AUDITING LEARNED MEAN-FIELD EQUILIBRIUM CLAIMS: CLAIM-CONDITIONED RESPONSE AND FLOW EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Mean-field game solvers return a policy together with a population flow, but standard endpoint evidence does not determine whether this pair supports an equilibrium claim. A high return need not imply response optimality; a best response to a declared flow need not generate that flow; and passive trajectories may not expose either failure. We instantiate the broader Claim-Aware Protocol Auditing (CAPA) pattern for MFGs as *MF-CAPA*, a post-hoc auditor that separates response optimality from population consistency, restricts evidence to executable trajectory interventions, and returns CERTIFY, REFUTE, or ABSTAIN. The central distinction is that auditing a fixed claim need not identify the game. For finite quotient models, we define a claim-conditioned trajectory-KL characteristic time, prove that audit characteristic time can be arbitrarily smaller than the time required for model identification, and derive an expected-time lower bound for decisively terminating audits and an almost-sure upper rate. An anytime mixture confidence sequence controls adaptive stopping, while the intervention interface and sound semantic bounders determine whether a verdict is supportable. In actual sequential trials, passive aliases yield only 3/200 and 1/200 decisive stops (3 and 0 wrong), versus 200/200 for each active condition. Across two separately family-wise-calibrated suites of 60 final-candidate/control cells using exact best-response probes and exact mean-field transitions, every decisive audit agrees with exact two-axis truth and all 90 single-axis/wrong-flow controls are refuted; a separate 360-cell study across three native MFGLib learners again yields no wrong decisive decision. Structured numerical relaxations illustrate robust-box verification, distinct from auditing an unknown true model. MF-CAPA provides an evidence contract for deciding which learned equilibrium declarations the available interventions support; it audits supplied candidates rather than introducing a new equilibrium solver.

Keywords: mean-field games, equilibrium auditing, sequential experimental design, trajectory interventions, claim-conditioned inference

1 INTRODUCTION

Mean-field games (MFGs) reduce a large population of strategic agents to the interaction between a representative agent and a population distribution (3; 14; 17). Scalable solvers based on fictitious play, online mirror descent, and deep reinforcement learning now return a policy together with a population flow (5; 24; 25). This pair is naturally read as an equilibrium declaration. Yet solver convergence or model-based exploitability is not, by itself, evidence that the supplied pair is supported under the observations and interventions available to an external auditor.

Selective prediction and conformal risk control provide public precedents for tolerance-controlled decisions with an abstention option (2; 8). MF-CAPA specializes this evidence interface to mean-field games: it adds the policy–flow equilibrium semantics and restricts evidence to executable MFG trajectories. It is therefore an auditor for a supplied claim, not a second equilibrium solver or a relabeling of exploitability.

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

Two broken equivalences create the problem. A policy selected for population return can admit a profitable representative-agent deviation. Conversely, a policy can be a best response to a declared flow even though executing that policy produces a different flow. Thus high return is not response optimality, and conditional response optimality is not population consistency. An MFE declaration requires both. Passive trajectories can also leave an unplayed deviation observationally aliased, so collecting more of the same on-policy data need not resolve either verdict.

We study the resulting audit problem: *when is a supplied MFE declaration auditable under a stated intervention interface?* We introduce MF-CAPA, which treats the candidate $c = (\pi, L)$, its tolerances, and the allowed trajectory experiments as an explicit audit contract. The auditor need only rule out models that reverse this fixed claim's verdict; same-verdict nuisance differences can remain unresolved. We show that the resulting audit characteristic time can be arbitrarily smaller than the time required to identify the observational model. The corresponding *trajectory audit information* maximizes, over executable experiment allocations, the worst KL separation from an opposite-verdict alternative. We give an expected-time change-of-measure lower bound and an anytime-valid finite-model procedure: a mixture likelihood-ratio confidence sequence protects the verdict under adaptive predictable interventions, while forced- exploration Track-and-Stop gives an almost-sure upper rate. If opposite verdicts are observationally aliased, their common probability of a decisive verdict is at most 2δ for a δ-correct audit.

Exact candidates isolate the two semantic failures; actual sequential trials test the role of admissible interventions. We train three native MFGLib solvers from five random initial policies and audit seven checkpoints using three independent audit seeds. We additionally train shared neural policy–flow models from five initializations without equilibrium labels in three-state Treasure Hunting and 9×9 Crowd Motion. Their outputs and single-axis controls test full three-way decision calibration: both separately calibrated 60-cell neural suites and all 360 native-solver cells avoid a wrong decisive state. Finally, structured examples test numerical relaxations of robust semantic properties under explicit oracle and arithmetic limitations.

Contributions.

1. We instantiate CAPA for MFE evaluation as a two-axis, three-decision audit contract for a supplied policy–flow pair, explicitly separating the two claim axes from admissible intervention evidence.
2. We prove that claim auditing can be arbitrarily easier than model identification, characterize finite-model auditability by an executable trajectory-KL characteristic time, and give an expected-time lower bound and an almost-sure upper rate for an anytime-valid quotient procedure.
3. We compose statistical evidence with sound deterministic semantic bounders at the contract level, distinguish this from numerical robust-box verification, and audit controlled failures and learned public-MFG candidates.

2 Related Work and Boundary

MFG solution and evaluation. Classical MFG theory studies existence, uniqueness, approximation, and the coupled optimal-response/population equations (3; 14; 17). Recent algorithms compute equilibria with fictitious play, online mirror descent, or regularized reinforcement learning (1; 5; 9; 12; 13; 15; 18; 19; 24; 25; 31–33); these works target equilibrium computation, convergence, or learning complexity, including single-path and continuous/heterogeneous extensions. MFGLib provides finite-game implementations and model-based exploitability (10). Native MFGLib exploitability constructs the policy-induced flow before evaluating responses. Our additional consistency question concerns an independently supplied declaration L, not a missing check in that policy-only API. Related work on MFG imitation introduces the Nash imitation gap for measuring the quality of a learned imitation policy (27); that object concerns imitation from an expert, whereas we audit a supplied policy–flow declaration under an explicit intervention interface. Fixed-confidence equilibrium identification has also been studied for noisy static zero-sum payoff queries (35). That setting actively observes payoff entries and identifies a pure equilibrium; it does not test a supplied MFG policy–flow pair under trajectory reachability constraints. We do not propose a solver or a new

exploitability metric. The object here is the evidential status of one supplied pair when the auditor has a restricted trajectory interface and may need to abstain.

We use the same evidence roles of certification, deviation-witness refutation, abstention, and claim-sound selection, while grounding them in executable MFG trajectories. Our on-policy example and oracle substitution specialize these principles; they are not independent general contributions. The MFG-specific object is the supplied policy–flow claim and its reduction to an executable trajectory decision problem.

Sequential experimental design. Adaptive experiment selection dates to Chernoff (4); Track-and-Stop gives asymptotically optimal fixed-confidence allocation in best-arm identification (7). Controlled sensing treats finite controls and simple hypotheses with sequential first-order optimality (22); answer-based pure exploration develops lower bounds and Track-and-Stop with multiple correct answers (6). Thus deciding a claim without identifying every model detail is not a new general principle. Time-uniform martingale methods make inference valid under optional stopping (11). These tools provide the sequential inference engine; the object of study is the induced decision problem. Alternatives are models that reverse a supplied MFE claim's response/flow verdict, same-verdict nuisance models need not be distinguished, and feasible allocations are trajectory occupancies rather than arbitrary primitive queries. Our finite-model result is a specialization to the MFG claim and observation interface, not a new general controlled-testing optimality theorem.

Semantic oracles. Linear relaxations and branch-and-bound support neural property verification (29; 34), while robust/parametric MDP methods study uncertain transition semantics (26; 30). We use such techniques only as substitutable deterministic bounders for the two audit axes. General ReLU verification and history-dependent planning remain hard (16; 23); MF-CAPA therefore claims exact-arithmetic soundness contracts for specified structures, not a universal solver. The numerical LP experiments do not supply formal floating-point certificates.

3 The MFE Audit Contract

Consider a finite-horizon MFG model M with finite state and action spaces. The candidate is $c = (\pi, L)$, where π is a representative-agent policy and $L = (L_t)_{t=0}^{H}$ is a declared population flow. Let $J_M(\beta; L)$ be the value of response policy β against the exogenous flow L, and let $\Gamma_M(\pi)$ be the flow generated by executing π in model M. Define the response-gap and flow-consistency tolerances as ϵ_1 and ϵ_2, respectively, and write

$$G_M(c) = \sup_{\beta}\{J_M(\beta; L) - J_M(\pi; L)\}, \tag{1}$$

$$C_M(c) = d(L, \Gamma_M(\pi))\,, \tag{2}$$

$$\rho_M(c) = \max\{G_M(c) - \epsilon_1,\ C_M(c) - \epsilon_2\}. \tag{3}$$

The response class and metric d are declared parts of the claim. With an indifference margin $\gamma > 0$, let

$$\mathcal{M}_0(c) = \{M : \rho_M(c) \le 0\}, \qquad \mathcal{M}_1(c; \gamma) = \{M : \rho_M(c) \ge \gamma\}. \tag{4}$$

No fixed-confidence method is required to terminate uniformly on the omitted boundary strip. A budget limit or unresolved strip returns ABSTAIN.

Observation interface. Let $\mathcal{B}$ be a finite library of executable trajectory experiments. An experiment can deploy a representative-agent deviation against L or roll out π to test population consistency. Under M, experiment b produces a trajectory $Y \sim P_M^b$. At episode t, B_t is predictable from past data and $Y_t \sim P_M^{B_t}$. This excludes a hidden primitive-query oracle: reachability and policy constraints are part of the information available to the auditor.

Definition 3.1 (Three-state audit)**.** An audit is δ-correct on $\mathcal{M}_0 \cup \mathcal{M}_1$ if the probability of ever returning REFUTE under $M \in \mathcal{M}_0$, or CERTIFY under $M \in \mathcal{M}_1$, is at most δ. Non-refutation and an empty confidence set are ABSTAIN, not certification. This risk definition allows immediate abstention. An audit is *decisively terminating* if its first decisive time τ is finite almost surely on every separated model; budget exhaustion is not a decisive stopping time.

This definition separates statistical evidence from exact semantic evaluation. A model-based dynamic program can compute Equations (1) and (2) exactly; a rollout channel may instead provide only a lower witness for a profitable deviation or a confidence band for the flow.

4 CLAIM-CONDITIONED AUDITABILITY

We state the finite quotient result used by the algorithm. The finite class $\mathcal{M} = \mathcal{M}_0 \cup \mathcal{M}_1$ has both verdict regions nonempty and $\mathcal{B}$ is finite; all trajectory laws have common support and finite log-likelihood-ratio second moments. Write $M \sim M'$ when $P^b_M = P^b_{M'}$ for every b. For the constructive rate, each equivalence class must have one verdict. Full model identifiability is not needed.

4.1 AN OBSTRUCTION AND A LOWER BOUND

Proposition 4.1 (On-policy obstruction)**.** *For every $\epsilon_1 > 0$, there are two one-stage finite MFGs with identical on-policy laws for c such that $M_0 \in \mathcal{M}_0(c)$ and $M_1 \in \mathcal{M}_1(c;\gamma)$. For any on-policy-only δ-correct audit, the common probability of a decisive verdict is at most 2δ. In particular, decisive almost-sure termination is impossible for $\delta < 1/2$.*

The proof is in Appendix A.2. The construction keeps the played action and flow identical and changes only the reward of an unplayed deviation. The policy's observed return and flow are therefore uninformative about the response verdict. An intervention that plays the hidden action breaks the alias.

For M outside the boundary strip, let $\mathcal{A}(M,c)$ contain all models with the opposite verdict and define the claim-discrimination information D^*_c and its characteristic time T^*_c by

$$D^*_c(M,c) = \sup_{w\in\Delta(\mathcal{B})} \inf_{M'\in\mathcal{A}(M,c)} \sum_{b\in\mathcal{B}} w_b \,\mathrm{KL}(P^b_M \| P^b_{M'}), \tag{5}$$
$$T^*_c(M,c) = 1/D^*_c(M,c).$$

Here $\Delta(\mathcal{B})$ is the simplex over full trajectory experiments; reachability is already encoded in P^b_M. Unlike model-identification complexity, Equation (5) ignores alternatives that preserve the claim verdict.

Proposition 4.2 (Audit can be easier than identification)**.** *Assume verdict-homogeneous observational classes. Let the identity-discrimination information $D^*_I(M)$ replace $\mathcal{A}(M,c)$ in Equation (5) by every $M' \not\sim M$. Then $D^*_I(M) \le D^*_c(M,c)$, and the ratio of characteristic times can be arbitrarily large.*

The proof is in Appendix A.3. Identification minimizes over a larger alternative set. A same-verdict nuisance model arbitrarily close on a diagnostic experiment makes the inequality strict; the appendix gives an explicit one-stage MFG.

The construction in Figure 1 illustrates why same-verdict nuisance differences do not enter the audit characteristic time.

Theorem 4.3 (Trajectory audit lower bound)**.** *Let $0 < \delta < 1/2$. Every δ-correct, decisively terminating audit with $\mathbb{E}_M\tau < \infty$ satisfies, for a nonempty $\mathcal{A}(M,c)$,*

$$\mathbb{E}_M[\tau] \;\ge\; \frac{\mathrm{kl}(1-\delta,\delta)}{D^*_c(M,c)}. \tag{6}$$

If $D^*_c(M,c) = 0$, no such audit has finite expected decisive time.

The proof in Appendix A.4 applies sequential change of measure to every opposite-verdict model and normalizes expected experiment counts.

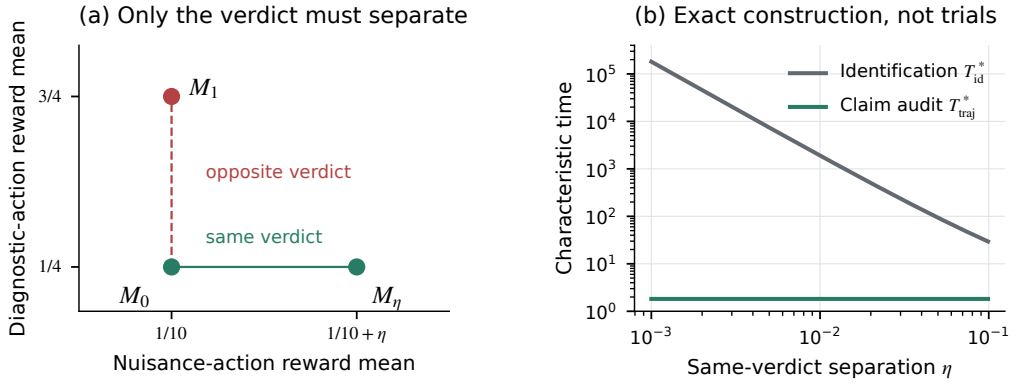


Figure 1: **Claim-conditioned information in the explicit construction.** In the three-model example from Appendix A, M_0 and M_η share a verdict but differ on a nuisance reward mean, while M_1 has the opposite response verdict. The right panel evaluates the construction's KL terms as η shrinks: the claim audit need only separate the opposite-verdict alternative, whereas full identification also resolves M_η. This is a theorem construction, not an empirical estimate.

4.2 Anytime-valid quotient Track-and-Stop

Let $L_t(M) = \prod_{s \le t} p_M^{B_s}(Y_s)$ and fix a positive prior q on $\mathcal{M}$. For each candidate model define the mixture likelihood-ratio process

$$E_t(M) = \sum_{M' \neq M} \frac{q(M')}{1 - q(M)} \frac{L_t(M')}{L_t(M)}, \qquad \mathcal{C}_t(\delta) = \{M : E_t(M) < 1/\delta\}. \tag{7}$$

The experiment-selection probabilities cancel in each ratio because B_t is predictable. The decisive time τ_δ is the first time the nonempty set $\mathcal{C}_t(\delta)$ has one verdict, and is infinite if this never occurs.

Theorem 4.4 (Anytime validity and rate)**.** *For any predictable adaptive experiment sequence, $\mathbb{P}_M\{M \in \mathcal{C}_t(\delta)\ \forall t\} \ge 1 - \delta$; hence the probability of ever returning a wrong verdict is at most δ. If forced exploration identifies the true observational equivalence class and target counts track an optimizer of Equation* (5)*, and $D_c^*(M, c) > 0$, then τ_δ is finite almost surely for each $0 < \delta < 1$. On a common infinite experiment stream with δ-independent allocation,*

$$\limsup_{\delta \downarrow 0} \frac{\tau_\delta}{\log(1/\delta)} \le T_c^*(M, c) \quad \textit{almost surely}. \tag{8}$$

The proof is in Appendix A.5. Ville's inequality for the unit-mean martingale $E_t(M)$ gives validity. Use one shared optimizer per observational class; forced exploration and tracking give log-likelihood growth at rate D_c^* against the hardest opposite verdict. The expected-time lower bound and almost-sure upper rate have the same information constant, but do not establish matching expected complexity: the latter needs additional tail control, not just finite second moments.

Algorithm 1 MF-CAPA quotient Track-and-Stop

Require: Candidate c, model class $\mathcal{M}$, experiments $\mathcal{B}$, risk δ

for $t = 1, 2, \ldots$ **do**
  Update likelihoods and $\mathcal{C}_t(\delta)$ using Equation (7)
  if $\mathcal{C}_t$ is nonempty and verdict homogeneous **then**
    return its verdict
  end if
  if some experiment has fewer than $\sqrt{t}/|\mathcal{B}|$ samples **then**
    execute a least-sampled experiment
  else
    choose an MLE $\widehat{M}_t$ and solve Equation (5)
    execute b maximizing $tw_b(\widehat{M}_t) - N_b(t)$
  end if
end for
return ABSTAIN if the external budget is exhausted

4.3 SUBSTITUTABLE SEMANTIC BOUNDERS

Let S be a nonempty model confidence set. Unknown-model certification requires $\forall M \in S : \rho_M(c) \leq 0$; refutation requires $\forall M \in S : \rho_M(c) \geq \gamma$. The violating axis may differ by model.

Proposition 4.5 (Oracle substitution)**.** *Suppose $G_M(c) \in [\underline{G}, \overline{G}]$ and $C_M(c) \in [\underline{C}, \overline{C}]$ for* every $M \in S$. *Certify if both upper bounds are within tolerance; refute if either lower bound is at least its tolerance plus γ; otherwise abstain. If S contains the true model, this rule cannot give a wrong decisive verdict.*

The proof is in Appendix A.6. A witness with $G_{M_0}(c) > \epsilon_1$ for one $M_0 \in S$ refutes the *robust* property that every model is safe, not the claim about the unknown true model. It lower-bounds $\sup_{M \in S} G_M$, not $\inf_{M \in S} G_M$. Mixed safe/unsafe confidence sets must remain unresolved.

5 EXPERIMENTS

The experiments test the need for both semantic axes, the effect of the intervention class on auditability, calibration along learned checkpoints, and the scope of structured numerical bounders. Unless noted, $\epsilon_1 = \epsilon_2 = 0.05$. Full protocols and raw-result paths are in Appendix B. The studies test distinct interfaces: finite-model sequential design, fixed-budget directional response bounds with simultaneous flow bands, and numerical robust-property relaxations. We do not claim that one benchmark simultaneously exercises every backend.

5.1 RESPONSE AND FLOW EVIDENCE

Table 3 isolates the broken equivalences. The return-selected candidate scores higher than the equilibrium but admits a response gain of 0.1604. The wrong-flow candidate is an exact response to its declaration, yet its policy induces a flow 0.5093 away. Neither failure is visible from the other axis.

Figure 2 shows the two failure modes in the same coordinates: neither is recoverable from the other scalar.

5.2 INTERVENTIONS AND AUDITABILITY

The candidate commits to a safe branch with Bernoulli(0.55) payoff. Responses may instead route to two terminal states with reach probabilities $(r, 1-r)$ and choose terminal actions. A single 0.68 payoff among 0.50 alternatives gives gap $\max\{0, 0.50 + 0.18r - 0.55\} = 0.094$ at $r = 0.8$. The flow residual is terminal total variation. This population-independent MFG has nine models and ten audit experiments, including the actual candidate rollout. Six truth/strategy conditions each use 200 trials at $\delta = 0.05$ and a 100,000-episode budget: 1,200 trials and 39,964,087 sampled episodes.

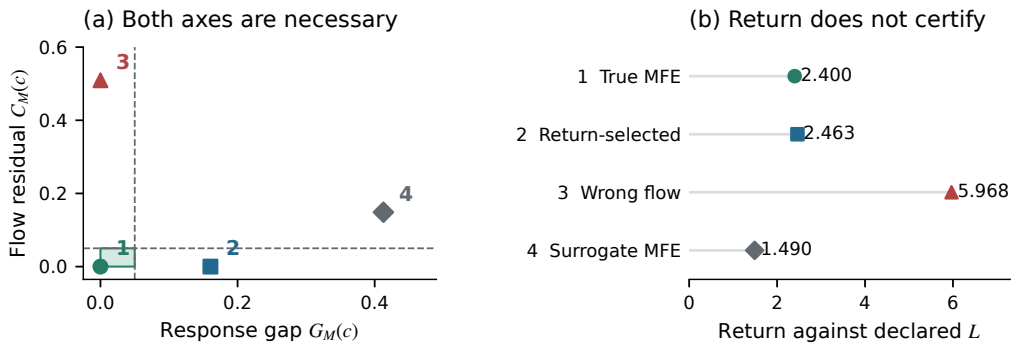


Figure 2: **Exact two-axis evidence from Table 3.** The shaded square is the conjunctive tolerance region. The return-selected candidate (2) fails response optimality, while the wrong-flow candidate (3) satisfies the declared-flow response condition but fails population consistency. Panel (b) shows why return ranking alone cannot replace either axis. Values are the retained exact controlled-audit results.

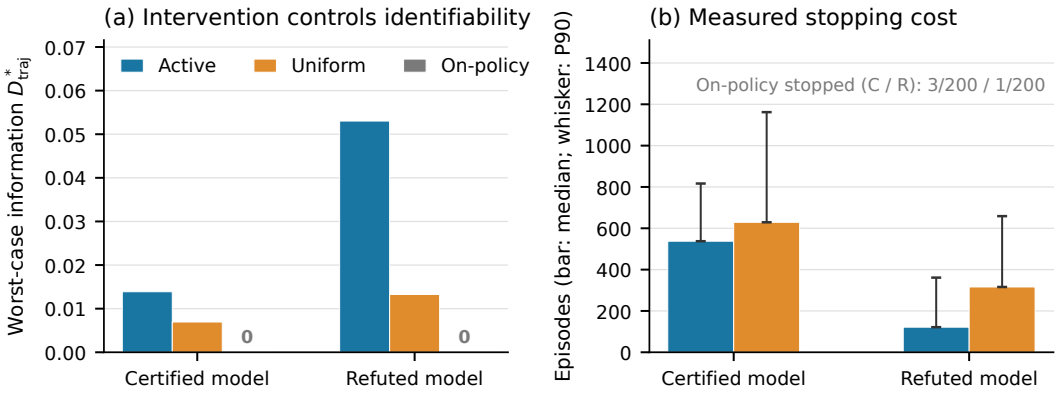


Figure 3: **Intervention-dependent auditability.** Trajectory information and stopping costs at $r = 0.8$. Active and uniform conditions each stop in 200/200 trials; passive conditions stop in only 3/200 and 1/200. Passive stopping quantiles concern rare true-model exclusions, not informative discovery. Full counts are in Table 5.

Active allocation reduces median episodes from 629 to 537.5 for certification and from 316 to 121.5 for refutation. Passive stops yield 3 and 0 wrong verdicts respectively; all four exclude the true model, including the single verdict-correct refutation. Thus passive nonidentifiability limits decisive power, not necessarily the observed error rate. CP95 intervals describe unconditional error frequency, not equality to the risk limit.

Reachability also changes truth: at $r = 0.55$, the gap is 0.049 and the claim is valid. Its only opposite model violates flow; the trajectory/primitive information ratio and cost multiplier are both one. Response-limited $1/r$ cost scaling applies above $r = 5/9$, not across this verdict change (Appendix B.2).

5.3 Claim-conditioned vs identity-aimed allocation

On the headline routing-reachability cell at model count k=8 under *same-verdict nuisance close*, identity-targeted and claim-adaptive allocation share one mixture e-process and one stopping rule: stop when the confidence set is nonempty and every model inside it carries the same verdict. Identity-targeted allocation does not require the set to shrink to a single model. On the evaluated routing trials with $k = 8$, the confidence set contained more than one model at stopping on $24/100$ trials under identity-targeted allocation and on $96/100$ trials under claim-adaptive allocation (set sizes $1, 2, 3, 4$

Table 1: Headline routing-reachability cell at model count k=8 under same-verdict nuisance close (n=100 trials per strategy; caps/failures retained in the denominator; shared $\delta = 0.05$ mixture e-process). All three rows use the same stopping rule—a nonempty, verdict-homogeneous confidence set—and differ only in the allocation. Means are trajectory counts to that shared stop; wrong rates use all trials as denominator. These empirical counts do not prove the risk theorem.

Strategy	Mean trajectories	Wrong verdicts
Identity-targeted	18530.78	$1/100$
Uniform allocation	656.26	$5/100$
Claim-adaptive	324.55	$5/100$

occurred on $76, 22, 1, 1$ trials, respectively, for identity-targeted allocation). Neither arm is a unique-model stopping rule. The strategies differ in the plug-in design of the maximum-likelihood model. The identity design separates every other model; the claim design separates only models whose verdict disagrees with that plug-in. Under this shared stop, mean trajectory counts are 18530.78 versus 324.55 (about $57\times$), with wrong-verdict rates $1/100$ and $5/100$ (Table 1). That factor is the cost of aiming samples at irrelevant model differences, not a change of stopping target and not a comparison with uniform allocation. On the same library and e-process, round-robin uniform allocation averages 656.26 trajectories ($5/100$ wrong), only about $2\times$ claim-adaptive allocation. The larger-library contrasts also vary nuisance separation, not model count alone; their full counts and censoring analysis are in Appendix B.4. Removing observationally identical alternatives does not eliminate separation: at $k = 32$ same-verdict, quotient stops on $4/100$ trials versus claim on $100/100$, with one and four errors; difficulty-axis, all-truth, and congestion controls are reported in Appendix B.6.

To isolate adaptation from a weak allocation baseline, we additionally run 2,400 audits in the nine-model intervention study, covering all eight observational classes. A fixed maximin design optimizes separation across all opposite-verdict pairs without knowing the true model; every strategy uses the same stopping rule. On the two safe truths, its mean cost is only 1.03–1.10 times the adaptive cost, with paired difference intervals including zero. On the six unsafe truths, the factor is 1.45–2.30. Thus the benefit of adaptation is more pronounced for finding violations than for certifying safety in this library (Appendix B.3).

5.4 Learned policy–flow checkpoints

We use MFGLib 0.2.0 Conservative Treasure Hunting ($H = 5$, three states, three actions). A 4,742-parameter network maps normalized time–state inputs through a shared two-layer tanh trunk to separate policy and declared-flow heads. Five random initializations optimize the differentiable public gap, a dense local Bellman-regret term, and 500 times the mean-squared McKean–Vlasov residual, without equilibrium labels, and are saved at steps $0, 25, 50, 75, 100, 200, 400$. The original MFGLib dynamic program and forward operator evaluate the detached outputs after training. These networks are controlled candidate generators trained against public MFGLib operators; they are not a claim of a native neural MFGLib solver or of arbitrary neural-MFG verification.

We audit each final joint candidate, both single-axis ablations, and a final-policy/initial-flow control under three independent audit seeds. Each cell uses 10,000 paired response and 20,000 flow trajectories. Because this public finite model supplies an exact dynamic-programming best-response probe, directional empirical-Bernstein bounds support response decisions; simultaneous Hoeffding bands handle flow. These are known-model simulator audits: even the flow sampler uses the exact implemented mean field. The symmetric response endpoints do not claim simultaneous two-sided coverage; the fixed-truth directional risk argument is in Appendix B.9. Risk 0.05 is divided across both channels and all 60 cells. The joint cells return 6/3/6 certify/refute/abstain: the invalid seed refutes, two valid seeds certify, and two valid seeds abstain. All 45 ablation/control cells refute, and no decisive state is wrong. The exact valid-seed count is 2/5 at threshold 0.03 and 4/5 throughout 0.05–0.10.

We then repeat the learned audit on independent MFGLib Crowd Motion with $H = 4$, a 9×9 spatial grid (81 states), and five actions. A shared 17,798-parameter spatial network is evaluated at 405 time–location inputs and emits policy and state-population-flow heads. Five fixed seeds train

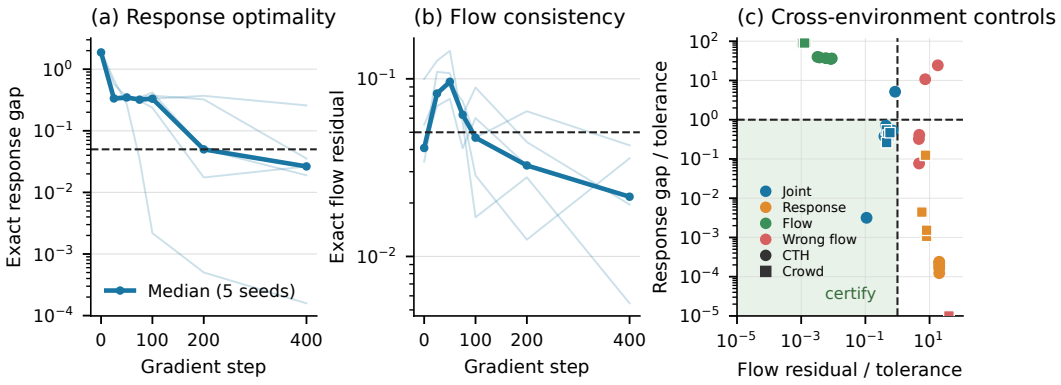


Figure 4: **A learned neural claim requires both axes across environments.** Panels (a–b) show five Treasure Hunting seeds; dashed lines mark its 0.05 tolerances. At step 400, joint training exactly certifies four seeds and refutes one. In (c), every final gap/residual is divided by its environment's declared tolerance; circles are Treasure Hunting and squares are 9×9 Crowd Motion. Color denotes joint, response-only, flow-only, or wrong-flow candidates, and the shaded square is the conjunctive certify set.

for 1,200 steps without equilibrium labels. All five detached outputs satisfy exact response/flow tolerances $0.05/0.02$; all 15 audit cells certify, while all 45 response-only, flow-only, and exact-response/wrong-flow cells refute. Each cell uses 100,000 response and 200,000 flow trajectories, with risk 0.05 divided over both channels and all 60 cells; no decisive state is wrong. Tightening the flow tolerance to 0.01 retains 3/5 exact certifications, and tightening response to 0.03 retains 5/5 at flow tolerance 0.02. As an independent solver-family check, native Fictitious Play, OMD, and Prior Descent add 360 separately calibrated cells: 207 refutations and 153 abstentions, with zero wrong decisions and no full-equilibrium certifications. Their checkpoint flows are induced by construction. Each study has its own 0.05 family risk, not a joint 0.05 guarantee over all studies.

5.5 Structured semantic backends

Table 2: Numerical structured-property experiments. Uppers concern worst-case response/flow over the domain. C/R/A denote certify/refute/abstain. Robust labels assume sound enclosures; the validated-proof gate abstains without formal arithmetic evidence. No unknown-model certificate is implied.

Backend	Stress or baseline	Numerical result	Robust	Validated
Factorable	rank-2 nonlinear	response upper 0.027, flow 0.020	C	A
Continuous	$17/9 \to 33/17$ grid	flow upper $0.0945 \to 0.0473$	A→C	A
History POMDP	memoryless gap 0	full-history witness 0.070	R	A
Global pricing	4^{30} policies	response upper 0.0398, flow 0.0111	C	A

The LP backends separate numerical, unknown-model, and validated decisions. The mixed box is numerically refuted but abstains for the unknown model, and every validated output abstains. Because SciPy/HiGHS lacks checked residuals and outward rounding, these runs do not establish Proposition 4.5's full contract. Continuous bounds assume declared smoothness and Lipschitz moduli; history-flow bounds assume action-independent transitions (Appendix C).

Limitations and conclusion. MF-CAPA audits a supplied declaration in a finite realizable model; it does not learn an equilibrium. The evidence gives no unknown-dynamics, universal, high-dimensional, formal-arithmetic, or finite-population guarantee. Certification requires opposite-verdict separation, anytime-valid stopping, and joint response/flow semantics; otherwise abstention is correct.

AI USE STATEMENT

Generative AI tools assisted with theoretical development, implementation, and experiment design, execution, and analysis. The authors take responsibility for the final content.

REFERENCES

[1] Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Learning mean-field games with discounted and average costs. *Journal of Machine Learning Research*, 24(17):1–59, 2023.

[2] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

[3] René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I–II*. Springer, 2018.

[4] Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959. doi: 10.1214/aoms/1177706205.

[5] Kai Cui and Heinz Koeppl. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 1909–1917. PMLR, 2021.

[6] Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. *arXiv preprint arXiv:1902.03475*, 2019. URL `https://arxiv.org/abs/1902.03475`.

[7] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory*, pp. 998–1027. PMLR, 2016.

[8] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[9] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[10] Xin Guo, Anran Hu, Matteo Santamaria, Mahan Tajrobehkar, and Junzi Zhang. MFGLib: A library for mean-field games. *arXiv preprint arXiv:2304.08630*, 2023.

[11] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

[12] Jiawei Huang, Niao He, and Andreas Krause. Model-based RL for mean-field games is not statistically harder than single-agent RL. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 19816–19870. PMLR, 2024.

[13] Jiawei Huang, Batuhan Yardim, and Niao He. On the statistical efficiency of mean-field reinforcement learning with general function approximation. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 289–297. PMLR, 2024.

[14] Minyi Huang, Roland P. Malhamé, and Peter E. Caines. Large population stochastic dynamic games: Closed-loop McKean–Vlasov systems and the Nash certainty equivalence principle. *Communications in Information and Systems*, 6(3):221–252, 2006.

[15] Noboru Isobe, Kenshi Abe, and Kaito Ariu. Last iterate convergence in monotone mean field games. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[16] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. *arXiv preprint arXiv:1702.01135*, 2017.

[17] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007. doi: 10.1007/s11537-007-0657-8.

[18] Mathieu Lauriere, Sarah Perrin, Sertan Girgin, Paul Muller, Ayush Jain, Theophile Cabannes, Georgios Piliouras, Julien Perolat, Romuald Elie, Olivier Pietquin, and Matthieu Geist. Scalable deep reinforcement learning algorithms for mean field games. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 12078–12095. PMLR, 2022.

[19] Lorenzo Magnino, Kai Shao, Zida Wu, Jiacheng Shen, and Mathieu Lauriere. Solving continuous mean field games: Deep reinforcement learning for non-stationary dynamics. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[20] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

[21] Garth P. McCormick. Computability of global solutions to factorable nonconvex programs: Part i—convex underestimating problems. *Mathematical Programming*, 10:147–175, 1976. doi: 10.1007/BF01580665.

[22] Sirin Nitinawarat, George Atia, and Venugopal V. Veeravalli. Controlled sensing for multihypothesis testing. *arXiv preprint arXiv:1205.0858*, 2013. URL https://arxiv.org/abs/1205.0858.

[23] Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987. doi: 10.1287/moor.12.3.441.

[24] Julien Perolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling up mean field games with online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 1028–1037, 2022.

[25] Sarah Perrin, Julien Perolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13199–13213, 2020.

[26] Tim Quatmann, Christian Dehnert, Nils Jansen, Sebastian Junges, and Joost-Pieter Katoen. Parameter synthesis for Markov models: Faster than ever. In *Automated Technology for Verification and Analysis*, pp. 50–67. Springer, 2016.

[27] Giorgia Ramponi, Pavel Kolev, Olivier Pietquin, Niao He, Mathieu Lauriere, and Matthieu Geist. On imitation in mean-field games. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[28] Naci Saldi, Serdar Yüksel, and Tamás Linder. Finite model approximations for partially observed Markov decision processes with discounted cost. *IEEE Transactions on Automatic Control*, 63(3):726–741, 2018.

[29] Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29909–29921, 2021.

[30] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[31] Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. Policy mirror ascent for efficient and independent learning in mean field games. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 39722–39754. PMLR, 2023.

[32] Muhammad Aneeq Uz Zaman, Alec Koppel, Sujay Bhatt, and Tamer Basar. Oracle-free reinforcement learning in mean-field games along a single sample path. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 10178–10206. PMLR, 2023.

[33] Fengzhuo Zhang, Vincent Y. F. Tan, Zhaoran Wang, and Zhuoran Yang. Learning regularized monotone graphon mean-field games. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[34] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[35] Yichi Zhou, Jialian Li, and Jun Zhu. Identify the Nash equilibrium in static games with random payoffs. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 4160–4169. PMLR, 2017.

A PROOFS

A.1 OBSERVATION MODEL AND ASSUMPTIONS

Fix a separated finite model class $\mathcal{M} = \mathcal{M}_0 \cup \mathcal{M}_1$ and a finite experiment library $\mathcal{B}$. Every experiment b has trajectory density p_M^b with respect to a common measure. We assume common support and a finite second moment for every log likelihood ratio. Let $M \sim M'$ exactly when $P_M^b = P_{M'}^b$ for every b. For the constructive rate, assume each equivalence class is verdict homogeneous and there is at least one opposite-verdict model. The obstruction and lower bound also apply without verdict homogeneity. Each new episode is conditionally sampled from $P_M^{B_t}$ given the past. Equivalently, one may expose the next unused sample from an independent iid stream for each experiment. Model and experiment sets are fixed, and the true model is in $\mathcal{M}$.

Write $v(M) = 0$ on $\mathcal{M}_0$ and $v(M) = 1$ on $\mathcal{M}_1$. The *decisive* time is the first return of $v \in \{0, 1\}$, with $\tau = \infty$ if there is no decisive return. Three-state correctness alone does not imply termination: immediate ABSTAIN has zero decision error. Theorem 4.3 additionally requires $\mathbb{P}_M(\tau < \infty) = 1$ on every separated model, and finite expectation at the model where the bound is used. An external episode budget B produces ABSTAIN when no verdict occurs by B; $\min\{\tau, B\}$ is not the decisive time in that theorem. We use $\mathrm{KL}(P\|Q) = \mathbb{E}_P \log(dP/dQ)$ and $\mathrm{kl}(p, q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))$, with the usual endpoint conventions. Characteristic times use $1/0 = \infty$.

A.2 ON-POLICY OBSTRUCTION

Proof of Proposition 4.1. Use a one-state, one-stage, population-independent MFG, which is a valid MFG subclass. The candidate always chooses action zero and the declared flow is the resulting degenerate flow. In both models, action zero has the same reward law and transition. In M_0, action one is no better than action zero; in M_1, it improves expected reward by $\epsilon_1 + \gamma$. The complete on-policy trajectory law is identical at every sample size, but the response verdict differs. The entire adaptive passive transcript, including any independent auditor randomization, has the same law under both models. Let p_C and p_R be its probabilities of certification and refutation. Correctness at M_1 gives $p_C \leq \delta$; correctness at M_0 gives $p_R \leq \delta$. Thus $\mathbb{P}(\text{decisive}) = p_C + p_R \leq 2\delta$. This allows always abstaining but excludes almost-sure decisive stopping when $\delta < 1/2$. An experiment that plays action one separates the models. □

A.3 AUDIT VERSUS MODEL IDENTIFICATION

Proof of Proposition 4.2. Verdict homogeneity gives $\mathcal{A}(M, c) \subseteq \{M' : M' \not\sim M\}$. The inequality follows for each allocation and then after taking suprema. For strict separation, use a one-state, one-stage population-independent MFG with three actions. The candidate action pays $1/2$ deterministically; the other two actions pay Bernoulli rewards. Their means in models (M_0, M_η, M_1) are respectively

$$(1/4, 1/10), \qquad (1/4, 1/10 + \eta), \qquad (3/4, 1/10), \qquad 0 < \eta < 1/10.$$

All flows are the same degenerate distribution. With $0 < \epsilon_1 < 1/4$ and $0 < \gamma \leq 1/4 - \epsilon_1$, the first two models are safe and the last is separated unsafe. The experiment library contains the three executable actions. At M_0, allocating entirely to the second action gives $D_c^* = \mathrm{kl}(1/4, 3/4) >$

0. Distinguishing M_η additionally requires the third action, so $D_I^* \le \mathrm{kl}(1/10, 1/10+\eta) \to 0$. Therefore T_I^*/T_c^* is unbounded. This is a concrete MFG reduction of the general distinction between answer and model identification, not a separate general testing principle. □

A.4 Sequential change of measure

Let $N_b(\tau) = \sum_{t \le \tau} \mathbf{1}\{B_t = b\}$.

Lemma A.1 (Stopped change of measure)**.** *For an integrable stopping time τ and event $A \in \mathcal{F}_\tau$,*

$$\sum_b \mathbb{E}_M[N_b(\tau)]\,\mathrm{KL}(P_M^b \| P_{M'}^b) \ge \mathrm{kl}(\mathbb{P}_M(A), \mathbb{P}_{M'}(A)).$$

Proof. Predictability of B_t makes the stopped expected log likelihood ratio equal to the left side by conditional expectation and Fubini. The moment and integrability assumptions justify optional stopping of the accumulated increments. Data processing of relative entropy through the map from the stopped history to $\mathbf{1}_A$ gives the binary relative entropy on the right. □

Proof of Theorem 4.3. Let A be the event that the auditor returns the verdict of M. For every $M' \in \mathcal{A}(M, c)$, decisive almost-sure termination and correctness give $\mathbb{P}_M(A) \ge 1-\delta$, whereas correctness gives $\mathbb{P}_{M'}(A) \le \delta$. Since $\delta < 1/2$, monotonicity of binary relative entropy yields a lower bound $\mathrm{kl}(1-\delta, \delta) > 0$. Apply the lemma. Zero expected time would make its left side zero and is impossible. If $0 < \mathbb{E}_M\tau < \infty$, define $w_b = \mathbb{E}_M N_b(\tau)/\mathbb{E}_M\tau$. Then $w \in \Delta(\mathcal{B})$. Minimizing over M' and upper-bounding the resulting minimum by the supremum in Equation (5) yields Equation (6). If $D_c^* = 0$, the same inequality contradicts finite $\mathbb{E}_M\tau$ for $\delta < 1/2$. □

For comparison, let an integrable terminal time T allow abstention. Write $Z_T = \perp$ for the terminal abstention decision and $a_M = \mathbb{P}_M(Z_T = \perp)$. Then only $\mathbb{P}_M(A) \ge 1 - \delta - a_M$ follows. If $a_M < 1 - 2\delta$, the same proof gives

$$\mathbb{E}_M T \ge \frac{\mathrm{kl}(1-\delta-a_M, \delta)}{D_c^*(M, c)}.$$

Without this separation of event probabilities the positive numerator is not justified; in particular, $a_M = 1$ gives no positive complexity bound.

A.5 Anytime validity and allocation tracking

Lemma A.2 (Adaptive likelihood-ratio martingale)**.** *For fixed $M' \ne M$, $L_t(M')/L_t(M)$ is a unit-mean nonnegative martingale under M for every predictable experiment sequence. Hence $E_t(M)$ in Equation* (7) *is also a unit-mean nonnegative martingale.*

Proof. Conditionally on $\mathcal{F}_{t-1}$ and $B_t = b$,

$$\mathbb{E}_M\left[\frac{p_{M'}^b(Y_t)}{p_M^b(Y_t)}\middle|\mathcal{F}_{t-1}\right] = \int \frac{p_{M'}^b(y)}{p_M^b(y)} p_M^b(y)\,dy = 1.$$

Finite positive mixtures preserve the martingale property and initial value. □

Lemma A.3 (Quotient consistency and tracking)**.** *Under the forced-exploration rule in Algorithm 1, an MLE eventually belongs to the true equivalence class almost surely. If every label in one class uses a shared optimizer w^* of Equation* (5)*, target-deficit tracking satisfies $N_b(t)/t \to w_b^*$ for every b.*

Here $N_b(t)$ counts completed episodes of experiment b after t episodes, and $w(\widehat{M}_t)$ is a fixed choice of optimizer, shared by all labels in the estimated observational class. Ties in estimation and sampling use fixed rules. The deficit is $tw_b(\widehat{M}_t) - N_b(t)$; the allocation does not use δ. Extra constraints beyond mixtures of the library would need a separate feasible-allocation and tracking argument.

Proof. Forced exploration makes every $N_b(t)$ diverge. For any $M' \not\sim M$, some experiment has positive KL divergence, and the strong law makes its accumulated log likelihood favor M without bound. A finite union eliminates all labels outside $[M]$. Labels inside $[M]$ have identical likelihoods and need not be separated. Once the target is fixed, the usual deficit-tracking argument bounds cumulative target-count discrepancies by a sublinear forced exploration term plus a constant, which proves convergence of proportions. □

Proof of Theorem 4.4. Ville's inequality applied to $E_t(M)$ gives $\mathbb{P}_M\{\sup_t E_t(M) \geq 1/\delta\} \leq \delta$. On the complementary event, the true model remains in every confidence set, so a nonempty homogeneous set cannot have the wrong verdict.

For the rate, quotient consistency and tracking imply, for each opposite M',

$$\frac{1}{t} \log \frac{L_t(M)}{L_t(M')} \longrightarrow \sum_b w_b^* \, \mathrm{KL}(P_M^b \| P_{M'}^b) \quad \text{a.s.}$$

Use the same infinite collection of experiment streams and the same δ-independent allocation rule for every δ, continuing the rule past any counterfactual stopping time. Fix a sample path on which the stated likelihood limits hold. For each opposite M', the M term in $E_t(M')$ is a positive constant times $L_t(M)/L_t(M')$. For any $0 < \eta < D_c^*$, all opposite labels are therefore excluded for all sufficiently large

$$t \geq \frac{\log(1/\delta) + O(1)}{D_c^* - \eta}.$$

Nonemptiness also matters. For a maximum-likelihood label $\widehat{M}_t$, every ratio $L_t(M')/L_t(\widehat{M}_t) \leq 1$, so $E_t(\widehat{M}_t) \leq 1 < 1/\delta$ for $0 < \delta < 1$. Thus the confidence set is nonempty at every time, even on a coverage-failure path. Exclusion of all opposite labels therefore implies decisive stopping, which is almost surely finite for each fixed risk under the rate assumptions. Letting $\eta \downarrow 0$ on the common stream proves Equation (8). This pathwise argument does not bound $\mathbb{E}_M \tau_\delta$ or establish uniform integrability of normalized stopping times. It is not a matching expected-time upper theorem. □

A.6 Oracle substitution

Proof of Proposition 4.5. On the event that the true model lies in the nonempty confidence set, its exact response and flow values lie in the simultaneous oracle intervals. Joint upper bounds below both tolerances therefore imply $\rho_M(c) \leq 0$. A valid lower bound at least either tolerance plus the margin implies $\rho_M(c) \geq \gamma$. Every remaining interval configuration abstains. Thus the deterministic rule is correct whenever coverage holds, and its error probability is no larger than the confidence-set noncoverage probability. □

More generally, a partition $S = \bigcup_j S_j$ may use different violating axes in different cells. Refutation is sound only if every nonempty S_j has a valid response or flow lower bound beyond its margin. Equivalently, one needs a certified lower bound on $\inf_{M \in S} \rho_M(c)$ of at least γ. Taking the largest feasible response value in S instead bounds $\sup_{M \in S} G_M(c)$ from below. That witness can disprove $\forall M \in S : \rho_M(c) \leq 0$, but leaves the unknown-model verdict unresolved when S contains both safe and unsafe models. Time-uniform statistical composition additionally requires that the true model remain in S_t simultaneously over all decision times, as in Equation (7).

B Experimental Protocols and Additional Results

B.1 Exact finite MFG controls

The controlled finite MFG uses an entropy-regularized response convention. The exact dynamic program evaluates all candidate response gaps and computes the maximum-coordinate discrepancy between declared and implemented flows. The paper-facing rows are a subset of the frozen exact-audit records. The complete table is shown in Table 4.

Table 3: Exact controlled audit. Returns are evaluated against each row's declared flow L. Within the restricted self-consistent stationary search, return optimization beats the MFE's return; this is not a ranking of all rows. A conditional best response can still declare an impossible flow.

Candidate	Return at L	Response gap	Flow residual	Verdict
True equilibrium	2.400	0.0000	0.0000	CERTIFY
Return-optimized candidate	2.463	0.1604	0.0000	REFUTE
Best response with inconsistent flow	5.968	0.0000	0.5093	REFUTE
Surrogate equilibrium	1.490	0.4127	0.1487	REFUTE

Table 4: Complete exact controlled-candidate audit. Return is evaluated against the row's declared flow, which need not be implementable.

Candidate	Return at L	Declared gap	Implemented gap	Flow residual	Status
true MFE	2.400	0.0000	0.0000	0.0000	CERTIFY
return optimized	2.463	0.1604	0.1604	0.0000	REFUTE
return greedy	2.308	0.2297	0.2297	0.0000	REFUTE
best response, wrong flow	5.968	0.0000	0.1964	0.5093	REFUTE
surrogate MFE	1.490	0.4127	0.1573	0.1487	REFUTE
uniform	2.200	1.2084	1.2084	0.0000	REFUTE

B.2 TRAJECTORY-FEASIBLE AUDITABILITY

The two-stage game has an initial safe-commitment action and two routing actions. The candidate takes safe, yielding a terminal Bernoulli(0.55) payoff independent of a binary terminal state with model-specific mass f. Its declaration puts mass 0.5 on each safe terminal state and zero on the routing states. With terminal total variation as the flow metric, $C_M(c) = |f - 0.5|$. The alternative initial actions reach routing states L/R with probabilities $(r, 1-r)$ or $(1-r, r)$, then choose one of two terminal actions with Bernoulli means $p_{L0}, p_{L1}, p_{R0}, p_{R1}$. There is no terminal safe action or additional reward. Rewards and transitions are population independent, a degenerate finite MFG rather than a test of nontrivial population coupling.

The response class includes safe and all eight deterministic routing policies; randomization cannot improve their maximum. Thus, writing $p_L^* = \max_a p_{La}$ and $p_R^* = \max_a p_{Ra}$,

$$G_M(c) = \max\{0,\ rp_L^* + (1-r)p_R^* - 0.55,\ (1-r)p_L^* + rp_R^* - 0.55\}.$$

The ten audit experiments comprise those eight routes, a nuisance sensor, and the actual candidate rollout, which jointly observes state and payoff. The sensor is auxiliary evidence, not a strategic action. Passive sampling uses only the candidate rollout, with probability one. It does not sample a routing policy in place of the candidate.

The nine labels consist of two identical baseline models with all terminal means, f, and nuisance mean equal to 0.5; a baseline nuisance twin with sensor mean 0.505; four models with one routed mean changed to 0.68; a flow alternative with $f = 0.68$; and a nuisance twin of the third routed alternative. For the baseline candidate $G = C = 0$. At $r = 0.8$ the four hidden-deviation models and the routed nuisance twin have $G = 0.094$, $C = 0$; the flow alternative has $G = 0$, $C = 0.18$. The finite family is separated from the 0.05 thresholds. Stable label names are not used as truth labels when reachability changes.

Active forced exploration and target-count tracking, uniform sampling of the ten experiments, and actual passive simulation share the likelihood rule Equation (7). Each truth/strategy condition uses 200 independent seeds, $\delta = 0.05$, and the original 100,000-episode budget at $r = 0.8$. The 1,200 trials sample 39,964,087 episodes, including 39,600,274 passive episodes. Passive batching evaluates every sequential prefix and stops at the first decisive one; unused prefetched randomness is not counted as data.

All four passive stops excluded the true model and retained only the flow-violating alternative. The certified-truth stops were all wrong; the hidden-deviation truth's one stop was verdict-correct despite retaining the wrong model. The corresponding non-stop counts were 197/200 and 199/200. Passive conditional stopping medians therefore describe rare coverage failures, not efficient discovery of

Table 5: Actual sequential trajectory audits. Wrong counts use all 200 trials as denominator; brackets are exact two-sided CP95 intervals for that unconditional error probability. Median/P90 condition on stopping.

Truth	Strategy	Stopped	Wrong / 200 (CP95)	Median	P90
Certified	active	200/200	9 [0.0208, 0.0837]	537.5	816.6
	uniform	200/200	8 [0.0174, 0.0773]	629.0	1162.0
	passive	3/200	3 [0.0031, 0.0432]	54.0	118.8
Refuted	active	200/200	5 [0.0082, 0.0574]	121.5	361.3
	uniform	200/200	1 [0.0001, 0.0275]	316.0	658.8
	passive	1/200	0 [0.0000, 0.0183]	49.0	49.0

the hidden deviation. Active/uniform wrong-decision rates are $0.045/0.040$ for certified truth and $0.025/0.005$ for refuted truth. Both passive CP95 intervals and the uniform-refutation interval exclude 0.05. This is compatible with an upper-risk guarantee, which does not assert equality to δ; Monte Carlo counts do not independently prove anytime correctness.

The certified model's active/uniform trajectory information is $0.01388024/0.00694012$, and the refuted model's is $0.05302218/0.01325554$. Passive information is zero for both default truths. For the hidden deviation at $r = 0.8$, primitive-query information is 0.06627772, so the trajectory/primitive ratio is 0.8 and the cost multiplier is 1.25. These information values are per episode, not machine runtimes.

Table 6: Reachability changes both information and claim truth. The information ratio is trajectory/primitive; the cost ratio is its reciprocal.

r	Best response	Gap	Truth	Information ratio	Cost ratio
0.55	0.599	0.049	CERTIFY	1.00	1.0000
0.65	0.617	0.067	REFUTE	0.65	1.5385
0.80	0.644	0.094	REFUTE	0.80	1.2500
0.95	0.671	0.121	REFUTE	0.95	1.0526

Here the candidate return stays 0.55. The response boundary is $r = 5/9$, safe at equality. At $r = 0.55$, all eight non-flow models are safe and the only opposite model violates flow. Its passive, trajectory, and primitive information all equal 0.06940120. Thus this point is a different decision problem, not evidence of $1/r$ scaling at fixed truth. On the unsafe side $r > 5/9$, the hidden-response comparison has trajectory information equal to r times primitive information. The complete candidate laws, label-specific semantics, designs, trials, and summaries are retained in the supplementary artifact.

B.3 FIXED ALLOCATION UNDER THE SAME STOPPING RULE

We compare claim-adaptive allocation, round-robin uniform allocation, and a truth-independent fixed design on the original nine-model, ten-experiment library at reach 0.8. The fixed design solves

$$\max_{w \in \Delta(\mathcal{B})} \min_{M, M' : v(M) \neq v(M')} \sum_b w_b \, \mathrm{KL}(P_M^b \| P_{M'}^b).$$

Its weights do not depend on observations. It uses the same forced-exploration and target-count tracking schedule as the adaptive arm, and all arms share the mixture confidence set and nonempty, verdict-homogeneous stopping rule. One representative of each observational class gives eight truths, each with 100 fresh seeds per strategy, risk 0.05, and a 10,000-episode cap. Seeds are paired across strategies within truth; different allocations still observe different trajectories. All 2,400 trials stop before the cap, using 958,327 episodes in total.

For the two safe truths, adaptive-minus-fixed mean differences have paired 2,000-resample bootstrap 95% intervals $[-76.94, 45.82]$ and $[-121.12, 1.42]$ episodes. All six unsafe-truth intervals are below zero; the flow case gives $[-164.43, -62.46]$. These are pointwise intervals, not simultaneous claims across truths. Wrong counts range from zero to four per 100 trials; pooled counts are 9/800, 8/800,

Table 7: Fresh allocation control. Entries are mean episodes, followed by wrong verdicts out of 100 in parentheses; capped trials would remain in each mean. The two exactly aliased baseline labels share one truth row.

Truth	Adaptive	Uniform	Fixed maximin
Safe baseline	539.36 (4)	757.41 (3)	556.28 (4)
Safe nuisance twin	558.24 (3)	721.52 (4)	616.49 (0)
Response 1	195.34 (0)	392.26 (0)	385.17 (0)
Response 2	170.02 (0)	390.74 (0)	391.08 (0)
Response 3	173.05 (0)	306.63 (0)	325.25 (0)
Response 4	174.04 (0)	358.97 (0)	317.61 (0)
Flow violation	251.98 (1)	749.20 (1)	364.45 (0)
Response 3 nuisance twin	182.31 (1)	362.02 (0)	343.85 (1)

and 5/800 for adaptive, uniform, and fixed allocation. Risk 0.05 is per trial, not family-wise over this experiment. A cell with four errors has one-sided CP95 upper bound 0.0892, so these counts do not empirically establish error below 0.05. The study tests allocation in the existing population-independent game, not learned unknown dynamics or the larger-library scaling experiment.

B.4 Larger-library allocation contrasts

Three other scored same-verdict cells show the same split—identity-aimed allocation much more expensive than claim-aimed allocation, with only a modest uniform-versus-adaptive gap—whereas under *different-verdict models close* the two plug-in designs are near parity (routing $k{=}8$: 1372.08 vs. 1759.99), as expected when opposite-verdict alternatives and all other models are nearly the same separation task. Four *boundary indistinguishable verdicts* cells leave the identity design unscored because a unique model label is undefined. The $k{=}16$ and $k{=}32$ runs keep the same stop and the same two plug-in objectives, but they are not a pure model-count contrast: the nearest same-verdict nuisance parameter moves from 0.52 at $k{=}4$ to 0.515, 0.505, and 0.502 at $k{=}8, 16, 32$, closer to the baseline value 0.5, while further opposite-verdict models are added as well. The $k = 32$ constructions also contain duplicate non-true observation laws: the routing libraries have 31 and 28 distinct classes in the same- and different-verdict regimes, and both congestion libraries have 31. The implemented identity objective compares model indices and falls back to uniform allocation at a plug-in model with an indistinguishable alternative. Thus these larger-library counts compare the implemented allocations, not a fully quotient-reduced identity baseline; the $k = 8$ headline has no such duplicates. On the routing same-verdict cell the means are 507.48 vs. 186014.69 at $k{=}16$ (about $367\times$; no trial hit the cap) and 459.42 vs. 1075194.57 at $k{=}32$ (about $2340\times$). Read these ratios as the sample cost of an allocation that keeps separating harder same-verdict nuisances, not as savings from switching the stopping target and not as an effect of model count alone. The first $k{=}32$ identity-aimed mean, 917764, was censored ($19/100$ trials hit $1{,}600{,}000$). The same cell and seed rerun with cap $6{,}400{,}000$ stops all 100 trials (median 893055; 19 exceed the old cap; none hit the new cap; $1/100$ wrong). A 2000-resample bootstrap 95% interval for that uncapped mean is $931{,}999.85$–$1{,}233{,}956.05$. The interval covers the mean trajectory count, not the ratio against 459.42. Uniform allocation remains only modestly above claim-adaptive allocation (706.41 and 720.60). Across $k \in \{4, 8, 16\}$, same-verdict cost ratios track the identify-versus-audit information ratio of the two designs (within about 0.7–1.4); that diagnostic compares allocation objectives, not stopping rules. Different-verdict cells stay near parity at every k. At flow $k{=}32$ the information ratio over-predicts the realized cost ratio, so the diagnostic is not treated as the cost. Wrong-verdict counts stay single-digit per hundred trials on these cells. These empirical counts sit beside observed error rates and do not prove the risk theorem; the large factors do not claim that the adaptive allocator dominates uniform allocation.

B.5 Identity allocation over observational classes

To test whether the duplicate-law fallback explains the allocation contrast, we replace the identity design's alternatives by $\{M' : M' \not\sim \widehat{M}_t\}$, where equivalence requires equality of the trajectory law under every experiment. The confidence process, model-index prior, verdict-homogeneous stopping rule, maximum-likelihood selection, forced exploration and tracking remain unchanged. Only the

allocation constraints change; this is not a comparison of quotient-level priors or stopping rules. At $k = 8$, no duplicate laws occur and the two identity allocations coincide. At $k = 32$, the corrected design no longer treats an observationally identical model as a separation target.

We use both routing regimes, $k \in \{8, 32\}$, four allocations, and 100 fresh trials per cell, with risk 0.05 per trial and a fixed cap of 20,000 trajectories. Random streams are independent across arms; the true model is the baseline of each construction. All 1,600 trials are retained in Table 8. Costs are means of the smaller of stopping time and the cap, including wrong decisions and abstentions; they are not estimates of uncensored mean stopping time.

At $k = 32$ in the same-verdict regime, claim and distinct-law identity allocation stop on $100/100$ and $4/100$ trials. Their pointwise 95% Wilson intervals for the decisive fraction are $[0.963, 1]$ and $[0.016, 0.098]$. Restricted mean costs are 531.23 and 19253.47; 2000-resample percentile-bootstrap intervals are $[484.19, 578.95]$ and $[18472.38, 19827.31]$. In the different-verdict regime, both stop on all 100 trials, with mean costs 1744.84 and 2941.02. Thus excluding identical laws does not remove the finite-budget contrast caused by close same-verdict alternatives in this construction. Uniform allocation also stops on all same-verdict trials, so the contrast does not establish a comparably large benefit over uniform allocation. The two identity arms at $k = 8$ implement the same design on different random streams; their count differences are sampling variation, not an algorithmic effect. The intervals are pointwise, not multiplicity-adjusted guarantees. Observed errors do not prove the per-trial risk bound, and the 96 censored distinct-law identity trials at $k = 32$ preclude an uncensored cost ratio. In particular, this experiment does not validate the historical $2340\times$ factor against the corrected identity baseline.

B.6 Separated difficulty axes

We vary within-verdict nuisance separation and the nearest opposite-verdict coordinate independently in one eight-model routing class, holding the truth, library, risk level, stopping rule, and 20,000-trajectory cap fixed. Each cell contains 4,000 trials over four fixed streams. The first axis leaves the claim-audit information number unchanged while quotient-identity decision rate rises from 6.2% to 100% as the nuisance models separate. On the second axis, claim allocation remains decisive in every trial and its restricted mean cost falls from 293 to 185. Quotient identity names the allocation objective; every arm uses the common verdict stop.

B.7 Robustness across all truths

We next fix the eight-model class and use each distinct model as truth: indices 0–3 are certified and 4–7 are refuted. Each truth–allocation cell contains 4,000 trials. Claim and uniform allocation stop in every trial for every truth. Claim restricted costs range from 94 to 244, uniform costs from 195 to 396, and quotient-identity costs from 622 to 11160 with decision rates from 94.0% to 100%. In all 16 pointwise unpaired claim-minus-comparator cost contrasts, the bootstrap interval is below zero; wrong-decision rates do not have a uniform ordering.

B.8 Congestion-family quotient allocation control

The corrected distinct-law allocation is also evaluated on the $k = 32$ flow-coupled congestion family. All arms retain the same model-index mixture process, verdict-homogeneous stopping rule, risk 0.05, and cap of 20,000 trajectories. Each row contains 100 fresh trials; costs include abstentions at the cap.

The same-verdict cell remains allocation-limited after exact duplicate laws are removed, while all three arms decide every different-verdict-close trial. These are restricted-budget results and do not estimate uncensored stopping times.

B.9 Fixed-budget directional decision risk

Condition on a candidate and its response probe chosen independently of the fresh audit episodes. The learned-checkpoint study uses known-model simulators: the neural response probe is an exact dynamic-programming best response to the declared flow, and the population sampler constructs its time-dependent transitions from the exact implemented mean field. Thus the paired gain mean is

Table 8: Fixed-budget routing control. Each row contains 100 trials. S denotes the same-verdict nuisance regime; D denotes the different-verdict close regime. Stops include wrong verdicts; all remaining trials abstain at the 20,000 cap. Restricted cost averages all trials.

Regime	Allocation	Stops	Wrong	Restricted cost
S, $k = 8$	Claim	100	3	397.02
	Uniform	100	3	664.93
	Identity, model indices	62	4	16408.37
	Identity, distinct laws	54	0	17112.37
D, $k = 8$	Claim	100	3	1430.77
	Uniform	98	3	5897.42
	Identity, model indices	100	4	1718.79
	Identity, distinct laws	100	2	1983.09
S, $k = 32$	Claim	100	4	531.23
	Uniform	100	1	719.20
	Identity, model indices	6	3	18874.93
	Identity, distinct laws	4	1	19253.47
D, $k = 32$	Claim	100	3	1744.84
	Uniform	94	1	8468.36
	Identity, model indices	100	4	2906.40
	Identity, distinct laws	100	5	2941.02

Table 9: Fixed-budget verdict auditing under isolated model-class difficulty interventions. Entries are decision percentage / mean $\min(T, 20000)$. Each row aggregates 4,000 trials over four fixed seeds; the supplement reports Wilson and bootstrap 95% intervals.

Axis	Level	Claim	Uniform	Quotient identity
		Decision % / restricted mean cost		
Within	0.005	100.0 / 245	100.0 / 393	6.2 / 19316
	0.01	100.0 / 242	100.0 / 395	8.2 / 18995
	0.02	100.0 / 242	100.0 / 397	93.5 / 11230
	0.04	100.0 / 252	100.0 / 403	100.0 / 3032
	0.08	100.0 / 256	100.0 / 435	100.0 / 974
Cross	0.655	100.0 / 293	100.0 / 465	92.7 / 11451
	0.66	100.0 / 277	100.0 / 452	93.7 / 11378
	0.68	100.0 / 243	100.0 / 400	94.5 / 11226
	0.72	100.0 / 204	100.0 / 345	94.1 / 11195
	0.76	100.0 / 185	100.0 / 322	94.5 / 11113

Wrong decisions ranged from 1.4% to 2.7% across all cells; exact counts and Wilson intervals are reported in the supplementary CSV. Quotient-identity denotes its allocation target; all arms use the same verdict stopping rule.

$\mu = G_M(c)$ for the neural studies. This is oracle-assisted probe construction, not certification from unknown-dynamics rollouts alone.

Let $X_1, \ldots, X_n$ be iid episode gains in the declared interval $[a, b]$, $n \geq 2$, with sample mean $\bar{X}$ and unbiased sample variance S_n^2. For a response-channel risk α, set

$$r_\alpha = \sqrt{\frac{2S_n^2 \log(2/\alpha)}{n}} + \frac{7(b-a)\log(2/\alpha)}{3(n-1)}, \qquad L_\alpha = \bar{X} - r_\alpha, \quad U_\alpha = \bar{X} + r_\alpha.$$

Theorem 4 of Maurer & Pontil (20), applied after rescaling X and separately to $-X$, gives

$$\mathbb{P}(L_\alpha > \mu) \leq \alpha, \qquad \mathbb{P}(U_\alpha < \mu) \leq \alpha.$$

These are two directional statements, not simultaneous two-sided coverage at level $1 - \alpha$. The latter would require splitting the risk between tails, replacing $\log(2/\alpha)$ by $\log(4/\alpha)$ in this construction.

The response rule certifies if $U_\alpha \leq \epsilon_1$ and refutes if $L_\alpha > \epsilon_1$. For a fixed truth $G_M(c) \leq \epsilon_1$, only false refutation is wrong and implies $L_\alpha > G_M(c)$. For $G_M(c) > \epsilon_1$, only false certification is wrong and implies $U_\alpha < G_M(c)$. Hence the response decision error is at most α without a two-tail

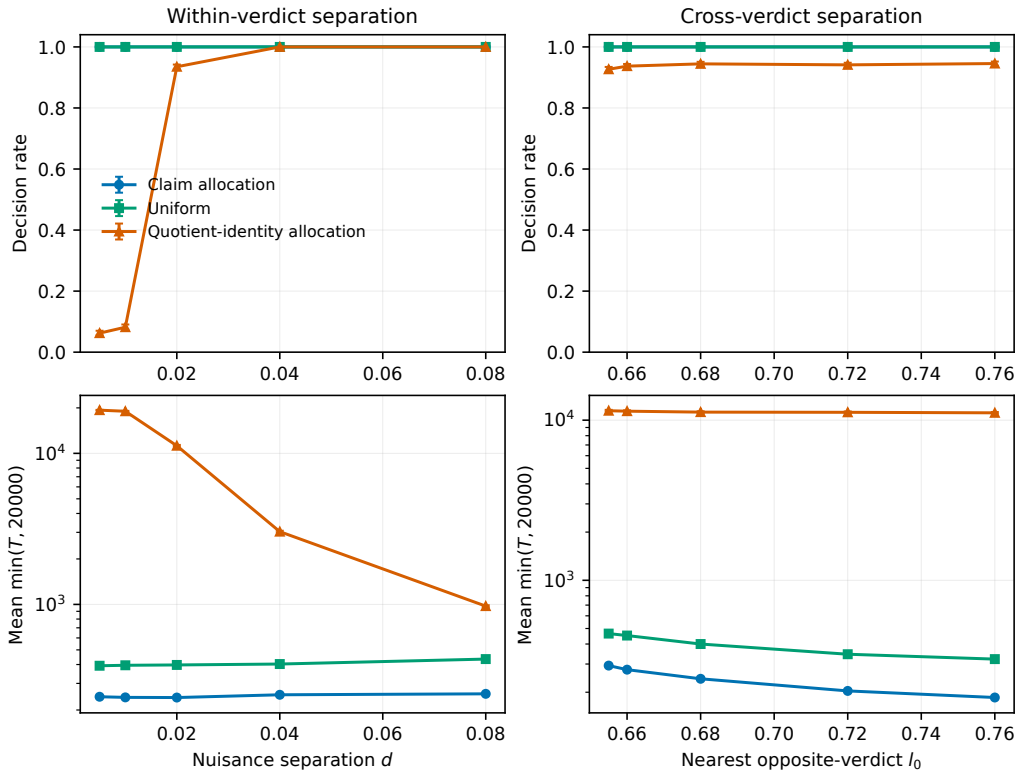


Figure 5: Fixed-budget response to isolated model-class separations. Error bars are pointwise 95% intervals. Costs are means of $\min(T, 20000)$, not uncensored identification times.

Table 10: Fixed-budget results for every truth in the eight-model class. Entries are decision percentage / mean $\min(T, 20000)$ / wrong decisions, from 4,000 trials per cell.

Index	True model	Verdict	Claim	Uniform	Quotient identity
0	truth_cert	CERTIFIED	100.0 / 244 / 114	100.0 / 396 / 76	94.0 / 11160 / 84
1	cert_nuisance_1	CERTIFIED	100.0 / 244 / 91	100.0 / 389 / 90	99.9 / 4429 / 54
2	cert_nuisance_2	CERTIFIED	100.0 / 244 / 96	100.0 / 383 / 77	100.0 / 1269 / 40
3	cert_nuisance_3	CERTIFIED	100.0 / 243 / 95	100.0 / 379 / 89	100.0 / 622 / 36
4	refute_nearest_l0	REFUTED	100.0 / 124 / 23	100.0 / 378 / 41	99.9 / 1871 / 52
5	refute_l1	REFUTED	100.0 / 94 / 33	100.0 / 196 / 15	99.7 / 1783 / 53
6	refute_r0	REFUTED	100.0 / 95 / 33	100.0 / 195 / 15	99.9 / 1425 / 32
7	refute_flow	REFUTED	100.0 / 123 / 62	100.0 / 389 / 61	99.5 / 3351 / 50

All arms use the same verdict stopping rule. Pointwise Wilson and bootstrap intervals, explicit CERTIFIED/REFUTED/correct counts, and unpaired arm-contrast intervals are supplied separately.

union bound: its direction is determined by the fixed truth, not selected from the samples. This study compares exact thresholds without an indifference strip; no sequential termination claim is made for it.

For K flow coordinates and n_F independent episodes, coordinatewise Hoeffding bounds and a union bound give simultaneous radius $h = \sqrt{\log(2K/\alpha)/(2n_F)}$. If $\widehat{C}$ is the maximum-coordinate discrepancy from the declared flow, then $|\widehat{C} - C_M(c)| \le h$ on this event. Correlation between coordinates within an episode does not invalidate the union bound. Likewise, shared randomness within a paired response episode does not invalidate iid sampling of gains across episodes.

For m predetermined candidate/seed cells, allocating $\alpha = 0.05/(2m)$ to each response and flow channel gives family-wise wrong-decision probability at most 0.05 by a union bound, conditionally

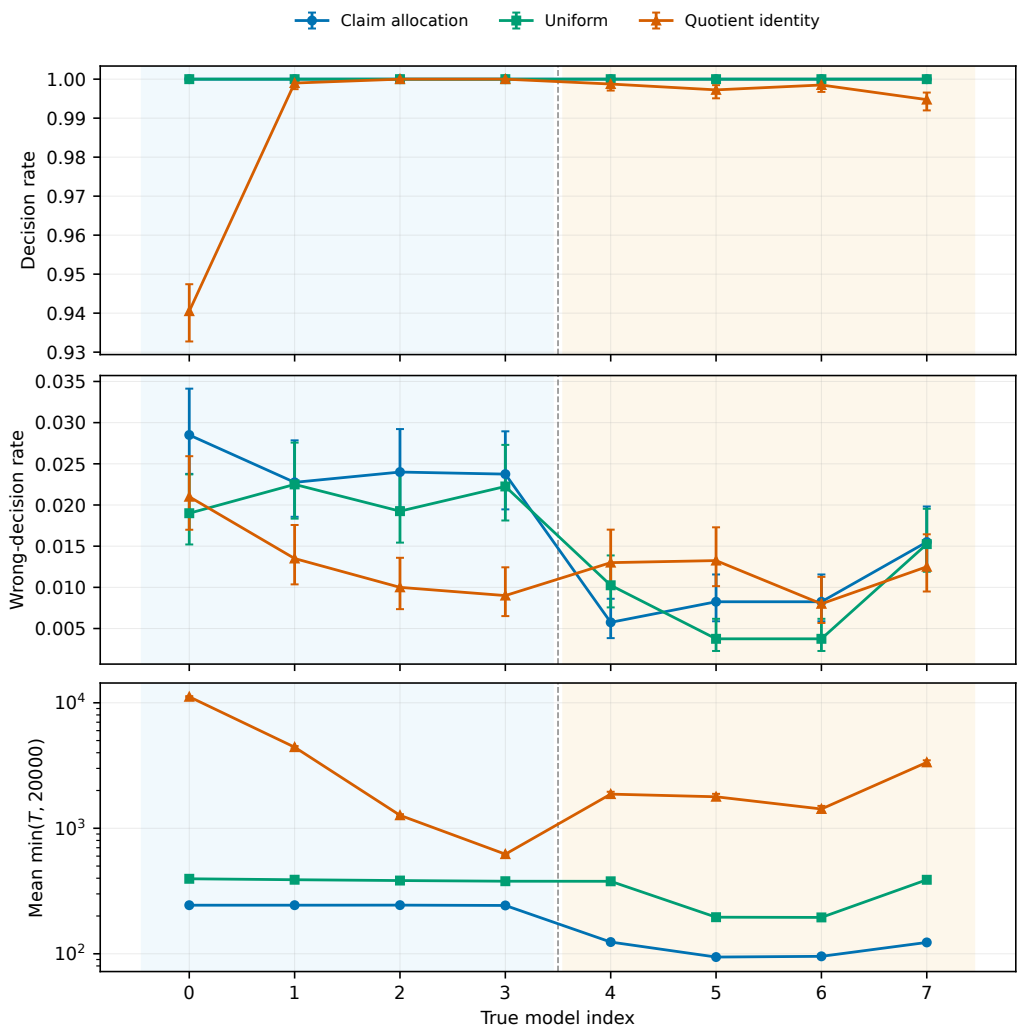


Figure 6: Fixed-budget results by discrete true-model index. Connecting lines only guide the eye and do not represent a continuous parameter trend. Error bars are pointwise 95% intervals, not simultaneous bands.

Table 11: Fixed-budget $k = 32$ congestion control.

Regime	Allocation	Stops	Wrong	Restricted cost
Same verdict	Claim	100	2	98.33
	Uniform	100	3	210.10
	Quotient identity	16	2	17961.77
Different verdict	Claim	100	2	1273.19
	Uniform	100	0	5742.28
	Quotient identity	100	2	2413.06

on the trained candidates and hence also unconditionally. If neither axis has a wrong decisive label, the conjunctive certify/disjunctive refute rule cannot have a wrong overall label. The native-solver response channel uses only L_α for a feasible probe whose mean is at most $G_M(c)$; it supports refutation but no response certification. Each of the three studies is separately calibrated; this does not give a joint 0.05 guarantee over their union. These fixed-budget bounds do not cover adaptive stopping or probe selection on the same audit episodes; the sequential design study instead uses Equation (7).

B.10 Learned neural policy–flow study

The neural candidate has a shared two-layer tanh trunk of width 64 and two heads. The policy head is normalized over actions at each time–state input; the flow head is normalized jointly over state–action coordinates at each time. Its 4,742 parameters are trained from seeds 11–15 for 400 gradient steps with learning rate 0.01. For parameters θ, the joint objective is

$$G_M(\pi_\theta, L_\theta) + 5\,\overline{r}_B(\theta) + 500\,\frac{1}{54}\|L_\theta - \Gamma_M(\pi_\theta)\|_2^2,$$

where $\overline{r}_B$ averages, over all six time steps and three states, the exact local gap between $\max_a Q_t(s,a)$ and the policy-weighted $Q_t(s,a)$. This dense term improves optimization coverage; only the original response gap and maximum-coordinate flow residual determine audit truth. Training uses differentiable implementations of the public Bellman and flow operators, checked elementwise against MFGLib to tolerance 10^{-6}; it does not use equilibrium labels. Exact MFGLib operators evaluate detached saved outputs. Response-only and flow-only objectives remove the corresponding other term. The wrong-flow control combines each final joint policy with its step-zero neural flow.

Audit seeds 201–203 use 10,000 bounded paired response episodes and 20,000 population-flow episodes. Family risk 0.05 is allocated over the response and flow channels and all 60 final-candidate/control cells. The two response endpoints support directional decision-risk control because the model supplies an exact dynamic-programming best-response probe (Appendix B.9); they are not a simultaneous two-sided $1-\alpha$ interval. The flow band is simultaneous over all 54 coordinates.

Table 12: Neural final-checkpoint truth and calibrated outcomes. C/R/A counts are over three fresh audit seeds.

Seed	Exact gap	Flow residual	Exact	Audit C/R/A
11	0.0002	0.0054	C	3/0/0
12	0.0264	0.0357	C	0/0/3
13	0.0354	0.0217	C	0/0/3
14	0.2583	0.0422	R	0/3/0
15	0.0189	0.0196	C	3/0/0

Each audit seed returns exactly 2 certifications, 16 refutations, 2 abstentions, and no wrong decisive state over the 20 candidates. Tightening a common response/flow tolerance from 0.05 to 0.03 changes exact final certification from 4/5 to 2/5; relaxing it to 0.075 or 0.10 leaves 4/5. Thus the mixed learner quality is not a single-seed or infinitesimal-threshold artifact.

B.11 Cross-environment Crowd Motion neural study

The second learned environment is MFGLib 0.2.0 Crowd Motion with horizon four, a 9×9 spatial grid, five actions, and the public environment seed seven. A shared two-layer tanh trunk of width 128 is evaluated on all 405 time–location inputs. Its policy head is normalized over actions and its state-flow head over the 81 locations at each time, for 17,798 parameters. Five fixed seeds train for 1,200 gradient steps at learning rate 0.005. The joint objective is

$$G_M(\pi_\theta, m_\theta) + 5\,\overline{r}_B(\theta) + 4000\,\frac{1}{405}\|m_\theta - \Gamma^s_M(\pi_\theta)\|_2^2.$$

Only the exact response gap and maximum state-flow discrepancy determine truth, with tolerances $\epsilon_1 = 0.05$ and $\epsilon_2 = 0.02$. The differentiable reward, transition, Bellman, and flow operators match MFGLib elementwise to 10^{-6}; training uses no equilibrium labels. Exact MFGLib dynamic programming and forward flow evaluate detached checkpoints at steps $0, 50, 100, 200, 400, 800, 1200$.

Audit seeds 301–303 use 100,000 paired fixed-flow response episodes and 200,000 policy-flow episodes per cell. Candidate-specific reward spans bound the paired gains; a simultaneous Hoeffding band covers all 405 time–location flow coordinates. Family risk 0.05 is divided over both channels and all 60 final-candidate/control cells.

Response-only training leaves state-flow residuals 0.115–0.162 despite gaps below 0.0064; flow-only training leaves gaps 4.501–4.506 despite residuals below 2.8×10^{-5}. Exact responses to

Table 13: Crowd Motion final-checkpoint truth and calibrated outcomes. C/R/A counts are over three fresh audit seeds.

Seed	Exact gap	Flow residual	Exact	Audit C/R/A
21	0.0185	0.0110	C	3/0/0
22	0.0275	0.0090	C	3/0/0
23	0.0202	0.0100	C	3/0/0
24	0.0131	0.0092	C	3/0/0
25	0.0234	0.0117	C	3/0/0

deliberately inconsistent flows have gap zero and residuals 0.801–0.802. All 45 such control cells refute. Every audit seed returns 5 certifications and 15 refutations with no abstention or wrong decisive state. Tightening ϵ_2 from 0.02 to 0.01 retains three exact-certified seeds; tightening ϵ_1 to 0.03 at $\epsilon_2 = 0.02$ retains all five.

B.12 Native MFGLib trained-checkpoint study

The environment is MFGLib 0.2.0 Conservative Treasure Hunting with horizon five and three states/actions. Learners are trained from random normalized policies at seeds 11–15. Checkpoints are fixed before auditing. Audit seeds 101–103 use 2,000 bounded paired response episodes and 10,000 population-flow episodes. Family risk 0.05 is allocated across both channels and all 360 cells. The paired-gain empirical-Bernstein lower bound supports response refutation only. The simultaneous Hoeffding flow band supports either flow decision. All candidate archives and row-level evidence are retained.

Table 14: First exact-certified checkpoint for each training seed.

Learner / seed	11	12	13	14	15
Fictitious Play	10	10	10	10	10
OMD	50	50	50	50	50
Prior Descent	50	50	20	50	50

All 315 valid checkpoint flows are flow-certified. The rollout response channel refutes large early gaps and abstains after deviations become too small; it never converts failure to refute into certification. Each of the 15 final policies is also paired with its seed-specific initialization flow and audited under three seeds. All 45 controls are response- and flow-refuted.

C Structured Semantic Oracles

The structured experiments concern the robust property that every semantic parameter in a declared domain S is safe. Define

$$G^{\max}(S) = \sup_{M \in S} G_M(c), \qquad C^{\max}(S) = \sup_{M \in S} C_M(c).$$

Feasible points give lower witnesses for these suprema; sound outer relaxations would give upper bounds. Such upper bounds within both tolerances also suffice for unknown-model certification if S is a valid confidence set. A violating feasible point, however, refutes only the robust property. It cannot replace a uniform lower bound in Proposition 4.5. These backend experiments do not construct a statistically calibrated confidence set or establish end-to-end unknown-model certification.

Separate decision fields. For compatibility with the released implementation, reported validation fields and verdicts remain numerical diagnostics over robust parameter boxes. A schema flag marks their claim semantics as robust-box diagnostics, and a second flag identifies worst-case suprema. Separate confidence-cell records store pointwise enclosures and surviving domains covering the root box. Uniform response lower bounds use an optimal-value lower bound minus a candidate-value upper bound; uniform flow lowers use distance from the declaration to candidate-mass intervals. These are not sampled maxima. The unknown-model rule requires every surviving cell to pass an

axis-wise condition, possibly on different axes, and never prunes a cell because an LP failed. Its tolerance-level numerical refutation would need an additional margin check to support the positive-γ conclusion of Proposition 4.5.

The conditional outcome is a numerical verdict; the validated verdict and the strict unknown-model verdict are ABSTAIN for every backend-generated result. No generated enclosure is marked as validated. A caller-supplied validation indicator would assert an external enclosure proof, not one produced by this implementation. Unknown, nonfinite, or reversed bounds and empty covers also yield ABSTAIN. The caller must establish both complete coverage of the surviving model set and its statistical validity.

Table 15: LP diagnostics under different claim semantics. Robust and conditional unknown-model columns are numerical only. Every validated verdict and strict unknown-model verdict is ABSTAIN. C/R/A denote certify/refute/abstain.

Case	Cells evaluated	Robust	Conditional unknown	Validated
Factorable safe	1	C	C	A
Factorable mixed	1	R	A	A
Global unsafe interior	5	R	A	A
Global safe lifted RLT	13	C	C	A
Global safe edge McCormick	83	C	C	A
Global stress	1	C	C	A

The same numerical enclosures support different conclusions under these two quantifiers. Neither robustly refuted row is a refutation of the unknown true model.

Numerical arithmetic boundary. The exact-arithmetic enclosure argument assumes valid outer support cuts and valid upper bounds for every maximization. The implementation uses ordinary SciPy/HiGHS floating-point LPs and successful solver objectives, without independently validated dual certificates, residual corrections, or directed outward rounding. Sparse constraint assembly retains all nonzero coefficients, but solver-internal tolerances and dropping remain unvalidated. An underestimated support value can remove feasible semantics before occupancy optimization begins. Sampled containment checks do not establish global containment, and threshold margins alone do not bound numerical errors. Accordingly, all LP-derived values and C/R/A labels below are numerical relaxation outputs under exact-arithmetic soundness assumptions, not formal floating-point certificates. Formal certification would require outward-safe enclosures and checked error budgets at every support and occupancy stage. No margin-completeness claim is made for the implemented floating-point unknown-model procedure.

Factorable neural semantics. A CROWN/LP relaxation represents a neural output. An acyclic graph of affine, multiplication, and ReLU nodes is relaxed with McCormick envelopes (21) and convex-hull ReLU constraints. Under exact arithmetic, valid LP support queries project the graph to an outer semantic polytope for the finite-state occupancy backend. A conditional margin-completeness argument additionally needs compact domains, exhaustive refinement, and enclosure widths converging to zero; solver success alone does not establish these conditions. The rank-two test uses $z_1 z_2$, z_1^2, and $\mathrm{ReLU}(z_1 - z_2)$ in both rewards and transitions. The safe case reports $G^{\max} \in [0.024, 0.027]$ and $C^{\max} \in [0.020, 0.020]$ numerically. The mixed box reports $G^{\max} \in [0.064, 0.070]$, which witnesses failure of robust safety, not true-model refutation. Indeed its latent point $(0, 0)$ has $(G, C) = (0.030, 0)$, while $(1, -1)$ has $(G, C) = (0.064, 0.020)$. At tolerances $(0.05, 0.05)$ both a safe and an unsafe model remain possible. Thus this box must remain unresolved as an unknown-model confidence set; neither $[0.064, 0.070]$ nor $[0.020, 0.020]$ encloses every model's axis value.

Continuous abstraction. This example has compact state and action spaces $[0, 1]$, deterministic candidate actions with Lipschitz constant L_π, and population-independent transitions. Rewards are L_r-Lipschitz and kernels are L_P-Lipschitz in Wasserstein-1 for the state–action sum metric; the terminal value has constant L_T. Cover radii h_x, h_a bound nearest-grid projection errors. The flow metric is $d(L, \Gamma) = \max_t W_1(L_t, \Gamma_t)$, not the maximum-coordinate metric of the exact and learned-checkpoint studies.

For completeness, let $\ell_H = \ell_H^\pi = L_T$, $E_H = E_H^\pi = L_T h_x$, and define backwards

$$\ell_t = L_r + L_P \ell_{t+1}, \qquad \ell_t^\pi = (1 + L_\pi)(L_r + L_P \ell_{t+1}^\pi),$$
$$E_t = E_{t+1} + L_r(h_x + h_a) + \ell_{t+1}\{L_P(h_x + h_a) + h_x\},$$
$$E_t^\pi = E_{t+1}^\pi + L_r h_\pi + \ell_{t+1}^\pi(L_P h_\pi + h_x), \qquad h_\pi = (1 + L_\pi)h_x + h_a.$$

These bound optimal and candidate value errors by Bellman induction: reward perturbation contributes L_r times the joint cover error and kernel transport contributes the next value's Lipschitz constant times its Wasserstein error. With $d_0 = h_x$, coupling candidate flows gives $d_{t+1} = L_P\{(1 + L_\pi)d_t + h_a\} + h_x$. Consequently,

$$G_c \le \overline{G}_g + E_0 + E_0^\pi, \qquad C_c \le \overline{C}_g + \max_t d_t + h_x,$$

where the final h_x transports the declared continuous flow to its grid. Both inflations vanish at fixed horizon under the declared moduli, consistent with finite-model approximation theory (28). The example uses $H = 3$, $(L_r, L_P, L_T, L_\pi) = (0.016, 0.2, 0, 0.6)$. A $17/9$ state/action grid returns nominal abstention with flow upper 0.094504; $33/17$ and $65/33$ grids give nominal certification with flow uppers 0.047252 and 0.023626. These recursions do not cover arbitrary nonsmooth policies or population-dependent kernels without additional moduli and propagation terms; numerical rounding is not formally enclosed.

Complete finite history. A finite POMDP response oracle conditions beliefs on every observation and maximizes over actions at every history. In the six-stage separation, a profitable final action must remember the first signal. Current-observation policies have gap zero, while complete-history dynamic programming visits 2,730 nodes and finds gap 0.07. The represented deterministic policy-tree class has approximately 8.51×10^{40} members. This is complete-history dynamic programming, not policy enumeration. The accompanying flow extraction requires action-independent transitions; it is not a generic controlled-POMDP flow certificate. The response witness concerns this specified known model and response class.

Global neural policy pricing. For rewards and transitions affine in one relaxed neural output, action occupancies represent every randomized Markov response at once. Shared occupancy–semantic products and RLT support cuts preserve one semantic vector through the Bellman/flow computation. A six-stage, five-state, four-action stress case contains $4^{30} = 1.1529 \times 10^{18}$ deterministic Markov policies. One semantic cell gives numerical worst-case response bracket $[0.037207, 0.039764]$ and flow bracket $[0.011071, 0.011071]$, using 61 completed occupancy LP calls without policy enumeration. That counter excludes transition-support solves, failed attempts, and validation work; it is not a total solver-call cost. The lower endpoints are feasible-point witnesses for suprema, not uniform lower enclosures for every model.

Computational boundary. General ReLU-semantic refutation contains ReLU reachability and is NP-hard (16); complete finite-horizon partially observed response pricing contains POMDP planning and is PSPACE-hard (23). On compact continuous domains without a known modulus, a finite-query verifier cannot exclude an unqueried narrow spike. Exact-arithmetic verification therefore depends on explicit structure and convergent outer bounds; the numerical examples do not establish a universal solver, a formal floating-point certificate, or finite-population Nash transfer.

D Reproducibility Statement

The artifact uses Python, NumPy, SciPy, PyTorch, and MFGLib 0.2.0. Result records retain raw rows, metadata, and summaries. The main figures are regenerated from the archived result tables. Separate procedures reproduce the learned neural Treasure Hunting study, the Crowd Motion study, and an independent native trained-checkpoint study. The supplementary artifact documents the regression tests and maps every paper-facing number to its source data and analysis. Its quick verification procedure replays saved policy and flow outputs and audit decisions without retraining; a separate full mode reruns the generators. The recorded replay environment does not reconstruct missing historical training-environment details, and bitwise retraining reproduction is not claimed.