LCIA Self-Play: The London Court of International Arbitration as a Benchmark for Verifiable Self-Evolving Scientific Agents

David Scott Lewis [1] **Haley Yi** [1]

Abstract

Self-evolving scientific agents are often evaluated where the training record is observable and correctness can be checked directly, but many scientific and institutional settings expose only selected public traces, contain strategic actors, and impose formal admissibility constraints. We introduce *LCIA Self-Play*, a synthetic benchmark inspired by the London Court of International Arbitration that models confidential arbitration as a partially observable hidden-type game with a deterministic legality gate over procedural actions. Agents improve through self-play, belief updating, episodic memory, and capability-preserving rehearsal while being evaluated on calibration, public-slice overfitting, hidden-type robustness, exploitability, violation probability, and retention. Across twelve pre-specified seeds, public-slice supervision preserves ranking signal but is badly miscalibrated on the hidden population (Brier 0.238 ± 0.001 to 0.254 ± 0.002; ECE 0.017 ± 0.002 to 0.120 ± 0.006), while verifier-gated self-play raises worst-hidden-type value over public-greedy play (0.626 ± 0.007 versus 0.510 ± 0.017) with zero invalid-action mass. Memory replay improves retained old-family value over recent-only adaptation (0.685 ± 0.008 versus 0.656 ± 0.009), making the contribution a reproducible stress test for constrained self-evolution rather than a legal prediction system.

1. Introduction

A scientific agent that can revise its own workflow, memory, and policy is useful only if its improvement is stable, inspectable, and constrained by the rules of the domain. Recent work on self-evolving agents studies capability erosion, runtime memory, test-time self-improvement, agentic tree search, and reflective learning (Yu et al., 2026; Zhang et al., 2026; Acikgoz et al., 2025; Yamada et al., 2025; Allard et al., 2026). The common assumption behind many evaluations is that the agent can observe enough of the target process to learn from outcomes. That assumption fails in settings where important evidence is private, public data are selected non-randomly, and actors strategically shape the observations that become visible.

International arbitration supplies a compact version of this difficulty. The London Court of International Arbitration is a mature arbitral institution with distinctive procedural features: institutional appointment and challenge mechanisms, express confidentiality norms, joinder and consolidation tools, emergency procedures, early determination, and an hourly-cost model rather than a purely ad valorem fee schedule (Gerbay, 2012; Collett et al., 2019; Veeder, 2016; Zaman & Sabharwal, 2014; Richman et al., 2021). These features are attractive computationally because they turn legal procedure into a structured decision process. They are also difficult computationally because confidentiality makes the public record a biased sample, not a supervised-learning substrate (Collins, 1995; Platte, 2002; Wang, 2023). A model trained only on visible awards, court challenges, or reported disputes learns a distribution conditioned on becoming public, not the hidden decision process itself.

We therefore treat LCIA-style arbitration as a benchmark for self-evolving agents rather than as a target for automated legal prediction. The benchmark abstracts a confidential proceeding as an imperfect-information game in which the tribunal and counterparty have hidden types, the agent observes only public or party-local signals, and every proposed procedural action must pass a formal admissibility gate. This gate rejects actions that violate simplified procedural preconditions, such as attempting consolidation where compatible arbitration agreements are absent or seeking emergency relief without urgency. The resulting environment tests whether an agent can improve while maintaining three properties that matter for scientific agents generally: robustness to missing-not-at-random evidence, strategic adaptation under partial observability, and capability preservation under long-horizon self-evolution.

[1]AIXC Research, Zaragoza, Spain. Correspondence to: David Scott Lewis <reports@aixcresearch.com>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

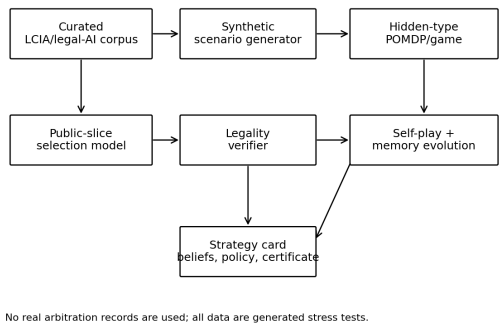


Figure 1. LCIA Self-Play architecture. A synthetic case generator emits hidden-type arbitration games; a self-play module computes policies; a verifier rejects inadmissible moves; memory and capability-preservation modules stabilize long-horizon adaptation.

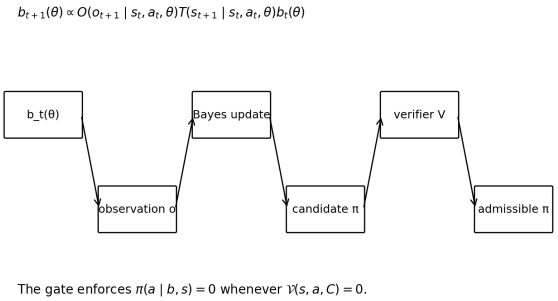


Figure 2. Belief-state update and legality gate. The policy sees a belief state, not the hidden type. The verifier gates actions before they can enter the self-evolution loop.

The paper makes four contributions. First, it gives a mathematical formulation of an LCIA-inspired POMDP/Bayesian-game benchmark. Second, it defines a verifier-gated policy interface that returns a "strategy card" rather than an ungrounded answer: beliefs, admissible actions, mixed policy, exploitability, and failure certificates. Third, it provides three reproducible synthetic experiments that isolate public-slice overfitting, self-play robustness, and capability-preserving evolution. Fourth, it releases code, pre-specified seeds, generated data, figures, and reproducibility metadata in the artifact bundle.

2. LCIA as an Imperfect-Information Benchmark

The legal features of LCIA arbitration map naturally onto machine-learning stressors. Confidentiality induces biased observation. Multiparty disputes create coupled action spaces. Joinder, consolidation, emergency relief, security for costs, early determination, and confidentiality orders create admissibility constraints. Party autonomy and tribunal discretion introduce hidden types: a tribunal may be efficiency-oriented, due-process-oriented, confidentiality-sensitive, cost-sensitive, or interventionist. The English-law and London-seat literature further emphasizes the role of supervisory courts, challenge standards, and enforcement conditions (Merkin & Flannery, 2019; Blackaby et al., 2015; Yalabk, 2021; Hill & Chong, 2010; Grierson, 2022).

We model the environment as

$$\mathcal{E}_{\text{LCIA}} = \langle S, A, T, R, \Omega, O, \Theta, C, \mathcal{V} \rangle, \tag{1}$$

where S is procedural state, A is the action set, T is state transition, R is utility, Ω is the observation alphabet, O is the observation kernel, Θ is the hidden-type space, C is a finite set of formal constraints, and $\mathcal{V} : S \times A \times C \to \{0, 1\}$ is an admissibility verifier. The agent maintains a belief over hidden types:

$$b_{t+1}(\theta) \propto O(o_{t+1} \mid s_t, a_t, \theta) T(s_{t+1} \mid s_t, a_t, \theta) b_t(\theta). \tag{2}$$

A policy is admissible when $\pi(a \mid b, s) = 0$ for every action with $\mathcal{V}(s, a, C) = 0$.

The benchmark is deliberately synthetic. It does not encode confidential awards, private party materials, or real parties. Its value is structural: it turns a real institutional bottleneck into a reproducible environment for studying learning under secrecy. This mirrors the broader evaluation concern in legal AI: benchmarks such as LegalBench and LEXTREME measure legal reasoning tasks, while hallucination evaluations show that retrieval-enhanced legal tools still fail in high-stakes settings (Guha et al., 2023; Niklaus et al., 2023; Magesh et al., 2025). LCIA Self-Play targets a different failure mode: public data can be accurate and still be non-representative.

3. Agents and Verifier-Gated Self-Evolution

The action set contains nine procedural strategies: settlement, early determination, consolidation, targeted disclosure, full-merits hearing, security for costs, confidentiality order, emergency application, and issue narrowing. The hidden-type set contains five stylized types: efficiency, due process, confidentiality, cost sensitivity, and interventionism. The legality gate is a deterministic checker over a simplified rule schema. For example, consolidation requires a multiparty or related-transaction state and a compatible-agreement flag; emergency action requires urgency; security for costs requires a funding-risk flag. The checker is intentionally small so that invalidity is inspectable rather than learned implicitly.

We compare agent families that separate observability, planning, robustness, and admissibility. Public-slice learners train on observed cases only. Greedy, posterior,

Table 1. Benchmark components. The design abstracts LCIA-style procedure into hidden types, observable state variables, legal-action preconditions, and evaluation metrics.

Component	Examples	Purpose in the benchmark
Hidden type θ	efficiency, due process, confidentiality, cost sensitivity, interventionism	Makes the decision process partially observed and prevents public-outcome training from being sufficient.
State features s	quantum, urgency, multiparty flag, agreement compatibility, funding risk, confidentiality sensitivity, factual complexity	Encodes the procedural posture available to the agent without revealing the tribunal or counterparty type.
Actions A	settlement, issue narrowing, early determination, disclosure, consolidation, emergency relief, confidentiality order, merits hearing	Provides a finite strategic action set with a mixture of ordinary, aggressive, and constraint-sensitive moves.
Verifier $\mathcal{V}$	urgency required for emergency relief; compatible agreements for consolidation; funding-risk flag for security for costs	Converts legal procedure into an executable admissibility gate.
Metrics	calibration, public-overfit gap, exploitability, worst-type value, violation probability, retained capability	Measures whether self-evolution improves robust reasoning rather than merely public-slice performance.

Algorithm 1 Verifier-gated self-play update

Initialize belief $b_0(\theta)$, regret tables r_A, r_B, memory M
for episode $e = 1, \ldots, E$ **do**
  Observe public/local signal o_t and update b_t
  $A_{\text{adm}} \leftarrow \{a \in A : \mathcal{V}(s_t, a, C) = 1\}$
  Compute mixed strategy π_t by regret matching on A_{adm}
  Execute or simulate $a_t \sim \pi_t$, observe reward and signal
  Update regrets with counterfactual values under sampled hidden type
  Store $(s_t, b_t, a_t, o_{t+1}, r_t, \mathcal{V})$ in episodic memory
  Apply capability-preserving rehearsal on prior scenario families
end for
Return strategy card $(b, A_{\text{adm}}, \pi, \epsilon_{\text{exploit}}, M)$

particle-belief, Dirichlet-robust, maximin, oracle, and regret-matching planners test alternative responses to POMDP uncertainty. Verifier-gated self-evolving agents combine self-play with memory and capability-preserving rehearsal, motivated by recent evidence that self-evolving agents can forget old capabilities unless preservation is explicit (Yu et al., 2026; Zhang et al., 2026; Allard et al., 2026). The design is also related to counterfactual regret minimization (Zinkevich et al., 2007; Brown & Sandholm, 2019), POMDP planning (Kaelbling et al., 1998), reinforcement learning (Sutton & Barto, 2018), and proof-carrying or argumentation-style verification interfaces (Dung, 1995).

4. Experiments

All experiments use generated data only and are repeated over twelve pre-specified seeds beginning at 12345. The point is not to estimate real LCIA outcomes. The point is to test failure modes that are hard to see in outcome-only benchmarks: public-slice calibration, hidden-type robustness, legality preservation, and retention under self-evolution. Appendix G gives the exact file layout and command.

4.1. Experiment 1: Public-Slice Overfitting

We generate 5,000 training matters and 3,000 test matters for each seed. Observable variables include quantum, multiparty structure, urgency, confidentiality, factual complexity, compatibility of agreements, funding risk, and novel issue status. Hidden types influence the synthetic outcome. A matter becomes public according to a missing-not-at-random selection model depending on quantum, novelty, public salience, outcome extremity, outcome, and hidden type. In the representative generated training set, 46.4% of matters are public. The public slice over-represents interventionist and due-process types and under-represents confidentiality-sensitive types.

Table 2 shows that AUC alone would miss the problem: the public-only model remains a plausible ranker but is substantially miscalibrated on the full hidden population. Registry-propensity weighting improves neither calibration nor full-population Brier score because the selection process also depends on unobserved type and outcome. An oracle-propensity diagnostic nearly recovers the full-oracle Brier

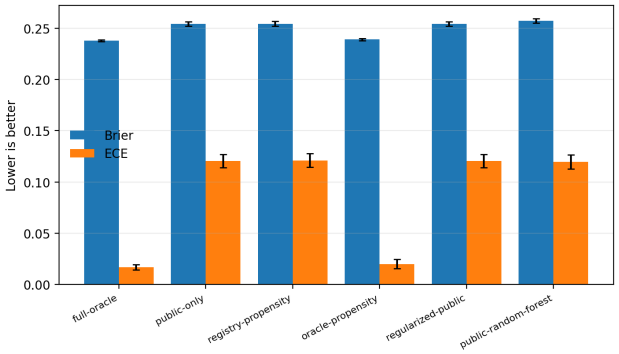


Figure 3. Experiment 1. Public-slice training preserves some rank discrimination but damages calibration and full-population performance. Bars show means with 95% confidence intervals over twelve seeds.

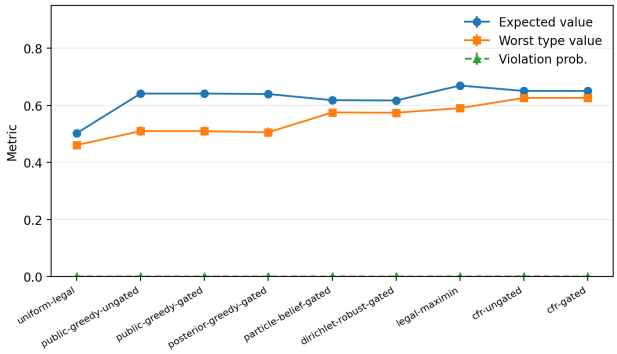


Figure 4. Experiment 2. Gated self-play improves worst-hidden-type robustness relative to public-greedy play while preserving admissibility. Error bars are 95% confidence intervals over seeds and scenario families.

Table 2. Experiment 1 full-population metrics, mean ± 95% CI over twelve seeds.

Model	AUC	Brier	ECE
Full-oracle	0.615±0.004	0.238±0.001	0.017±0.002
Oracle-propensity	0.609±0.004	0.239±0.001	0.020±0.005
Public-only	0.606±0.005	0.254±0.002	0.120±0.006
Reg.-public	0.606±0.005	0.254±0.002	0.120±0.006
Registry-IPW	0.606±0.006	0.254±0.002	0.121±0.007
Public RF	0.588±0.007	0.257±0.002	0.119±0.007

score, confirming that the failure is structural rather than an implementation artifact. This is the benchmark's first lesson: a scientific agent that learns only from public traces can look competent under ranking metrics while losing calibrated belief. A selection-strength sweep in Appendix N confirms the mechanism: public-only ECE is low when visibility depends weakly on hidden terms, but rises above 0.12 once outcome and hidden-type selection terms are active.

4.2. Experiment 2: Hidden-Type Self-Play

We next instantiate a finite hidden-type game for five scenario families. Each family has a type-specific payoff matrix and a legality mask over actions. The public-greedy policy optimizes a biased public hidden-type prior. Additional baselines include uniform legal play, ungated greedy play, posterior-greedy planning from weak procedural signals, particle-belief and Dirichlet-robust planners (Silver & Veness, 2010; Iyengar, 2005; Nilim & El Ghaoui, 2005; Rockafellar & Uryasev, 2000), legal maximin planning, an oracle true-prior diagnostic, ungated CFR, and verifier-gated CFR. We report expected value, worst-hidden-type value, exploitability proxy, entropy, and violation probability.

The result is a tradeoff rather than a single winner. Legal maximin obtains the highest average value in this synthetic payoff family, but its worst-hidden-type value is lower than CFR. Public-greedy play obtains high public-prior value, but its worst-hidden-type value is 0.510 ± 0.017. Gated self-play raises worst-hidden-type value to 0.626 ± 0.007 and keeps invalid-action probability at zero. The particle-belief and Dirichlet-robust baselines improve over uniform legal play but remain below gated CFR on worst-type value (0.575 ± 0.006 and 0.574 ± 0.006). The oracle true-prior greedy policy has the highest true expected value (0.689 ± 0.008) but lower worst-type value (0.565 ± 0.011), illustrating that correct-average optimization is not the same as robust hidden-type play. This is the intended stress: a self-evolving agent must decide whether to optimize public success, average payoff, adversarial robustness, or procedural safety, and the benchmark exposes those choices rather than collapsing them into one scalar.

4.3. Experiment 3: Capability-Preserving Evolution

Finally, we run a five-family curriculum over scenario families. Recent-only adaptation updates on the current family. Gated adaptation masks invalid actions. We then separate memory and capability preservation with explicit ablations: memory-only replay, CPE-only drift control, memory+CPE, and uncertainty-aware memory+CPE. The uncertainty-aware variant prioritizes sparse or high-variance family memories under partial observability. This experiment mirrors the stability-plasticity problem in self-evolving agents: new skills should not erase old ones (Yu et al., 2026; Zhang et al., 2026).

At the end of the curriculum, recent-only adaptation retains old-family value 0.656 ± 0.009 and has a small but nonzero invalid-action probability. The ablation table shows that memory replay supplies the largest retention gain in this configuration: memory-only reaches 0.685 ± 0.008, memory+CPE reaches 0.673 ± 0.004, and CPE-only reaches 0.646 ± 0.007. The uncertainty-aware memory variant is competitive but not uniformly dominant, which is useful: the benchmark can expose when prioritizing uncertainty helps, when it merely reallocates rehearsal budget, and

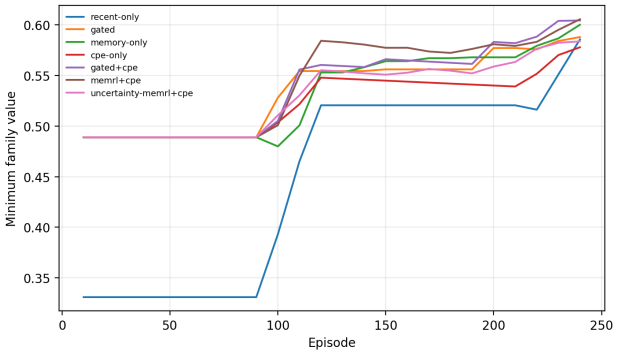


Figure 5. Experiment 3. Memory and capability-preservation ablations expose retention tradeoffs while the legality gate keeps invalid-action probability at zero.

when ordinary memory is already sufficient. This is the second intended stress: a self-evolving agent should not be rewarded only for its latest adaptation if it has silently lost prior procedural competence.

5. Benchmark Tasks and Metrics

The benchmark is organized around three tasks. The first task is *public-slice inference*. The agent receives a visible subset of synthetic cases and must estimate outcome probabilities on the full hidden population. This task intentionally creates a trap: ordinary discrimination metrics can remain stable while calibration collapses on the hidden population. We report AUC for ranking quality, Brier score for proper probabilistic loss, expected calibration error for bin-wise calibration, and a public-overfit gap measuring the difference between full-population and public-slice performance. The public-overfit gap is important because a model can appear reliable when evaluated only on public observations while failing on the population it is implicitly supposed to reason about.

The second task is *strategic self-play*. Each scenario family defines a payoff matrix indexed by procedural actions and hidden types. The agent does not know the hidden type at decision time; it only has a belief. Public-greedy play chooses the action with the highest expected value under a biased public prior. Gated self-play computes a mixed strategy over admissible actions using regret matching. We report true expected value, worst-type value, exploitability proxy, entropy, and violation probability. For family f with payoff matrix M_f, true prior p_f, legal mask L_f, and row policy π, the reported single-agent exploitability proxy is

$$\epsilon_f(\pi) = \max_{a:L_f(a)=1} e_a^\top M_f p_f - \pi^\top M_f p_f, \tag{3}$$

and violation probability is

$$v_f(\pi) = \sum_{a:L_f(a)=0} \pi(a). \tag{4}$$

Violation is measured before gate repair for ungated baselines and after masking for gated agents. These metrics deliberately pull in different directions. A deterministic public-greedy action can have high public-prior value and low entropy, while a mixed self-play policy can have lower public-prior value but better worst-type robustness and lower exploitability. The reported exploitability proxy is the one-step best-response gap $\max_{a\in A_{\mathrm{adm}}} e_a^\top M p_{\mathrm{true}} - \pi^\top M p_{\mathrm{true}}$ under the full hidden-type prior. Violation probability is measured before execution as policy mass on actions rejected by the verifier, $\sum_{a:\mathcal{V}(s,a,C)=0} \pi(a \mid s)$.

The third task is *self-evolution under retention*. The agent is exposed to a curriculum of scenario families. It updates its action preferences from observed rewards and memory. A recent-only agent can become specialized to the most recent family. A capability-preserving agent rehearses previous families and penalizes destructive parameter drift. We report current-family value, mean value across families, worst-family value, old-family retention, and invalid-action probability. The retention metric is central because self-evolving systems can improve on the current distribution while silently degrading earlier competencies. This is the same stability-plasticity tension that appears in lifelong agent adaptation, except here it is coupled to a hard admissibility constraint.

6. Strategy Cards and Failure Ledgers

The output of an LCIA Self-Play agent is not a predicted award, a legal conclusion, or a textual recommendation. The output is a *strategy card*. A strategy card contains: the current belief over hidden types; the admissible action set; the rejected actions with typed verifier failures; the mixed policy; expected value under the public prior and under the synthetic full prior; worst-type value; exploitability; relevant memory items; and a short failure ledger describing where the policy is fragile. This artifact is intentionally more verbose than a single answer because hidden-type reasoning cannot be audited from an answer alone.

The failure ledger is useful for both evaluation and debugging. In Experiment 1, the ledger records whether an observed performance improvement is public-slice-specific. In Experiment 2, it records whether a policy gains value by exploiting a biased prior or by improving robust play. In Experiment 3, it records whether a curriculum update improved current-family performance at the cost of retained capability. The ledger also records whether the verifier rejected a move because of a missing precondition rather than because the model merely assigned low value to that move. This separation matters: low value is a strategic judgment; inadmissibility is a formal constraint.

This strategy-card interface also supports human-AI collab-

Table 3. Primary metrics and what they diagnose.

Metric	Task	Diagnostic role
AUC	Public-slice inference	Measures ranking quality; useful but insufficient under selection bias.
Brier score	Public-slice inference	Penalizes miscalibrated probabilities on the full hidden population.
Expected calibration error	Public-slice inference	Detects confidence misalignment that AUC can hide.
Public-overfit gap	Public-slice inference	Difference between public-slice and full-population reliability.
Worst-type value	Self-play	Robust value against the least favorable hidden type.
Exploitability proxy	Self-play	Distance from a locally stable mixed strategy in the finite hidden-type game.
Violation probability	Self-play and evolution	Frequency of invalid actions before or after the verifier gate.
Old-family retention	Evolution	Preserved performance on earlier scenario families after later adaptation.

oration. A human expert can inspect the belief distribution, challenge the state abstraction, change the legal precondition schema, or adjust the payoff design without treating the agent as an opaque advisor. The agent's role is to expose the consequences of assumptions under a reproducible game model. That is a narrower but safer role than autonomous legal decision-making. It is also the role that makes the benchmark relevant beyond arbitration: many scientific domains require agents to reason from partial evidence, maintain a record of admissible operations, and preserve earlier skills while adapting to new evidence.

7. Interpretation of the Three Experiments

The three experiments should be read together rather than as independent leaderboards. Experiment 1 shows why a public record is not enough. The public-only classifier retains a plausible AUC relative to the full-oracle classifier, which could be mistaken for modest degradation. Yet its calibration is much worse on the full population. This is precisely the failure expected when public observations are selected by outcome, salience, and hidden type. In other words, public data can support a competent ranker while failing as a probabilistic model of the hidden population.

Experiment 2 shows why calibration alone is not enough. Even if beliefs are reasonably calibrated, an agent must choose actions in a strategic environment. The public-greedy policy is strong under the biased public prior, but its worst-type value is much lower than gated self-play. This is not a contradiction. It means that a public-prior objective and a hidden-type robustness objective select different policies. The benchmark makes that tradeoff visible: public-greedy play maximizes a narrow average, while self-play spreads probability mass across actions that are more stable under type uncertainty.

Experiment 3 shows why one successful adaptation is not enough. The recent-only agent continues to perform reasonably on the current scenario family, but capability-preserving variants retain more old-family value and remove invalid actions. This is a more stringent standard for self-evolution. A self-evolving agent should not simply chase the latest distribution; it should expose the cost of adaptation, maintain a memory of prior competencies, and keep invalid actions out of the policy support. The verifier removes one class of unsafe drift by construction, but the capability-preservation term is still needed because valid actions can nevertheless become strategically brittle.

8. Why Arbitration Is a Useful Scientific-Agent Stress Test

The London Court of International Arbitration is not included because legal prediction is the target application. It is included because its institutional structure makes several general scientific-agent problems concrete. Confidential proceedings resemble unpublished negative results or proprietary experiments: the public record is informative but selected. Tribunal discretion resembles latent evaluator preference in peer review: the agent observes outcomes but not the full decision rule. Procedural preconditions resemble admissible transformations in formal mathematics or experimental science: some moves are not merely bad; they are invalid. Multiparty disputes resemble distributed scientific collaborations in which actors have different incentives and partial information.

This analogy is useful because it resists over-simplified evaluation. A benchmark built only from public labels would reward the public-only classifier in Experiment 1. A bench-

mark built only from average payoff would reward public-greedy play in Experiment 2. A benchmark built only from final-current performance would understate capability erosion in Experiment 3. LCIA Self-Play instead requires a multi-metric evaluation: calibration, robustness, legality, exploitability, and retention. These are not cosmetic metrics; they correspond to distinct failure modes that appear when agents are allowed to change themselves over long horizons.

The legal abstraction also helps define boundaries. The benchmark can be run without confidential data. It can be audited because the state variables, hidden types, payoffs, and constraints are generated by code. It can be varied systematically by changing selection severity, hidden-type priors, payoff matrices, or verifier strictness. It therefore functions as a middle ground between toy matrix games and real high-stakes deployment: richer than a generic POMDP, but safer and more reproducible than an empirical study of private disputes.

The design also makes a deliberate methodological choice: it treats legal procedure as a source of formal structure, not as a source of authority. A real arbitral rule is written for parties and tribunals, while a benchmark rule is written for machines and evaluators. The two should not be confused. The benchmark preconditions are simplified enough to be executable and transparent, which is precisely why they are useful for testing self-evolving agents. When an action is rejected, the agent receives an explicit reason. When a hidden-type belief shifts, the strategy card records the shift. When capability retention falls, the trace identifies which scenario family was damaged. These observability properties are easier to obtain in a synthetic benchmark than in a real legal system, and they are the reason the environment is suitable for controlled scientific evaluation.

A second design choice is to keep the benchmark small enough for exact inspection. Larger environments could add natural-language pleadings, document production, arbitrator appointment histories, or multi-round settlement bargaining. Those additions may be useful later, but they would make it harder to determine whether a failure came from extraction, strategic learning, memory, or the verifier. The present version isolates the core loop: biased observations, hidden types, legal masks, regret-minimized play, memory, and capability preservation. This modularity makes the benchmark extensible. A theorem-proving variant can replace the finite verifier with a proof-checked transition relation. A natural-language variant can prepend an autoformalization task. A multi-agent variant can let claimant, respondent, tribunal, and institution agents co-evolve under separate information partitions.

Proof-carrying variant. The finite gate can be lifted into a proof assistant by defining procedural states and actions as inductive types, admissibility as an inductive relation, and the executable gate as a Boolean decision procedure with a soundness theorem: if the gate returns true, the corresponding admissibility proposition is derivable. The strategy card would then carry not only a rejected-action certificate but also a proof object or proof hash for each authorized action. Appendix R sketches this variant in Lean-style pseudocode.

A third design choice is to avoid a single success metric. The benchmark reports calibration, hidden-type robustness, exploitability, validity, and retention because each metric exposes a different failure: public success can hide full-population miscalibration; high average value can hide brittleness; zero invalid actions can coexist with weak strategy; and new-family adaptation can erase old skills. Selection bias is therefore part of the environment rather than a nuisance variable, and admissibility is separated from preference: the policy may prefer an action that the verifier rejects, while an admissible action may still be strategically poor.

9. Related Work

This work sits at the intersection of arbitration procedure, legal-AI evaluation, and self-evolving scientific agents. Arbitration scholarship supplies the institutional structure: confidentiality, appointment mechanisms, challenge standards, joinder, consolidation, emergency relief, early disposition, and London-seat supervision (Turner & Mohtashami, 2009; Lew et al., 2001; Park, 2012; Franck et al., 2015; Gaukrodger & Gordon, 2012; Franck, 2005). Legal-AI research supplies benchmark and reliability concerns, including legal reasoning benchmarks, case-based retrieval, legal hallucination tests, fairness, and automation risks (Wiratunga et al., 2024; Gray et al., 2025; Hindi et al., 2025; Katz et al., 2024; Alhasan, 2025; Feldman et al., 2015; Gans-Combe, 2022; McKay, 2019). Scientific-agent work supplies the self-evolution loop: autonomous hypothesis generation, agentic tree search, runtime memory, test-time self-improvement, reflective learning, peer-review diagnostics, and science-of-science evaluation (Yamada et al., 2025; Eger et al., 2025; Li et al., 2025b;a; Garikaparthi et al., 2026; Huang, 2026; Shao et al., 2025; Fortunato et al., 2018; Rubin, 2023).

The closest conceptual relation is not to legal prediction but to benchmarks that expose whether agents can reason under incomplete evidence. ResearchGym and grounded mini-research loops test end-to-end research behavior (Garikaparthi et al., 2026; Huang, 2026); peer-review diagnostics test whether agents can evaluate scientific claims without over-weighting superficial signals (Li et al., 2025b; Ebadi et al., 2025; Garg et al., 2025). LCIA Self-Play contributes a complementary environment in which the missing evidence is structural, the opponent is strategic, and admissibility is formally gated.

Table 4. Failure modes intentionally separated by the benchmark.

Failure mode	What a weak evaluation would miss	What LCIA Self-Play measures
Public-record overfitting	AUC remains stable on visible cases.	Full-population Brier score, calibration error, and public-overfit gap.
Strategic brittleness	Average expected value looks high.	Worst-hidden-type value and exploitability proxy.
Procedural invalidity	Invalid actions are hidden inside a policy score.	Verifier rejection rate and typed failure certificates.
Capability erosion	Latest-family value improves after adaptation.	Retained old-family value and worst-family value.
Memory noise	More memory appears to help in aggregate.	Memory-only, CPE-only, Memory+CPE, and uncertainty-aware memory are compared against gated and recent-only adaptation.

10. Limitations

The benchmark is synthetic. It does not claim to model the real LCIA, predict awards, advise parties, or automate arbitral decision-making. The legal gate encodes simplified preconditions, not the full LCIA Rules, English arbitration law, or any enforceability doctrine. The experiments are diagnostic and should be read as controlled probes of learning dynamics. Stronger future versions should add richer natural-language-to-state extraction, proof-assistant checked rule fragments, adversarial public-record generators, and human evaluation of strategy-card interpretability.

11. Conclusion

LCIA Self-Play turns a confidentiality-heavy institutional setting into a reproducible scientific-agent benchmark. The central lesson is that self-evolution should be evaluated as a constrained process, not merely as an improvement in final score. Public-slice observations can support apparently competent prediction while degrading full-population calibration. Strategic policies can maximize a biased average while losing hidden-type robustness. Adaptive agents can improve on the newest family while eroding older capabilities. A benchmark that measures only one of these dimensions rewards incomplete forms of improvement.

The London Court of International Arbitration is useful here because it makes the evaluation problem concrete. Confidentiality creates biased evidence. Multiparty procedure creates coupled actions. Emergency relief, consolidation, security for costs, confidentiality orders, and early determination create preconditions that can be checked before optimization. These features yield a compact test of whether an agent can maintain beliefs, policies, memories, and admissibility certificates at the same time. The point is not to automate arbitration. The point is to build a stress test for any scientific agent that must reason behind a partial public record while respecting explicit constraints.

Impact Statement

This paper presents a benchmark and simulation framework for studying self-evolving agents under confidentiality, strategy, and formal constraints. The intended use is evaluation and scientific analysis, not legal advice or real arbitration automation. The main positive impact is a safer way to test agents in a high-stakes structural setting without using confidential data. The main risk is overinterpretation: synthetic results could be mistaken for empirical claims about real arbitral outcomes. The paper mitigates this by using generated data, explicit disclaimers, reproducible code, and a strategy-card interface that foregrounds uncertainty and admissibility.

References

Acikgoz, E. C., Qian, C., Ji, H., Hakkani-Tur, D., and Tur, G. Self-improving llm agents at test-time, 2025. URL http://arxiv.org/abs/2510.07841v1.

Alhasan, T. K. Integrating ai into arbitration: Balancing efficiency with fairness and legal compliance. *Conflict Resolution Quarterly*, 2025. doi: 10.1002/crq.21470. URL https://doi.org/10.1002/crq.21470.

Allard, M.-A., Teinturier, A., Xing, V., and Viaud, G. Experiential reflective learning for self-improving llm agents, 2026. URL http://arxiv.org/abs/2603.24639v2.

Blackaby, N., Partasides, C., Redfern, A., and Hunter, M. *Redfern and Hunter on International Arbitration*. Oxford University Press eBooks, 2015. doi: 10.1093/law/9780198714248.001.0001. URL https://doi.org/10.1093/law/9780198714248.001.0001.

Bommasani, R. et al. On the opportunities and risks of foundation models. *arXiv (Cornell University)*, 2021. doi: 10.48550/arxiv.2108.07258. URL http://arxiv.org/abs/2108.07258.

Booeshaghi, A. S., Luebbert, L., and Pachter, L. Science should be machine-readable. *biorxiv*, 2026. doi: 10.64898/2026.01.30.702911. URL https://doi.org/10.64898/2026.01.30.702911.

Brown, C. E., Kaliszyk, C., and Urban, J. Agent hunt: Bounty based collaborative autoformalization with llm agents. *arXiv (Cornell University)*, 2026. doi: 10.48550/arxiv.2603.06737. URL https://openalex.org/W7134861061.

Brown, N. and Sandholm, T. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019. doi: 10.1126/science.aay2400.

Collett, M. J., Davies, J. C., and McRae, B. The london court of international arbitration (lcia) rules. In *Oxford University Press eBooks*. Oxford University Press, 2019. doi: 10.1093/law/9780198784807.003.0019. URL http://dx.doi.org/10.1093/law/9780198784807.003.0019.

Collins, M. Privacy and confidentiality in arbitration proceedings. *Arbitration International*, 1995. doi: 10.1093/arbitration/11.3.321. URL https://doi.org/10.1093/arbitration/11.3.321.

Cuniberti, G. and Gonzalez, M. J. J. S. Choice of law in latin american arbitration: Some empirical evidence and reflections on the latin american market for contracts. *THEMIS Revista de Derecho*, 2020. doi: 10.18800/themis.202001.002. URL https://doi.org/10.18800/themis.202001.002.

Dung, P. M. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2): 321–357, 1995. doi: 10.1016/0004-3702(94)00041-X.

Ebadi, S., Nejadghanbar, H., Salman, A. R., and Khosravi, H. Exploring the impact of generative ai on peer review: Insights from journal reviewers. *Journal of Academic Ethics*, 2025. doi: 10.1007/s10805-025-09604-4. URL https://doi.org/10.1007/s10805-025-09604-4.

Eger, S. et al. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *ArXiv.org*, 2025. doi: 10.48550/arxiv.2502.05151. URL https://doi.org/10.48550/arxiv.2502.05151.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact, 2015. URL https://doi.org/10.1145/2783258.2783311.

Fortunato, S., Bergstrom, C. T., Borner, K., Evans, J. A., Helbing, D., Milojevic, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., and Barabasi, A.-L. Science of science. *Science*, 2018. doi: 10.1126/science.aao0185. URL https://doi.org/10.1126/science.aao0185.

Franck, S. D. The legitimacy crisis in investment treaty arbitration: Privatizing public international law through inconsistent decisions. *eYLS (Yale Law School)*, 2005. URL https://digitalcommons.wcl.american.edu/facsch_lawrev/1582.

Franck, S. D., Freda, J., Lavin, K., Lehmann, T. A., and van Aaken, A. The diversity challenge: Exploring the "invisible college" of international arbitration. *eYLS (Yale Law School)*, 2015. URL https://digitalcommons.wcl.american.edu/facsch_lawrev/963.

Gans-Combe, C. Automated justice: Issues, benefits and risks in the use of artificial intelligence and its algorithms in access to justice and law enforcement. In *Research Ethics Forum*. Springer, 2022. doi: 10.1007/978-3-031-15746-2_14. URL https://doi.org/10.1007/978-3-031-15746-2_14.

Garg, M. K., Prasad, T. S. L., Singhal, T., Kirtani, C., Mandal, M., and Kumar, D. Reviewer: An evaluation framework for ai-generated reviews, 2025. URL https://doi.org/10.18653/v1/2025.findings-emnlp.1120.

Garikaparthi, A., Patwardhan, M., and Cohan, A. Researchgym: Evaluating language model agents on real-world ai research. *ArXiv.org*, 2026. URL https://openalex.org/W7130482342.

Gaukrodger, D. and Gordon, K. Investor-state dispute settlement. *OECD working papers on international investment*, 2012. doi: 10.1787/5k46b1r85j6f-en. URL https://doi.org/10.1787/5k46b1r85j6f-en.

Gerbay, R. The london court of international arbitration (lcia). *SSRN Electronic Journal*, 2012. URL https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2346843_code1885414.pdf?abstractid=2346843andmirid=1.

Gray, M., Zhang, L., and Ashley, K. D. Generating case-based legal arguments with llms, 2025. URL https://doi.org/10.1145/3709025.3712216.

Grierson, J. Two brief comments on the law commissions proposed reform of the arbitration act 1996. *Journal of International Arbitration*, 2022. doi: 10.54648/joia2022033. URL https://doi.org/10.54648/joia2022033.

Guha, N. et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *SSRN Electronic Journal*, 2023. doi: 10.2139/ssrn.4583531. URL https://doi.org/10.2139/ssrn.4583531.

Hill, J. and Chong, A. *International Commercial Disputes*. Hart Publishing, 2010. doi: 10.5040/9781849468558. URL https://doi.org/10.5040/9781849468558.

Hindi, M., Mohammed, L., Maaz, O., and Alwarafy, A. Enhancing the precision and interpretability of retrieval-augmented generation (rag) in legal technology: A survey. *IEEE Access*, 2025. doi: 10.1109/access.2025.3550145. URL https://doi.org/10.1109/access.2025.3550145.

Huang, H. Towards grounded autonomous research: an end-to-end llm mini research loop on published computational physics. *ArXiv.org*, 2026. URL https://openalex.org/W7154656351.

Iyengar, G. N. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Katz, D., Bommarito, M. J., Gao, S., and Arredondo, P. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A Mathematical Physical and Engineering Sciences*, 2024. doi: 10.1098/rsta.2023.0254. URL https://doi.org/10.1098/rsta.2023.0254.

Lew, J. D. M., Mistelis, L., and Kroll, S. *Comparative International Commercial Arbitration*. Kluwer Law International, 2001. URL http://bvbr.bib-bvb.de:8991/F?func=serviceanddoc_library=BVB01andlocal_base=BVB01anddoc_number=014678062andsequence=000001andline_number=0001andfunc_code=DB_RECORDSandservice_type=MEDIA.

Li, C., Lai, Z., An, D., Hu, J., and Wen, Z. Advancing mathematical research via human-ai interactive theorem proving. *arXiv (Cornell University)*, 2025a. doi: 10.48550/arxiv.2512.09443. URL https://doi.org/10.48550/arxiv.2512.09443.

Li, J., Li, Y., Hu, X., Gao, M., and Wan, X. Where do llms go wrong? diagnosing automated peer review via aspect-guided multi-level perturbation, 2025b. URL https://doi.org/10.1145/3746252.3761274.

Liu, Q. et al. Evomdt: a self-evolving multi-agent system for structured clinical decision-making in multi-cancer. *NPJ digital medicine*, 2026. doi: 10.1038/s41746-025-02304-8. URL https://pubmed.ncbi.nlm.nih.gov/41501128/.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. Hallucinationfree? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 2025. doi: 10.1111/jels.12413. URL https://doi.org/10.1111/jels.12413.

McKay, C. Predicting risk in criminal procedure: actuarial tools, algorithms, ai and judicial decision-making. *Current Issues in Criminal Justice*, 2019. doi: 10.1080/10345329.2019.1658694. URL https://doi.org/10.1080/10345329.2019.1658694.

Merkin, R. and Flannery, L. *Merkin and Flannery on the Arbitration Act 1996*. Routledge, 2019. doi: 10.4324/9781315692616. URL https://doi.org/10.4324/9781315692616.

Motwani, S., Du, C., Petrov, O., Davis, C., Torr, P., Papania-Davis, A., and Yan, W. Autoor: Scalably post-training llms to autoformalize operations research problems. *ArXiv.org*, 2026. doi: 10.48550/arxiv.2604.16804. URL https://openalex.org/W7155245008.

Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Sturmer, M., and Chalkidis, I. Lextreme: A multi-lingual and multi-task benchmark for the legal domain, 2023. URL https://doi.org/10.18653/v1/2023.findings-emnlp.200.

Nilim, A. and El Ghaoui, L. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Park, W. W. Arbitrator integrity: The transient and the permanent. *Oxford University Press eBooks*, 2012. doi: 10.1093/law/9780199657131.003.0002. URL https://doi.org/10.1093/law/9780199657131.003.0002.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

Peter, A. Climate change and supply chain arbitrations: Impact of eu law on the bri and non-eu entities. *Arbitration.*, 2023. doi: 10.54648/amdm2023028. URL http://dx.doi.org/10.54648/amdm2023028.

Platte, M. When should an arbitrator join cases? *Arbitration International*, 2002. doi: 10.1023/a:1014252924905. URL https://doi.org/10.1023/a:1014252924905.

Ravi, N., Ying, K., Nesterov, V., Krishnan, R., Uskuplu, E., Xia, B., Aswedige, J., and Nashold, L. Formalproofbench: Can models write graduate level math proofs that are formally verified? *arXiv (Cornell University)*, 2026. URL https://openalex.org/W7148178980.

Richman, L. M., Prado, M. C. R. D., and Udow, H. Early dismissal of claims under the 2020 lcia arbitration rules. *BCDR International Arbitration Review*, 2021. doi: 10.54648/bcdr2021037. URL http://dx.doi.org/10.54648/bcdr2021037.

Rockafellar, R. T. and Uryasev, S. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

Rubin, M. Questionable metascience practices. *Journal of Trial and Error*, 2023. doi: 10.36850/mr4. URL https://doi.org/10.36850/mr4.

Shao, E., Wang, Y., Qian, Y., Pan, Z., Liu, H., and Wang, D. Sciscigpt: Advancing human-ai collaboration in the science of science, 2025. URL http://arxiv.org/abs/2504.05559v3.

Silver, D. and Veness, J. Monte-carlo planning in large pomdps. In *Advances in Neural Information Processing Systems*, 2010.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Turner, P. and Mohtashami, R. *A Guide to the LCIA Arbitration Rules*. Oxford University Press eBooks, 2009. doi: 10.1093/law/9780199234431.001.0001. URL https://doi.org/10.1093/law/9780199234431.001.0001.

Veeder, V. V. The new 2014 lcia rules. In *Oxford University Press eBooks*. Oxford University Press, 2016. doi: 10.1093/law/9780198783206.003.0004. URL https://doi.org/10.1093/law/9780198783206.003.0004.

Wang, J. Joinder mechanism in international commercial arbitration: A trend in the digital age? *International Journal for the Semiotics of Law - Revue internationale de Semiotique juridique*, 2023. doi: 10.1007/s11196-023-10068-1. URL https://doi.org/10.1007/s11196-023-10068-1.

Wiratunga, N. et al. Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering. In *Lecture Notes in Computer Science*. Springer, 2024. doi: 10.1007/978-3-031-63646-2_29. URL https://doi.org/10.1007/978-3-031-63646-2_29.

Yalabk, F. T. The impact of the seat of arbitration on judicial-interference: Do sections 67, 68 and 69 of the english arbitration act 1996 regarding challenges of awards make london an attractive hub? *Annales de la Faculte de Droit dIstanbul*, 2021. doi: 10.26650/annales.2021.70.0009. URL https://doi.org/10.26650/annales.2021.70.0009.

Yamada, Y. et al. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *ArXiv.org*, 2025. doi: 10.48550/arxiv.2504.08066. URL https://doi.org/10.48550/arxiv.2504.08066.

Yu, Y., Yuan, X., Jin, H., Liu, H., Yu, Y., and Wang, H. Do self-evolving agents forget? capability degradation and preservation in lifelong llm agent adaptation, 2026. URL http://arxiv.org/abs/2605.09315v1.

Zaman, R. and Sabharwal, D. Vive la difference? convergence and conformity in the rules reforms of arbitral institutions: The case of the lcia rules 2014. *Journal of International Arbitration*, 2014. doi: 10.54648/joia2014033. URL https://doi.org/10.54648/joia2014033.

Zhang, S. et al. Memrl: Self-evolving agents via runtime reinforcement learning on episodic memory, 2026. URL http://arxiv.org/abs/2601.03192v2.

Zhang, T. Toward weight-level self-improving agents with meta-knowledge discovery, 2025. URL https://doi.org/10.36227/techrxiv.175744083.37752625/v1.

Zhang, T. J., Yang, Y., Huang, Y., Lu, S., Scholkopf, B., and Jin, Z. Collective intelligence: On the promise and reality of multi-agent systems for ai-driven scientific discovery. *Preprints.org*, 2025. doi: 10.20944/preprints202508.1640.v3. URL https://doi.org/10.20944/preprints202508.1640.v3.

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*, 2007.

A. Corpus Use and Scope

The source corpus used to design this benchmark consists of curated LCIA materials, LCIA bibliography files, AI-for-law evaluation and legal reasoning bibliographies, self-evolving-agent bibliographies, metascience bibliographies, human-AI collaboration bibliographies, and a prior multilateral legal-reasoning design note. The corpus was used as a design substrate, not as empirical data. The LCIA files motivate confidentiality, multiparty procedure, institutional appointment, London-seat supervision, emergency relief, early determination, and fee/cost structure. The self-evolving-agent files motivate memory, capability preservation, test-time adaptation, and agentic scientific loops. The legal-AI and metascience files motivate reliability, hallucination, public-record bias, reproducibility, and human-AI evaluation.

The benchmark uses only synthetic scenarios. It deliberately avoids confidential arbitral records and does not infer real tribunal behavior. This separation is central: the benchmark asks whether an agent can reason under the same *structure* of missing information without claiming access to the missing information itself.

B. Formal Environment Details

The environment state is a vector

$$s = (q, m, u, c, f, g, r, n, \tau), \tag{5}$$

where q is normalized quantum, m is multiparty structure, u is urgency, c is confidentiality sensitivity, f is factual complexity, g is compatibility of arbitration agreements, r is funding risk, n is novel issue status, and τ is procedural time. The hidden type $\theta \in \Theta$ is not observed by the agent. Public observations are generated by a selection process

$$P(P = 1 \mid X, Y, \theta) = \sigma(\alpha + \beta^\top X + \gamma Y + \delta_\theta + \eta |P(Y = 1 \mid X, \theta) - 1/2|). \tag{6}$$

so $P(X, Y \mid P = 1) \neq P(X, Y)$ whenever γ, δ_θ, or η is nonzero.

The verifier is a predicate over state-action pairs. In the current implementation it checks a finite set of preconditions:

$$\mathcal{V}(s, \texttt{consolidate}, C) = 1[g = 1 \wedge m = 1], \tag{7}$$

$$\mathcal{V}(s, \texttt{emergency}, C) = 1[u > u_0], \tag{8}$$

$$\mathcal{V}(s, \textsc{security-for-costs}, C) = \mathbf{1}\{r_{\text{funding}}(s) = 1\}, \tag{9}$$

$$\mathcal{V}(s, \texttt{early_det}, C) = 1[f < f_0 \vee n = 0]. \tag{10}$$

The thresholds are synthetic and configurable. Rejected actions return a typed failure certificate: failed predicate, required variable, observed value, and local repair hint.

Algorithm 2 Legality gate

Require: state s, action a, constraint set C
for predicate $c_i \in C[a]$ **do**
  if $c_i(s) = 0$ **then**
    return $(0, \textsc{Failure}(a, i, s))$
  end if
end for
return $(1, \textsc{Certificate}(a, C[a]))$

C. Experiment 1 Details

Experiment 1 isolates the observability problem. The full-oracle learner is not a deployable baseline; it is a diagnostic upper bound trained on the hidden full population. The public-only learner represents the common failure mode: using visible decisions as though they were representative. The selection-weighted learner estimates public-observation propensity from observable covariates and applies inverse-propensity weights. It fails because the observation mechanism also depends on hidden type and outcome. This mirrors the identifiability problem in confidential settings: without assumptions about the unobserved selection process, public records cannot identify the hidden population distribution.

D. Experiment 2 Details

Each scenario family defines an action-by-type payoff matrix. Counterfactual regret matching is run for 350 iterations per family in the released experiment script. For a policy π in family f, the legal best-response gap is

$$\epsilon_{\text{BR}}(\pi; f) = \max_{a:\, L_f(a)=1} \sum_\theta p_f(\theta) M_f(a, \theta) - \sum_a \pi(a) \sum_\theta p_f(\theta) M_f(a, \theta). \tag{11}$$

For CFR diagnostics we additionally report the symmetric matrix-game residual

$$\epsilon_{\text{CFR}}(x, y) = \max_i e_i^\top M y - x^\top M y + x^\top M y - \min_j x^\top M e_j. \tag{12}$$

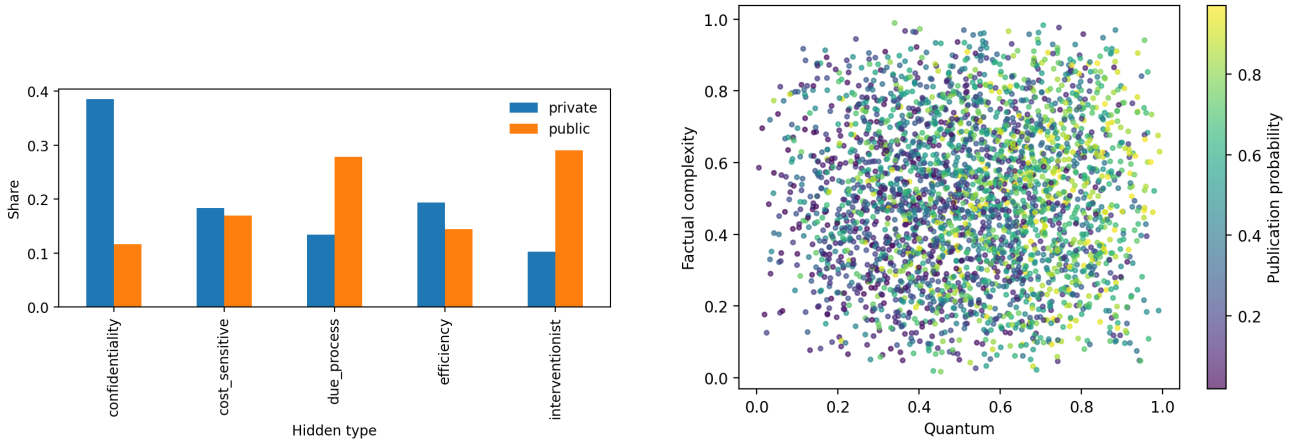
Violation probability is measured as

$$v(\pi, s) = \sum_a \pi(a \mid s) \mathbf{1}\{\mathcal{V}(s, a, C) = 0\}, \tag{13}$$

with ungated baselines evaluated before verifier repair and gated baselines evaluated after action masking. The public-

Table 5. How curated source families informed the benchmark design.

Source family	Design signal	Benchmark translation
LCIA institution and rules	Confidentiality, joinder, consolidation, appointment, challenge, emergency and early determination tools	Hidden observations, admissible action masks, procedural-state variables
Comparative arbitration	Institutional differences in cost, scrutiny, multiparty handling, and seat effects	Scenario families and cost/efficiency trade-offs
AI legal evaluation	Hallucination, benchmark fragility, legal reasoning task design	Calibration, public-slice gap, strategy-card reporting
Self-evolving agents	Capability erosion, runtime memory, self-improvement, agentic tree search	Memory+CPE experiment and retention metrics
Metascience and collaboration	Replication, research-front detection, peer-review diagnostics	Reproducibility bundle, perturbation and failure-ledger framing



(a) Hidden-type mix distortion.

(b) Selection probability surface.

Figure 6. Experiment 1 appendix visuals. The public sample distorts both hidden-type frequencies and the success-probability distribution.

greedy policy is deterministic and optimized against a biased public prior. The particle-belief planner samples hidden types from the posterior and applies a payoff-variance penalty. The Dirichlet-robust planner samples plausible hidden-type priors and optimizes a lower payoff quantile. The gated CFR policy optimizes a mixed strategy over admissible actions. For gated policies, best responses in the exploitability proxy are restricted to legal actions; for ungated policies, exploitability is computed on the unconstrained action simplex and violation probability separately reports illegal mass. Violation is counted from the policy before any verifier repair for ungated baselines and after verifier masking for gated baselines. The important distinction is robustness: public-greedy play can be strong under the average prior while weaker against the least favorable hidden type.

The unrounded CSV files store violation mass before rounding. In this release, the ungated CFR policy assigns a small average illegal mass ($4.44 \times 10^{-4} \pm 8.3 \times 10^{-5}$), which rounds to 0.000 in coarse tables; all verifier-gated policies have exactly zero post-mask violation mass. The violation metric should therefore be read as a support diagnostic for the policy distribution, not as a claim about a final executed legal act.

E. Experiment 3 Details

Experiment 3 uses a ten-phase curriculum in which five scenario families are presented twice: commodity confidentiality, finance multiparty, sanctions urgency, construction supply, and low-value complexity. Each phase lasts 35 episodes. Agents maintain action-scoring weights. Recent-only adaptation updates only on the current family. The gated agent masks invalid actions. Capability-preserving evolution adds rehearsal updates on previous families and a small drift penalty toward a reference parameter snapshot.

F. Additional Algorithms

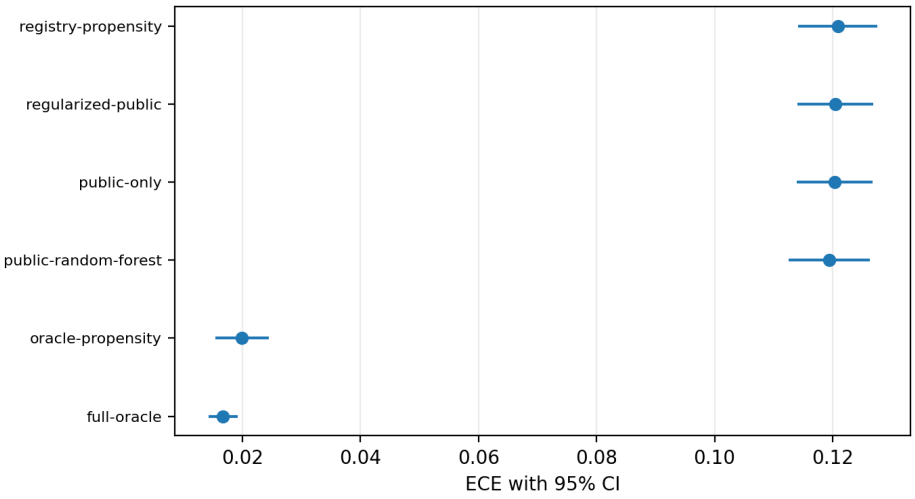
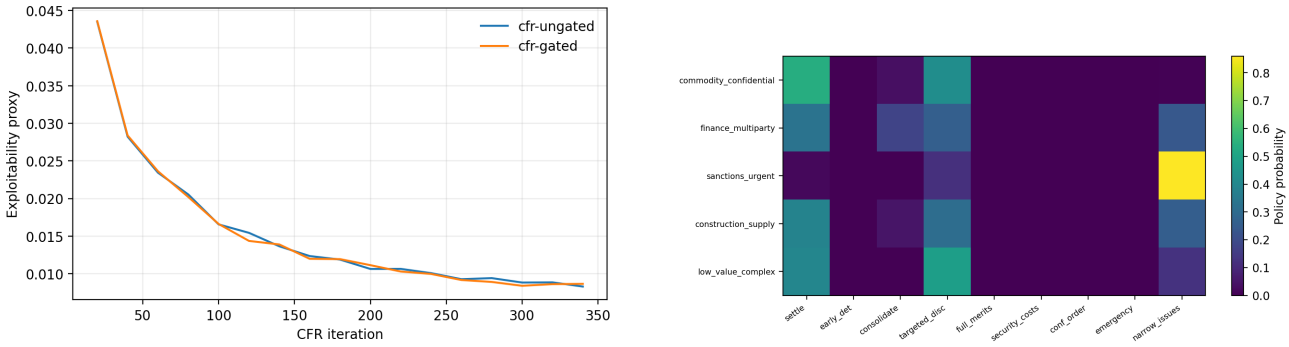


Figure 7. Expected calibration error by hidden type. Public-slice training produces type-local calibration failures even when aggregate AUC remains stable.



(a) Regret convergence.

(b) Gated policy heatmap.

Figure 8. Experiment 2 learning dynamics and learned action distributions.

G. Reproducibility Package

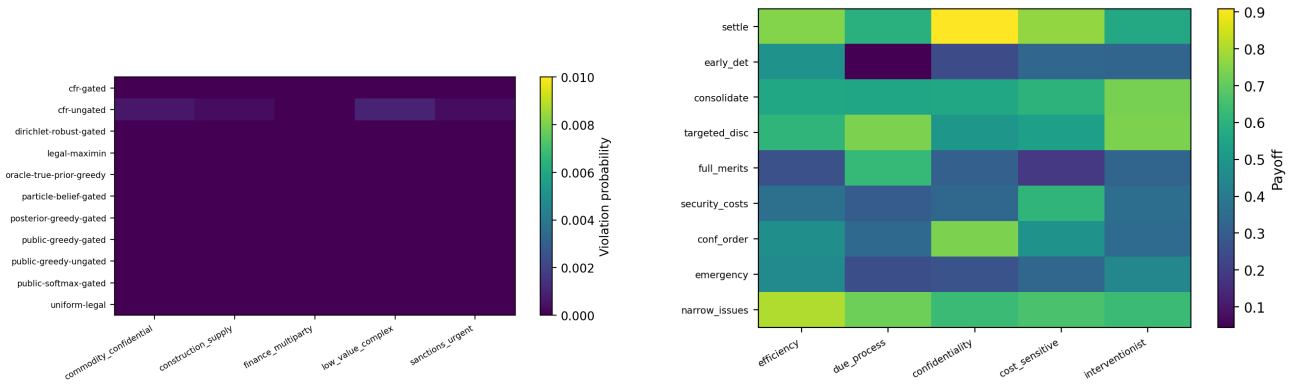
The released bundle contains the LaTeX source, Markdown source, bibliography, style file, all generated figures, CSV data, JSON summaries, and Python code. The experiments can be reproduced with:

```
python experiments/run_all.py
```

The generated files are deterministic under the pre-specified seed list $12345, 12362, 12379, \ldots, 12532$. The implementation uses NumPy, pandas, matplotlib, and scikit-learn. No network access or proprietary data are required.

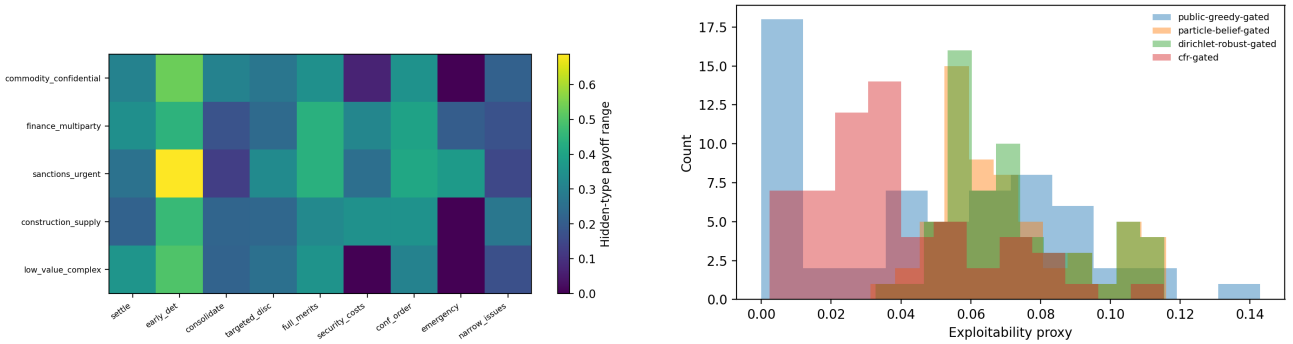
H. Additional Discussion

The benchmark intentionally separates three questions that are often conflated. First, can an agent extract a state representation from text? This version does not test that; it supplies synthetic state variables. Second, can an agent compute robust policies under hidden types? Experiment 2 tests a finite version of that question. Third, can an agent evolve without degrading old capabilities or violating constraints? Experiment 3 tests this directly. Future work can connect the layers by adding a natural-language front end and proof-assistant-checked rule fragments. Until then, LCIA Self-Play is best understood as a controlled intermediate benchmark between legal-text evaluation and fully embodied scientific-agent research.



(a) Violation heatmap. *(b)* Payoff matrices.

Figure 9. Experiment 2 admissibility and payoff structure.



(a) Action value by hidden type. *(b)* Exploitability distribution.

Figure 10. Sensitivity and final exploitability diagnostics.

Algorithm 3 Capability-preserving evolution

Initialize parameters W_0 and reference $\bar{W} \leftarrow W_0$
for episode $e = 1, \ldots, E$ **do**
  Sample current family f_e and observe state feature $\phi(s)$
  Choose admissible action by softmax over $W\phi(s)$
  Update W toward the best observed action in f_e
  for rehearsal family f in memory **do**
    Apply small update toward best admissible action for f
  end for
  $W \leftarrow (1-\lambda)W + \lambda\bar{W}$
end for

Algorithm 4 Public-slice calibration diagnostic

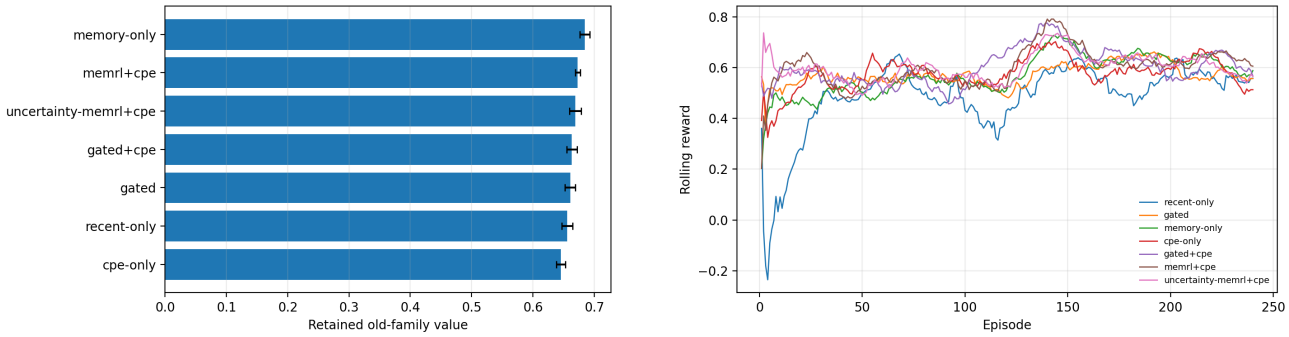
Generate full synthetic population $D = \{(x_i, \theta_i, y_i, p_i)\}_{i=1}^n$
Select public subset $D_P = \{i : p_i = 1\}$ using MNAR mechanism
Train h_F on D and h_P on D_P
Evaluate AUC, Brier score, ECE on the full test population
Report public-overfit gap $Brier(D) - Brier(D \mid P = 1)$

Algorithm 5 Strategy-card construction

Require: state s, belief b, policy π, verifier $\mathcal{V}$, memory M
$A_{adm} \leftarrow \{a : \mathcal{V}(s, a, C) = 1\}$
$\epsilon \leftarrow$ exploitability proxy of π over hidden-type game
$F \leftarrow$ list of rejected actions with typed failure certificates
$H \leftarrow$ retrieved high-utility memory records relevant to (s, b)
return $(b, A_{adm}, \pi, \epsilon, F, H)$

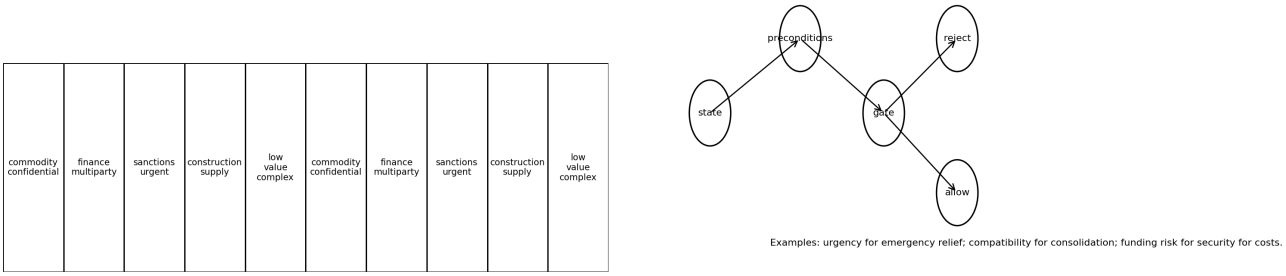
agent	true_expected_value	worst_hidden_type_value	violation_probability
cfr-ungated	0.650	0.626	0.000
cfr-gated	0.650	0.626	0.000
legal-maximin	0.669	0.590	0.000
particle-belief-gated	0.618	0.575	0.000
dirichlet-robust-gated	0.617	0.574	0.000
oracle-true-prior-greedy	0.689	0.565	0.000
public-softmax-gated	0.592	0.557	0.000
public-greedy-ungated	0.641	0.510	0.000
public-greedy-gated	0.641	0.510	0.000
posterior-greedy-gated	0.639	0.505	0.000
uniform-legal	0.502	0.461	0.000

Figure 11. Aggregate Experiment 2 metric table rendered as a figure for visual inspection.



(a) Final retained capability. *(b)* Rolling realized payoff.

Figure 12. Experiment 3 retention and realized-payoff diagnostics.



(a) Curriculum phases. *(b)* Constraint automaton.

Figure 13. Experiment 3 curriculum and an example finite-state view of procedural constraints.

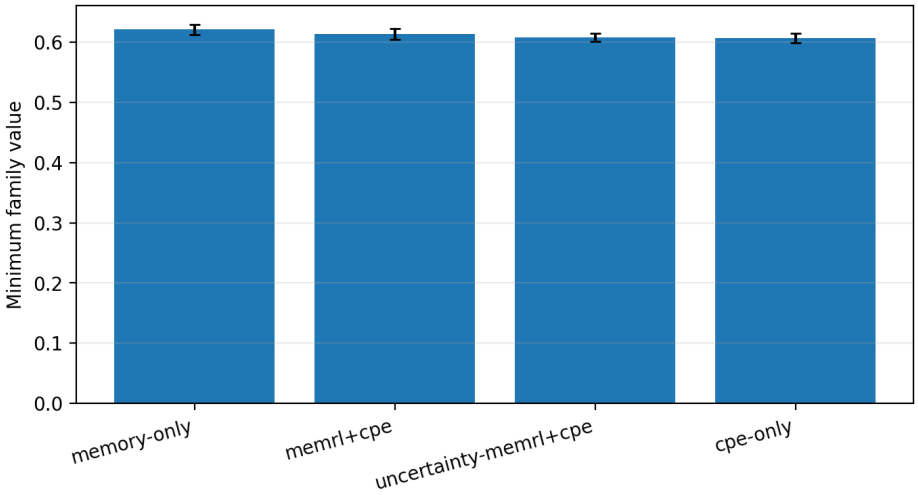


Figure 14. Memory and capability-preservation ablation. Memory-only replay, CPE-only drift control, memory+CPE, and uncertainty-aware memory+CPE are evaluated separately so that retention gains cannot be attributed to a single opaque adaptation bundle.

Table 6. Primary generated files in the artifact bundle.

Path	Contents
`experiment script`	Data generation, three experiments, all figures
`data/e1_metrics.csv`	Public-slice calibration metrics
`data/e1_selection_...csv`	MNAR selection-strength sensitivity metrics
`data/e2_metrics.csv`	Self-play and legality metrics
`data/e3_eval.csv`	Capability-retention evaluation trace
`figures/*.png`, `figures/*.pdf`	Main and appendix visuals

I. Synthetic Generator Parameterization

The generator produces a full synthetic population before selecting the public subset. This is essential because the experiment must know the hidden population in order to measure the public-slice error. A real analyst of confidential arbitration data would not know this population. The oracle population is therefore a diagnostic device, not a deployment assumption.

The selection process creates three forms of public-record distortion. First, higher-quantum cases are more likely to be visible. Second, cases with outcomes far from the decision boundary are more likely to be visible. Third, hidden types have different visibility offsets. This means that inverse-propensity weighting based only on observable variables is not guaranteed to repair the bias: the selection mechanism depends on variables that are either unobserved or endogenous to the hidden decision process.

J. Verifier Rule Schema and Certificates

The verifier is implemented as a small deterministic predicate checker. It is intentionally not a learned model. Its job is not to decide which action is strategically best but to remove invalid moves from the action support and to return typed failure information. This design follows a simple principle: learned policies may rank actions, but formal preconditions should be checked by explicit code whenever the relevant variables are available.

A certificate is a JSON-serializable record containing five fields: action, decision, failed predicate, observed state value, and repair hint. In the current implementation, certificates are generated but not stored as separate files for every episode; aggregate failure rates are stored in the experiment tables. A future proof-assistant version could replace the predicate checker with a formally verified finite-state transition system.

Algorithm 6 Typed failure-certificate construction

Require: state s, action a, predicates $C[a]$
$failures \leftarrow []$
for $(name, predicate, repair)$ in $C[a]$ **do**
  if $predicate(s) = 0$ **then**
    append $(a, name, s, repair)$ to $failures$
  end if
end for
if $failures$ is empty **then**
  return ADMISSIBLE$(a, C[a])$
else
  return REJECTED$(a, failures)$
end if

Table 7. State variables used by the synthetic generator. All variables are generated from public code and are not derived from confidential materials.

Variable	Type	Interpretation
`quantum`	continuous $[0, 1]$	Normalized amount in dispute. Higher values increase salience and public-selection probability.
`urgency`	continuous $[0, 1]$	Proxy for emergency-relief suitability. Emergency action is admissible only above a threshold.
`complexity`	continuous $[0, 1]$	Factual and legal complexity. High complexity lowers the appeal of early determination.
`confid.`	continuous $[0, 1]$	Sensitivity of private information. High values increase payoff for confidentiality orders and settlement.
`multiparty`	binary	Whether more than two parties or related contracts are present.
`compatible`	binary	Whether arbitration agreements are compatible enough for consolidation in the simplified rule schema.
`funding_risk`	binary	Whether cost/security concerns make security for costs admissible.
`novel_issue`	binary	Whether an issue is legally novel. Novelty reduces early-determination value.
`public`	binary	Whether the synthetic case appears in the public slice. Generated by the MNAR selection mechanism.

Table 8. Hidden type definitions. Types should be understood as stylized evaluator or environment tendencies, not claims about real arbitrators.

Type	Payoff tendency	Observable signature
Efficiency	Rewards issue narrowing, early determination, and settlement where complexity is low.	Shorter trajectories, higher value for narrowing actions.
Due process	Penalizes aggressive narrowing when complexity or novelty is high.	Lower value for early determination in difficult cases.
Confidentiality	Rewards confidentiality orders and private settlement in sensitive disputes.	Higher payoff when confidentiality feature is high.
Cost sensitivity	Rewards settlement, security for costs, and proportional procedures.	Higher payoff when quantum and funding risk interact.
Interventionism	Rewards emergency relief, disclosure, and active case management when conditions support them.	Higher payoff for targeted procedural moves.

Table 9. Simplified admissibility schema used in the experiments.

Action	Required condition	Failure certificate
`settle`	always admissible	none
issue narrowing	always admissible	none
full merits	always admissible	none
targeted disclosure	complexity above a low threshold or interventionist signal	failed disclosure
early determination	low complexity or non-novel issue	failed early determination
`consolidate`	multiparty and compatible-agreement flags true	failed consolidation
security for costs	funding-risk flag true	failed security for costs
confidentiality order	confidentiality sensitivity above a low threshold	failed confidentiality
`emergency`	urgency above threshold	failed emergency

K. Experiment 1 Full Protocol

The public-slice experiment proceeds in five stages. First, the generator samples hidden types and observable state variables. Second, a synthetic outcome is sampled using both state variables and hidden type. Third, a public-selection indicator is generated from quantum, outcome salience, outcome extremity, and type visibility. Fourth, six baselines are trained: a full-oracle model, a public-only model, a regularized public model, a registry-propensity inverse-weighted model, an oracle-propensity diagnostic, and a public-slice random forest. Fifth, all models are evaluated on the full hidden test population. The full test population is used only for evaluation.

The representative public subset contains approximately 46.4% of generated cases. Its outcome mean is higher than the full-population outcome mean because successful, salient, and high-visibility matters are more likely to become public. The public subset also over-represents interventionist and due-process hidden types and under-represents confidentiality-sensitive types. This distorted mix explains why a public-only learner can retain plausible AUC while becoming badly calibrated: the decision boundary remains partly learnable, but the probability scale is shifted.

L. Experiment 2 Full Protocol

The self-play experiment uses five scenario families. Each family has a payoff matrix $M_f \in \mathbb{R}^{|A|\times|\Theta|}$ and a legality mask $L_f \in \{0,1\}^{|A|}$. Public-greedy play maximizes $a^\top M_f \hat{p}_{public}$ over legal actions. Additional baselines include uniform legal play, softmax public-greedy play, posterior-greedy play from weak procedural signals, particle-belief and Dirichlet-robust planners (Silver & Veness, 2010; Iyengar, 2005; Nilim & El Ghaoui, 2005; Rockafellar & Uryasev, 2000), legal maximin planning, an oracle true-prior greedy diagnostic, ungated CFR, and verifier-gated CFR. CFR-gated performs regret matching over legal actions only. CFR-ungated computes regrets over all actions and is evaluated for violation probability. In Table 11, exploitability is a one-step legal best-response gap, not a global POMDP-equilibrium theorem:

$$\epsilon(\pi; f) = \max_{a:\, L_f(a)=1} e_a^\top M_f p_f - \pi^\top M_f p_f. \tag{14}$$

For ungated policies, violation probability is measured before verifier repair as

$$v(\pi; f) = \sum_{a:\, L_f(a)=0} \pi(a), \tag{15}$$

and for gated policies it is measured after masking. This protocol separates strategic brittleness from procedural invalidity.

Table 10. Experiment 1 full metrics, mean ± 95% CI over twelve seeds.

Model	AUC	Brier	ECE	Worst-type Brier
Full-oracle	0.615±0.004	0.238±0.001	0.017±0.002	0.263±0.004
Oracle-propensity	0.609±0.004	0.239±0.001	0.020±0.005	0.263±0.005
Public-only	0.606±0.005	0.254±0.002	0.120±0.006	0.276±0.005
Reg.-public	0.606±0.005	0.254±0.002	0.120±0.006	0.276±0.005
Registry-IPW	0.606±0.006	0.254±0.002	0.121±0.007	0.276±0.005
Public RF	0.588±0.007	0.257±0.002	0.119±0.007	0.280±0.004

Table 11. Experiment 2 metrics, mean ± 95% CI over twelve seeds and five scenario families.

Agent	True EV	Worst type	Exploitability	Violation prob.
CFR-gated	0.650±0.007	0.626±0.007	0.038±0.006	0.000±0.000
CFR-ungated	0.650±0.007	0.626±0.007	0.038±0.006	$4.44\times10^{-4} \pm 8.3\times10^{-5}$
Legal maximin	0.669±0.008	0.590±0.008	0.019±0.007	0.000±0.000
Particle-belief gated	0.618±0.006	0.575±0.006	0.071±0.005	0.000±0.000
Dirichlet-robust gated	0.617±0.006	0.574±0.006	0.072±0.005	0.000±0.000
Oracle true-prior greedy	0.689±0.008	0.565±0.011	0.000±0.000	0.000±0.000
Public softmax gated	0.592±0.005	0.557±0.005	0.097±0.005	0.000±0.000
Public-greedy ungated	0.641±0.013	0.510±0.017	0.048±0.010	0.000±0.000
Public-greedy gated	0.641±0.013	0.510±0.017	0.048±0.010	0.000±0.000
Posterior-greedy gated	0.639±0.014	0.505±0.016	0.049±0.010	0.000±0.000
Uniform legal	0.502±0.003	0.461±0.004	0.187±0.006	0.000±0.000

Table 12. Curriculum families used in Experiment 3.

Family	Dominant stressor	Intended competence
Commodity confidentiality	high confidentiality sensitivity	Preserve private-information reasoning and settlement value.
Finance multiparty	multiparty structure and agreement compatibility	Distinguish valid consolidation from invalid aggregation.
Sanctions urgency	urgency and enforcement risk	Use emergency and protective moves only when preconditions hold.
Construction supply	high complexity and many related contracts	Avoid brittle early narrowing when factual complexity is high.
Low-value complexity	low quantum but high procedural work	Balance proportionality and due process.

M. Worked Strategy-Card Example

The following example illustrates the intended output format. It is not produced from a real dispute. It is a synthetic case with moderate quantum, high confidentiality sensitivity, low urgency, a multiparty flag, and compatible agreements.

```
state:
  quantum: 0.64
  urgency: 0.18
  complexity: 0.41
  confidentiality: 0.83
  multiparty: 1
  compatible: 1
  funding_risk: 0
belief:
  efficiency: 0.21
  due_process: 0.18
  confidentiality: 0.34
  cost_sensitive: 0.15
  interventionist: 0.12
admissible_actions:
  settle, issue_narrow, disclosure,
  early_det, consolidate,
  confidentiality, full_merits
rejected_actions:
  emergency: urgency below threshold
  costs: funding_risk false
policy:
  confidentiality: 0.36
  settle: 0.25
  consolidate: 0.17
  issue_narrow: 0.12
  full_merits: 0.10
robustness:
  exp_value_full: 0.69
  worst_type: 0.63
  exploitability: 0.028
ledger:
  confidentiality-heavy policy;
  emergency relief rejected by verifier;
  consolidation permitted.
```

This representation allows a human evaluator to inspect the assumptions without treating the mixed policy as a legal command. The most important information is often in the rejected actions and the hidden-type belief, not in the top-ranked action.

N. Ablations and Sensitivity Axes

The code exposes four major sensitivity axes. First, public-selection severity can be increased by raising the outcome-salience coefficient or type-visibility offsets. This increases the public-overfit gap. Second, hidden-type overlap can be increased by making payoff matrices more similar across types. This reduces the value of robust self-play because hidden-type risk is less severe. Third, verifier strictness can be increased by raising admissibility thresholds. This shrinks the action set and can increase exploitability if the remaining actions are too narrow. Fourth, capability-preservation strength can be increased by raising rehearsal frequency or drift damping. This improves retention but can reduce current-family adaptation.

Selection-bias sensitivity. Table 14 reports the numeric calibration sweep underlying Figure 15. Public-only and registry-propensity learners become sharply miscalibrated once visibility becomes type/outcome dependent, while the full-oracle diagnostic remains stable. Registry-propensity weighting does not solve the problem because the registry-observed covariates do not identify the hidden selection terms.

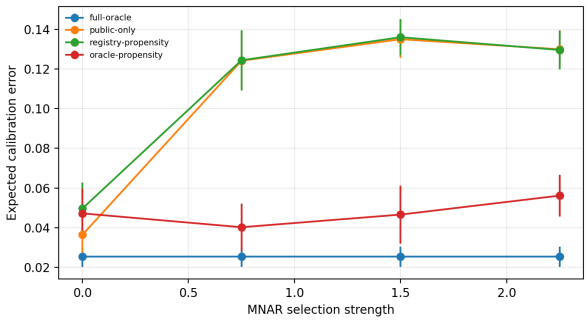


Figure 15. Selection-bias sensitivity. Full-population expected calibration error rises for public-only and registry-propensity learners as public visibility becomes more outcome- and type-dependent. Error bars are 95% confidence intervals over twelve seeds.

O. Non-Goals and Misuse Boundaries

The benchmark has clear non-goals. It does not predict real LCIA awards. It does not model the full LCIA Rules. It does not advise parties, counsel, tribunals, or institutions. It does not infer arbitrator behavior from confidential data. It does not rank arbitral institutions. It does not attempt to replace human legal judgment. The use of LCIA-style structure

Table 13. Sensitivity axes and expected qualitative effects.

Axis	Increase means	Expected effect
Selection severity	More outcome/type-dependent public visibility	Larger calibration gap and larger mismatch between public and full priors.
Type separation	Payoffs differ more sharply by hidden type	Greater value for mixed robust policies; higher penalty for public-greedy specialization.
Verifier strictness	More actions rejected by preconditions	Lower violation probability but potentially higher exploitability.
Rehearsal strength	More updates on previous families	Higher retention and lower current-family specialization.
Memory selectivity	More aggressive filtering of past records	Lower noise in adaptation but possible loss of useful minority cases.

Table 14. Selection-strength sweep, expected calibration error with 95% confidence intervals.

Strength	Full-oracle	Public-only	Registry-IPW
0.00	0.025	0.036	0.050
0.75	0.025	0.124	0.124
1.50	0.025	0.135	0.136
2.25	0.025	0.130	0.130

is methodological: confidentiality, strategic behavior, and procedural constraints create an unusually clean test of self-evolving agents under partial observability.

P. Figure Inventory

The main body contains five visuals: the system architecture, belief-state/verifier diagram, MNAR dashboard, self-play dashboard, and evolution dashboard. The appendix contains sixteen additional visuals: hidden-type mix, selection surface, calibration heatmap, regret convergence, policy heatmap, violation heatmap, payoff matrices, retention bars, rolling payoff, curriculum phases, action-type sensitivity, constraint automaton, exploitability histogram, a rendered Experiment 2 metric table, an uncertainty-aware memory comparison, and a selection-strength sensitivity sweep. Each figure is generated from the same experiment script and is included in both PDF and PNG form for inspection.

Q. Artifact Checklist

Table 16 records the reproducibility components included in the release bundle. All listed items are synthetic, generated, or source-level artifacts; no confidential arbitration materials are included.

R. Proof-Carrying Variant

The current verifier is executable code over a finite schema. A closer fit to formal mathematics is a proof-carrying gate: every authorized action is accompanied by a proof that the action belongs to an inductively defined admissibility relation. In a Lean-style prototype, the benchmark would define finite procedural states, actions, and admissibility constructors:

```
inductive Action
| cons | emerg | secCosts | earlyDet
| confid | disclose | settle | merits | narrow

structure State where
  multi : Bool
  compat : Bool
  urgent : Nat
  fundRisk : Bool
  complex : Nat
  novel : Bool

inductive Admissible : State -> Action -> Prop
| cons_ok (s) : s.multi = true ->
    s.compat = true -> Admissible s Action.cons
| emerg_ok (s) : s.urgent >= 7 ->
    Admissible s Action.emerg
| sec_ok (s) : s.fundRisk = true ->
    Admissible s Action.secCosts
| early_ok (s) : s.complex <= 4 -> s.novel = false ->
    Admissible s Action.earlyDet

-- executable checker plus proof obligation
theorem gate_sound :
  gate s a = true -> Admissible s a := by
  cases a <;> simp [gate, Admissible]
```

A proof-carrying strategy card would replace the present Boolean certificate with $(a, \text{decision}, \text{proof_hash}, \text{failed_constructor})$. For authorized actions, the proof hash identifies a checkable derivation of `Admissible s a`; for rejected actions, the failed constructor identifies the missing precondition. This extension would test a stronger agent property than the current benchmark: not merely whether the agent avoids invalid moves after masking, but whether every proposed move can be accompanied by a machine-checkable admissibility witness. It also connects the benchmark

Table 15. Threat model and mitigations.

Risk	Failure mode	Mitigation in this package
Legal overclaiming	Synthetic policy mistaken for legal advice	Explicit non-deployment boundary and generated data only.
Hidden-data inference	Benchmark assumed to reveal confidential awards	No confidential data, no real parties, no empirical LCIA outcome claims.
Metric gaming	Agent optimizes one metric while failing others	Strategy card reports calibration, robustness, legality, exploitability, and retention together.
Verifier brittleness	Hard gate encodes an oversimplified rule	Gate is disclosed, small, configurable, and treated as a benchmark component.
Capability erosion	Self-evolution improves new cases while forgetting old cases	Experiment 3 directly measures retained old-family value.

Table 16. Reproducibility checklist for the artifact bundle.

Item	Status	Notes
Pre-specified seed list	yes	Twelve seeds beginning at 12345 are used by the experiment script.
Synthetic data generation	yes	All data are generated locally; no confidential data.
Experiment code	yes	One Python script runs all three experiments.
Raw generated data	yes	CSV and JSON files are included.
Main figures	yes	Five main-body figures in PDF and PNG.
Appendix figures	yes	Sixteen appendix figures in PDF and PNG.
Bibliography source	yes	BibTeX file generated from curated bibliographies plus foundational methods references.

directly to autoformalization and proof-agent evaluation, since natural-language procedural facts could first be translated into a formal state and then checked against the inductive rule schema.

S. Future Extensions

The most valuable extension is a proof-checked rule fragment. A Lean, Rocq, or Isabelle implementation could represent admissibility as an inductive relation over procedural states and verify that the gate never authorizes actions outside the relation. A second extension is a natural-language state extractor that maps synthetic procedural narratives to the benchmark state vector. That would connect the current finite game to autoformalization and machine-readable scientific artifacts, a direction related to recent work on autoformalized optimization, theorem-proving agents, and machine-readable science (Motwani et al., 2026; Brown et al., 2026; Booeshaghi et al., 2026). A third extension is an adversarial public-record generator that learns to select public cases in ways that maximize calibration error while preserving apparent ranking quality. A fourth extension is a human evaluation protocol for strategy-card interpretability.

The benchmark can also be transferred to non-legal scientific domains. Confidential arbitration can be replaced by unpublished experiments, proprietary laboratory logs, private peer-review records, or incomplete clinical-study reports. The same structure applies whenever public observations are biased, actors are strategic, and a formal admissibility layer constrains what an agent may do. This is why the LCIA setting is useful: it gives a concrete institutional name to a general problem in self-evolving scientific reasoning.