SolarChain-Eval: A Physics-Constrained Benchmark for Trustworthy Economic Agents in Decentralized Energy Markets

Shilin Ou[1,†], Yifan Xu[1,†], Luyao Zhang[1,*]

[1]Duke Kunshan University, Kunshan, Jiangsu, China

[†]Joint first authors; [*]Corresponding author: lz183@duke.edu

KDD 2026 Jeju, Korea August 9–13, 2026

Motivation and Research Questions

Central claim. Evaluating agentic AI in decentralized energy markets cannot stop at scalar reward. A policy can increase apparent market utility while backing invalid generation, injecting artificial liquidity, or producing unstable governance actions.

Research questions

- How do autonomous market policies trade off economic utility, physical safety, stability, and fairness against static and heuristic baselines?
- Do reward-maximizing agents exploit invalid generation and create artificial liquidity when physics penalties are removed?
- Can an LLM Planner/Auditor provide auditable risk mitigation without retraining the RL controller?

Benchmark contribution

- A Gymnasium-compatible hourly market-governance MDP with physical PV upper bounds and FDIA labels.
- Joint reporting of utility, physics safety, market stability, action smoothness, spatial fairness, and auditability.
- Structured Planner/Auditor traces that record triggers, revisions, action deltas, and rationales.

The benchmark treats trustworthiness as a joint property of economic payoff, physical feasibility, and explainable intervention, so a policy is not considered reliable merely because its reward is high.

Data and Evaluation Matrix

Table 1. Benchmark data and policy coverage.

Component	Scale	Evaluation use
Cities	Beijing, Shanghai, Chengdu, Shenzhen, Hangzhou	Spatial fairness and heterogeneous weather
Nodes	50	Distributed energy prosumers
Period	2026-04-01 to 2026-04-30	720 hourly market states
Generation	36,000 records	PV capacity, verified generation, FDIA labels
P2P trades	1,185 records	Market clearing and token burning
Rollouts	90 per policy across 3 seeds	Mean and paired comparisons
Agentic logs	12,960 steps	Planner/Auditor traces
Policies	Static, Random, Myopic, PPO, SAC, DQN	Baseline and learned governance

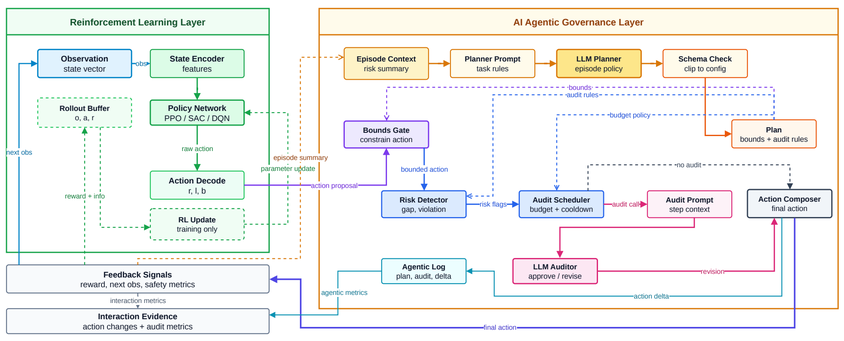
The data matrix links three layers of evidence: physical solar feasibility, market execution, and agentic oversight. *SolarChain-Eval* therefore tests whether higher reward is backed by valid supply, stable liquidity, and traceable governance.

- **Physical layer:** weather, PV capacity, verified generation, and FDIA labels.
- **Market layer:** liquidity, slippage, P2P trading, token burn, and unmatched demand.
- **Agentic layer:** Planner bounds, Auditor triggers, revisions, action deltas, and rationales.

Paired comparisons reuse matched seeds and episodes, so policy differences are interpreted as governance effects rather than weather or city-composition artifacts.

Evaluation settings. The poster integrates the main benchmark, a no-physics reward ablation, and deployment-time LLM auditing into one evaluation matrix.

Method: RL plus LLM Audit



Evaluation-time Planner/Auditor wrapper.

State and action.

$$s_t = [\sin\tau_t, \cos\tau_t, G_t^v, G_t^r, P_t^{\max}, \Delta_t, L_t, p_t, \nu_t, \sigma_t, \alpha_{t-1}, \ell_{t-1}],$$
$$a_t = (\alpha_t, \ell_t, b_t), \qquad \alpha_t + \ell_t \le 0.98.$$

Reward.

$$R_t = M_t - \lambda_d D_t - \lambda_j J_t - \lambda_u U_t - \lambda_p \Phi_t - \lambda_f F_t,$$
$$\Phi_t = V_t + A_t, \qquad A_t = \beta_t X_t.$$

Market transition.

$$M_t = \min(L_t + \ell_t(G_t^v + \beta_t X_t), \widetilde{Q}_t),$$
$$L_{t+1} = \max(L_t + \ell_t(G_t^v + \beta_t X_t) - M_t, 0).$$

Audit trigger.

$$H_t = \mathbb{I}[\nu_t > \tau_\nu \vee \Delta_t^{audit} < \tau_\Delta \vee \sigma_t > \tau_\sigma],$$
$$S_t = \mathbb{I}[\kappa_t > \tau_\kappa], \qquad E_t = H_t \vee S_t,$$
$$I_t^{audit} = E_t \wedge \left[H_t \vee (N_{e,t}^{audit} < B_e \wedge t - t_e^{last} > c_e)\right].$$

- PPO/SAC control continuous actions; DQN uses a $5^3 = 125$ action grid.
- Planner bounds action ranges and audit policy at the episode level.
- Auditor reviews risky eligible steps and returns `approve` or schema-validated `revise`.
- Sanitizer clips actions and enforces $\alpha_t + \ell_t \le 0.98$.

Logged outputs include the raw RL action, bounded action, audited final action, trigger type, action delta, and natural-language rationale.

Execution contract. The LLM layer never directly controls the market: every proposed action is bounded by the Planner, revised only under explicit triggers, and clipped again before environment execution.

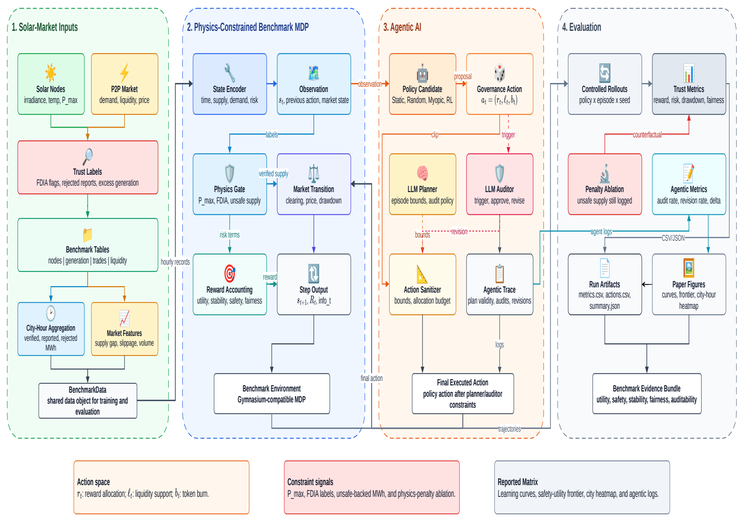
Trustworthiness dimensions.

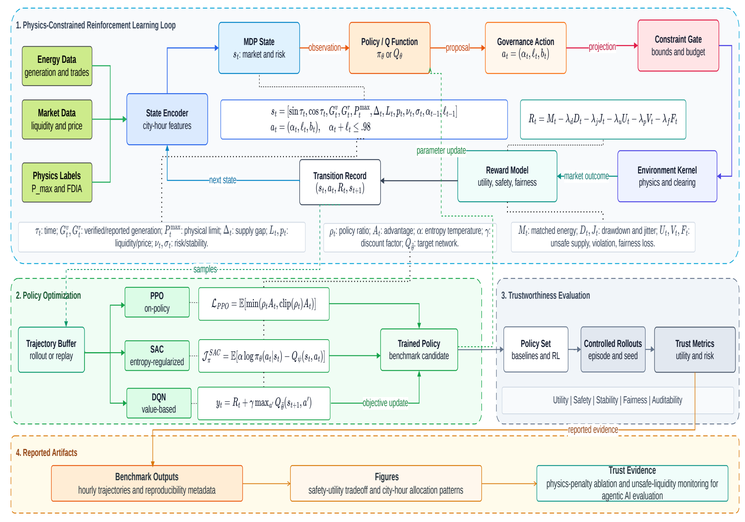
Dimension	Evidence
Utility	Reward and trading volume
Physical safety	Violation rate and artificial liquidity
Market stability	Slippage, liquidity drawdown, and jitter
Auditability	Audit rate, revision rate, action delta, rationale

Why the wrapper is auditable. The benchmark stores raw, bounded, and audited actions, allowing later analysis to separate controller competence from governance intervention.

Policy-to-evidence linkage. Each intervention links what RL proposed, what the Auditor changed, and how the executed action affected market clearing. Failures can therefore be assigned to reward design, controller instability, or audit rules.

Benchmark Pipeline and RL Formulation





End-to-end benchmark workflow.

- Hourly states aggregate weather, PV capacity, reported generation, verification labels, liquidity, and demand.
- Evaluation joins market utility with physical validity and auditable intervention traces.

Pipeline output. Each rollout emits aligned records for market utility, physical risk, stability, fairness, and audit traces, enabling paired comparisons across RL, RL+LLM, and no-physics settings.

RL environment and metrics.

- Physics-constrained reward prevents reward-only exploitation of invalid generation.
- LLM oversight is evaluated at deployment time, not during RL training.

Evaluation Result

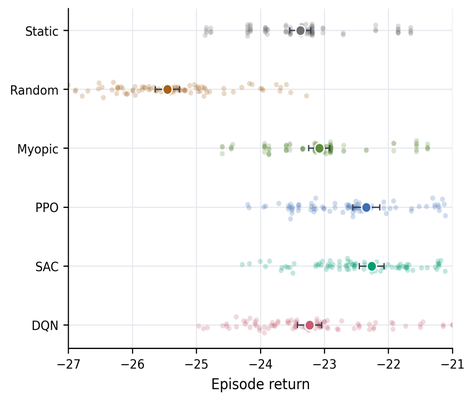
Table 2. Main benchmark under the physics-constrained reward.

Policy	Reward	Volume	Violation	Slip.	Art. liq.
Static	-23.38	0.514	0.4874	0.0140	0.1803
Random	-25.46	0.480	0.4771	0.0148	0.1137
Myopic	-23.08	0.456	0.4779	0.0144	0.1009
PPO	-22.35	0.470	0.4520	**0.0134**	0.1453
SAC	**-22.26**	**0.521**	**0.4461**	0.0178	**0.0686**
DQN	-23.23	0.491	0.4688	0.0148	0.1298

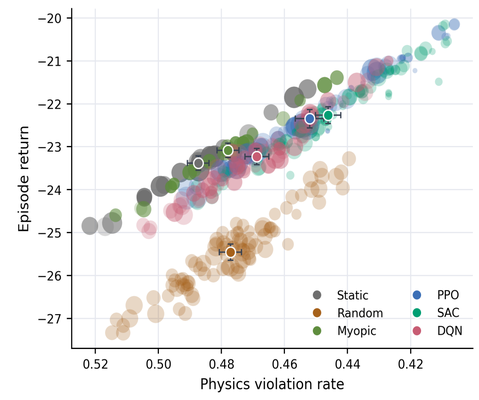
Main finding. Learned policies improve reward over non-RL baselines, but each policy occupies a different trustworthiness point. SAC leads reward, volume, physical validity, and artificial liquidity; PPO gives the lowest slippage.

Table 3. Learned-policy means.

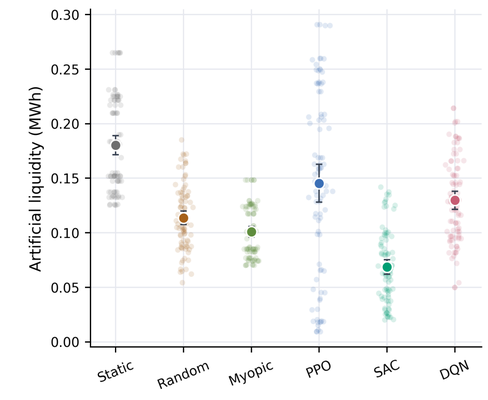
Setting	Reward	Violation	Art. liq.
RL	-22.61	0.4557	0.1146
RL+LLM	-22.76	0.4626	0.1181
RL no-physics	0.28	0.4858	**0.2012**
LLM no-physics	0.26	0.4843	**0.1901**



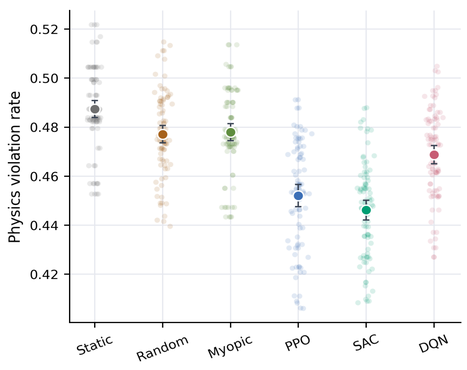
Reward distribution.



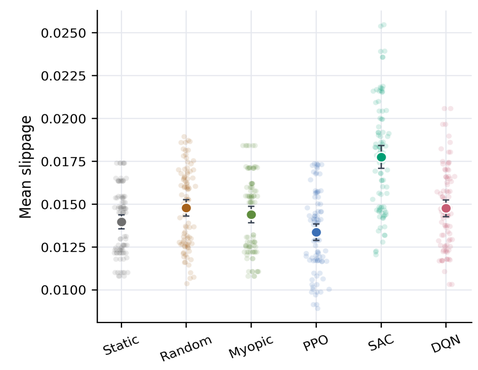
Utility-safety frontier.



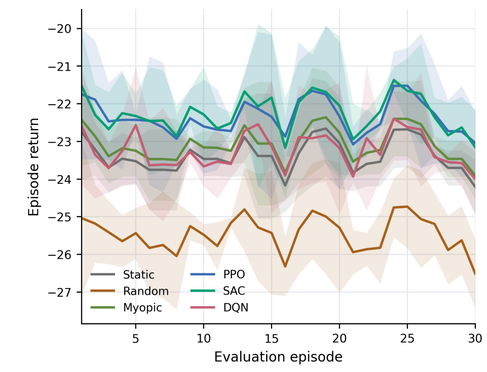
Artificial liquidity.



Physics violation.



Mean slippage.



Reward trajectory.

Result synthesis. Reward gains are interpreted together with physical validity and market execution. The no-physics rows show why scalar utility alone is insufficient: reward can rise while artificial liquidity and violation exposure worsen.

Evaluation implication. The benchmark should be read as a multi-objective stress test: a deployable policy must improve market utility, avoid backing suspicious supply, preserve execution stability, and expose its interventions through traceable audit logs.

How to read the panel set. Table 2 gives policy-level averages, while the six panels check distributional spread, utility-safety shape, artificial liquidity, constraint violations, slippage, and temporal reward stability. SAC's average advantage is meaningful because it also reduces artificial liquidity; PPO's lower slippage marks a deployment tradeoff rather than a universal winner.

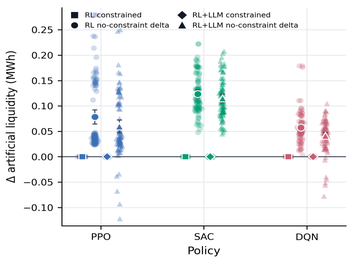
Benchmark message. Reward is trusted only when it is backed by feasible supply, stable execution, and a traceable intervention record.

Constraint and Audit Diagnostics

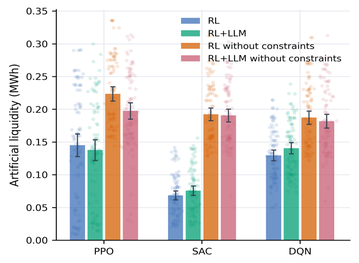
Table 4. Physics-penalty ablation.

Setting	ΔViolation	ΔArt. liq.
RL PPO	0.0311	0.0785
RL SAC	0.0412	**0.1238**
RL DQN	0.0182	0.0576
LLM PPO	0.0225	0.0598
LLM SAC	0.0344	0.1149
LLM DQN	**0.0081**	**0.0413**

Removing the physics penalty inflates scalar utility while increasing artificial liquidity. LLM governance reduces some unsafe exposure, but it cannot fully repair reward misspecification.



Risk deltas.



Liquidity by setting.

Table 5. Audit trigger profile.

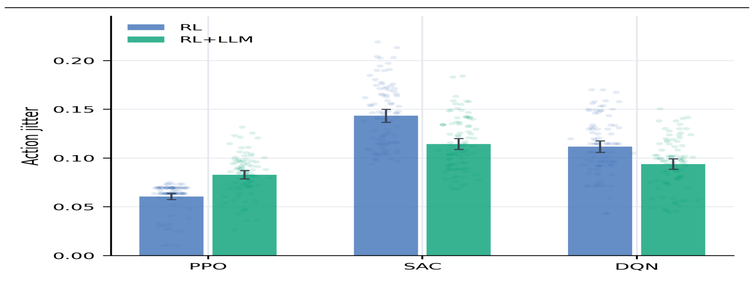
Policy	Phys.	Gap	Slip.	Jitter
DQN	0.120	0.440	0.168	0.230
PPO	0.121	0.440	0.171	0.171
SAC	0.120	0.441	0.137	0.211

Diagnostic takeaway. Gap and physics triggers dominate safety reviews, while jitter flags unstable controller behavior. The largest artificial-liquidity increase appears when SAC loses the physics penalty, indicating reward misspecification rather than random rollout noise.

Conclusions and Acknowledgments

Table 6. LLM Planner/Auditor effect.

Policy	ΔReward	ΔJitter	Audit	Rev.
PPO	-0.094	0.0223	0.344	0.961
SAC	-0.062	**-0.0290**	0.342	0.966
DQN	-0.277	**-0.0179**	0.372	0.958



Action stability under RL and RL+LLM.

Conclusions

- Trustworthy agentic AI evaluation requires physical constraints and transparent intervention logs.
- RL improves governance utility, but reward-only optimization can back invalid supply.
- LLM auditing supports risk control, but it is not a substitute for correct reward design.

Acknowledgments. Shilin Ou is grateful for the support from the Summer Research Scholar Program at Duke Kunshan University, supervised by Prof. Luyao Zhang.

GitHub **Hugging Face** **arXiv**