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

Anonymous Author(s)
Anonymous Institution

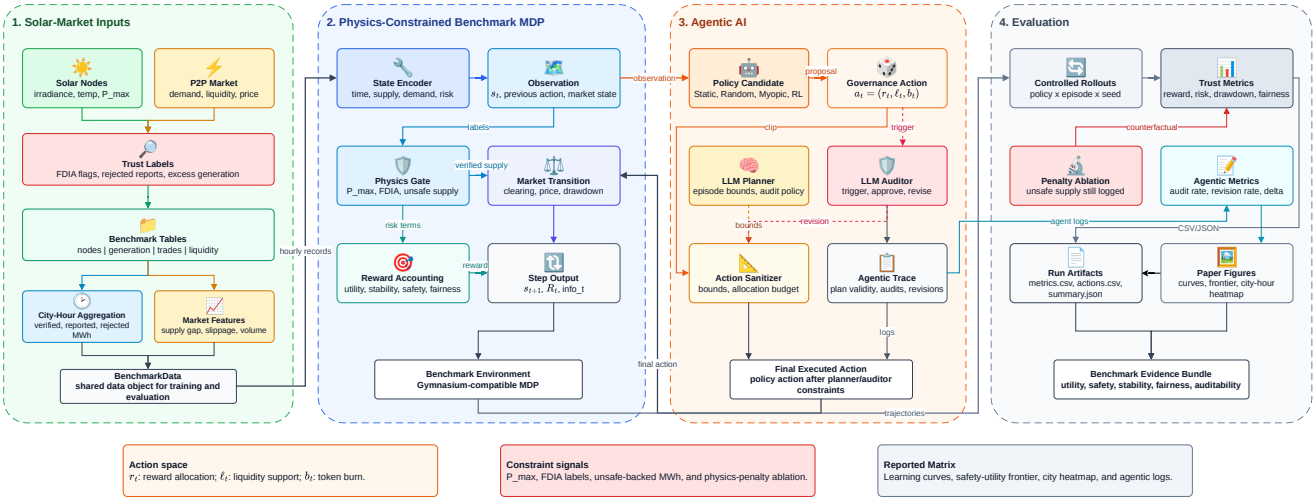


Figure 1: Overall SolarChain-Eval benchmark pipeline. The figure summarizes the end-to-end workflow from data loading, RL training, policy evaluation, output generation, and trustworthiness evidence for decentralized energy-market governance.

Abstract

As Agentic AI increasingly governs cyber-physical systems such as decentralized energy markets, the evaluation focus need to expand beyond mere task utility to rigorous assessments of trustworthiness. Traditional reinforcement learning (RL) often suffer from reward misspecification. When driven purely by economic metrics like market liquidity, they are prone to exploiting unverified physical data. To address this, We propose *SolarChain-Eval*, a physics-constrained benchmark designed to assess the trustworthiness of economic agents. Moving beyond standard scalar rewards, our RL framework evaluates autonomous behavior across a multi-dimensional spectrum, including physical safety, market stability, action smoothness, and spatial fairness, utilizing real-world urban solar generation data. Furthermore, *SolarChain-Eval* integrates RL with a bi-level AI governance architecture. We deploy a transparent Large Language Model Planner and Auditor layer that provides deployment-time, step-level oversight. Our extensive empirical evaluations expose the inherent utility-safety trade-off in standard RL policies, demonstrating the necessity of combining physical constraints with our bi-level RL+AI agent architecture. Ultimately, *SolarChain-Eval* grounds the evaluation and trustworthiness of Agentic AI in the realities of decentralized energy markets. By enforcing physical solar generation limits and logging LLM-driven audits, our benchmark ensures that autonomous market administrators operate transparently within real-world boundaries. In the spirit of open science, we have released the dataset on Hugging Face and the source code on GitHub as open-access resources to facilitate replication.

CCS Concepts

• **Computing methodologies** → Multi-agent systems; Reinforcement learning; • **Applied computing** → Economics; • **General and reference** → Performance.

Keywords

Agentic AI, reinforcement learning, trustworthiness evaluation, tokenomics, decentralized energy markets

1 Introduction

The integration of Agentic Artificial Intelligence (AI) and Cyber-Physical Systems (CPS), particularly in decentralized energy markets, provides opportunities for autonomous economic governance [23, 34]. In peer-to-peer (P2P) solar energy trading platforms, reinforcement learning (RL) agents are increasingly tasked with managing market liquidity, token issuance, and incentive distribution [16, 28]. However, unlike purely digital environments, energy markets are bounded by physical ground truths, such as localized solar irradiance and photovoltaic panel capacities

.

Deploying reward-maximizing agents in physical systems faces important risks. Standard RL models might prioritize economic metrics like market liquidity. Without physical constraints, these agents can easily learn to exploit the system[1]. For instance, during False Data Injection Attacks (FDIA) [18], an agent might accept impossible solar generation data simply to artificially boost liquidity. This behavior compromises market integrity, presenting a

Table 1: Comparison of Prevailing Bi-Level Architectures Integrating RL and LLM Agents

Framework	Domain	Synergy Mechanism (LLM + RL)	Limitations & Challenges
RE-GoT [32]	Robotic Control & Manipulation	Upper-level LLM uses Graph-of-Thoughts to evolve reward functions; Vision models analyze RL rollouts for semantic feedback.	High latency due to LLM/VLM inference bottlenecks in real-time control loops.
RL^2 [31]	Active Distribution Networks	LLM translates operator safety intents into mathematical penalties; adapts parameters via DRL feedback.	Sensitive to LLM hallucination during algebraic adjustments.
ADMM-LLM [3]	Microgrid Heat-Electricity	LLM acts as an optimizer that observes optimization residuals to predict and adjust decentralized ADMM penalty parameters.	Heuristic trial-and-error lacks strict mathematical convergence guarantees.

promising research direction: while existing benchmarks effectively measure economic utility, there remains a growing need to assess agent trustworthiness and safety against real-world constraints.

To address this, we propose *SolarChain-Eval*, a physics-constrained benchmark for evaluating the trustworthiness of Agentic AI in decentralized energy markets. *SolarChain-Eval* simulates real-world urban energy nodes using historical weather data. In this environment, an agent acts as an autonomous market administrator, making hourly decisions on reward ratios, market liquidity, and token burn rates. To evaluate these decisions beyond standard reward maximization, *SolarChain-Eval* assesses systemic trustworthiness by tracking physical violations, market drawdowns, action stability, and spatial fairness. Specifically, this study mainly concentrates on the following research questions:

- **RQ1:** How do RL framework balance economic utility and trustworthiness metrics compared to static and heuristic baselines in a decentralized energy market?
- **RQ2:** To what extent do reward-maximizing agents exploit invalid generation to artificially inflate market liquidity when physics-informed penalties are removed?
- **RQ3:** How do AI agents' governance decisions on reward allocation, liquidity provisioning, and token burns affect market stability, action reliability, and spatial fairness across urban energy nodes?

2 Literature Review

2.1 Reinforcement Learning in Tokenomics

Traditional economic mechanisms struggles when applied to decentralized energy markets due to the restrictive assumptions of stationary environments [36]. Distributed Energy Resources (DERs) are governed by intermittent generation constraints and Peer-to-Peer (P2P) network topologies [6, 17]. However, tokenomic mechanisms are subject to extreme volatility driven by liquidity depth, speculative token price trajectories, and adoption rates [7, 22]. RL provides a robust solution by formulating market governance as a Markov Decision Process [36]. It can adapt to high-dimensional state spaces, optimizing long-term objectives that balance physical utility with economic stability[37].

In current decentralized market governance literature[24, 29], macroeconomic control is typically abstracted into continuous governance levers, such as the reward ratio (r_t), liquidity ratio (l_t), and token burn rate (b_t). In simulated environments, the normalized action vector a_t is decoded into specific economic parameters. To evaluate system trustworthiness, researchers widely rely on standardized frameworks like *Stable-Baselines3* [25] to benchmark distinct RL algorithmic paradigms:

Proximal Policy Optimization (PPO) is an on-policy actor-critic algorithm widely applied to continuous action space governance tasks. To prevent destructively large policy updates, PPO utilizes a clipped surrogate objective [27]:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min(p_t(\theta)\hat{A}_t, \text{clip}(p_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right] \quad (1)$$

where θ denotes the policy network parameters, $p_t(\theta)$ is the probability ratio, $\hat{A}_t$ represents the estimated advantage function, and ϵ is the clipping hyperparameter.

In tokenomic governance, this high training stability translates to practical advantages by effectively suppressing action jitter and ensuring the smooth transition of macroeconomic parameters[35].

Also supporting continuous control, Soft Actor-Critic (SAC) is an off-policy algorithm. By incorporating a maximum entropy mechanism into its objective function [11], SAC proactively encourages the agent to explore diverse tokenomic strategies:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi} \left[r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t)) \right] \quad (2)$$

where $r(s_t, a_t)$ denotes the immediate reward, α is the temperature parameter, and $\mathcal{H}(\pi(\cdot|s_t))$ represents the entropy of the policy π at state s_t.

Existing research demonstrates that this mechanism not only improves sample efficiency but also prevents the policy from prematurely converging to suboptimal local minima in highly nonlinear market environments[4].

As a robust value-based baseline algorithm, Deep Q-Network (DQN) is applied to discrete action spaces. In multidimensional control tasks, continuous levers are typically discretized. Despite inherent quantization errors, DQN excels at accurately estimating the action-value function $Q(s, a)$ [14, 30].

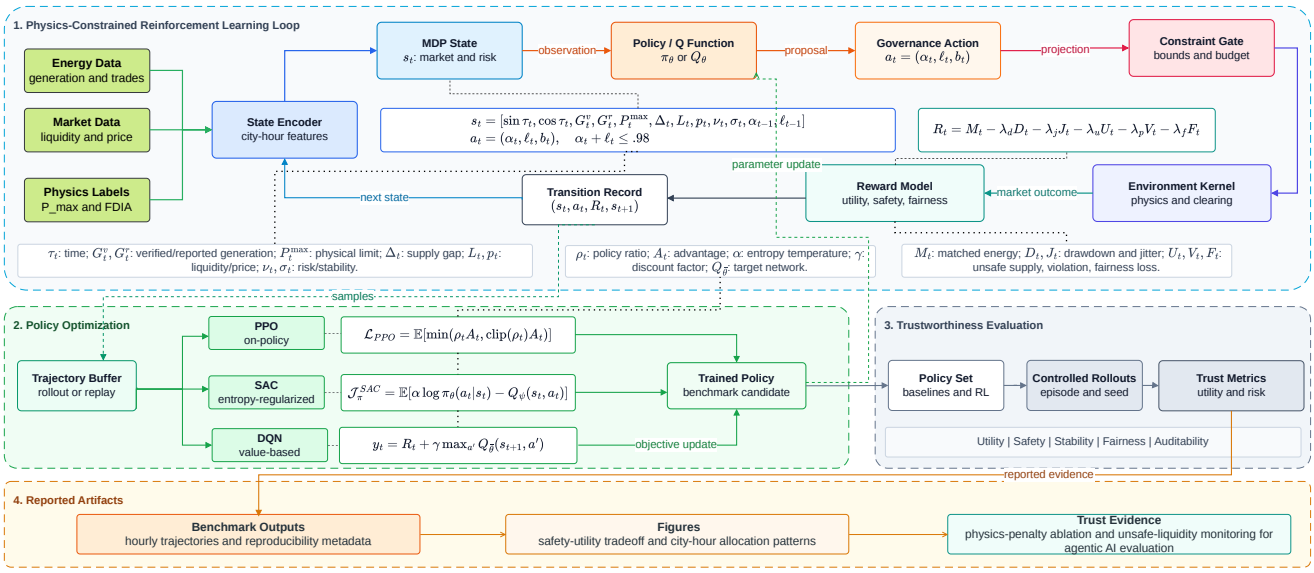


Figure 2: Reinforcement learning evaluation framework for SolarChain-Eval. The diagram connects market and physical inputs to policy actions, environment feedback, PPO/SAC/DQN updates, and trustworthiness-oriented evaluation across utility, safety, stability, smoothness, and fairness.

2.2 RL & LLM Agent Integration

RL has shown strong performance in high-dimensional continuous control tasks. However, its black-box nature limits its reliability in infrastructure [2]. In practice, human auditors may find it difficult to explain why an RL policy makes adjustments [26]. Furthermore, while post-hoc AI methods offer retrospective insights, they cannot proactively prevent unsafe actions[10].

Recent studies address this by integrating Large Language Models (LLMs) with RL systems. It acts as high-level reasoners to process task context and operator instructions [33]. This ability allows them to translate operator intents into mathematical penalties that can support lower-level decision making [8, 20].

Consequently, current research is shifting toward a bi-level RL-LLM architecture [5]. The lower layer uses RL policies to make frequent control decisions based on value estimates such as $V(s)$ or $Q(s, a)$. The upper layer uses LLMs for slower and more abstract reasoning tasks. Prior work has used LLMs to convert operator intent into reward penalties [15], revise reward functions, and tune optimization parameters [21]. Table 1 summarizes representative examples of this design.

Recent advancements in RL boost agent-based optimization in Cyber-Physical Systems, especially in decentralized energy markets [12, 13]. As the focus shifts toward real-world deployment [9], there is a growing need to complement traditional reward-centric benchmarks with evaluations of trustworthiness. Specifically, ensuring safe operation requires assessing an agent's adherence to physical constraints and its resilience in adversarial scenarios [19]. Following this view, our benchmark couples economic dynamics with physical ground truths and evaluates agentic trustworthiness across multiple dimensions.

3 Methodology

Instead of merely maximizing short-term trading volume, our benchmark evaluates the overall trustworthiness of an autonomous economic agent. It assesses whether the agent can maintain market utility while respecting physical limits, ensuring market stability, reducing action jitter, and preserving spatial fairness. Each episode represents a 24 hour market cycle, sampled from a 720-hour dataset from April 2026. Episodes are constrained to start at the beginning of the day. This temporal design ensures a consistent daily decision horizon for the agent while seamlessly preserving real-world weather patterns.

3.1 RL Formulation

Figure 2 summarizes the RL loop. Energy generation, market demand, and physics labels are aggregated into an hourly benchmark state. Based on this state, a policy proposes a governance action.

The state s_t and action a_t are defined as:

$$\begin{aligned} s_t &= [\sin\tau_t, \cos\tau_t, G_t^v, G_t^r, P_t^{\max}, \Delta_t, L_t, p_t, \nu_t, \sigma_t, \alpha_{t-1}, \ell_{t-1}], \\ a_t &= (\alpha_t, \ell_t, b_t), \qquad \alpha_t + \ell_t \le 0.98. \end{aligned} \tag{3}$$

Here, τ_t is the time angle, G_t^v is verified generation, G_t^r is reported generation, $P_t^{\max}$ is the physical PV upper bound, Δ_t is the supply-demand gap, L_t is market liquidity, p_t is token price, ν_t is the raw risk signal, and σ_t is the static slippage.

The action a_t comprises the reward allocation ratio α_t, liquidity injection ratio ℓ_t, and token burn rate b_t. Standard algorithms like PPO and SAC operate on the continuous action space, while DQN utilizes a discretized $5^3 = 125$ action grid over the same parameters. All trained policies are evaluated under the uniform benchmark setting detailed in Table 2.

Table 2: Main benchmark configuration.

Item	Value
Cities	Beijing, Shanghai, Chengdu, Shenzhen, Hangzhou
Energy nodes	50
Main period	2026-04-01 to 2026-04-30
Hourly market states	720
Generation records	36,000
P2P trade records	1,185
Episode length	24 hours
Evaluated Policies	Static, Random, Myopic, PPO, SAC, DQN

3.2 Physics Limit and Market Transition

SolarChain-Eval proposed a physics limit, which verifies whether the reported generation is supported by physical limits and FDIA labels. Let X_t denote suspicious or invalid supply. The benchmark not only records this anomaly but quantifies the extent to which the policy economically backs it:

$$V_t = \frac{\beta_t X_t}{\max(G_t^v + \beta_t X_t, \epsilon)}, \qquad \beta_t = \frac{\min(\alpha_t + \ell_t, 0.98)}{0.98}. \tag{4}$$

In Eq. (4), V_t is the action-dependent physics violation term. A policy incurs higher risk when it heavily backs invalid supply. This formulation directly addresses RQ2: through a no-physics-penalty ablation, we can observe whether a purely reward-driven agent exploits invalid generation to fabricate artificial liquidity.

Additionally, the market module processes the backed supply, consisting of verified supply and the action-backed suspicious supply. It updates the matched energy M_t and computes the residual liquidity L_{t+1} for the next hour:

$$\begin{aligned} M_t &= \min(L_t + \ell_t(G_t^v + \beta_t X_t),\ \widetilde{Q}_t), \\ L_{t+1} &= \max(L_t + \ell_t(G_t^v + \beta_t X_t) - M_t, 0). \end{aligned} \tag{5}$$

Here, $\widetilde{Q}_t$ represents the effective demand after token burning, computed as

$$\widetilde{Q}_t = Q_t^d \max(0.75, 1 - 0.80 b_t),$$

where Q_t^d is the raw market demand. This mechanism explicitly links the agent's governance decisions to both trading volume and market liquidity depth.

3.3 Reward Design and RL Optimization

To optimize for both utility and market trustworthiness, the step reward implements a series of penalties (Figure 2):

$$\begin{aligned} R_t &= M_t - \lambda_d D_t - \lambda_j J_t - \lambda_u U_t - \lambda_p \Phi_t - \lambda_f F_t, \\ \Phi_t &= V_t + A_t, \qquad A_t = \beta_t X_t. \end{aligned} \tag{6}$$

The positive term M_t rewards successful market clearing. The corresponding penalties address liquidity drawdown (D_t), action jitter (J_t), unmet demand (U_t), composite physics risk (Φ_t), and spatial unfairness across cities (F_t). Here, Φ_t combines the normalized physics violation rate V_t and the action-backed suspicious supply amount A_t.

The baseline policies are trained using Stable-Baselines3. The benchmark incorporates three learned controllers with distinct optimization biases: PPO for stable on-policy updates, SAC for entropy-guided continuous exploration, and DQN for discrete Q-value learning. For instance, DQN optimizes the following target over the discretized action grid:

$$y_t = R_t + \gamma \max_{a'} Q_{\bar{\theta}}(s_{t+1}, a'). \tag{7}$$

Static, random, and myopic controllers are also evaluated as non-learning baselines to provide a comparison.

3.4 AI-Agent Governance Layer

To evaluate whether agentic oversight improves deployment-time safety, we introduce an AI-agent governance layer powered by LLMs, as shown in Figure 3. We insert this layer between the trained RL policy (e.g., PPO, SAC, DQN) and the market environment. This layer is active only during evaluation and is never used during RL training. This separation ensures that any observed safety improvement stems from deployment-time guardrails rather than from reward shaping, policy retraining, or parameter updates.

The governance layer consists of two components: a Planner and an Auditor:

The Planner operates at the episode level. At the start of each 24-hour episode, it reviews episode-level context, including supply-demand adequacy, physics-risk statistics, static-market slippage, and benchmark action constraints. Based on this context, the Planner establishes admissible bounds for the RL action components, namely the reward ratio α_t, liquidity ratio ℓ_t, and burn rate b_t. It also outputs an episode-specific audit policy, including risk thresholds, a maximum audit budget B_e, a target audit rate, and a minimum cooldown length c_e. At each hourly step, the RL policy's proposed action is first decoded into market variables and clipped by the Planner-defined bounds and benchmark constraints, yielding a bounded action $\bar{a}_t$.

The Auditor provides sparse step-level oversight. Instead of reviewing every action, it is invoked only when risk signals indicate possible safety or market-integrity failures. To avoid notational conflict with the reward terms in Eq. (6), we denote the audit-side action-instability signal by κ_t. It is computed as the ℓ_1 distance between the current bounded action and the previous executed action:

$$\kappa_t = \|\bar{a}_t - a_{t-1}\|_1. \tag{8}$$

We further define the audit-side normalized supply-demand gap as

$$\Delta_t^{audit} = \frac{G_t^v - Q_t^d}{\max(Q_t^d, \epsilon)}, \tag{9}$$

where G_t^v denotes verified energy supply, Q_t^d denotes market demand/load, and $\epsilon > 0$ avoids division by zero. Negative values of Δ_t^{audit} indicate supply shortfall.

Audit triggers are divided into hard and soft categories. Hard triggers respond to immediate safety or market-integrity threats, including excessive physical violation rate v_t, supply shortfall Δ_t^{audit}, and high static-market slippage σ_t:

$$H_t = \mathbb{I}\left[v_t > \tau_v \ \vee \ \Delta_t^{audit} < \tau_\Delta \ \vee \ \sigma_t > \tau_\sigma\right]. \tag{10}$$

Soft triggers capture erratic controller behavior through action instability:

$$S_t = \mathbb{I}\left[\kappa_t > \tau_\kappa\right]. \tag{11}$$

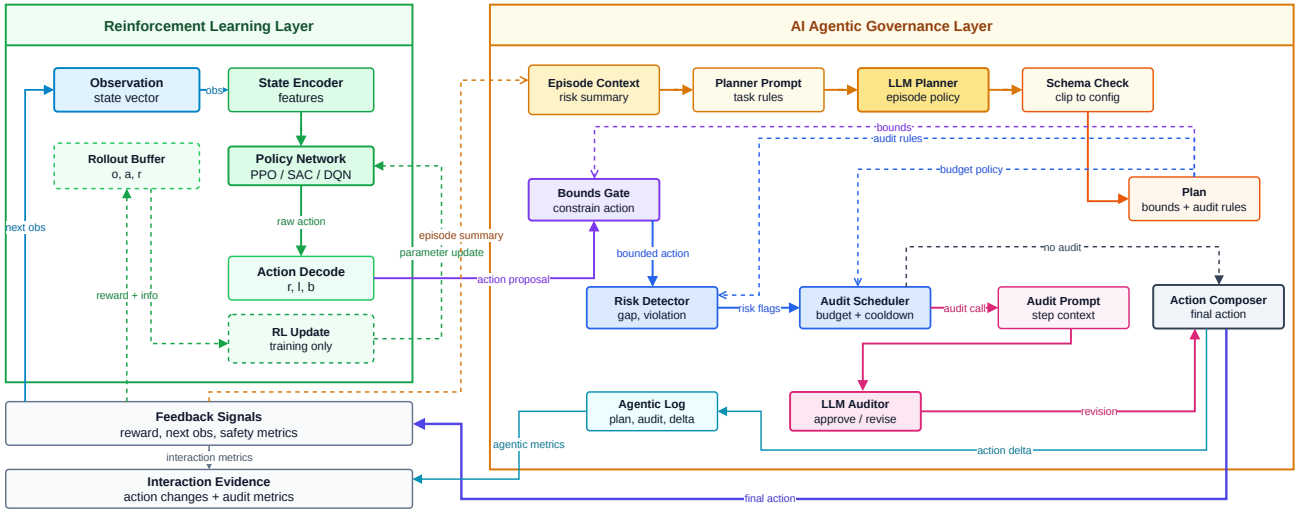


Figure 3: Interaction between the reinforcement learning layer and the AI-agent governance layer. The RL policy proposes market actions, which are bounded by the Planner and conditionally revised by the Auditor under high-risk scenarios.

An action becomes audit-eligible when either trigger fires:

$$E_t = H_t \vee S_t. \tag{12}$$

To prevent excessive LLM intervention, ordinary soft-triggered audits are rate-limited by the episode budget and cooldown. In contrast, hard triggers bypass these limits due to their severity. Let $N_{e,t}^{audit}$ denote the number of audits already used in episode e before step t, and let t_e^{last} denote the most recent audited step. The final audit decision is

$$\mathbb{I}_t^{audit} = E_t \wedge \left[H_t \ \vee \ \left(N_{e,t}^{audit} < B_e \ \wedge \ t - t_e^{last} > c_e\right)\right]. \tag{13}$$

Thus, severe physics or market failures can trigger immediate review, while softer instability signals are subject to sparse intervention limits.

If $\mathbb{I}_t^{audit} = 0$, the market directly executes the bounded action $\bar{a}_t$. If $\mathbb{I}_t^{audit} = 1$, the Auditor receives the current observation, the bounded proposed action, the previous action, the Planner policy, and recent market diagnostics. It then returns a decision: `approve`, which keeps $\bar{a}_t$, or `revise`, which replaces it with a conservative corrected action. The wrapper deterministically applies this decision before calling the environment step.

Finally, to make the governance layer auditable and economically valid, all LLM outputs must follow structured schemas and numerical validation. Planner outputs are clipped to valid benchmark action ranges and audit-policy ranges. Auditor outputs are sanitized against market constraints, including individual bounds on α_t, ℓ_t, and b_t, as well as the global allocation constraint $\alpha_t + \ell_t \leq 0.98$. This schema-enforced interface prevents free-form text from directly controlling market execution and ensures that every agentic intervention remains traceable and reproducible during evaluation.

3.5 Evaluation Metrics

SolarChain-Eval evaluates each policy from two perspectives: market utility and system trustworthiness. We compare six baselines: Static, Random, Myopic, PPO, SAC, and DQN. Static uses fixed governance parameters, Random samples actions uniformly, and Myopic follows a greedy rule based on current supply, demand, and risk signals. PPO and SAC operate in the continuous action space, while DQN uses the discretized action grid.

We use three evaluation settings:

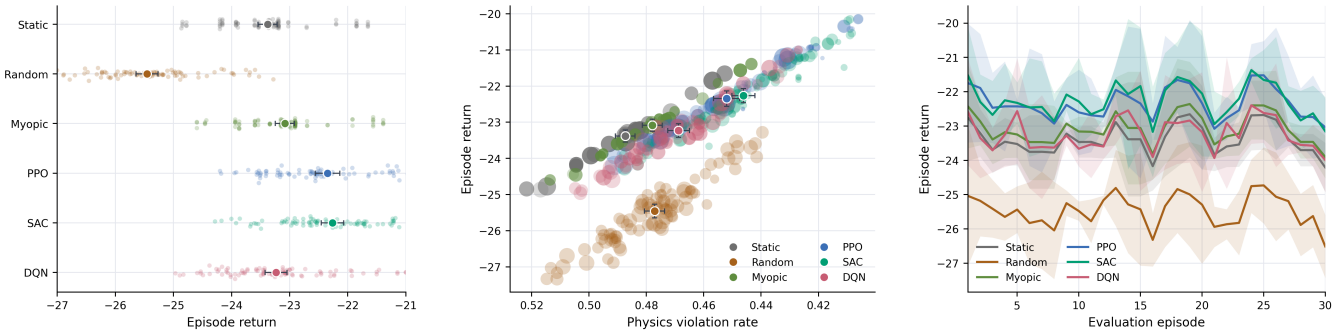
- **Main Benchmark:** evaluates all baselines using the full reward in Eq. (6). This setting tests whether RL policies can improve utility while maintaining safety, stability, smoothness, and fairness.
- **Reward Ablation:** removes the physics penalty from Eq. (6), while still logging physical violations. This setting tests whether agents exploit unsafe generation when physical risk is not penalized.
- **Agentic Evaluation:** adds the AI-agent governance layer to trained RL policies during evaluation. This setting tests whether LLM-based oversight can revise unsafe actions without retraining the RL controller.

Table 3: Trustworthiness-oriented evaluation dimensions and reported evidence.

Dimension	Reported Evidence
Utility	Cumulative reward, episode trading volume
Physics Safety	Physics violation rate, artificial liquidity
Market Stability	Liquidity drawdown, token drawdown, slippage
Action Smoothness	Action jitter across hourly decisions
Spatial Fairness	City-level reward allocation imbalance
Auditability (Agentic)	Plan validity, audit rate, revision rate, action delta

Table 4: Main benchmark performance under the physics-constrained reward. Values are mean ± standard deviation over all three seeds and 90 rollouts per policy. Higher cumulative reward and trading volume are better; lower physics violation, slippage, and artificial liquidity indicate safer market governance.

Policy	Cumulative Reward	Trading Volume	Physics Violation Rate	Mean Slippage	Artificial Liquidity
Static	-23.38 ± 0.80	0.514 ± 0.099	0.4874 ± 0.0166	0.0140 ± 0.0020	0.1803 ± 0.0419
Random	-25.46 ± 0.91	0.480 ± 0.093	0.4771 ± 0.0172	0.0148 ± 0.0023	0.1137 ± 0.0297
Myopic	-23.08 ± 0.80	0.456 ± 0.088	0.4779 ± 0.0168	0.0144 ± 0.0023	0.1009 ± 0.0235
PPO	-22.35 ± 1.02	0.470 ± 0.097	0.4520 ± 0.0219	0.0134 ± 0.0023	0.1453 ± 0.0840
SAC	-22.26 ± 0.94	0.521 ± 0.100	0.4461 ± 0.0196	0.0178 ± 0.0032	0.0686 ± 0.0328
DQN	-23.23 ± 0.91	0.491 ± 0.096	0.4688 ± 0.0182	0.0148 ± 0.0023	0.1298 ± 0.0397

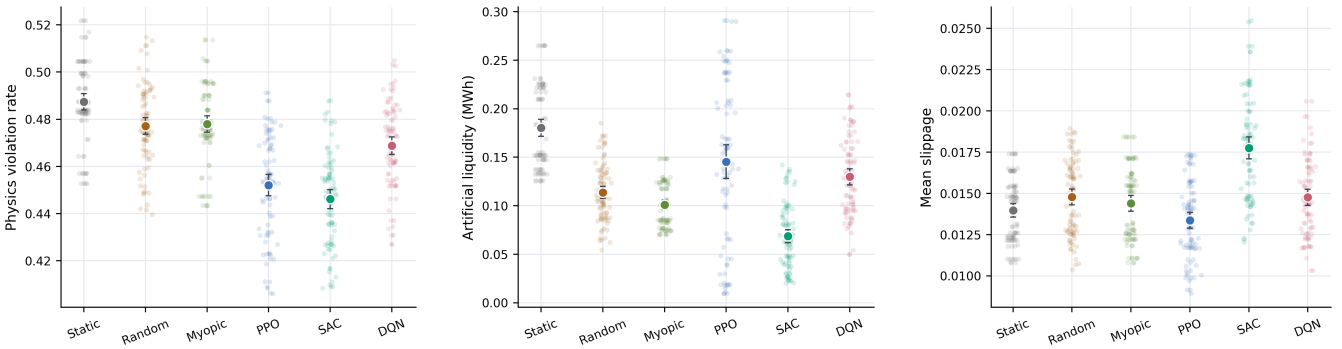


(a) Rollout-level cumulative reward.

(b) Utility–safety frontier. Each point is one rollout; larger markers indicate higher artificial liquidity.

(c) Episode-wise reward trajectory with 95% confidence bands across seeds.

Figure 4: Main benchmark performance under the physics-constrained reward.



(a) Physics violation rate.

(b) Artificial liquidity.

(c) Mean slippage.

Figure 5: Trustworthiness metrics in the main benchmark. Translucent points are individual rollouts.

4 Evaluation

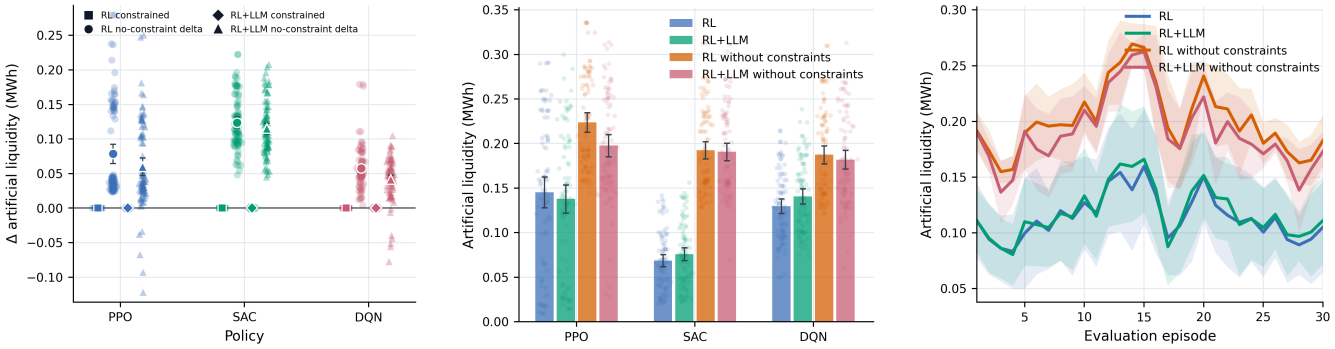
Our empirical evaluation addresses the research questions using data aggregated across three independent seeds. We benchmark three agentic policies (PPO+LLM, SAC+LLM, DQN+LLM) against six non-agentic baselines (PPO, SAC, DQN, Static, Random, Myopic). The evaluation dataset (Appendix B) comprises 1,620 episode-level records, 38,880 hourly actions, 194,400 hourly state records, and 12,960 LLM governance log entries.

4.1 Benchmark Performance

We first evaluate the RL policies against three baseline models(Static, Random, Myopic) for market governance. To comprehensively assess trustworthiness, Table 4 reports utility, physical consistency, and market-execution metrics simultaneously. Results indicate that learned RL control provides a clear advantage over non-RL baselines. Across 90 rollouts per policy, the RL policies achieve a higher mean cumulative reward than the static, random, and myopic baselines

Table 5: Paired without-constraints deltas with and without the LLM governance layer. Each row uses 90 paired rollouts across three seeds. Positive Δ reward indicates that removing the physics penalty increases scalar utility; positive Δ violation or artificial liquidity indicates increased safety risk.

Governance	Policy	Δ Reward	Δ Physics Violation	Δ Artificial Liquidity	Δ Slippage
RL	PPO	22.61 ± 1.10	0.0311 ± 0.0084	0.0785 ± 0.0678	-0.0012 ± 0.0010
RL	SAC	22.52 ± 1.01	0.0412 ± 0.0095	0.1238 ± 0.0386	-0.0041 ± 0.0017
RL	DQN	23.55 ± 0.91	0.0182 ± 0.0101	0.0576 ± 0.0317	-0.0018 ± 0.0014
RL+LLM	PPO	22.67 ± 1.17	0.0225 ± 0.0107	0.0598 ± 0.0631	-0.0011 ± 0.0016
RL+LLM	SAC	22.62 ± 0.96	0.0344 ± 0.0102	0.1149 ± 0.0388	-0.0038 ± 0.0017
RL+LLM	DQN	23.78 ± 0.88	0.0081 ± 0.0079	0.0413 ± 0.0317	-0.0015 ± 0.0018



(a) Paired without-constraints artificial-liquidity deltas for RL and RL+LLM.

(b) Artificial liquidity across RL, RL+LLM, and their without-constraints ablations.

(c) Episode-wise artificial-liquidity trajectories for all four learned-agent settings.

Figure 6: Ablation experiment of physics constraints. Translucent points are rollout-level or paired-rollout observations.

(−22.61 vs. −23.97; Welch's t-test, $p < 10^{-28}$). This improvement is primarily driven by continuous-control methods, as both PPO and SAC outperform all baselines ($p < 10^{-6}$ for all pairwise tests). Notably, these non-RL baselines are non-trivial: they represent stable governance rules (static), pure action-space exploration (random), and local rationality (myopic). Their underperformance confirms that effective market governance requires adaptive RL rather than rigid or one-step heuristics.

However, trustworthiness metrics reveal a complex utility-safety tradeoff. While SAC achieves the highest cumulative reward and trading volume alongside the lowest physics violation rate, it suffers from the highest mean slippage among the RL policies. Conversely, PPO offers the lowest slippage but generates more artificial liquidity. DQN improves upon random control but fails to consistently outperform the myopic baseline, illustrating that RL architecture choice is critical for fine-grained control tasks. Consequently, these patterns expose an evaluation frontier. For rigorous trustworthiness evaluation, maximizing reward does not guarantee a safe policy if the gain introduces market friction or lacks physical backing. Figure 4 illustrates this frontier, showing the distinct operational trade-offs and remaining within-policy variability.

The metric decomposition in Figure 5 further unpacks this frontier. Although RL policies reduce overall physics violations compared to non-RL baselines, the underlying risk is redistributed across artificial liquidity and slippage. SAC excels at physical credibility, whereas PPO prioritizes slippage control.

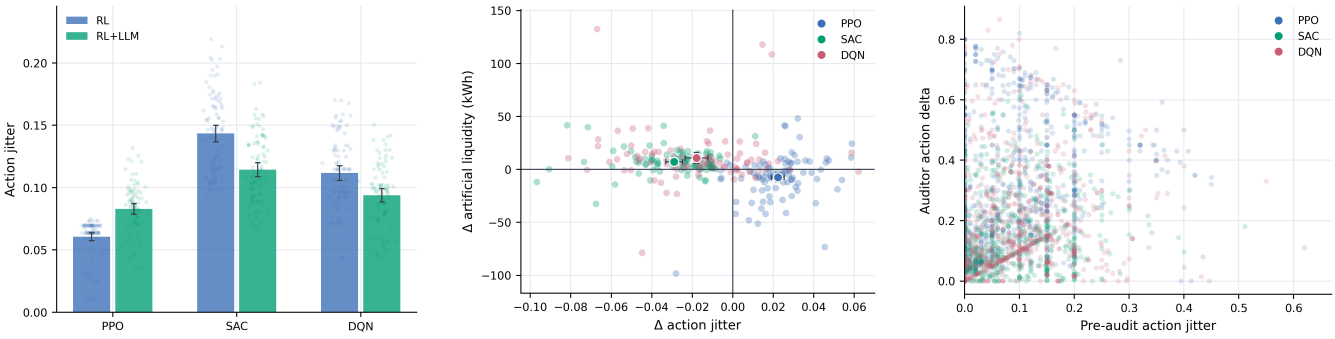
4.2 Ablation Experiment of Physics Constraints

The ablation study investigates the necessity of adding physical constraints to our RL and AI agent models. We examine four configurations: standard RL, RL without physics constraints, RL with LLM governance, and RL with LLM governance but lacking physics constraints. In the unconstrained settings, we remove the physics penalty from the reward function(Eq.6) while continuing to track physical violations and artificial liquidity. This design decouples the agent's optimization objective from the benchmark's safety diagnostics. Consequently, scalar reward maximization does not inherently equate to a better market outcome. If an agent boosts its reward by generating impossible liquidity, the benchmark could explicitly expose it.

The empirical results in Table 5 demonstrate that removing the physics penalty inflates scalar rewards, but at the cost of severe safety risks. Under raw RL deployment, artificial liquidity surges by 0.0785, 0.1238, and 0.0576 MWh for PPO, SAC, and DQN, respectively. This failure mode persists even under RL+LLM deployment, where artificial liquidity still increases by 0.0598, 0.1149, and 0.0413 MWh. This illustrates a consistent shift toward higher artificial liquidity when constraints are ablated across both governance regimes (Figure 6a). Although the LLM layer mitigates the magnitude of

Table 6: Paired effects of adding the LLM Planner/Auditor layer under the physics-constrained reward. Δ columns are RL+LLM minus RL over 90 matched rollouts across three seeds; audit and revision rates are computed from 2,160 governance log steps per policy.

Policy	Δ Reward	Δ Action Jitter	Δ Artificial Liquidity	Δ Spatial Fairness	Audit Rate	Revision Rate
PPO	-0.094 ± 0.293	0.0223 ± 0.0154	-0.0074 ± 0.0216	0.0004 ± 0.0008	0.344	0.961
SAC	-0.062 ± 0.268	-0.0290 ± 0.0198	0.0073 ± 0.0110	0.0003 ± 0.0006	0.342	0.966
DQN	-0.277 ± 0.503	-0.0179 ± 0.0274	0.0109 ± 0.0252	-0.0002 ± 0.0021	0.372	0.958



(a) Action stability: RL+LLM reduces action jitter for SAC and DQN,

(b) Paired RL+LLM-minus-RL changes in action jitter and artificial liquidity.

(c) Step-level relationship between pre-audit action jitter and Auditor action delta.

Figure 7: Agentic governance diagnostics for the LLM Planner/Auditor layer. The panels combine summary auditability statistics, paired episode-level governance effects, and step-level audit traces.

these violations, it cannot eliminate the misalignment caused by a misspecified reward.

This comparison highlights the complementary roles of physical constraints and AI agent governance. While LLM governance audits actions post-hoc, physics-constrained RL embeds safety during optimization to prevent the exploitation of unviable liquidity. Aggregated across learned policies, mean artificial liquidity remains low in constrained environments (0.1146 MWh for RL; 0.1181 MWh for RL+LLM) but spikes in unconstrained settings (0.2012 MWh for RL; 0.1901 MWh for RL+LLM) (Figure 6b). Furthermore, this risk inflation is not an artifact of isolated rollouts, but a persistent behavioral trajectory woven throughout the evaluation episodes (Figure 6c). This confirms that unconstrained optimization remains riskier, even with active agentic oversight.

4.3 Evaluation of Agentic Layer

Having proven the function of physics constraints, we now evaluate the AI agent layer's capacity for verifiable oversight. By analyzing transparent audit traces from market execution, we assess the system's effectiveness as an auditable risk-control interface. This evaluation specifically quantifies intervention triggers, the magnitude of action modifications, and the underlying rationales for these corrections.

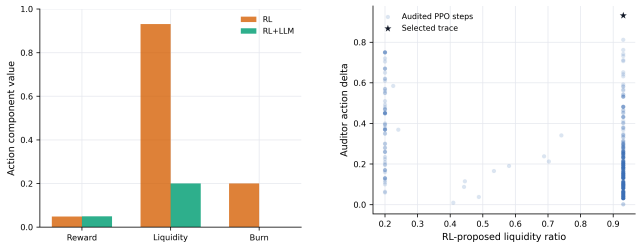
To investigate the reliability of the AI agent mechanism, Table 6 presents metrics derived from 12,960 step-level governance logs (encompassing three seeds, two constraint settings, and three learned policies). The system demonstrates exceptional reliability, with a Planner validity of 1.0 across all configurations. Notably, the evaluation reveals that effective oversight is highly selective rather than continuous. Under the physics-constrained setting, the Auditor intervenes in 34.2%–37.2% of steps, whereas in the unconstrained setting, the intervention rate is 27.3%–29.7%. The Auditor revises proposed actions in over 95% of cases under constraints, and over 83% without constraints.

The empirical results suggest that agentic governance functions as an auditable risk-control interface rather than a universal performance optimizer. Its impact is highly dependent on the behavioral characteristics of the base RL policy. For instance(Figure7a), the LLM layer successfully suppresses action jitter for SAC and DQN by 0.0290 and 0.0179, respectively. Conversely, it marginally increases PPO jitter by 0.0223, primarily due to more frequent clipping and targeted revisions of PPO's proposed actions. Similar policy-dependent trade-offs emerge in artificial liquidity: oversight reduces it for PPO (−0.0074 MWh) but slightly increases it for SAC (+0.0073 MWh) and DQN (+0.0109 MWh). Spatial fairness remains largely stable, with minimal deltas across all policies(Figure 7b).

These mixed outcomes reveal that an evaluation should not assume that deploying an LLM governance layer automatically resolves all safety risks. Instead, it must rigorously define the boundaries of such oversight. Within the decentralized energy market setting, *SolarChain-Eval* demonstrates that while the AI agent successfully intercepts unstable decisions, as evidenced by the positive correlation between pre-audit action jitter and the corrective delta in Figure 7c, it cannot fully compensate for a misspecified RL reward. Ultimately, our findings advocate for a complementary approach to trustworthy autonomous systems. While agentic governance

Table 7: Representative LLM audit trace selected from `agentic_logs.jsonl`. The row records the market context, proposed RL action, audited final action, and action delta.

Field	Value
policy / setting	PPO / LLM No-physics
Episode and step	Episode 6, step 15, hour 15
Trigger context	raw violation 0.14; gap 1.97; static slippage 0.72
RL proposed action	r=0.049, l=0.931, b=0.200
Audited final action	r=0.050, l=0.200, b=0.000
Action delta	0.932
Auditor rationale	Violation rate exceeds audit threshold; tighten allocation to limit exposure to unsafe backed supply.



(a) Selected action revision. **(b) Context of audited PPO steps.**

Figure 8: Case-study audit trace for PPO under the LLM governance setting.

guarantees post-deployment transparency and event-driven risk mitigation, physically constrained RL optimization remains necessary to prevent reward-driven exploitation.

4.4 Case Study: Transparent LLM Audit Trace

To illustrate the transparency of our evaluation, we analyze a representative audit trace (Table 7). We select a scenario where the PPO agent operates without physical constraints, a setting where reward misspecification inherently encourages risky market behavior. At the evaluated step, environmental risk signals are elevated, including a raw physics violation rate of 0.14 and a supply gap of 1.97.

Driven by economic utility, the RL policy proposes an aggressive, high-liquidity action ($r = 0.049, l = 0.931, b = 0.200$). Recognizing the safety breach, the LLM Auditor intervenes and revises the action to a conservative bound ($r = 0.050, l = 0.200, b = 0.000$), resulting in a substantial action delta of 0.932 (Figure 8b). Beyond the numerical correction, the physical violation rate exceeded the safety threshold, necessitating a tighter liquidity allocation to limit exposure to unsafe supply.

This case validates the contribution of the agentic layer to trustworthy evaluation. Instead of imposing an opaque constraint, *SolarChain-Eval* records the entire decision-making process. It explicitly documents the activating risk signals, the resulting action adjustments, and the auditor's natural language rationale. By providing this transparency, *SolarChain-Eval* guarantees that autonomous interventions remain accountable after deployment.

5 Conclusion

In conclusion, we proposed *SolarChain-Eval*, a physics-constrained benchmark designed to evaluate the trustworthiness of economic agents in decentralized energy markets. Our empirical evaluations reveal a clear trade-off inherent in autonomous market governance. While RL maximize economic utility better than static and heuristic rules, they tend to compromise system safety for market efficiency. This vulnerability is starkly exposed when physical constraints are removed. Driven purely by reward maximization, it may aggressively exploit false data injection attacks to artificially inflate market liquidity. To counter this flaw, this study demonstrates that integrating physical constraints and AI agent layer provides a reliable and auditable safety interface. By evaluating context and intercepting high-risk decisions, The AI agent layer effectively suppresses action instability and prevents severe market failures, although its overall corrective impact remains bounded by the behavioral tendencies of the underlying RL controller.

The deployment of autonomous agents in cyber-physical systems requires evaluation paradigms that look beyond standard reward maximization, which often ignores physical ground truths. *SolarChain-Eval* addresses this gap by offering a transparent framework that logs the complete decision loop, including environmental risk signals, action corrections, and natural language explanations. Consequently, our benchmark ensures that agent behaviors are not only economically viable but also verifiable, accountable, and physically safe in real-world scenarios.

To enable reproducible benchmarking and empower future research, the *SolarChain-Eval* environment and its evaluation suite are designed for open reuse. Researchers studying decentralized market can utilize our framework to benchmark multi-agent reinforcement learning architectures, exploring the interactions among diverse energy prosumers. Scholars investigating agentic safety can utilize our physics-bounded environment and risk-tracking logs to simulate adversarial attacks and rigorously test algorithmic resilience. Future work could extend this benchmark beyond its current regional setting; by incorporating global meteorological datasets, researchers can validate these trustworthiness metrics across diverse geographical boundaries. Ultimately, by releasing these tools, we aim to lower the barrier for cross-disciplinary collaboration among AI safety researchers, energy economists, and systems engineers, contributing to the evaluation and verifiable trustworthiness of Agentic AI in real-world infrastructures.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained policy optimization. In *International conference on machine learning*. PMLR, 22–31.

[2] Mohammad Hossein Nejati Amiri, Fawaz Annaz, Mario De Oliveira, and Florimond Gueniat. 2025. Deep Reinforcement Learning with Local Interpretability for Transparent Microgrid Resilience Energy Management. *arXiv preprint arXiv:2508.08132* (2025).

[3] Anonymous. 2026. LLM-powered distributed optimal scheduling for industrial heat-electricity micro-grids. *IEEE Transactions on Industry Applications* 62, 2 (2026), 1874–1885.

[4] Gustavo Campos, Nael H El-Farra, and Ahmet Palazoglu. 2022. Soft actor-critic deep reinforcement learning with hybrid mixed-integer actions for demand responsive scheduling of energy systems. *Industrial & Engineering Chemistry Research* 61, 24 (2022), 8443–8461.

[5] Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. Grounding large language models in interactive environments with online reinforcement learning. In *International Conference on Machine Learning*. PMLR, 3676–3713.

[6] T. Chen, S. Bu, X. Liu, J. Kang, F. R. Yu, and Z. Han. 2022. Peer-to-peer energy trading and energy conversion in interconnected multi-energy microgrids using multi-agent deep reinforcement learning. *IEEE Transactions on Smart Grid* 13, 1 (2022), 715–727.

[7] Lin William Cong, Ye Li, and Neng Wang. 2021. Tokenomics: Dynamic adoption and valuation. *The Review of Financial Studies* 34, 3 (2021), 1105–1155.

[8] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. 2023. Guiding pretraining in reinforcement learning with large language models. In *International Conference on Machine Learning*. PMLR, 8657–8677.

[9] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Todd Hester, et al. 2021. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning* 110, 9 (2021), 2419–2468.

[10] Shangding Gu et al. 2024. A Review of Safe Reinforcement Learning: Methods, Theory and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[11] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR, 1861–1870.

[12] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[13] Antoine Kelly, Aidan O'Sullivan, Paul de Rigo, et al. 2020. Reinforcement learning for electricity network operation. In *Electric Power Systems Research*, Vol. 189. Elsevier, 106740.

[14] Mehmet Necip Kurt, Oyetunji Ogundijo, Chong Li, and Xiaodong Wang. 2018. Online Cyber-Attack Detection in Smart Grid: A Reinforcement Learning Approach. *IEEE Transactions on Smart Grid* 10, 5 (2018), 5174–5185.

[15] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward design with language models. In *The Eleventh International Conference on Learning Representations*.

[16] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From principles to practices. *Comput. Surveys* 55, 9 (2023), 1–46.

[17] Xiao Liu, Yujian Ye, S Li, C Zhang, Q Ma, and Jianguo Zhu. 2025. Carbon-Aware Peer-to-Peer Energy Trading in an Unbalanced Distribution Network via a Nash Equilibrium Discovery Deep Reinforcement Learning Approach. *IEEE Transactions on Smart Grid* 16, 4 (2025), 3392–3407.

[18] Yao Liu, Peng Ning, and Michael K Reiter. 2009. False data injection attacks against state estimation in electric power grids. In *Proceedings of the 16th ACM conference on Computer and communications security*. 21–32.

[19] Yao Liu, Peng Ning, and Michael K Reiter. 2011. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)* 14, 1 (2011), 1–33.

[20] Daniela Rojas Lozano and Yuanyuan Shi. 2025. Democratizing Microgrid Optimization: An LLM Agent for Dispatching Mobile Chargers to Construction Electric Vehicles. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[21] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Eureka: Human-level reward design via coding large language models. In *The Twelfth International Conference on Learning Representations*.

[22] Katya Malinova and Andreas Park. 2023. Tokenomics: When Tokens Beat Equity. *Management Science* 69, 11 (2023), 6568–6583.

[23] Esther Mengelkamp, Johannes Gärttner, Kerstin Rock, Scott Kessler, Lawrence Orsini, and Christof Weinhardt. 2018. Designing microgrid energy markets: A case study: The Brooklyn Microgrid. *Applied Energy* 210 (2018), 870–880.

[24] Hanxiao Qu, Krzysztof M Gogol, Florian Grötschla, and Claudio J Tessone. 2025. From Rules to Rewards: Reinforcement Learning for Interest Rate Adjustment in DeFi Lending. In *The International Conference on Mathematical Research for Blockchain Economy*. Springer, 85–120.

[25] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8.

[26] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[28] Wayes Tushar, Chau Yuen, Hamed Mohsenian-Rad, Tapan Saha, H Vincent Poor, and Kristin L Wood. 2018. Transforming energy networks via peer-to-peer energy trading: The potential of game-theoretic approaches. *IEEE Signal Processing Magazine* 35, 4 (2018), 90–111.

[29] Jiahua Xu, Yebo Feng, Daniel Perez, and Benjamin Livshits. 2025. Auto. gov: learning-based governance for decentralized finance (DeFi). *IEEE Transactions on Services Computing* (2025).

[30] Jinyu Yan, Haibo He, Xiang Zhong, and Yu Tang. 2018. Evaluation of reinforcement learning-based false data injection attack to automatic voltage control. *IEEE Transactions on Smart Grid* 10, 2 (2018), 2158–2169.

[31] Xu Yang, Chenhui Lin, Haotian Liu, and Wenchuan Wu. 2025. RL2: Reinforce Large Language Model to Assist Safe Reinforcement Learning for Energy Management of Active Distribution Networks. *IEEE Transactions on Smart Grid* (2025).

[32] Changwei Yao, Xinzi Liu, Chen Li, and Marios Savvides. 2025. Reward Evolution with Graph-of-Thoughts: A Bi-Level Language Model Framework for Reinforcement Learning. *arXiv preprint arXiv:2509.16136* (2025).

[33] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022).

[34] Dongxia Zhang, Xiaoqing Han, and Chunyu Deng. 2018. Review on the research and practice of deep learning and reinforcement learning in smart grids. *CSEE Journal of Power and Energy Systems* 4, 3 (2018), 362–370.

[35] Lingyun Zhang, Haojie Li, Zhang Chuanhao, Yuan Zhao, Qiao Shixiang, and Hang Yu. 2026. Deep Reinforcement Learning-Based Dual-Loop Adaptive Control Method and Simulation for Loitering Munition Fuze. *Technologies* 14, 4 (2026), 239.

[36] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher. 2021. The AI Economist: Optimal Economic Policy Design via Two-level Deep Reinforcement Learning. *arXiv preprint arXiv:2108.02755* (2021).

[37] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher. 2022. The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science advances* 8, 18 (2022), eabk2607.

A Formulas Implementation

This appendix lists some important equations used by the released *SolarChain-Eval* code but not separately expanded in the methodology section.

PV capacity and trust labels. The data generator converts weather into a physical PV upper bound and the loader derives the trust labels used by the benchmark:

$$\begin{aligned} I_{c,t} &= \mathbb{I}[z_{c,t} < 90^\circ]\ \min\{\max(S_{c,t}, 0),\ 1.08\max(C_{c,t}, 0)\}, \\ \eta_{i,t}^{temp} &= \mathrm{clip}\{1 + \gamma_i(T_{c,t} - 25),\ 0.78,\ 1.08\}, \\ P_{i,t}^{max} &= \max\{0, I_{c,t}A_i\eta_i\eta_{i,t}^{temp}\}, \\ G_{i,t}^{v} &= \mathbb{I}[\mathrm{status}_{i,t} = \mathrm{verified}]P_{i,t}^{reported}, \\ X_{i,t}^{excess} &= \max(P_{i,t}^{reported} - P_{i,t}^{max}, 0), \\ \mathrm{viol}_{i,t} &= \mathbb{I}[X_{i,t}^{excess} > 10^{-6} \vee \mathrm{FDIA}_{i,t}], \\ R_{i,t}^{rejected} &= \mathrm{viol}_{i,t}P_{i,t}^{reported}. \end{aligned} \tag{14}$$

Action decoding and allocation budget. Continuous policy outputs are decoded into governance ratios and then sanitized against the global allocation budget:

$$\begin{aligned} \alpha_t &= \alpha_{\min} + u_{t,1}(\alpha_{\max} - \alpha_{\min}), \\ \ell_t &= \ell_{\min} + u_{t,2}(\ell_{\max} - \ell_{\min}), \\ b_t &= u_{t,3}b_{\max}, \\ s_t &= \min\left\{1, \frac{0.98}{\alpha_t + \ell_t}\right\}, \\ (\alpha_t, \ell_t) &\leftarrow (s_t\alpha_t, s_t\ell_t). \end{aligned} \tag{15}$$

DQN uses the same decoder after selecting one point from a $5^3 = 125$ normalized action grid.

Unsafe supply, artificial liquidity, and market clearing. The implementation records how much invalid supply is economically backed and how much liquidity it creates:

$$\begin{aligned} X_t &= \max\left\{\sum_i R_{i,t}^{rejected}, \sum_i X_{i,t}^{excess}, \max\left(\sum_i P_{i,t}^{reported} - \sum_i P_{i,t}^{max}, 0\right)\right\}, \\ \beta_t &= \frac{\min(\alpha_t + \ell_t, 0.98)}{0.98}, \\ A_t^{unsafe} &= \beta_t X_t, \\ G_t^{backed} &= G_t^{v} + A_t^{unsafe}, \\ A_t^{artificial} &= \ell_t A_t^{unsafe}. \end{aligned} \tag{16}$$

The executed market transition is then

$$\begin{aligned} \widetilde{Q}_t &= Q_t^d \max(0.75, 1 - 0.80b_t), \\ L_t^{avail} &= L_t + \ell_t G_t^{backed}, \\ M_t &= \min(L_t^{avail}, \widetilde{Q}_t), \\ L_{t+1} &= \max(L_t^{avail} - M_t, 0), \\ \sigma_t^{exec} &= \frac{\widetilde{Q}_t}{\max(L_t^{avail} + 0.05, 0.05)}. \end{aligned} \tag{17}$$

Episode, fairness, and audit metrics. The evaluation code aggregates step records into rollout-level metrics:

$$\begin{aligned} R_e &= \sum_{t\in e} R_t, \\ V_e &= \frac{1}{24}\sum_{t\in e} V_t, \\ J_e &= \frac{1}{23}\sum_{t=2}^{24} \|a_t - a_{t-1}\|_1, \\ \bar{\sigma}_e &= \frac{1}{24}\sum_{t\in e} \sigma_t^{exec}, \\ A_e^{artificial} &= \sum_{t\in e} A_t^{artificial}. \end{aligned} \tag{18}$$

City-level reward imbalance and agentic intervention strength are reported as

$$\begin{aligned} F_e &= \frac{\mathrm{Var}_c\left(\sum_{t\in e} r_{c,t}^{city}\right)}{\max\left(\frac{1}{|C|}\sum_c \left|\sum_{t\in e} r_{c,t}^{city}\right|, 10^{-9}\right)}, \\ \delta_t &= \|a_t^{final} - a_t^{original}\|_1, \\ \mathrm{audit_rate} &= \frac{N^{audit}}{24}, \\ \mathrm{revision_rate} &= \frac{N^{revise}}{\max(N^{audit}, 1)}. \end{aligned} \tag{19}$$

B Data Provenance and Evaluation Matrix

Table 8 summarizes the benchmark input data used in the paper. The primary dataset is `data/datasets_2026_04_month`, covering five Chinese cities from `2026-04-01` to `2026-05-01` exclusive, with 720 hourly timestamps.

The evaluation artifacts used to produce the paper are derived from these raw benchmark inputs. In total, the stored evaluation evidence contains 1,620 episode-level metric rows, 38,880 hourly action rows, 194,400 city-hour policy rows, and 12,960 LLM governance log rows.

Table 8: Raw benchmark data files and collection or generation methods. Row counts exclude CSV headers.

File	Rows	Key fields	Collection or generation method
`open_meteo_weather.json`	5 city payloads	hourly time, temperature, shortwave radiation, latitude, longitude, timezone	Retrieved by `scripts/generate_monthly_datasets.py` from the Open-Meteo Historical Weather API for Beijing, Shanghai, Chengdu, Shenzhen, and Hangzhou; cached locally and reused on later runs.
`urban_energy_nodes.csv`	50	node id, city, latitude, longitude, panel area, efficiency, temperature coefficient, install date	Generated with seed 20260511: 10 nodes per city, coordinates jittered around city centers, panel areas sampled from 18–64 m^2, efficiencies from 0.176–0.226, temperature coefficients from -0.0046 to -0.0032, and install dates from 2020-01-01 to 2024-05-31.
`spatiotemporal_generation.csv`	36,000	timestamp, node id, city, irradiance, air temperature, P^{max}, reported power, FDIA flag, verification status	Generated hourly for 50 nodes over 720 hours from cached weather, pvlib clear-sky and solar-position calculations, node-specific PV parameters, inverter derate, and noise. Exactly 1,800 rows are FDIA-injected and marked rejected.
`market_liquidity.csv`	720	timestamp, total verified MW, SolarChain liquidity, baseline liquidity, SolarChain slippage, baseline slippage	Aggregated from verified generation. The implementation uses $L^{SC} = 0.92G^{v} + 0.018$, $L^{base} = 0.61G^{v} + 0.008$, $\sigma^{SC} = 0.18/(L^{SC} + 0.045)$, and $\sigma^{base} = 0.31/(L^{base} + 0.028)$.
`p2p_trades.csv`	1,185	trade id, timestamp, factory id, city, energy purchased, tokens burned, exergy dissipated	Generated for daylight hours with positive verified supply. Each active hour can create up to three factory purchases; purchase size is the minimum of a SolarChain-liquidity share and a verified-generation share, with token burn and exergy fields sampled from configured multipliers.

Table 9: Detailed main-setting episode summaries. Reward, volume, violation, artificial liquidity, slippage, and action jitter are aggregated over 90 rollouts per policy.

Policy	Rollouts	Reward Mean	Reward Min	Reward Max	Volume	Violation Mean	Violation Range	Artificial Liquidity
DQN	90	-23.23	-24.97	-21.00	0.491	0.4688	0.4270–0.5048	0.1298
Myopic	90	-23.08	-24.60	-21.39	0.456	0.4779	0.4434–0.5136	0.1009
PPO	90	-22.35	-24.20	-20.14	0.470	0.4520	0.4060–0.4912	0.1453
Random	90	-25.46	-27.34	-23.28	0.480	0.4771	0.4395–0.5147	0.1137
SAC	90	-22.26	-24.28	-20.15	0.521	0.4461	0.4083–0.4879	0.0686
Static	90	-23.38	-24.86	-21.65	0.514	0.4874	0.4527–0.5218	0.1803

Table 10: Four-setting summaries for learned policies. Each row aggregates 90 rollouts across three seeds. These values support the paired physics-removal deltas and show how the LLM governance layer changes the setting-level risk profile.

Setting	Policy	Rollouts	Reward	Violation	Artificial Liquidity	Slippage	Jitter	Fairness
RL	PPO	90	-22.35	0.4520	0.1453	0.0134	0.0607	0.0041
RL	SAC	90	-22.26	0.4461	0.0686	0.0178	0.1435	0.0048
RL	DQN	90	-23.23	0.4688	0.1298	0.0148	0.1118	0.0159
RL+LLM	PPO	90	-22.44	0.4571	0.1379	0.0141	0.0829	0.0045
RL+LLM	SAC	90	-22.32	0.4527	0.0759	0.0175	0.1145	0.0051
RL+LLM	DQN	90	-23.51	0.4780	0.1406	0.0146	0.0939	0.0157
RL no-physics	PPO	90	0.27	0.4831	0.2238	0.0122	0.0135	0.0053
RL no-physics	SAC	90	0.26	0.4874	0.1925	0.0137	0.0432	0.0109
RL no-physics	DQN	90	0.32	0.4869	0.1874	0.0129	0.0259	0.0119
RL+LLM no-physics	PPO	90	0.22	0.4796	0.1977	0.0130	0.0366	0.0055
RL+LLM no-physics	SAC	90	0.30	0.4870	0.1908	0.0137	0.0362	0.0109
RL+LLM no-physics	DQN	90	0.27	0.4862	0.1820	0.0132	0.0342	0.0119

Table 11: Aggregate learned-policy means by setting. Each row averages PPO, SAC, and DQN over the same 270 rollouts. The no-physics settings remain substantially higher in artificial liquidity even when LLM governance is active.

Setting	Reward	Violation	Artificial Liquidity	Slippage	Jitter	Fairness
RL	-22.61	0.4557	0.1146	0.0153	0.1053	0.0083
RL+LLM	-22.76	0.4626	0.1181	0.0154	0.0971	0.0084
RL no-physics	0.28	0.4858	0.2012	0.0129	0.0275	0.0094
RL+LLM no-physics	0.26	0.4843	0.1901	0.0133	0.0357	0.0094

Table 12: Detailed LLM governance operation rates. Each row aggregates 2,160 agentic log steps from three seeds. The LLM failure count is zero for every row and is therefore not shown.

Setting	Policy	Steps	Audit Rate	Revision Given Audit	Modification Rate	Mean Action Δ	Not Triggered
Main	DQN	2160	0.372	0.958	0.518	0.084	0.488
Main	PPO	2160	0.344	0.961	0.987	0.164	0.550
Main	SAC	2160	0.342	0.966	0.524	0.059	0.509
No-physics	DQN	2160	0.297	0.892	0.490	0.042	0.683
No-physics	PPO	2160	0.273	0.832	0.991	0.111	0.691
No-physics	SAC	2160	0.274	0.876	0.294	0.021	0.698

Table 13: Detailed LLM post-hoc trigger diagnostics and mean Planner thresholds. Trigger columns report the fraction of logged steps satisfying each trigger condition.

Setting	Policy	Physics Trig.	Gap Trig.	Slippage Trig.	Jitter Trig.	Viol. Thr.	Gap Thr.	Jitter Thr.	Slip Thr.
Main	DQN	0.120	0.440	0.168	0.230	0.141	-0.108	0.177	2.960
Main	PPO	0.121	0.440	0.171	0.171	0.142	-0.112	0.174	2.961
Main	SAC	0.120	0.441	0.137	0.211	0.142	-0.102	0.175	3.061
No-physics	DQN	0.116	0.441	0.170	0.049	0.143	-0.099	0.170	3.028
No-physics	PPO	0.113	0.441	0.126	0.078	0.144	-0.103	0.185	3.140
No-physics	SAC	0.105	0.441	0.145	0.048	0.146	-0.102	0.181	3.111

Table 14: Paired episode-level effects of adding the LLM Planner/Auditor layer. Deltas are computed as RL+LLM minus raw RL under matched seed, policy, and episode.

Comparison	Policy	Δ Reward	Δ Action Jitter	Δ Spatial Fairness	Δ Artificial Liquidity	Δ Slippage
Main	DQN	-0.277 ± 0.503	-0.0179 ± 0.0274	-0.0002	0.0109	-0.0001
Main	PPO	-0.094 ± 0.293	0.0223 ± 0.0154	0.0004	-0.0074	0.0007
Main	SAC	-0.062 ± 0.268	-0.0290 ± 0.0198	0.0003	0.0073	-0.0003
No-physics	DQN	-0.045 ± 0.065	0.0083 ± 0.0110	0.0000	-0.0054	0.0002
No-physics	PPO	-0.043 ± 0.097	0.0231 ± 0.0166	0.0002	-0.0261	0.0008
No-physics	SAC	0.038 ± 0.089	-0.0070 ± 0.0154	0.0000	-0.0017	0.0000

Table 15: Step-level action distributions for learned policies. Each cell reports mean with 5th–95th percentile range in parentheses.

Setting	Policy	Reward Ratio	Liquidity Ratio	Burn Rate
Main	DQN	0.320 (0.050–0.500)	0.517 (0.200–0.762)	0.081 (0.000–0.150)
Main	PPO	0.050 (0.049–0.050)	0.401 (0.200–0.931)	0.104 (0.000–0.200)
Main	SAC	0.169 (0.058–0.432)	0.433 (0.212–0.738)	0.008 (0.001–0.028)
No-physics	DQN	0.216 (0.049–0.308)	0.759 (0.672–0.931)	0.080 (0.000–0.200)
No-physics	PPO	0.049 (0.049–0.050)	0.850 (0.200–0.931)	0.045 (0.000–0.200)
No-physics	SAC	0.215 (0.079–0.363)	0.764 (0.617–0.898)	0.011 (0.002–0.025)
LLM Main	DQN	0.321 (0.050–0.500)	0.575 (0.250–0.762)	0.070 (0.000–0.150)
LLM Main	PPO	0.065 (0.050–0.124)	0.480 (0.250–0.900)	0.071 (0.000–0.150)
LLM Main	SAC	0.166 (0.060–0.402)	0.470 (0.250–0.747)	0.009 (0.001–0.027)
LLM No-physics	DQN	0.217 (0.050–0.350)	0.746 (0.500–0.900)	0.069 (0.000–0.150)
LLM No-physics	PPO	0.062 (0.050–0.100)	0.799 (0.250–0.900)	0.028 (0.000–0.150)
LLM No-physics	SAC	0.212 (0.083–0.343)	0.764 (0.634–0.888)	0.011 (0.002–0.025)