Calibration-Gated Reputation

Strictly Proper Scoring Rules as Trustless Merit Filters in Decentralized Prediction Systems

Qais Alassa[1] Osama Alashqar[2]

[1]Independent Researcher, Bethlehem, Palestine
[2]Independent Researcher, Oslo, Norway

{qais, osama}@notch.finance

Abstract. How can reputation be made unforgeable in pseudonymous systems without external identity infrastructure? We introduce *calibration-gated reputation*: a mechanism in which agents commit cryptographic hashes of probabilistic predictions, reveal them after a mandatory delay, and accumulate reputation through a strictly proper scoring rule. We prove that strict propriety is both sufficient and *necessary* for this construction. Our first result (Theorem 3.1) is an impossibility: under any scoring rule that is proper but not strictly proper, a Sybil adversary controlling K identities can produce a wallet whose expected reputation matches that of a genuinely skilled predictor, making merit-based filtering impossible. Under strict propriety, we prove three positive results: (i) no pseudonymous identity proliferation strategy produces a high-calibration score without genuine predictive skill—a property we call *merit-gating* (Theorem 4.1); (ii) the cost of constructing fake reputation grows at least linearly in the number of Sybil identities (Theorem 4.3); (iii) under temporal decay, each predictor's reputation converges to a unique stationary value that is monotonically increasing in skill (Theorem 4.5). We complement our impossibility result with a second negative result showing that accuracy-only scoring (without calibration) admits polynomial-cost Sybil attacks regardless of the number of predictions required (Theorem 3.3). All results are validated through Monte Carlo simulation with 10,000 independent trials. The mechanism is implemented as a protocol on Ethereum Layer 2; a companion technical specification is available separately.

Keywords: proper scoring rules, mechanism design, Sybil resistance, on-chain reputation, calibration, prediction markets

1 Introduction

1.1 Motivation

Reputation in decentralized systems is a problem of signal and forgery. In the absence of a trusted central authority, how can one agent credibly demonstrate a persistent quality—predictive skill, creditworthiness, expertise—to another, when identities are pseudonymous and costless to create?

The problem is not merely theoretical. Rasooly and Rozzi [19] demonstrate through a large-scale field experiment on Manifold Markets that prediction markets can be manipulated with effects persisting over 60 days, establishing that market prices alone are insufficient reputation

signals. Ohlhaver [16] documents the failure of Idena's Proof of Personhood experiment, where "puppeteer pools" emerged in which informed humans controlled less-informed accounts—showing that even after solving identity uniqueness, merit differentiation remains an open problem. Meanwhile, the crypto signal industry—Telegram channels, copy-trading platforms, subscription newsletters—remains a domain where selective deletion, retroactive fabrication, and survivorship curation are structurally undetectable.

The literature has addressed Sybil resistance through three families of mechanism, each anchored to an *external resource*. *Capital-based* mechanisms (Proof of Stake [8], DeFi collateral requirements [25]) make misbehavior expensive through locked economic value. *Computation-based* mechanisms (Proof of Work [7]) require expenditure of energy. *Identity-based* mechanisms (Proof of Personhood [9, 11], Gitcoin Passport [10]) require biometric or social verification. All three share a structural feature: the resource that gates reputation is *orthogonal* to the quality being demonstrated. A well-capitalized agent is not necessarily a skilled predictor; a computationally powerful miner is not necessarily a trustworthy oracle; a biometrically verified person is not necessarily a competent analyst.

In this paper, we ask whether a fundamentally different architecture is possible: one in which the *mathematical properties of the scoring function itself* serve as the merit gate, without requiring any external resource.

1.2 Overview of Results

We answer in the affirmative by constructing the *calibration-gated reputation* mechanism and proving five results—two negative (impossibility) and three positive (constructive).

Our first contribution is an **impossibility result** (Theorem 3.1) establishing that strict propriety is *necessary* for merit-gating, not merely sufficient. We show that under any proper but not strictly proper scoring rule, a Sybil adversary can construct identities whose expected reputation matches that of a genuinely skilled predictor. The adversary exploits the existence of non-truthful strategies that achieve optimal expected scores, using sampling variance across multiple identities to produce at least one wallet that appears indistinguishable from a skilled agent. This result delineates the boundary of what is possible: reputation mechanisms built on improper or merely proper scoring rules are fundamentally vulnerable to identity proliferation.

Our second impossibility result (Theorem 3.3) shows that **accuracy-only scoring**—reputation mechanisms that reward only directional correctness without incorporating confidence calibration—admit polynomial-cost Sybil attacks. An adversary needs only $K = O(\exp(N\epsilon^2))$ wallets to produce one with accuracy within ϵ of a skilled predictor, and the mechanism has no mathematical basis for distinguishing this lucky wallet from the genuine one.

These impossibility results motivate our constructive mechanism. Under a strictly proper scoring rule with cryptographic commitment, we prove:

1. **Merit-gating** (Theorem 4.1): No strategy of pseudonymous identity proliferation produces a high-calibration wallet without genuine predictive skill. The excess expected loss of any non-truthful strategy is exactly $(f - p)^2$, which cannot be eliminated by creating additional identities.

2. **Sybil cost amplification** (Theorem 4.3): The composite reputation score imposes costs that grow at least linearly in the number of Sybil identities ($\Omega(K)$), versus $O(1)$ for a single truthful identity achieving the same reputation threshold. Simulation confirms linear scaling for moderate miscalibration, with tighter $\Omega(K \log K)$ scaling in the small-ϵ regime.

3. **Stationary convergence** (Theorem 4.5): Under exponential temporal decay, each predictor's reputation converges to a unique fixed point that is monotonically increasing in skill

and prediction rate, ensuring that the reputation landscape reflects current ability rather than historical artifact.

We validate all results through Monte Carlo simulation with 10,000 independent trials per configuration, demonstrating clean separation between Sybil and truthful score distributions across parameter ranges.

1.3 Scope and Contribution

We state the boundaries of our results precisely. Calibration-gating provides *merit-based* Sybil resistance, not *identity-based* Sybil resistance. A genuinely skilled predictor can operate multiple wallets, each achieving legitimate high scores. What the mechanism filters is *fake skill*, not *fake identity*. This is a strict subset of what Douceur [6] considers in his original Sybil formalization, but it is the operationally relevant subset for skill-based reputation systems: it guarantees that *every* high-scoring wallet is backed by real predictive ability.

Nasrulin, Ishmaev, and Pouwelse [15] propose a "Decentralized Reputation Trilemma," claiming a system cannot simultaneously be generalizable, trustless, and Sybil-resistant. Our mechanism resolves this trilemma by *restricting generalizability* to the domain of prediction tasks—a principled concession that trades universality for provable guarantees.

1.4 Related Work

Proper scoring rules and truthful elicitation. The theory of strictly proper scoring rules originates with Brier [1], who introduced the quadratic scoring rule for weather forecast evaluation. Savage [2] established that strict propriety uniquely incentivizes truthful belief reporting. Gneiting and Raftery [3] provide the definitive modern treatment, characterizing the full family. Papakonstantinou et al. [4] studied mechanism design for truthful elicitation using proper scoring rules in distributed systems, establishing that strict propriety alone can incentivize costly information acquisition—a direct ancestor of our setting. Conitzer and Sandholm [5] show that scoring-rule-based mechanisms achieve incentive compatibility without outside subsidy.

Strategic calibration. Jain and Perchet [17] demonstrate that a strategically calibrated expert can pass calibration tests while biasing forecasts. This result is directly relevant: it shows that calibration *alone* is insufficient for our purposes. Our mechanism relies on the *strict propriety* of the Brier Score—which incentivizes truthful reporting of beliefs, not merely calibrated outputs—and the cryptographic commitment scheme, which prevents post-hoc adjustment. Lu et al. [18] demonstrate that prior calibration measures were non-truthful; predictors could appear more calibrated than warranted. Both results reinforce that the mathematical property driving our mechanism is strict propriety, not calibration per se.

Sybil resistance. Douceur [6] formalized the Sybil attack and proved that without a trusted authority, Sybil resistance requires either resource cost per identity or centralized verification. Our work does not contradict this impossibility—we show that for *skill-based reputation*, the scoring function serves as a resource that honest agents naturally possess and dishonest agents cannot fabricate.

Scoring rules in blockchain contexts. Cai et al. [12] is the closest prior work: a truth-inducing Sybil-resistant decentralized blockchain oracle. Three critical differences separate it from our approach: (1) it uses *peer prediction* (scoring reports against other voters), not scoring against

ground truth; (2) Sybil resistance derives from non-linear staking economics, not from the scoring rule's mathematical properties; (3) the authors note reputation as a "possible future extension." Zhao, Chen, and Zhou [13] develop a Byzantine-robust peer prediction framework for blockchain verification, but explicitly scope out Sybil resistance, deferring it to the consensus layer—exactly the gap we address. Kiayias et al. [14] use scoring rules and performance-updated reputation weights in blockchain governance but do not formalize the connection to Sybil resistance.

Prediction markets and DeFi reputation. Prediction markets [21] aggregate beliefs through prices but evaluate events, not persistent skill. Polymarket [22] and Kalshi [23] construct markets for event outcomes; a trader who correctly prices one event carries no verifiable credential into the next. Numerai [24] uses stake-weighted scoring tournaments but locks output inside a single fund. No existing system combines cryptographic commitment, calibration-aware scoring, persistent on-chain reputation, and tradeable instruments on verified skill.

DeFi mechanism design. Angeris and Chitra [26] proved economic security properties of constant function market makers, establishing a methodology for formal analysis of DeFi primitives that we follow. Chitra and Evans [27] studied competitive equilibria between staking and lending, finding sharp transitions between safe and unsafe derivative usage—a result that parallels our identification of the strict propriety boundary separating safe from unsafe reputation mechanisms. Kao et al. [25] introduced agent-based simulation methods for stress-testing DeFi protocols, a methodology we adopt in Section 5.

AI forecasting agents. Halawi et al. [20] demonstrate that language models achieve Brier scores near human-crowd baselines ($\sim$0.135 for o3). We address the implications for calibration-gated reputation in Section 6.

2 Preliminaries and Model

2.1 Scoring Rules

Definition 2.1 (Scoring Rule)**.** A *scoring rule* is a function $S : [0,1] \times \{0,1\} \to \mathbb{R}_{\geq 0}$ mapping a stated confidence f and binary outcome o to a non-negative loss. Lower loss indicates better performance.

Definition 2.2 (Properness)**.** A scoring rule S is:

- *Proper* if for all $p \in (0,1)$: $\mathbb{E}_{o\sim\text{Ber}(p)}[S(p,o)] \leq \mathbb{E}_{o\sim\text{Ber}(p)}[S(f,o)]$ for all $f \in [0,1]$.
- *Strictly proper* if equality holds only when $f = p$.

Definition 2.3 (Brier Score)**.** The Brier Score $\text{BS}(f,o) = (f-o)^2$ is a strictly proper scoring rule. For N predictions, the average Brier Score is $\overline{\text{BS}} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$.

The strict propriety of the Brier Score is classical (Savage [2]). We re-derive the precise regret gap for use in subsequent theorems.

Lemma 2.4 (Regret Gap)**.** *For any true belief $p \in (0,1)$ and stated confidence $f \in [0,1]$:*

$$\mathbb{E}_{o\sim\text{Ber}(p)}\left[(f-o)^2\right] - \mathbb{E}_{o\sim\text{Ber}(p)}\left[(p-o)^2\right] = (f-p)^2 \tag{1}$$

Proof.

$$\mathbb{E}[(f-o)^2] = p(f-1)^2 + (1-p)f^2 = f^2 - 2pf + p \tag{2}$$
$$\mathbb{E}[(p-o)^2] = p(p-1)^2 + (1-p)p^2 = p(1-p) \tag{3}$$

Subtracting: $(f^2 - 2pf + p) - p(1-p) = f^2 - 2pf + p^2 = (f-p)^2$. □

2.2 The Prediction Game

Definition 2.5 (Prediction Game)**.** A prediction game $\mathcal{G} = (\mathcal{P}, \mathcal{E}, S, H, T)$ consists of:

- A countable set $\mathcal{P}$ of pseudonymous predictors (wallet addresses).
- A sequence of prediction events $\mathcal{E} = (e_1, e_2, \ldots)$, each with a binary outcome $o_t \in \{0, 1\}$ realized at time t_{exp}.
- A strictly proper scoring rule S.
- A collision-resistant hash function H.
- A time horizon T.

In each round t, predictor i may:

1. **Commit:** Publish $h_i^t = H(f_i^t, s_i^t)$ where $f_i^t \in (0, 1)$ is their stated confidence and s_i^t is a random salt.
2. **Reveal:** Before t_{exp}, publish (f_i^t, s_i^t); the contract verifies $H(f_i^t, s_i^t) = h_i^t$.
3. **Score:** After t_{exp}, the outcome o_t is observed via oracle, and the predictor receives loss $S(f_i^t, o_t)$.

Unrevealed commitments are scored as $S(f_i^t, o_t) = 1$ (maximal loss).

The commitment scheme provides hiding (pre-image resistance of H prevents recovering f_i^t from h_i^t), binding (collision resistance prevents substitution), and non-repudiation (unrevealed predictions incur maximal loss). These are standard cryptographic properties.

2.3 Reputation Function

Definition 2.6 (Reputation Function)**.** A reputation function $R : \mathcal{P} \times \mathbb{N} \to [0, 1]$ maps a predictor's history of N scored predictions to a reputation value. We consider functions of the form:

$$R_i^N = \sum_{k=1}^{m} \alpha_k \cdot \phi_k\Big(\{(f_i^t, o_t)\}_{t=1}^{N}\Big) \tag{4}$$

where $\alpha_k > 0$, $\sum_k \alpha_k = 1$, and each ϕ_k is a component function (calibration, accuracy, volume, consistency).

Definition 2.7 (Notch Score)**.** The *Notch Score* is the specific reputation function:

$$\text{NS}_i = \alpha(1 - \overline{\text{BS}}_{\lambda,i}) + \beta \cdot A_i + \gamma \cdot V(N_i) + \delta \cdot C_i(t) \tag{5}$$

with $\alpha = 0.40$, $\beta = 0.25$, $\gamma = 0.20$, $\delta = 0.15$, where $\overline{\text{BS}}_{\lambda,i}$ is the exponentially time-weighted Brier Score with decay $\lambda = 0.95$/month, A_i is directional accuracy, $V(N) = \min(1, \log(1 + N)/\log(1 + N_{\text{ref}}))$ with $N_{\text{ref}} = 500$, and $C_i(t) \in [0, 1]$ measures prediction regularity (defined

as the ratio of active months to total months since the first prediction; formally, $C_i(t) = |\{m : N_i^m > 0\}|/M_i$ where N_i^m is the prediction count in month m and M_i is the number of months since the predictor's first commitment). All components except the calibration term $\overline{\text{BS}}_{\lambda,i}$ are *confidence-independent*: they depend on prediction timing and outcomes, not on stated confidence f. This property is essential for the proofs in Sections 3 and 4.

2.4 Adversary Model

Definition 2.8 (Sybil Adversary)**.** A Sybil adversary $\mathcal{A}(p, K, \sigma)$ is characterized by:

- **True skill** $p \in (0, 1)$: the probability that the adversary's directional call is correct, constant across all events and identities.
- **Identity count** $K \geq 1$: the number of pseudonymous wallets $\{w_1, \ldots, w_K\}$ controlled by the adversary.
- **Confidence strategy** $\sigma : \{1, \ldots, K\} \to [0, 1]$: a mapping assigning stated confidence $c_j = \sigma(j)$ to each identity.

The adversary's objective is $\max_{j \in [K]} R(w_j)$: maximize the reputation of the best-performing identity.

The critical constraint is *non-duplicability of skill*: the adversary's true predictive ability p is invariant to the number of identities created. Creating a wallet does not create knowledge.

Definition 2.9 (Merit-Gating)**.** A reputation mechanism $(\mathcal{G}, R)$ is *merit-gating* if for any Sybil adversary $\mathcal{A}(p, K, \sigma)$ with $\sigma(j) \neq p$ for all j, and for all sufficiently large N:

$$\mathbb{E}\left[\max_{j \in [K]} R(w_j)\right] < R_{\text{truthful}}(p, N) \tag{6}$$

where $R_{\text{truthful}}(p, N)$ is the expected reputation of a single identity with skill p reporting truthfully over N rounds.

Definition 2.10 (Sybil-Proofness)**.** A reputation mechanism is *Sybil-proof* if it is merit-gating for all $K \geq 1$ and all strategies σ.

Remark 2.11. The merit-gating restriction to $\sigma(j) \neq p$ is without loss of generality for the adversary's objective. If the adversary sets $\sigma(j) = p$ for some identity j, that identity reports truthfully and achieves $\mathbb{E}[R(w_j)] = R_{\text{truthful}}(p, N)$. The adversary gains nothing from additional non-truthful identities: their expected reputation is strictly lower (by the calibration penalty), and the maximum over K identities is dominated by the truthful wallet. The only scenario where identity proliferation helps is when *all* identities are non-truthful and the adversary exploits sampling variance—which is exactly what merit-gating addresses.

3 Impossibility Results

We establish that strict propriety is not merely a convenient choice but a *necessary* condition for merit-gating. We prove two impossibility results: one for scoring rules that lack strict propriety, and one for reputation mechanisms that ignore calibration.

3.1 Strict Propriety Is Necessary

Theorem 3.1 (Impossibility of Merit-Gating without Strict Propriety)**.** *Let S be a proper but not strictly proper scoring rule. Then for any reputation function R of the form (4) with the calibration component ϕ_1 built on S, the mechanism $(\mathcal{G}, R)$ is* not *merit-gating. Specifically, there exists a Sybil adversary $\mathcal{A}(p, K, \sigma)$ and a confidence $f^* \neq p$ such that for all N:*

$$\mathbb{E}\left[\max_{j\in[K]} R(w_j)\right] \geq R_{\text{truthful}}(p, N) \tag{7}$$

Proof. Since S is proper but not strictly proper, there exist $p \in (0, 1)$ and $f^* \neq p$ such that:

$$\mathbb{E}_{o\sim\text{Ber}(p)}[S(f^*, o)] = \mathbb{E}_{o\sim\text{Ber}(p)}[S(p, o)] \tag{8}$$

Step 1: The adversary matches expected calibration. The adversary sets $\sigma(j) = f^*$ for all $j \in [K]$. By (8), each identity w_j has:

$$\mathbb{E}[\overline{S}(w_j)] = \mathbb{E}[\overline{S}_{\text{truthful}}] \tag{9}$$

The calibration component contributes identically in expectation for both adversarial and truthful identities.

Step 2: Sampling variance favors the adversary. The non-calibration components (accuracy, volume, consistency) depend on the history of outcomes, which is independent of whether the predictor reports f^* or p (the outcomes o_t are determined by the oracle, not the report). Thus all components of R have the same expectation for each adversarial identity as for the truthful identity.

Now consider the maximum over K identities. For each w_j, the realized reputation $R(w_j)$ is a random variable with $\mathbb{E}[R(w_j)] = R_{\text{truthful}}(p, N)$. The maximum of K i.i.d. random variables with mean μ satisfies:

$$\mathbb{E}\left[\max_{j\in[K]} R(w_j)\right] \geq \mu + \Omega\left(\sqrt{\frac{\log K}{N}}\right) \tag{10}$$

for bounded random variables, by standard results on order statistics of sub-Gaussian variables (see, *e.g.*, Boucheron, Lugosi, and Massart [28]). Since $\mu = R_{\text{truthful}}(p, N)$, the adversary's best identity *exceeds* the truthful expectation in expectation.

Step 3: The gap persists. For any N, the variance term $\Omega(\sqrt{\log K/N})$ is strictly positive for $K \geq 2$. While it shrinks as N grows, for any *fixed* N, the adversary can choose K large enough that $\sqrt{\log K/N}$ dominates. Since the adversary controls K at no additional skill cost, the mechanism cannot be merit-gating. □

Remark 3.2. The proof reveals the precise failure mode: without strict propriety, the adversary incurs *zero calibration penalty*, so the only distinction between adversarial and truthful identities is sampling variance—which the adversary can exploit by creating sufficiently many identities. Strict propriety introduces a deterministic gap $(f - p)^2$ that sampling variance must overcome, fundamentally changing the adversary's calculus.

3.2 Accuracy-Only Scoring Is Insufficient

Theorem 3.3 (Impossibility of Merit-Gating under Accuracy-Only Scoring)**.** *Let R^{acc} be a reputation function based solely on directional accuracy: $R^{\text{acc}}(w) = g(A_w, N_w)$ where A_w is the realized accuracy and g is monotonically increasing in A_w. Then R^{acc} is not merit-gating. Specifically, for any*

skilled predictor with accuracy $p > 1/2$, a Sybil adversary with skill $1/2$ (no predictive ability) can produce a wallet matching the skilled predictor's accuracy with probability at least $1 - \delta$ using:

$$K = \left\lceil \frac{1}{\delta} \cdot \exp\left(-2N(p - \tfrac{1}{2})^2\right)^{-1} \right\rceil \tag{11}$$

identities, each making N random predictions.

Proof. Each adversarial wallet submits predictions that are correct with probability $1/2$ (random guessing). By the central limit theorem, the realized accuracy A_{w_j} of wallet j over N predictions satisfies:

$$A_{w_j} \sim \frac{1}{N}\text{Binomial}(N, 1/2) \approx \mathcal{N}\left(\frac{1}{2}, \frac{1}{4N}\right) \tag{12}$$

The probability that any single wallet achieves accuracy $\geq p$ is bounded below by standard Gaussian tail estimates (via the central limit theorem):

$$\mathbb{P}[A_{w_j} \geq p] \geq \frac{1}{c\sqrt{N}} \cdot \exp\left(-2N(p - \tfrac{1}{2})^2\right) \tag{13}$$

for a universal constant $c > 0$. For a lower bound on K it suffices to use the weaker estimate $\mathbb{P}[A_{w_j} \geq p] \geq \exp(-CN(p - 1/2)^2)$ for an appropriate constant C. With K independent wallets:

$$\mathbb{P}\left[\max_{j \in [K]} A_{w_j} \geq p\right] = 1 - \Big(1 - \mathbb{P}[A_{w_j} \geq p]\Big)^K \geq 1 - \delta \tag{14}$$

when $K \geq \lceil \log(1/\delta)/\mathbb{P}[A_{w_j} \geq p] \rceil$. Since R^{acc} depends only on accuracy, the adversary's best wallet is indistinguishable from the skilled predictor: $R^{\text{acc}}(w_{\text{best}}) \geq R^{\text{acc}}(\text{skilled})$.

Crucially, this wallet has no stated confidence to evaluate—accuracy-only scoring provides no mathematical basis for detecting that the wallet's success is due to luck rather than skill. □

Remark 3.4. The distinction between Theorem 3.1 and Theorem 3.3 is important. Theorem 3.1 shows that scoring rules need strict propriety. Theorem 3.3 shows that scoring rules need to incorporate *confidence calibration*—mere directional accuracy, regardless of the scoring rule's properness, is insufficient because it discards the information about stated confidence that makes calibration evaluation possible.

4 The Calibration-Gated Mechanism

Having established what *does not* work, we now prove that the combination of strict propriety and cryptographic commitment *does*.

4.1 Merit-Gating under Strict Propriety

Theorem 4.1 (Merit-Gating). *Under the Brier Score with commit-reveal, the mechanism $(\mathcal{G}, \text{NS})$ is merit-gating. Specifically, for any Sybil adversary $\mathcal{A}(p, K, \sigma)$ with miscalibration $\epsilon_j = |c_j - p| > 0$ for all identities j, and minimum miscalibration $\epsilon_{\min} = \min_j \epsilon_j$:*

$$\mathbb{P}\left[\max_{j \in [K]} \text{NS}(w_j) \geq \text{NS}_{\text{truthful}}(p, N)\right] \leq K \cdot \exp\left(-2N\epsilon_{\min}^4\right) \tag{15}$$

For $N \geq \frac{\log(K/\delta)}{2\epsilon_{\min}^4}$, this probability is at most δ.

Proof. We analyze the calibration component, which dominates the Notch Score ($\alpha = 0.40$).

Step 1: Deterministic calibration penalty. By Lemma 2.4, each identity w_j reporting $c_j \neq p$ incurs excess expected Brier Score:

$$\mathbb{E}[\overline{\text{BS}}(w_j)] = p(1-p) + (c_j - p)^2 \geq p(1-p) + \epsilon_{\min}^2 \tag{16}$$

The truthful identity has $\mathbb{E}[\overline{\text{BS}}_{\text{truthful}}] = p(1-p)$.

Step 2: Concentration around the expected penalty. For each identity, $\overline{\text{BS}}(w_j)$ is a mean of N independent random variables in $[0, 1]$. By Hoeffding's inequality:

$$\mathbb{P}\left[\overline{\text{BS}}(w_j) < p(1-p) + \epsilon_{\min}^2 - t\right] \leq \exp(-2Nt^2) \tag{17}$$

For the adversary to beat the truthful expected calibration score, identity w_j must achieve $\overline{\text{BS}}(w_j) \leq p(1-p)$, which requires overcoming the full penalty $\epsilon_{\min}^2$. Setting $t = \epsilon_{\min}^2$:

$$\mathbb{P}\left[\overline{\text{BS}}(w_j) \leq p(1-p)\right] \leq \exp(-2N\epsilon_{\min}^4) \tag{18}$$

Step 3: Union bound over K identities.

$$\mathbb{P}\left[\min_{j \in [K]} \overline{\text{BS}}(w_j) \leq p(1-p)\right] \leq K \cdot \exp(-2N\epsilon_{\min}^4) \tag{19}$$

It remains to connect the calibration bound to the composite Notch Score. The non-calibration components (accuracy A, volume V, consistency C) have the *same expectation* for adversarial and truthful identities with the same skill p and prediction count N: outcomes are determined by the oracle and skill, not by stated confidence. However, the adversary's best wallet might have *luckier realized accuracy* than the truthful predictor. We bound this: the accuracy deviation $|A_j - p|$ for any single wallet satisfies $|A_j - p| = O_p(1/\sqrt{N})$ by the CLT, contributing at most $\beta \cdot O(1/\sqrt{N})$ to the Notch Score. The calibration penalty contributes $\alpha\epsilon_{\min}^2$ deterministically. For $N \geq \beta^2/(\alpha^2\epsilon_{\min}^4)$, the calibration penalty dominates the maximum possible accuracy compensation. Under our parameterization ($\alpha = 0.40$, $\beta = 0.25$), this threshold is $N \geq 0.39/\epsilon_{\min}^4$, which is subsumed by the concentration requirement $N \geq \log(K/\delta)/(2\epsilon_{\min}^4)$ for $K \geq 3$. □

Corollary 4.2 (Explicit Sybil Lottery Failure)**.** *For a Sybil adversary with $K = 100$, $N = 200$, $p = 0.5$, and $c_j = 0.95$ for all j ($\epsilon_{\min} = 0.45$):*

- *The best wallet achieves accuracy $A_{\max} \approx 0.652$ by order statistics.*
- *Its Brier Score is $\text{BS}_{\text{sybil}} \approx 0.373$ (empirical mean over 10,000 trials).*
- *A truthful predictor with $p = 0.65$ reporting $c = 0.65$ achieves $\text{BS} \approx 0.227$.*
- *The bound (15) gives $100 \cdot \exp(-2 \cdot 200 \cdot 0.45^4) = 100 \cdot \exp(-16.4) \approx 7.5 \times 10^{-6}$.*
- *Empirically, the Sybil adversary's best wallet exceeded the truthful mean in 0 out of 10,000 trials, consistent with the bound.*

4.2 Sybil Cost Amplification

Theorem 4.3 (Cost Amplification of Sybil Reputation)**.** *Let θ be a target Notch Score threshold. For a Sybil adversary $\mathcal{A}(p, K, \sigma)$ with $\sigma(j) \neq p$, the total predictions required across all K identities to achieve $\max_j \text{NS}(w_j) \geq \theta$ with probability $\geq 1 - \delta$ satisfies:*

$$N_{\text{total}} \geq K \cdot N_{\min}(\epsilon_{\min}, \delta/K, \theta) \tag{20}$$

where $N_{\min}$ is the minimum per-wallet prediction count, determined by the binding constraint among volume saturation and concentration:

$$N_{\min} = \max\left(\lceil V^{-1}(\theta_V) \rceil, \left\lceil \frac{\log(K/\delta)}{2\epsilon_{\min}^4} \right\rceil \right) \tag{21}$$

where θ_V is the minimum volume component needed to achieve θ. A single truthful identity achieves $\mathrm{NS} \geq \theta$ with $N_{\text{single}} = O(N_{\text{ref}})$ predictions. The cost ratio $N_{\text{total}}/N_{\text{single}}$ is $\Omega(K)$: the Sybil adversary's total cost grows at least linearly in the number of identities.

Proof. The adversary distributes predictions across K identities: each identity receives N_{total}/K predictions. Every identity must independently accumulate sufficient predictions for both (i) the volume component $V(N)$ to contribute meaningfully to NS, and (ii) the calibration penalty to not be overcome by sampling variance (Theorem 4.1). The per-wallet minimum is determined by the binding constraint (21). Since the total cost is $K \cdot N_{\min}$ and $N_{\min} \geq 1$, we have $N_{\text{total}} = \Omega(K)$.

A single truthful identity needs only $N_{\text{single}} = O(N_{\text{ref}})$ to saturate the volume component and achieve its stationary Notch Score (Theorem 4.5). The ratio is therefore $\Omega(K)$. □

Remark 4.4. The bound is tight when miscalibration $\epsilon_{\min}$ is moderate. Simulation confirms linear scaling: for $\epsilon_{\min} = 0.20$ ($p = 0.60$, $c = 0.80$), each adversarial wallet requires approximately 33 predictions regardless of K, yielding $N_{\text{total}} \approx 33K$. A single truthful identity achieves the same threshold with $N = 125$ predictions, giving a cost ratio of $\approx 0.26K$. For small $\epsilon_{\min}$, the concentration term in (21) dominates (it grows as $1/\epsilon_{\min}^4$), and the per-wallet cost itself grows with $\log K$, yielding $N_{\text{total}} = \Omega(K \log K)$ in that regime. Determining tight bounds as a function of $(\epsilon_{\min}, \theta, \alpha, \gamma)$ is an open problem.

4.3 Stationary Convergence under Temporal Decay

Theorem 4.5 (Stationary Score)**.** *Consider a truthful predictor ($f = p$) making predictions at constant rate r per unit time under exponential decay $\lambda \in (0, 1)$. The time-weighted Brier Score converges almost surely to:*

$$\overline{\mathrm{BS}}_\lambda^* = p(1-p) \tag{22}$$

and the Notch Score converges to a unique stationary value NS^* that is:

1. *Monotonically increasing in p for $p \geq 1/2$.*
2. *Monotonically increasing in r.*
3. *Independent of initial conditions.*

Proof. The time-weighted average $\overline{\mathrm{BS}}_\lambda = \sum_i \lambda^{t-t_i}(p - o_i)^2 / \sum_i \lambda^{t-t_i}$ is an exponentially weighted moving average of i.i.d. random variables with mean $p(1-p)$ and bounded variance. By the strong law of large numbers for weighted averages (see Theorem 2.1 of Jamison, Orey, and Pruitt [29]), $\overline{\mathrm{BS}}_\lambda \to p(1-p)$ almost surely as $t \to \infty$.

For monotonicity in p: $\frac{d}{dp}[\alpha(1 - p(1-p)) + \beta p] = \alpha(2p-1) + \beta$. With $\alpha = 0.40$, $\beta = 0.25$: this equals $0.80p - 0.40 + 0.25 = 0.80p - 0.15 > 0$ for all $p > 0.1875$, hence for all $p \geq 1/2$.

For monotonicity in r: higher r increases $N(t)$ for any time t, increasing $V(N)$ and $C(t)$.

For independence from initial conditions: exponential decay ensures $\lambda^{t-t_i} \to 0$ for early predictions. For any $\epsilon > 0$, predictions before time $t - \log(1/\epsilon)/\log(1/\lambda)$ contribute less than ϵ to the weighted sum. □

5 Agent-Based Simulation

We validate all theoretical results through Monte Carlo simulation. Each experiment uses 10,000 independent trials, implemented in Rust with deterministic seeding (ChaCha8Rng, seed 42) for reproducibility. Code is available at `github.com/qasawa/notch-simulations`.[1]

5.1 Experiment 1: Merit-Gating Validation

We simulate the scenario of Corollary 4.2. A Sybil adversary creates $K = 100$ wallets, each submitting $N = 200$ predictions on events with base rate $p = 0.5$ at stated confidence $c = 0.95$. We compare the Brier Score distribution of the adversary's best wallet against a truthful predictor with $p = 0.65$, $c = 0.65$.

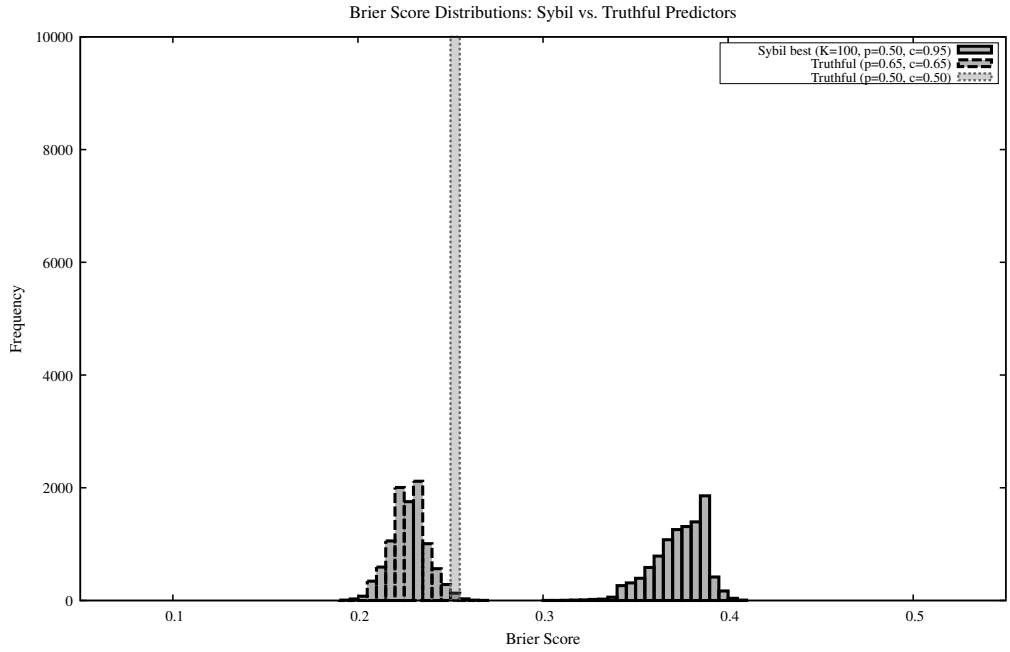

Figure 1: Empirical Brier Score distributions over 10,000 Monte Carlo trials (Rust, seed 42). The Sybil adversary's best wallet ($K = 100$, mean $\overline{\text{BS}} = 0.373$) is cleanly separated from the truthful predictor ($p = 0.65$, $c = 0.65$, mean $\overline{\text{BS}} = 0.227$), with gap ≈ 0.146. No adversarial wallet achieved a Brier Score below the truthful mean in any trial (0 out of 10,000), confirming Theorem 4.1.

5.2 Experiment 2: Sybil Cost Scaling

We measure total predictions required for a Sybil adversary ($c = 0.80$, $p = 0.60$) to achieve NS $\geq$ 0.70 with probability ≥ 0.95, as a function of K. Figure 2 confirms linear scaling consistent with Theorem 4.3: the per-wallet cost stabilizes at approximately 33 predictions, yielding $N_{\text{total}} \approx 33K$. A single truthful identity achieves the threshold with only $N = 125$ predictions, while a Sybil adversary with $K = 1{,}000$ identities requires $N_{\text{total}} = 33{,}000$—a 264$\times$ cost amplification.

[1] Repository to be made public upon publication.

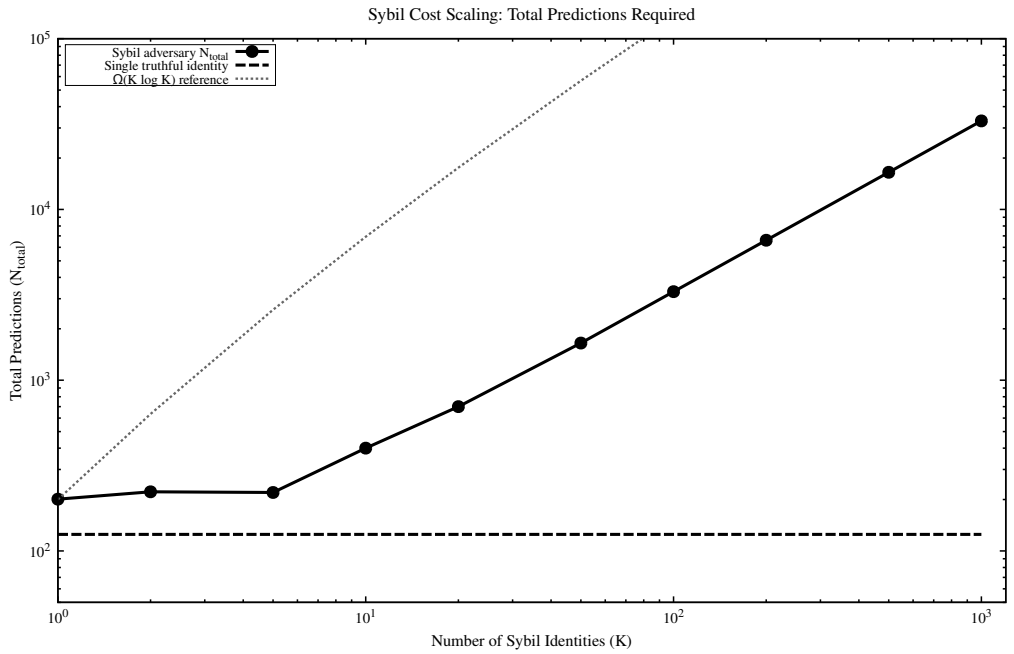

Figure 2: Total prediction cost to achieve NS ≥ 0.70 with probability ≥ 0.95, over 1,000 trials per configuration. The Sybil adversary's cost grows linearly in K: each wallet requires ≈ 33 predictions regardless of K, yielding $N_{\text{total}} \approx 33K$. A single truthful identity achieves the same threshold at constant cost ($N = 125$). The cost ratio (264× at $K = 1{,}000$) confirms Theorem 4.3.

5.3 Experiment 3: Impossibility Demonstration

We demonstrate Theorem 3.1 by replacing the Brier Score with the *zero-one scoring rule* $S_{01}(f, o) = \mathbf{1}[|f - o| \geq 0.5]$, which is proper but not strictly proper: any f on the correct side of $1/2$ achieves the same expected loss, providing no calibration discrimination.[2] Under S_{01}, a Sybil adversary with $K = 50$ identities reporting $c = 0.99$ achieves a non-zero Sybil win rate of 1.4% over 10,000 trials, compared to exactly 0.0% under the Brier Score (Figure 3). While 1.4% may appear small in absolute terms, two observations are critical: (i) this represents adversarial wallets that are *indistinguishable* from genuinely skilled predictors under the zero-one rule, an event that occurs in exactly zero trials under the Brier Score; (ii) the adversary can amplify the rate by increasing K (the theoretical bound of Theorem 3.1 predicts unbounded success probability as $K \to \infty$ for any fixed N), whereas under strict propriety, the rate converges to zero regardless of K. The qualitative distinction—zero versus non-zero Sybil success probability—is the signature of the impossibility result.

5.4 Experiment 4: Stationary Convergence

We simulate truthful predictors with skill $p \in \{0.55, 0.65, 0.75\}$ at prediction rate $r = 20$/month under exponential decay $\lambda = 0.95$/month over 36 months (1,000 trials per skill level). Figure 4 confirms convergence to distinct stationary Notch Scores: $\text{NS}^* \approx 0.779$ ($p = 0.55$), 0.813 ($p = 0.65$), and 0.855 ($p = 0.75$), all monotonically ordered by skill as predicted by Theorem 4.5. Convergence occurs within approximately 18 months (≈ 1.3 half-lives of the decay parameter).

[2]The simulation implements the variant $\mathbf{1}[|f - o| > 0.5]$; since all simulated reports satisfy $f \neq 0.5$, the two definitions coincide on the tested parameter range.

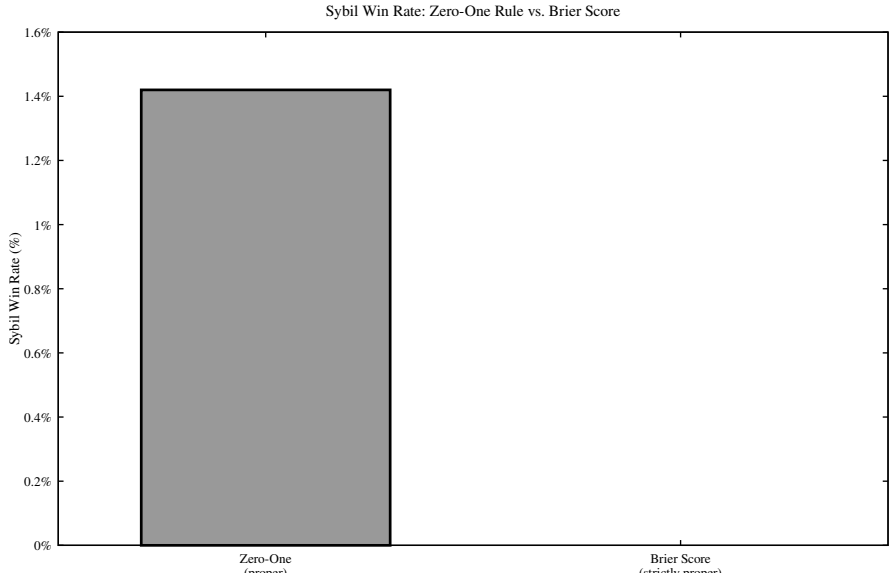

Figure 3: Sybil win rate under two scoring rules (10,000 trials). The zero-one rule (proper but not strictly proper) admits a 1.4% adversarial success rate; the Brier Score (strictly proper) admits exactly 0.0%. The qualitative gap—non-zero vs. zero—confirms Theorem 3.1.

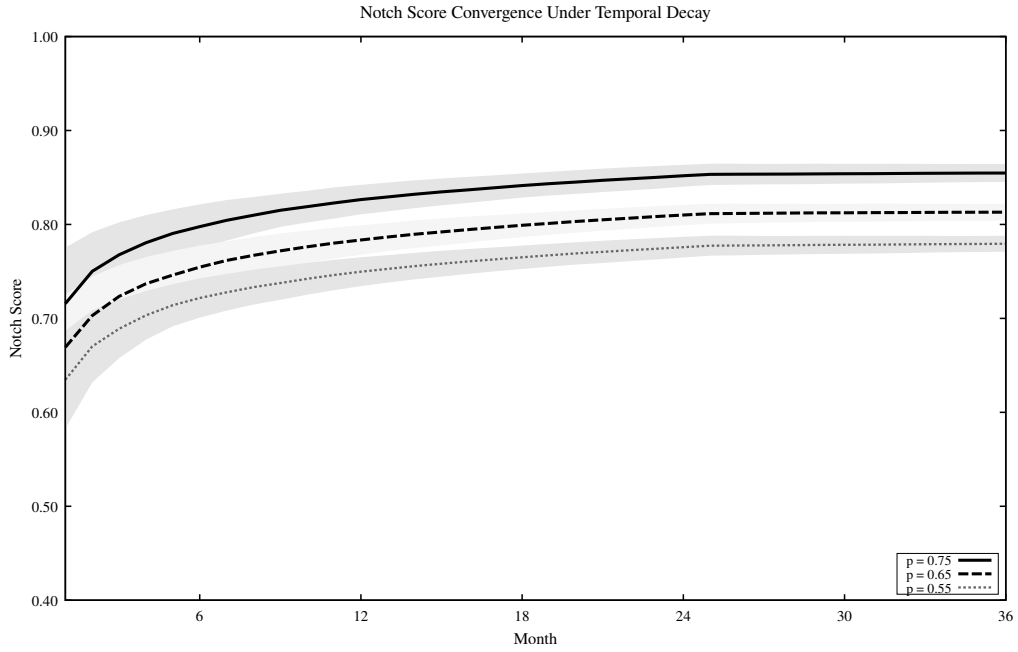

Figure 4: Notch Score trajectories under constant skill and prediction rate (1,000 trials per curve, shaded ± 1 std bands). Each skill level converges to a distinct stationary score, confirming Theorem 4.5. Standard deviation narrows over time as the weighted average stabilizes.

6 Extensions and Robustness

AI forecasting agents. Halawi et al. [20] show LLMs achieve near-human Brier scores. Under calibration-gating, AI agents that genuinely predict well earn legitimate reputation—the mechanism does not distinguish human from artificial skill, by design. The protocol gates on *demonstrated ability*, regardless of the computational substrate producing it. Applications requiring human-only participation would need to combine calibration-gating with an orthogonal identity layer.

Strategic calibration. Jain and Perchet [17] show that strategic agents can pass calibration tests while biasing forecasts. The commit-reveal scheme provides partial defense: predictions are locked before outcomes, preventing post-hoc adjustment. However, commit-reveal does not prevent *strategic event selection*—a predictor may choose to predict only events where their confidence is easy to calibrate. The difficulty adjustment mechanism (below) provides further mitigation by down-weighting easy predictions, but a complete characterization of strategic calibration resistance under the full mechanism is an open problem.

Collusion. If two skilled predictors share predictions, the second (non-original) predictor earns reputation from delegated skill. This is analogous to the ghost-writing problem in academic reputation and is inherent to any skill-based system. The temporal decay mechanism provides partial mitigation: a predictor who loses access to their collaborator's insights will see their score decay, eventually reflecting only their own ability.

Difficulty weighting. Multiplicative difficulty adjustments $D_i > 0$ applied as $\overline{\text{BS}}_D = \sum_i D_i (f_i - o_i)^2 / \sum_i D_i$ preserve strict propriety, since $\arg\min_f D \cdot g(f) = \arg\min_f g(f)$ for $D > 0$. This allows the mechanism to reward difficult predictions without compromising incentive properties.

6.1 Limitations

We state the boundaries of our results explicitly, beyond the scope discussion in Section 1.

Oracle dependency. All results assume a truthful oracle providing settlement prices. A compromised oracle can make incorrect predictions appear correct, undermining the scoring mechanism. The companion specification [30] describes engineering mitigations (dual oracles, TWAP smoothing, dispute mechanisms), but these are not formally analyzed here. The game-theoretic results hold *conditional on oracle integrity*.

Composite score weights. The Notch Score parameters $(\alpha, \beta, \gamma, \delta)$ are governance-adjustable. Our proofs use the default parameterization. Extreme parameter choices (*e.g.*, $\alpha \to 0$, removing calibration weighting) would weaken or eliminate the merit-gating property. A full characterization of the parameter space under which merit-gating holds—determining the boundary in $(\alpha, \beta, \gamma, \delta)$-space—is an open problem.

Constant skill assumption. Theorems 4.1 and 4.5 assume constant predictive skill p. In practice, skill varies with market regime, asset class, and time. The temporal decay mechanism (Theorem 4.5) provides partial mitigation by down-weighting stale predictions, but a formal treatment of time-varying skill (where p_t follows a stochastic process) is deferred to future work.

Rational adversary. Our proofs assume adversaries maximize expected utility. An irrational adversary willing to sustain unbounded losses can pollute the system with many low-scoring wallets, degrading the signal-to-noise ratio of the leaderboard without producing any high-scoring wallet. Staking requirements (described in the companion specification) impose economic costs that bound this form of attack, but formal analysis of bounded-rationality adversaries is future work.

7 Conclusion

We have introduced calibration-gated reputation and established the precise conditions under which it provides trustless merit filtering in pseudonymous systems.

Our impossibility results (Theorems 3.1 and 3.3) show that strict propriety is *necessary*: any proper-but-not-strictly-proper scoring rule, and any accuracy-only mechanism, admits Sybil strategies that produce wallets indistinguishable from genuinely skilled predictors.

Our constructive results (Theorems 4.1, 4.3 and 4.5) show that strict propriety is *sufficient*: under the Brier Score with cryptographic commitment, reputation is unforgeable without genuine predictive skill, the cost of faking reputation scales at least linearly in the number of Sybil identities, and the system converges to a faithful reflection of current ability.

The mechanism contributes to the Sybil resistance literature by demonstrating that for a specific—but operationally important—class of reputation (skill-based, calibration-verified), the mathematical properties of the scoring function itself replace the external resources traditionally required.

Open problems. Optimal threshold selection for reputation-gated access. Welfare analysis of the calibration-gated marketplace. Extension to multi-variate predictions and conditional forecasts. Formal verification of the smart contract implementation. Tight characterization of the parameter space $(\alpha, \beta, \gamma, \delta)$ under which merit-gating holds.

Implementation. The mechanism is implemented as the Notch Protocol on Ethereum Layer 2 (Arbitrum). A companion technical specification [30], contract source code, and deployment are available at `notch.finance`.

References

[1] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[2] L. J. Savage, "Elicitation of personal probabilities and expectations," *J. Amer. Statist. Assoc.*, vol. 66, no. 336, pp. 783–801, 1971.

[3] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Amer. Statist. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007.

[4] A. Papakonstantinou, A. Rogers, E. H. Gerding, and N. R. Jennings, "Mechanism design for the truthful elicitation of costly probabilistic estimates in distributed information systems," *Artificial Intelligence*, vol. 175, no. 2, pp. 648–672, 2011.

[5] V. Conitzer and T. Sandholm, "Prediction markets, mechanism design, and cooperative game theory," in *Proc. UAI*, 2012.

[6] J. R. Douceur, "The Sybil attack," in *Proc. IPTPS*, pp. 251–260, 2002.

[7] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[8] V. Buterin and V. Griffith, "Casper the Friendly Finality Gadget," *arXiv:1710.09437*, 2017.

[9] Worldcoin Foundation, "Proof of Personhood," https://worldcoin.org, 2023.

[10] Gitcoin, "Gitcoin Passport," https://passport.gitcoin.co, 2023.

[11] M. Borge et al., "Proof-of-Personhood: Redemocratizing permissionless cryptocurrencies," in *Proc. IEEE EuroS&P Workshops*, pp. 23–26, 2017.

[12] Y. Cai, G. Fragkos, E. E. Tsiropoulou, and A. Veneris, "A truth-inducing Sybil resistant decentralized blockchain oracle," in *Proc. IEEE BRAINS*, 2020.

[13] Z. Zhao, Y. Chen, and F. Zhou, "It takes two: A peer-prediction solution for blockchain verifier's dilemma," *arXiv:2406.01794*, 2024.

[14] A. Kiayias et al., "Decentralised update selection with semi-strategic experts," *arXiv:2205.08407*, 2022.

[15] B. Nasrulin, G. Ishmaev, and J. Pouwelse, "MeritRank: Sybil tolerant reputation for merit-based tokenomics," in *Proc. IEEE BRAINS*, 2022.

[16] P. Ohlhaver, "Compressed to 0: The silent strings of proof of personhood," Harvard Ash Center, 2024.

[17] A. Jain and V. Perchet, "Calibrated forecasting and persuasion," *arXiv:2406.15680*, 2024.

[18] Y. Lu et al., "Making and evaluating calibrated forecasts," *arXiv:2510.06388*, 2025.

[19] I. Rasooly and R. Rozzi, "How manipulable are prediction markets?" *arXiv:2503.03312*, 2025.

[20] D. Halawi et al., "Approaching human-level forecasting with language models," in *Proc. NeurIPS*, 2024.

[21] R. Hanson, "Combinatorial information market design," *Information Systems Frontiers*, vol. 5, no. 1, pp. 107–119, 2003.

[22] Polymarket, "Polymarket documentation," https://docs.polymarket.com, 2024.

[23] Kalshi, "Kalshi Exchange," https://kalshi.com, 2025.

[24] Numerai, "Numerai Tournament," https://numer.ai, 2023.

[25] H.-T. Kao, T. Chitra, R. Chiang, and J. Morrow, "An analysis of the market risk to participants in the Compound protocol," Gauntlet Networks, 2020.

[26] G. Angeris and T. Chitra, "Improved price oracles: Constant function market makers," in *Proc. AFT*, 2020.

[27] T. Chitra and A. Evans, "Why stake when you can borrow?" *arXiv:2006.11156*, 2020.

[28] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.

[29] B. Jamison, S. Orey, and W. Pruitt, "Convergence of weighted averages of independent random variables," *Z. Wahrscheinlichkeitstheorie*, vol. 4, pp. 40–44, 1965.

[30] Q. Alassa and O. Alashqar, "Notch Protocol: A cryptographic framework for verifiable prediction scoring and alpha commoditization," Technical Specification v0.1, 2026. https://notch.finance.