native budgets are those in Table 3; proportions describe this recorded sample.

For each rank test, the statistic is the absolute deviation of the first group's rank sum from $n(N+1)/2$, with average ranks for ties, $n = 10$, and $N = 20$. All $\binom{20}{10}$ label allocations are enumerated. The paired sensitivity instead uses the absolute mean of ten seed-label differences and enumerates all 2^{10} sign assignments. Comparisons include numerical tolerances of 10^{-10} and 10^{-12}, respectively. Within each 18-test family, sorted raw probabilities $p_{(j)}$ receive adjusted values $\widetilde{p}_{(j)} = \min\{1, \max_{i \leq j}(19 - i)p_{(i)}\}$ (Holm, 1979).

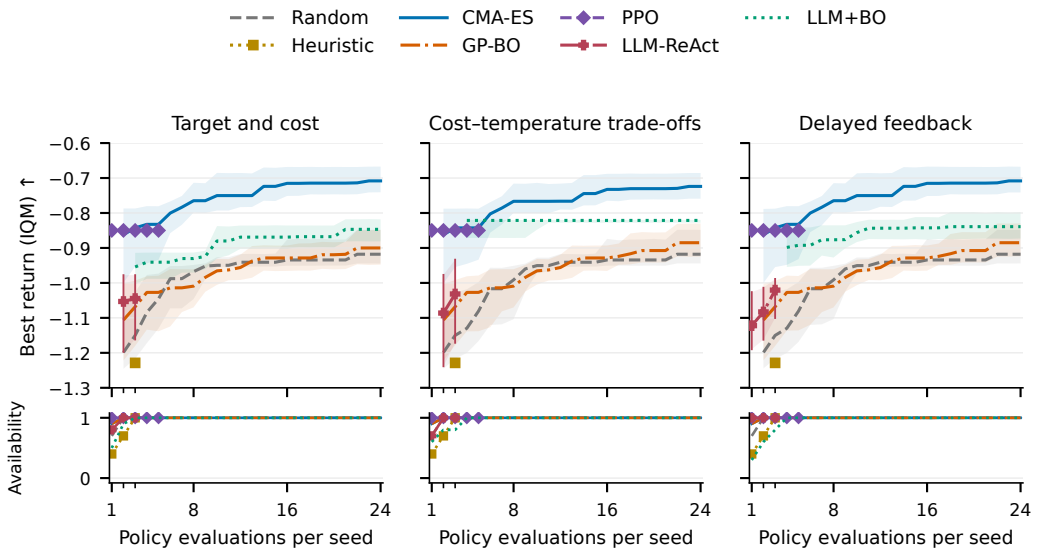

Figure 5: Evaluation progress for all evaluated methods across the three settings. Upper panels show the IQM of best complete return across ten seeds with saved pointwise 95% seed-bootstrap intervals; short traces also show vertical intervals. Lower panels show the fraction of seeds with a complete policy. The horizontal axes count policy evaluations per seed, including failures. Traces stop at each system's actual budget, and undefined all-seed estimates remain unplotted. PPO traces describe evaluation of mock-pretrained policies, not native learning; pretraining and LLM inference costs are excluded.

Table 6: Post-hoc return contrasts between the delayed-feedback and target-and-cost settings at each system's allocated budget, using ten matched seed labels per system. Positive Δ denotes higher best complete return under delayed feedback. Values are displayed to three decimals; sign counts use unrounded differences with a 10^{-12} tolerance for floating-point roundoff, not an equivalence test. The contrasts do not isolate a causal delay effect.

System	Median Δ	Min Δ	Max Δ	Higher / tied / lower
Random	0.000	0.000	0.000	0 / 10 / 0
Heuristic	0.000	0.000	0.000	0 / 10 / 0
CMA-ES	0.000	0.000	0.000	0 / 10 / 0
GP-BO	0.000	0.000	+0.099	2 / 8 / 0
PPO	0.000	0.000	0.000	0 / 10 / 0
LLM-ReAct	-0.014	-0.185	+0.228	4 / 0 / 6
LLM+BO	+0.016	-0.157	+0.155	6 / 1 / 3

Descriptive delayed-feedback contrasts. The median paired-label return difference Δ_s is zero for random search, the heuristic, CMA-ES, GP-BO, and PPO (Table 6). LLM-ReAct has median -0.014, with four positive and six negative differences, spanning $[-0.185, 0.228]$. Its target-reaching count nevertheless increases from six to eight seeds, illustrating that median return change and target discovery answer different questions. LLM+BO has median $+0.016$, spanning $[-0.157, 0.155]$. Its schedule search does not use stepwise feedback, so the latter contrast is not evidence of a within-episode delay response.

Delay timing and action records. For a seven-step delayed-feedback episode, the agent receives observation sources $(0, 1, 2, 3, 4, 5, 6)$, where zero denotes the reset state. Rewards from steps one to five arrive one step late; the terminal call returns the queued rewards from steps six and seven. The task metrics use the actual trajectory record. Pairing same-numbered seed cells across the target-and-cost and delayed-feedback settings gives 1/0/2/2/1/10/10 cells with different executed four-coordinate arrays for Random/Heuristic/CMA-ES/GP-BO/PPO/LLM-ReAct/LLM+BO. Comparing only the

active carbon-price coordinate gives 1/0/2/2/0/10/10; comparison of decoded price sequences gives the same counts. A shorter failed prefix also constitutes a different sequence, so these counts do not measure deliberate changes to full proposed schedules.

The action records distinguish changes in effective control from changes in inactive API coordinates. For LLM-ReAct, every matched seed cell contains at least one same-indexed attempt with a different executed carbon-price sequence. PPO emits a zero carbon-price coordinate in all 150 native evaluations: its single matched cell with a changed four-coordinate action array differs only in inactive controls. Thus, the PPO trace difference does not indicate a changed native carbon-price intervention. Differences in executed sequence length can also arise from failed prefixes; these paired-label comparisons describe the recorded behavior rather than isolate a causal effect of delay.

Cost–temperature point sets. The 834 points comprise 179/10/194/190/50/26/185 complete evaluations in method order. Every successful attempt in the trade-off setting contributes its recorded cost proxy and temperature; none is selected by its score or Pareto membership. Failed attempts contribute no complete coordinate pair. Repeated policies remain recorded separately, although their markers can overlap. The exported record preserves all successful and failed attempts, enabling alternative inspection without pretending that duplicate points are independent physical-policy observations.

D Runtime Evidence and Status Semantics

The original worker statuses are 2,493 `success`, 716 `solver_failure`, and one `api_failure`. The post-collection diagnostic taxonomy reclassifies known execution exceptions, giving 2,493 successes, 714 numerical failures, and three infrastructure failures. These are two labelings of the same 3,210 attempts. The main tables retain the full grid. The recorded infrastructure sensitivity excludes three entire cells: CMA-ES seed 9 in the trade-off setting, CMA-ES seed 3 under delayed feedback, and LLM-ReAct seed 0 in the trade-off setting; it does not replace the main endpoint summaries.

The runtime audit first checked all 2,493 recorded successful trajectories. It found 57 trajectories outside the declared observation space, affecting 43 method–view–seed cells. These were interface-contract violations, not newly introduced physical rejection thresholds. The audited data were preserved without clipping temperatures or deleting those records.

In a matched native diagnostic, two fixed policies were each executed with the archived residual-checking path and with a diagnostic omission of that path. The policy, restart, configuration, native library, and single-thread execution convention were held fixed. For the zero-price policy, recorded 2060 temperature changed from $2.256^\circ\mathrm{C}$ to $2.291^\circ\mathrm{C}$. For the selected GP-BO policy, it changed from $-0.6625^\circ\mathrm{C}$ to $1.964^\circ\mathrm{C}$, with CO_2 changing from 113.7 to 467.5 ppm. The native CSV contained the anomaly, and the preserved worker completed without recovery, so the example cannot be explained solely as a JSON-reading or replay error. The omission mode was diagnostic and did not establish a replacement convergence test.

The audit additionally identified an extra quantity-scale factor in the old residual ratio and fail-open recovery/read paths, including substitution of previous-year climate observations. Subsequent repairs were versioned separately. A clean adaptive search under a repaired measurement process cannot be reconstructed by replacing only an old winner's return: earlier feedback also determines later proposals. The source-linked diagnostic summary shipped with the manuscript identifies the preserved records and the scope of these observations.

The selected-policy diagnostic changes the endpoint climate penalty because the two readings straddle T_{ref}, although both satisfy T_{goal}. More generally, affected measurements can alter rewards, target qualification, and the policies included in minimum-target-cost and nondominance summaries. Their effect on the relative performance of the evaluated methods has not been established. A physically validated comparison requires a qualified runtime and new evaluations. For adaptive search, it also requires rerunning proposal generation with the validated feedback: rescoring selected policies cannot recover the search histories that different earlier measurements would have produced.

E REUSE REQUIREMENTS

Replay verifies the arithmetic from the selected records; saved bootstrap intervals and permutation tests are retained without reestimation. It does not rerun the native simulator. The archive omits native binaries, initialization artifacts, and trained PPO checkpoints; native re-execution requires those inputs, their dependencies, and applicable access permissions. Runtime and measurement validation are prerequisites for interpreting a new comparison physically. Broader controls and horizons beyond the present carbon-price study through 2060 would additionally require extending the interface and evaluation design.