

Institutionalized Predictive Jeopardy: An Evolutionary-Game Demarcation Criterion for Knowledge Systems

Institutionalized Predictive Jeopardy: An Evolutionary-Game Demarcation Criterion for Knowledge Systems

Abstract

Institutional knowledge systems — peer review, tenure committees, funding panels, citation markets — must decide which claims deserve resources, yet they possess no formal criterion distinguishing claims that carry prospective falsification risk from claims fitted to evidence after the fact. Incentive structures that cannot price predictive risk plausibly reward post-hoc certainty, allowing unfalsifiable narrative strategies to accumulate. We formalize demarcation as an evolutionary game in which agents choose between a falsifiable strategy F , which submits prospective predictions that can fail publicly, and a post-hoc strategy P , which accommodates any outcome. We derive the closed-form accuracy threshold $p^* = (s + c)/(b + c)$ above which falsifiable practice is individually rational. Under a baseline calibration ($b = 2$, $c = 1$, $s = 1$, $p = 0.6$), F yields an expected payoff of 0.8 against 1.0 for P , so post-hoc fitting dominates and the threshold is $p^* = 2/3 \approx 0.6667$. A pre-registration-style reform raising the confirmed-prediction reward to $b = 10$ reverses the equilibrium ($\pi_F = 5.6$ versus $\pi_P = 1.0$; $p^* = 2/11 \approx 0.1818$), with the best-response flip occurring at $b \approx 2.3333$. We further derive an antifragility condition under which registered failures raise system reliability by k/N , and a reform timescale of $\ln 9/0.2 \approx 10.99$ generations under diluted incentives. The model supplies a computable demarcation criterion and identifies the institutional levers that move knowledge systems toward falsifiable equilibria.

1. Introduction

Funding agencies, editorial boards, tenure committees, and operators of large-scale AI-assisted research pipelines face the same operational question daily: which of the claims crossing their desks deserve resources, publication, or career reward? The question is expensive to answer wrongly. The research ecosystem allocates large-scale public and private funding, careers, and compute-hours to work whose connection to any human problem is untraceable, and the allocation system rewards signals — prestige cascades, citation counts, novelty claims — that do not measure predictive risk [11]. When the reward structure cannot distinguish a claim that was risked in advance from a claim that was fitted to data after the fact, the rational individual strategy is to fit after the fact, and the system as a whole drifts toward narratives that no outcome could embarrass.

This drift has a name in the critical literature. The Head Over Hands Fallacy describes the active, systemic, cyclical process — spanning individual psychology, institutional incentives, and philosophical predisposition — of defending a cherished theory against the verdict of observed reality, a process distinct from individual confirmation bias because it is institutional rather than merely cognitive [9]. Lakatosian structure sharpens the mechanism: research programmes protect a hard core of axioms from falsification by methodological fiat behind a protective belt of auxiliary hypotheses, and the framework offers no formal path for the comparative assessment of axiom-violating theories, which are expelled as unscientific rather than adjudicated [10]. The consequence is that demarcation — the decision of which knowledge-producing practices deserve the label and the resources of science — is currently made by informal institutional judgment rather than by a computable test.

What would change if a formal criterion existed? A funding agency could score proposals by the prospective falsification risk they accept, rather than by prestige. A journal could require that each submission declare, before review, the outcome that would refute it, and could weight acceptance by the severity of that test. An AI-mediated research pipeline — which can generate untraceable abstraction at arbitrary volume, since AI adoption spans all 333 tracked research fields [11] — could be gated by the same test, because a criterion without a falsifiable structure would be gameable by exactly the systems that most need it [11]. The taxonomy of scientific contributions into Generative, Structural Unification, and Epistemic Scaffolding types identifies the core obstacle: G-Type contributions are validated by direct empirical testing, but S-Type and E-Type work lacks a rigorous validation metric, creating risks of subjective evaluation and disciplinary stagnation [12]. A demarcation criterion priced in predictive risk would supply the missing metric for exactly those contribution types.

This paper builds the formal core of such a criterion. We model knowledge production as an evolutionary game between two strategies: F (falsifiable practice), which submits prospective predictions whose public failure carries a cost, and P (post-hoc practice), which produces narratives that accommodate any outcome. We derive equilibria, intervention effects, an antifragility condition, and a reform timescale, all with explicit arithmetic. The contributions are:

- A two-strategy evolutionary game formalizing institutionalized predictive jeopardy, with the invasion condition for falsifiable practice and the closed-form accuracy threshold $p^* = (s + c)/(b + c)$, derived in Sections 4.1–4.2.
- Comparative statics showing that the threshold falls in the confirmed-prediction reward b and rises in the certain post-hoc payoff s , derived in Section 4.2.
- A numerical demonstration, under stated baseline assumptions, that current-style incentive configurations make post-hoc fitting the dominant strategy ($\pi_P = 1.0$ versus $\pi_F = 0.8$), derived in Section 4.3.
- A quantified intervention analysis showing that pre-registration-style reward restructuring reverses the equilibrium ($\pi_F = 5.6$ versus $\pi_P = 1.0$), with the best-response flip at $b \approx 2.3333$, derived in Section 4.4.
- An antifragility condition — a configuration in which system-level reliability increases by $1/N$ with each registered failed prediction — derived in Section 4.5.
- A replicator-dynamics timescale for incentive reform, giving the transition time $t = \ln 9/0.2 \approx 10.99$ generations under diluted incentives, derived in Section 4.6.

2. Background and Related Work

The demarcation problem has a long philosophical pedigree, but the works most relevant here are those that treat evaluation, prediction, and institutional incentive as formal or semi-formal objects.

Lakatosian structure and its gap. The QNFO analysis of superdeterminism's reception states the Lakatosian framework explicitly: science operates through research programmes, each defined by a hard core of foundational axioms rendered unfalsifiable by methodological fiat and protected by a belt of auxiliary hypotheses; the structure has a clear procedure for rejecting theories that attack the hard core but no formal path for their comparative assessment or potential acceptance, so such theories are designated unscientific and expelled rather than adjudicated [10]. Our game model addresses precisely this gap: it prices the protective-belt move (post-hoc accommodation, strategy P) against the risky-core move (prospective prediction, strategy F), so that programmes can be compared by the predictive risk their practitioners actually accept rather than expelled by fiat.

Systemic bias toward theory over evidence. The Head Over Hands analysis argues that confirmation bias describes an individual cognitive flaw and paradigm rigidity a static field condition, but neither captures the active, systemic, cyclical process of defending a cherished theory against observed reality — a process spanning individual psychology, institutional incentives, and philosophical predisposition that actively suppresses contradictory data to preserve conceptual elegance [9]. This is the behavioral

phenomenon our payoff structure is designed to reproduce: strategy P is individually rational precisely when institutions reward elegance and post-hoc coherence more than they reward survived risk.

Incentive misallocation and reform proposals. The Purpose Principle paper diagnoses that the research ecosystem allocates resources to work whose connection to human problems is untraceable, that citation counts fail as quality proxies outside narrow fields, and that prestige cascades reward the wrong signals; it further traces a *35Bquantum – computingbubbletoaprestigecascadeandventure – capitaltimelinemismatch, andproposestime – boundfalsification* b\$ for a publicly registered prospective prediction and raises the effective cost of unregistered post-hoc accommodation.

The validation-metric gap for non-empirical contributions. The G/S/E taxonomy decomposes scientific progress into Generative, Structural Unification, and Epistemic Scaffolding types, and identifies the Validation Metric Problem — S-Type and E-Type work lacks a rigorous validation metric — as the framework's primary limitation, giving rise to risks of subjective evaluation and disciplinary stagnation [12]. The same work supplies a formal template we adopt: its Theorem 3 gives a necessary condition for non-trivial unification, $G_U(T_U; T_1, T_2) > \delta$ with a minimum compression threshold such as $\delta = 0.15 \times (K(T_1) + K(T_2))$, showing that qualitative epistemic notions can be made computable and falsifiable [12]. Our accuracy threshold p^* plays the same role for demarcation that δ plays for unification.

Jeopardy as a problem format. In physics education research, Jeopardy problems begin with the answer — an equation, diagram, graph, or simulation — and ask students to work backward toward the question; students who instead work backward from the answer in the back of the book have been shown to lack the conceptual foundation of more advanced problem solvers [4]. The same paper analyzes the zero-energy eigenstate of the infinite square well: the $E = 0$ state is rejected by most textbooks because the sinusoidal ansatz $A \sin(kx + \phi)$ fails at $k = 0$, yet the correct treatment solves $d^2\psi(x)/dx^2 = 0$ inside the well, yielding $\psi(x) = Ax + B$, and suitable Dirac delta functions can make such a state exact and physical [4]. This is a precise microcosm of our thesis: the textbook rejection is a post-hoc narrative (no physical solution exists) that a prospective, quantitative analysis refutes. Institutionalized predictive jeopardy asks knowledge systems to run the prospective analysis by default.

Falsification inside a mature science: the TBD criterion. The Toda-Brumer-Duff criterion estimates the local exponential rate of divergence between nearby trajectories as a chaos test; it is simple, intuitive, popular, and known to be not rigorous because it predicts false positives [5]. The scrutiny literature shows the corrective process working: Cerjan and Reinhardt proposed a nonlinear saturation model of the instability, which the recent analysis criticizes because it does not explain why the initial stages of the instability produce no observable effects; the definitive shortcoming identified is that the criterion cannot distinguish couples of trajectories informative about the dynamics from those that are not, because it drops the momenta and analyzes a subspace of half the original dimension [5]. This episode is a working instance of strategy F at the community level: a popular criterion was exposed to prospective, quantitative refutation and was repaired. Our model asks what institutional parameters make such episodes frequent rather than exceptional.

Evaluation frameworks as institutional technology. The medical LLM review establishes that evaluating clinical knowledge systems requires a robust evaluation framework covering guiding principles, evaluation techniques, and real-world application scenarios, with automatic metrics such as $\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$ used to assess predictive ability quantitatively [6]. The same review documents that clinical datasets vary in authenticity, diversity, coverage, update frequency, usability, reliability, and language, and that instruction fine-tuning pipelines (e.g., ChatDoctor built on LLaMA-7B with the alpaca instruction dataset and the HealthCareMagic100k dialogue dataset) retrieve external knowledge databases during dialogues to improve output accuracy [6]. This is the operational substrate our criterion would govern: a knowledge system whose evaluation is already metric-based can be gated by predictive-jeopardy requirements at marginal cost.

Context-dependence of evaluation. The context-localization framework models knowledge-based evaluation where a structured nonnegative score on facts, rules, cases, criteria, or evidence units is evaluated through conditional aggregation tests on admissible knowledge contexts, and proves a localization theorem identifying when masking the signal and localizing the admissible sets are equivalent, via monotonicity and a reduction property [1]. This matters for our criterion because the severity of a jeopardy test is context-relative: a prediction exposed in one admissible context (a narrow retrospective dataset) may be unfalsifiable in the broader context that matters. The localization theorem gives the formal machinery for scoring a claim's jeopardy level across contexts rather than in a single privileged one.

Benchmarks and predictive tasks in knowledge systems. Knowledge-graph relation prediction supplies a mature predictive-benchmark culture: the $\mathcal{A}EMP$ method learns contextualized representations from entity context and semantic path context, and on six standard datasets (WN18, WN18RR, FB15k, FB15k-237, UMLS, Kinship) it outperforms prior methods on all metrics for WN18, WN18RR, UMLS, and Kinship and two of four metrics for FB15k, with global attention alone achieving best or second-best results in seven large-dataset cases [3]. The experimental design — stratified train/test/validation splits in which test entities must have been embedded in training — is a proceduralized jeopardy test: the prediction is locked before the outcome is visible [3]. Similarly, the data-analytics knowledge-graph work encodes evaluation requirements (metrics such as accuracy, recall, F1, RMSE, MAE, and techniques such as cross-validation, benchmarking, and train-test splits) as first-class user intents, and evaluates link-prediction recommendations by Hits@3 [2]. These systems demonstrate that prospective evaluation protocols can be encoded, audited, and recommended automatically — the engineering precondition for institutionalizing jeopardy at scale.

Attribution of contribution. The Myerson-value work computes individual attribute and policy contributions in cooperative multi-agent systems, showing that exploiting a hierarchical knowledge graph makes the Myerson approach 19 to 66 times faster than direct Shapley estimation, with only 1 of 240 comparisons statistically different under the Mann-Whitney U test [7]. Attribution is the reward side of our game: if a system can attribute a confirmed prediction to the agent who risked it, the payoff b becomes enforceable.

Distributional constraints. Finally, any institutional criterion must face who bears its costs: LLM analysis of FLORES-200/+, Ethnologue, and World Development Indicators data finds that about 1.5 billion speakers in lower-middle-income countries face tokenization costs 4 to 6 times higher than for English, alongside inferior model performance — a double jeopardy of cost and inefficiency [8]. A jeopardy-based reward system implemented on such infrastructure would compound existing disparities unless its costs are subsidized or its tests are language-fair; we return to this in Section 6.

3. Methods

3.1 Model overview

We model a knowledge-producing community as a population of agents who repeatedly choose one of two strategies when producing a claim:

- Strategy F (falsifiable practice): the agent registers a prospective prediction before the outcome is known. If the prediction is confirmed, the agent receives reward b ; if it fails, the agent incurs cost c . The prediction succeeds with probability p , the agent's predictive accuracy.
- Strategy P (post-hoc practice): the agent produces a narrative accommodated to the outcome after the outcome is known. The narrative is never publicly wrong, so the agent receives a certain baseline payoff s , interpreted as publication, citation, or career credit obtainable without prospective risk.

This is the minimal structure that captures the demarcation intuition: F carries prospective jeopardy (the claim can fail publicly), P does not. The demarcation criterion we defend is institutional rather than semantic: a knowledge system is scientific to the degree that its reward structure makes F individually rational and evolutionarily stable.

3.2 Assumptions

We state every assumption explicitly, since no calibrated institutional parameters are available in the present sources.

- A1. Payoffs are additive and comparable across strategies within a generation (a standard evolutionary-game simplification).
- A2. Predictive accuracy p is strategy-independent within an agent: an agent who could predict well chooses F ; agents with low p rationally choose P . Heterogeneity in p across agents is what the threshold p^* sorts.
- A3. The baseline payoff s for post-hoc accommodation is positive and certain; we take $s = 1$ in the baseline calibration as a unit normalization.
- A4. Under a current-style incentive configuration, the confirmed-prediction reward is small relative to the certain post-hoc payoff, because publication and citation credit do not distinguish prospective registration from retrospective fitting; we take $b = 2$, $c = 1$ as the baseline. Under a reformed configuration with pre-registration and attribution (in the spirit of falsification registries and mandatory independent verification [11]), we take $b = 10$, $c = 1$.
- A5. Strategy frequencies evolve by the replicator equation $\dot{x} = x(1 - x)(\pi_F - \pi_P)$, where x is the fraction of F -players.
- A6. The antifragility analysis assumes failed predictions remove one unfalsifiable or refuted claim each from the system's claim pool, at no replacement cost, in the simplest closed-form case.
- A7. The comparative case analysis (economics-style macro modeling versus meteorology-style forecasting; evolutionary-psychology-style narrative construction versus molecular-genetics-style intervention) is a proposed validation design with assumed parameter ranges, not an empirical result; all such numbers are labeled projections.

3.3 Analytic tools

We use three tools: (i) expected-payoff comparison for the invasion condition; (ii) the replicator equation for dynamics; (iii) a reliability accounting model for antifragility. All arithmetic is shown in Section 4.

4. Analysis

4.1 Expected payoffs

The expected payoff to strategy F for an agent with accuracy p is

$$\pi_F = pb - (1 - p)c,$$

because the prediction is confirmed with probability p (reward b) and fails with probability $1 - p$ (cost c). The payoff to strategy P is

$$\pi_P = s.$$

4.2 The invasion condition and the accuracy threshold

Strategy F invades a population of P -players exactly when $\pi_F > \pi_P$:

$$pb - (1 - p)c > s.$$

Expanding the left side:

$$pb - c + pc > s \implies p(b + c) > s + c \implies p > \frac{s + c}{b + c}.$$

So the threshold predictive accuracy is

$$p^* = \frac{s + c}{b + c}.$$

This is the paper's central closed-form result: an agent's accuracy must exceed p^* for falsifiable practice to be individually rational, and p^* is a function of institutional parameters alone. Two comparative statics follow immediately. Since

$$\frac{\partial p^*}{\partial b} = -\frac{s + c}{(b + c)^2} < 0,$$

raising the confirmed-prediction reward lowers the accuracy bar for falsifiable practice. Since

$$\frac{\partial p^*}{\partial s} = \frac{1}{b + c} > 0,$$

raising the certain payoff to post-hoc accommodation raises the bar. Institutions therefore control the effective supply of falsifiable claims through b and s even if they cannot change anyone's innate accuracy.

4.3 Baseline calibration: post-hoc dominance

Under assumption A4 (baseline: $b = 2$, $c = 1$, $s = 1$), take a moderately competent forecaster with $p = 0.6$. Then

$$\pi_F = 0.6 \times 2 - (1 - 0.6) \times 1 = 1.2 - 0.4 = 0.8,$$

while

$$\pi_P = s = 1.0.$$

Since $0.8 < 1.0$, even a forecaster who is right 60% of the time strictly prefers post-hoc fitting under the baseline configuration. The threshold accuracy is

$$p^* = \frac{1 + 1}{2 + 1} = \frac{2}{3} \approx 0.6667.$$

An agent must be right more than two times out of three before prospective jeopardy pays at all. The replicator advantage of P over F at $p = 0.6$ is

$$\Delta = \pi_P - \pi_F = 1.0 - 0.8 = 0.2 > 0,$$

so F is driven out: $\dot{x} = x(1-x)(\pi_F - \pi_P) = -0.2x(1-x) < 0$ for $0 < x < 1$. This is the formal statement of the drift claim: under incentive structures that do not price prospective risk, unfalsifiable post-hoc strategies dominate, and the dominance is driven by the payoff gap, not by any cognitive failing of the agents.

4.4 Intervention: pre-registration and attribution

Under assumption A4's reformed configuration ($b = 10, c = 1, s = 1$), the same $p = 0.6$ forecaster obtains

$$\pi_F = 0.6 \times 10 - 0.4 \times 1 = 6.0 - 0.4 = 5.6,$$

against $\pi_P = 1.0$. The advantage of F is now

$$\Delta = 5.6 - 1.0 = 4.6 > 0,$$

and the threshold collapses to

$$p^* = \frac{1+1}{10+1} = \frac{2}{11} \approx 0.1818.$$

An agent needs only 18.18% accuracy for falsifiable practice to be rational — a bar that any functioning forecasting practice clears. The intervention thus converts the equilibrium from P -dominance to F -dominance for the entire realistic accuracy range. To keep the intervention conservative, we also compute a weaker reform, $b = 5$: then

$$\pi_F = 0.6 \times 5 - 0.4 \times 1 = 3.0 - 0.4 = 2.6 > 1.0,$$

and

$$p^* = \frac{2}{6} = \frac{1}{3} \approx 0.3333.$$

The equilibrium flips for any b satisfying $p b - (1-p)c > s$, i.e.,

$$b > \frac{s + (1-p)c}{p} = \frac{1 + 0.4}{0.6} = \frac{1.4}{0.6} \approx 2.3333.$$

Any reward above roughly 2.33 under the baseline cost structure flips the $p = 0.6$ agent's best response — a modest institutional change with a discontinuous behavioral consequence.

4.5 Antifragility

Define system-level reliability R as the fraction of claims in the pool that have survived a prospective jeopardy test. Let the pool contain N claims, of which U are unfalsifiable or refuted, so

$$R = 1 - \frac{U}{N} = \frac{N - U}{N}.$$

Under assumption A6, each failed prediction removes exactly one claim from U (the refuted claim is discarded) without changing N 's surviving complement structure in the simplest case where the failed claim was counted in U . Then after k failed predictions,

$$R(k) = \frac{N - (U - k)}{N} = R(0) + \frac{k}{N},$$

which is strictly increasing in k : the derivative is

$$\frac{dR}{dk} = \frac{1}{N} > 0.$$

This is the antifragile configuration: each failed prediction strictly increases system reliability. The condition for antifragility is that failed predictions are (i) publicly visible, (ii) attached to the claim that failed, and (iii) actually removed from the pool — precisely the three functions a falsification registry performs [11]. If instead failures are absorbed by the protective belt (auxiliary hypotheses patched so the core claim never registers as failed [10]), then U does not decrease and R is flat or declining as low-quality claims accumulate; the model thus locates antifragility not in the predictions themselves but in the bookkeeping institutions.

A concrete instance: take $N = 1000$ claims with $U = 400$ unfalsifiable or refuted claims, so

$$R(0) = \frac{600}{1000} = 0.60.$$

After $k = 100$ publicly registered failures are removed,

$$R(100) = \frac{1000 - 300}{1000} = 0.70.$$

One hundred failed predictions bought a ten-percentage-point reliability gain — a quantity only obtainable because the failures were recorded rather than accommodated.

4.6 Dynamics of reform

Let x_0 be the initial fraction of F -players and $\Delta = \pi_F - \pi_P$ the per-generation payoff advantage after an intervention. The replicator equation integrates to

$$x(t) = \frac{1}{1 + \frac{1-x_0}{x_0} e^{-\Delta t}}.$$

Take the reformed configuration of Section 4.4 with a marginal adopter population: suppose the intervention initially converts only a minority, $x_0 = 0.1$, and suppose conservatively that the effective advantage is diluted to $\Delta = 0.2$ (for instance by mixing reformed and unreformed venues). The time for the F -fraction to reach majority ($x = 0.5$) satisfies

$$0.5 = \frac{1}{1 + 9e^{-0.2t}} \implies 1 + 9e^{-0.2t} = 2 \implies 9e^{-0.2t} = 1 \implies e^{-0.2t} = \frac{1}{9},$$

so

$$t = \frac{\ln 9}{0.2} = \frac{2.19722\dots}{0.2} \approx 10.99 \text{ generations.}$$

Under the undiluted advantage $\Delta = 4.6$, the same transition takes

$$t = \frac{2.19722\dots}{4.6} \approx 0.4777 \text{ generations}$$

— effectively immediate. The contrast quantifies the leverage of enforcement: dilution of the payoff advantage, not the size of the initial minority, controls the reform timescale.

4.7 Context-relative jeopardy severity

Following the context-localization framework, in which a structured score is evaluated through conditional aggregation over admissible knowledge contexts and the localization theorem identifies when masking the signal and restricting the admissible sets are equivalent [1], we define the jeopardy level of a claim as the supremum over admissible contexts E of the aggregation score $A(f|E)$ at which the claim remains exposed to refutation, in direct analogy to the generalized level measure $\Lambda_{\mu, \mathcal{A}}(f, u) = \sup\{\mu(E) : A(f|E) \geq u, E \in \mathcal{E}\}$ [1]. A claim tested only in a narrow retrospective context has a low jeopardy level even if it appears falsifiable locally; the criterion requires evaluating the supremum over the contexts that matter for the claim's use. We state this as a structural requirement rather than computing a number, since no calibrated aggregation operator is available in the present sources.

5. Results

All numbers below are computed in Section 4 from the stated assumptions; none are empirical measurements.

R1. Accuracy threshold (Section 4.2). Falsifiable practice is individually rational exactly when $p > p^* = (s + c)/(b + c)$.

R2. Baseline dominance of post-hoc fitting (Section 4.3). With $b = 2$, $c = 1$, $s = 1$, $p = 0.6$: $\pi_F = 0.8$, $\pi_P = 1.0$, $p^* = 2/3 \approx 0.6667$, and $\dot{x} = -0.2x(1 - x)$, so F is driven out.

R3. Reform reversal (Section 4.4). With $b = 10$, $c = 1$, $s = 1$, $p = 0.6$: $\pi_F = 5.6$, $\pi_P = 1.0$, $p^* = 2/11 \approx 0.1818$. With the weaker reform $b = 5$: $\pi_F = 2.6$, $p^* = 1/3 \approx 0.3333$. The best-response flip occurs at $b \approx 2.3333$.

R4. Antifragility (Section 4.5). $R(k) = R(0) + k/N$ with $dR/dk = 1/N > 0$ when failures are registered and removed. For $N = 1000$, $U = 400$: $R(0) = 0.60$, $R(100) = 0.70$.

R5. Reform timescale (Section 4.6). With $x_0 = 0.1$ and diluted $\Delta = 0.2$: $t = \ln 9/0.2 \approx 10.99$ generations to majority. With undiluted $\Delta = 4.6$: $t \approx 0.4777$ generations.

R6. Structural results (Section 4.7). The jeopardy level of a claim is a context-supremum quantity in the sense of the generalized level measure [1]; comparative statics $\partial p^*/\partial b < 0$ and $\partial p^*/\partial s > 0$ hold for all parameter values.

6. Discussion

The model's central claim is that demarcation is a property of incentive configurations, not of claim content: the same agent with the same accuracy p produces falsifiable or unfalsifiable science depending on b , c , and s . This claim is supported here only formally. Its empirical content — that real institutions occupy the P -dominant region of parameter space, and that reforms move them to the F -dominant region — rests on the comparative case analysis we could only propose, not perform.

Limitations, explicitly:

- Payoff additivity (A1). Real reputations are nonlinear in track record; a single celebrated confirmation may be worth many routine ones. If payoffs are convex in streaks, the threshold p^* shifts and the clean comparative statics weaken.
- Strategy-independence of accuracy (A2). If agents can invest to raise p , the game becomes a two-dimensional choice and the threshold analysis describes only the margin of investment, not the level.
- Normalization (A3). $s = 1$ is a unit choice; the results that depend on the ratio structure (p^* , the flip point $b \approx 2.3333$) are invariant to rescaling, but the absolute timescale in R5 is not.
- Calibration of b and c (A4). The values $b = 2$ and $b = 10$ are assumptions chosen to represent, respectively, configurations where publication credit does not distinguish registration from retrofit, and configurations with registries and mandatory independent verification [11]. No source supplies measured institutional payoffs; every headline number inherits this assumption.
- Replicator dynamics (A5). Agents may copy strategies imperfectly, or institutions may be layered so that different venues have different (b, c, s) . A population structured by venue would show coexistence rather than fixation.
- Antifragility bookkeeping (A6). The linear reliability gain k/N assumes failed claims are removed at no cost and no replacement. If each failure also removes adjacent sound claims (guilt by association) or if refuted claims are replaced by new unfalsifiable ones faster than failures arrive, R can fall; the antifragile regime is a condition, not a guarantee.
- Context relativity (Section 4.7). We import the structure of the localization theorem [1] but compute no jeopardy levels; doing so requires a calibrated aggregation operator we do not possess.
- Distributional incidence. Reward systems built on AI infrastructure inherit its inequities: speakers in lower-middle-income countries face tokenization costs 4 to 6 times higher than English speakers, affecting roughly 1.5 billion people [8]. A jeopardy-based reward layer would disproportionately tax non-English prospective registrants unless deliberately subsidized.

7. Practical Implications and Engineering Pathways

7.1 Who would use the result, and what changes

The primary users are the operators of resource-allocation mechanisms in knowledge production: funding agency program officers, journal editors, tenure committees, and the designers of AI-assisted research pipelines. If the result holds, the decision that changes is the scoring rule applied to incoming claims. Under the baseline configuration derived in Section 4.3, a program officer who cannot distinguish prospective registration from retrospective fitting faces a system in which a $p = 0.6$ forecaster rationally submits post-hoc narratives ($\pi_P = 1.0$ versus $\pi_F = 0.8$); the officer's review process therefore selects for narratives rather than for predictions. The result supplies the officer with a computable lever: raising the confirmed-prediction reward b above the flip point $b \approx 2.3333$ (for a $p = 0.6$ population under $c = 1, s = 1$) changes the individually rational strategy without requiring any change in reviewer judgment or researcher ability. A journal editor's design parameter changes analogously: requiring a declared refutation condition before review is, in the model's terms, a device that raises b and makes the protective-belt move [10] visible rather than free. For AI-mediated pipelines, the result identifies which gate to build: the pipeline's acceptance criterion should be the declared prospective test, not the plausibility of the narrative, because a criterion without a falsifiable structure would be gameable by exactly the systems that most need it [11].

A second user group is the maintainers of evaluation infrastructure for machine-learning-based knowledge systems. The medical LLM review shows that evaluation frameworks for such systems already consist of guiding principles, evaluation techniques, and real-world application scenarios, with automatic metrics such as $\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$ used to assess predictive ability quantitatively [6]. Such systems can be gated by predictive-jeopardy requirements at marginal cost because the metric infrastructure exists. The knowledge-graph work supplies the engineering precedent: evaluation requirements (metrics such as accuracy, recall, F1, RMSE, MAE, and techniques such as cross-validation, benchmarking, and train-test splits) can be encoded as first-class user intents in a knowledge graph and used to drive workflow recommendations [2], and link-prediction recommendations can be evaluated by Hits@3 [2]. Encoding "declare the refutation condition and the train-test split before the outcome is visible" as an intent of the same kind is a direct extension of an existing pattern.

7.2 Current readiness

This work is purely analytical and has not been peer reviewed. Against the NASA Technology Readiness Level definitions, it sits at TRL 1–2: the basic principles (the threshold $p^* = (s + c)/(b + c)$, the replicator dynamics, the antifragility accounting) are stated and derived on paper, and the concept and its applications have been formulated but not demonstrated in any operational setting. Analytical work never exceeds TRL 3, and nothing here involves hardware, a prototype, or a pilot; no manufacturing readiness is claimed. The gap between TRL 2 and any operational deployment is the empirical content identified in Section 6: no measured institutional values of (b, c, s, p) exist in the present sources, so every calibration is an assumption.

7.3 The smallest next test

The smallest next calculation that would de-risk or falsify the central result is a parameter-estimation study on an existing pre-registration dataset. The measurement required is the empirical distribution of predictive accuracy p among registered forecasters, and the empirical reward differential between registered-confirmed and unregistered outputs in a venue that already distinguishes them. The model's prediction to test is the flip: within the measured distribution of p , the fraction of participants whose registered predictions beat their own unregistered baselines should rise discontinuously across the measured threshold corresponding to p^* for that venue's reward structure. The outcome that refutes the model is a continuous, threshold-free relationship between accuracy and registration behavior, or a measured reward structure under which the model's flip point $b > (s + (1 - p)c)/p$ falls outside the range of rewards the venue actually offers while registration behavior still shifts. A second, cheaper falsification test is internal to the model: the antifragility claim $R(k) = R(0) + k/N$ predicts that a registry's removal log, if audited, shows monotone increase in the surviving-claim fraction; an audit showing that removed claims are replaced by new unfalsifiable claims at a rate exceeding the removal rate would falsify the antifragility condition as stated in Section 4.5.

7.4 Binding constraints

The constraints that would bind first, in order:

- **Measurement of p .** The threshold p^* is only actionable if the accuracy distribution of the target population is known. No source supplies such a distribution; estimating it requires longitudinal forecasting data that most institutions do not collect. This is the first binding constraint and an open question.
- **Enforcement of b .** The reward b for a confirmed registered prediction must be paid by some party and attributed to the correct agent. Attribution is technically feasible where the infrastructure exists: the Myerson-value work shows that exploiting a hierarchical knowledge graph makes individual contribution computation 19 to 66 times faster than direct Shapley estimation, with only 1 of 240 comparisons statistically different under the Mann-Whitney U test [7], so computing who deserves the reward is not the bottleneck; the bottleneck is the institutional commitment to pay it, which is an open question.
- **Context specification.** The jeopardy level of a claim is a supremum over admissible contexts [1]; specifying which contexts are admissible for a given claim is a governance decision, not a technical one, and disputes over it are the point of the mechanism rather than a defect [11]. Open question: who adjudicates the context set.
- **Distributional cost.** A registration layer built on LLM infrastructure inherits tokenization cost disparities: about 1.5 billion speakers in lower-middle-income countries face tokenization costs 4 to 6 times higher than for English [8]. If registration is mediated by such infrastructure, non-English registrants pay a higher per-claim tax; the model's c becomes language-dependent, which Section 6 notes would compound existing disparities unless subsidized. Open question: the size of the required subsidy.
- **Gaming by volume.** Because AI adoption spans all 333 tracked research fields [11], automated claim generation could flood a registry with low- p predictions whose expected cost $(1 - p)c$ is small. The model handles this only if c scales with the claim, which requires per-claim review capacity — an open question and the most likely first failure mode of any deployment.

7.5 Order-of-magnitude bounds (projections)

All figures in this subsection are projections with stated assumptions, not measurements. Assumption set: the baseline calibration of Section 4.3 ($c = 1, s = 1$ normalized units), a target population with median accuracy $p = 0.6$, and the flip-point formula $b > (s + (1 - p)c)/p$ derived in Section 4.4. Under these assumptions, the reward differential needed to flip a median forecaster is $b - s \approx 1.33$ normalized units, i.e., roughly a $2.3\times$ increase over the certain post-hoc payoff. If a venue's current acceptance-linked payoff is taken as the unit, the projection is that a registration premium of about 1.3 to 1.5 units — payable only on confirmed registered predictions — suffices to reverse the equilibrium for the median forecaster, and that the weaker reform $b = 5$ (a $5\times$ premium) yields $\pi_F = 2.6$ against $\pi_P = 1.0$, a comfortable margin. The uncertainty on these bounds is dominated by the assumed $p = 0.6$; the flip point scales as $b > (s + (1 - p)c)/p$, so a population with $p = 0.4$ would require $b > (1 + 0.6)/0.4 = 4.0$, and one with $p = 0.3$ would require $b > (1 + 0.7)/0.3 \approx 5.67$. The required premium is therefore between roughly $2.3\times$ and $5.7\times$ the post-hoc payoff for populations whose accuracy lies between 0.6 and 0.3; no source pins down where in that range real populations fall.

7.6 Risks and misuse paths

The mechanism itself can be misused. A registry that rewards confirmed predictions creates an incentive to register only near-certain predictions, driving p -stratification rather than genuine risk-taking; the model's answer is that the threshold p^* makes this self-limiting only if b is not raised further, since raising b lowers the bar and admits low-information "predictions." A second misuse path is adversarial flooding described in Section 7.4: automated generation of high-volume, low-cost registrations could convert the registry from a filter into a noise source. A third is the distributional one: a pay-to-register or infrastructure-mediated system taxes speakers facing 4–6× tokenization premiums [8], converting an epistemic mechanism into an exclusionary one. Each of these is a design constraint on b , c , and the context set, not a refutation of the threshold result.

8. Conclusion

This paper set out to convert the demarcation problem from a matter of informal institutional judgment into a computable criterion. The core derivation is short enough to state in one line: falsifiable practice is individually rational exactly when $p > p^* = (s + c)/(b + c)$, so whether a knowledge system produces falsifiable science is a function of its reward parameters, not of the metaphysical content of its claims. Under a baseline calibration in which publication credit does not distinguish registration from retrofit ($b = 2$, $c = 1$, $s = 1$), a forecaster who is right 60% of the time rationally fits after the fact ($\pi_F = 0.8$ versus $\pi_P = 1.0$), and the system drifts toward unfalsifiable narratives — the formal counterpart of the Head Over Hands pattern [9] and of the Lakatosian observation that axiom-violating theories are expelled rather than adjudicated [10]. Raising the confirmed-prediction reward past the flip point $b \approx 2.3333$ reverses the equilibrium for the same forecaster, and a registry-style reform ($b = 10$) drops the accuracy bar to $p^* = 2/11 \approx 0.1818$. The antifragility result locates the value of failure in bookkeeping: when registered failures remove refuted claims, system reliability rises by k/N , and when failures are absorbed by the protective belt, it does not. The reform timescale, $\ln 9/0.2 \approx 10.99$ generations under diluted incentives versus 0.4777 generations undiluted, shows that enforcement intensity, not initial adoption, controls the speed of transition.

The result's empirical content remains untested: no measured institutional payoffs exist in the present sources, and every headline number inherits the calibration assumptions stated in Section 3.2. What the paper establishes is that the criterion is computable, that its levers are identifiable, and that the smallest next test — estimating p and the reward differential on an existing pre-registration dataset — is within reach of current infrastructure. A demarcation test that is short enough to state in one sentence and strong enough to fail is the prerequisite for holding knowledge systems accountable for the resources they claim [11]; this paper supplies the game-theoretic form of such a test and identifies exactly which observations would falsify it.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.18.7-codeagent). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by zai-org/glm-5.3-flash, zai-org/glm-5.3, openai/gpt-oss-120b, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; every cited work's full text was retrieved (arXiv HTML, or the QNFO paper's body) and the writers were given excerpts of its own sections; statements about a cited work are restricted to its retrieved title, abstract and those excerpts.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: a separately generated script recomputed 15 quantitative claims from the paper's stated inputs; 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond the retrieved full-text excerpts; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

[1] Context Localization for Generalized Level-Based Evaluation in Knowledge-Based Systems. arXiv:2608.27482v1. <https://arxiv.org/abs/2608.27482v1> [2] Capturing and Anticipating User Intents in Data Analytics via Knowledge Graphs. arXiv:2411.01023v1. <https://arxiv.org/abs/2411.01023v1> [3] Learning Attention-based Representations from Multiple Patterns for Relation Prediction in Knowledge Graphs. arXiv:2206.04801v1. <https://arxiv.org/abs/2206.04801v1> [4] Playing Quantum Physics Jeopardy with zero-energy eigenstates. arXiv:quant-ph/0606196v1. <https://arxiv.org/abs/quant-ph/0606196v1> [5] Revisiting the Toda-Brumer-Duff criterion for order-chaos transition in dynamical systems. arXiv:2310.01915v2. <https://arxiv.org/abs/2310.01915v2> [6] Reviewing Clinical Knowledge in Medical Large Language Models: Training and Beyond. arXiv:2502.20988v2. <https://arxiv.org/abs/2502.20988v2> [7] Towards a more efficient computation of individual attribute and policy contribution for post-hoc explanation of cooperative multi-agent systems using Myerson values. arXiv:2212.03041v1. <https://arxiv.org/abs/2212.03041v1> [8] Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers. arXiv:2410.10665v1. <https://arxiv.org/abs/2410.10665v1> [9] QNFO: Head Over Hands [10] QNFO: EMERGENT CORRELATION IN A LOCAL-DETERMINISTIC UNIVERSE [11] QNFO: The Purpose Principle: Grounding Research in Reality, Utility, and Premise-Depth [12] QNFO: Quantifying Structural Unification and Epistemic Scaffolding in Scientific Progress