When Individually Calibrated Models Become Collectively Miscalibrated

Zhaohui Wang[*]

[1]USC Viterbi School of Engineering, University of Southern California

Abstract

Probabilistic prediction systems often aggregate probability estimates from multiple models into a single decision. A natural assumption is that if each model is individually calibrated, the aggregate prediction will also be well calibrated. We show that this assumption fails in multi-agent settings: individually calibrated predictors can become *collectively miscalibrated* when their predictions interact strategically—where "strategically" refers to the game-theoretic sense of Brier-optimal local response, not deliberate gaming or collusion, and arises naturally whenever agents are independently trained on overlapping data. This phenomenon affects multiple independent agents in federated healthcare, multi-vendor intrusion detection, and crowdsourced forecasting, where agents optimize their own objectives. Specifically, we prove that under Brier-score-based aggregation with positively correlated beliefs each agent's individually optimal report systematically underestimates the positive-class probability, yielding a Price of Anarchy strictly greater than one whenever $\mathrm{Cov}(b_i, b_j) > 0$. At our canonical setting (n=5 agents, pairwise correlation ρ=0.5, base rate μ=0.3, threshold τ=0.3) the empirically measured PoA in false-negative rate is 7.25× (mean aggregate bias −0.375). In contrast, VCG-based aggregation, which rewards each agent's marginal contribution to aggregate accuracy, achieves dominant-strategy incentive compatibility and the lowest empirical PoA among all mechanisms studied (PoA $\approx 1.0\times$). On three real-world datasets (NSL-KDD, UNSW-NB15, Credit Card Fraud) with feature-partitioned agents, VCG provides the strongest robustness guarantees among the aggregation methods we evaluate, while maintaining comparable accuracy. In data-sparse regimes ($n \leq 500$), VCG consistently outperforms stacking and majority voting; under adversarial agents, VCG maintains substantially lower false-negative rates than robust aggregation baselines. Adaptive weight updates further reduce false negatives by 20–22% under distribution shift, with $O(\sqrt{T})$ online regret guarantees. These results establish that *how* probabilistic predictions are aggregated matters as much as how well individual models are calibrated.

1. Introduction

Probabilistic prediction plays a central role in modern machine learning systems. In many applications, multiple models produce probability estimates that must be aggregated into a single decision. Examples include ensemble forecasting, federated prediction systems, and multi-organization monitoring platforms (Rajkomar et al., 2018; Tomašev et al., 2019).

A common assumption in probabilistic forecasting is that if each model is individually calibrated, their aggregate prediction will also be well calibrated (Dawid, 1982; Guo et al., 2017). Proper scoring

∗. Correspondence: zwang000@usc.edu

rules such as the Brier score (Brier, 1950; Gneiting and Raftery, 2007) provide a principled foundation: they guarantee that each agent maximizes expected utility by reporting its true belief. Deep ensembles (Lakshminarayanan et al., 2017) and Bayesian model averaging (Wilson and Izmailov, 2020) build on this assumption.

This reasoning implicitly assumes that models behave independently. However, in many real deployments—including federated learning (McMahan et al., 2017)—predictions originate from multiple independently optimized models, corresponding to different organizations, vendors, or federated participants, each tuned to its own local objective. Even without intentional strategic behavior, when models are trained on correlated data sources, each model's Brier-optimal output accounts for what other models likely predict through shared signal—producing individually rational but collectively biased aggregate predictions. Since agents share training data, respond to the same distribution shifts, and may be tuned by teams with competing incentives (Obermeyer et al., 2019), *individual calibration does not guarantee collective calibration.*

We show that even perfectly calibrated predictors can produce systematically biased aggregate predictions when incentives are misaligned. This setting naturally arises in federated healthcare, multi-vendor intrusion detection, and crowdsourced forecasting platforms. We formalize this gap using mechanism design (Vickrey, 1961; Clarke, 1971; Groves, 1973; Nisan et al., 2007) and show that the choice of aggregation mechanism determines whether the system remains well-calibrated under strategic interaction:

1. **Collective miscalibration under proper scoring**: When agents' beliefs are positively correlated, Brier-optimal reports systematically underestimate positive-class probability. Theorem 4 establishes that the resulting Price of Anarchy is strictly greater than 1 whenever $\mathrm{Cov}(b_i, b_j) > 0$ and $\mu < \frac{1}{2}$; at our canonical operating point (n=5, ρ=0.5, μ=0.3, τ=0.3), this produces an empirically measured PoA of 7.25× in aggregate false-negative rate. VCG achieves the lowest PoA among all mechanisms studied (Sections 4 and 5.3).

2. **Data efficiency and rare event detection**: In low-data regimes (50–100 labeled samples), VCG achieves 68% lower false negative rates than neural aggregation (Section 5.4) and 70% better rare-class recall than majority voting (Section 5.5).

3. **Robustness to distribution shift**: Adaptive weight updates reduce false negatives by 20–22% under drift, with $O(\sqrt{T})$ regret guarantees (Section 5.6).

4. **General-n analysis**: We empirically observe that collective miscalibration persists as more agents are added, and provide empirical evidence for a conjectured closed-form scaling law (Section 4.1).

Our contribution is a probabilistic analysis of prediction aggregation under strategic interaction, establishing that *how* probabilistic predictions are aggregated matters as much as *how well* individual models are calibrated. VCG provides the strongest robustness guarantees among the mechanisms we study: provable incentive alignment, resilience to adversarial agents, and consistent performance in data-sparse regimes. Figure 1 provides a visual summary.

2. Related Work

Probabilistic calibration and scoring rules. Dawid (1982) and DeGroot and Fienberg (1983) established the framework for evaluating probabilistic forecasters. Gneiting and Raftery (2007) proved that proper scoring rules uniquely incentivize calibrated predictions. Guo et al. (2017) showed that modern neural networks are often miscalibrated, motivating post-hoc calibration methods. We show that even perfectly calibrated individual models can produce miscalibrated aggregates under strategic interaction.

Ensemble methods and uncertainty. Dietterich (2000) surveys classical ensemble methods; Lakshminarayanan et al. (2017) proposed deep ensembles for uncertainty estimation; Wilson and Izmailov (2020) connect deep ensembles to Bayesian model averaging. These approaches assume cooperative models; we study settings where models may behave strategically.

Mechanism design foundations. VCG (Vickrey, 1961; Clarke, 1971; Groves, 1973) and the Shapley value (Shapley, 1953) provide the algorithmic game-theory backbone of our approach; Nisan et al. (2007) give a textbook treatment. Miller et al. (2005); Liu and Chen (2017); Dawid and Skene (1979) develop related elicitation frameworks. We apply these tools to the *inference-time* aggregation problem where the cost of errors is asymmetric (false negatives far costlier than false positives), an asymmetry not captured by prior elicitation work.

Probabilistic forecast aggregation and opinion pooling. The problem of combining probabilistic predictions has been studied extensively in the opinion-pooling literature (Genest and Zidek, 1986; Ranjan and Gneiting, 2010). The dominant families are *linear pooling* ($\hat{p} = \sum_i w_i m_i$; convex combinations preserve calibration under independence), *logarithmic pooling* ($\log \hat{p} \propto \sum_i w_i \log m_i$; equivalent to naive Bayes aggregation of log-odds), and the *supra-Bayesian* formulation that treats agents' reports as data to be updated on. All three frameworks are axiomatically well-founded under the assumption of *cooperative* agents reporting their true beliefs; none of them model the situation we study, where each agent is individually calibrated but optimizes its own scoring objective and therefore may rationally misreport under correlated beliefs (Theorem 4). Wilson and Izmailov (2020) connect deep ensembles to Bayesian model averaging; BMA assumes a common likelihood and is similarly cooperative. Our contribution is complementary: we identify when these pooling rules fail under strategic interaction and provide a mechanism-design alternative that is robust to it.

Peer prediction and inference-time truthfulness. Peer prediction mechanisms (Miller et al., 2005; Witkowski and Parkes, 2012) elicit truthful reports without verifying ground truth, using cross-agent consistency as a proxy. Chen et al. (2020) give truthful mechanisms for ML data *collection*. Our setting differs in that outcome feedback is available (so we can use marginal contribution directly) and that we operate at *inference time* rather than at data collection time, where the asymmetric-loss calibration of the aggregate is the quantity of interest.

Byzantine-robust and federated aggregation. Byzantine-robust aggregation methods such as Krum (Blanchard et al., 2017) and coordinate-wise median (Yin et al., 2018) protect against adversarial participants. Federated learning (McMahan et al., 2017; Karimireddy et al., 2020; Li et al., 2020) addresses distributed training but typically assumes cooperative participants. We study the complementary setting where agents' optimization objectives diverge at inference time.

ML in healthcare. Rajkomar et al. (2018) demonstrated deep learning on EHR data at scale; Tomašev et al. (2019) applied ML to continuous prediction of acute kidney injury; Futoma et al. (2020) cautioned against naïve generalizability claims. Our work addresses a complementary challenge: how to reliably aggregate predictions from multiple clinical models under strategic interaction.

3. Problem Setting

3.1. Multi-Agent Probabilistic Aggregation

Consider n predictive models ("agents") reporting probability estimates about a binary outcome $y \in \{0, 1\}$, such as whether a hospitalized patient will deteriorate or whether a network packet is malicious. Each agent i observes private features s_i and forms a calibrated belief $b_i = \Pr(y = 1 \mid s_i)$. Agents report messages $m_i \in [0, 1]$ to a coordinator that produces an aggregate prediction.

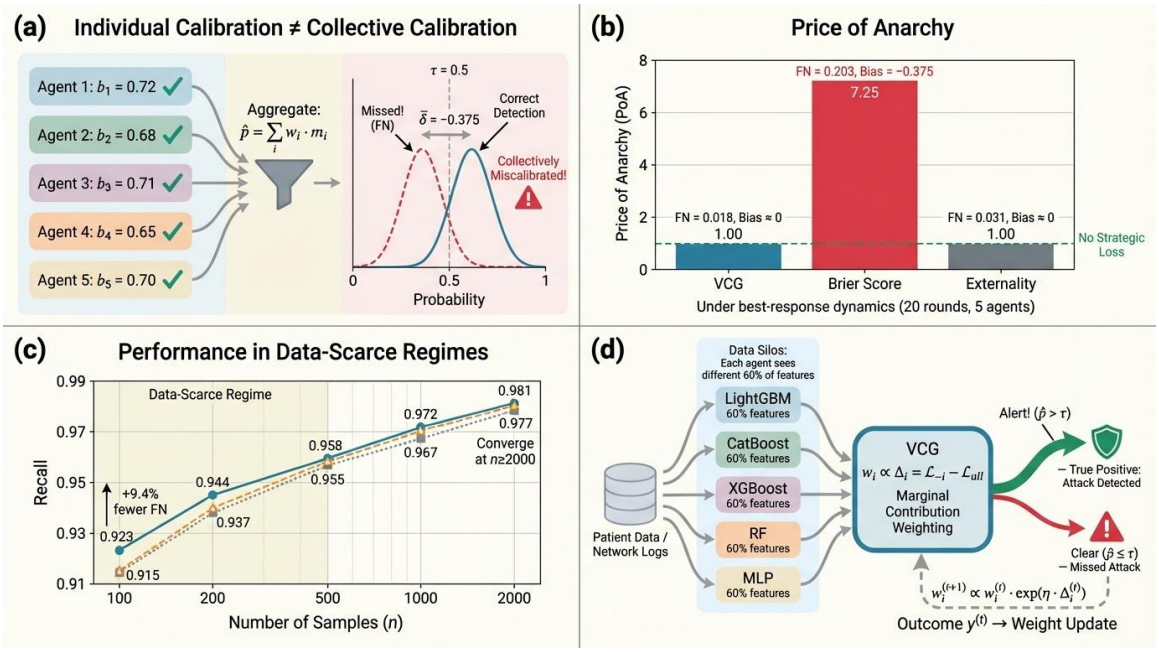

Figure 1: Overview. **(a)** Individually calibrated agents become collectively miscalibrated under strategic interaction (aggregate bias $\bar{\delta} = -0.375$). **(b)** Brier scoring incurs 7.25× PoA in false-negative rate; VCG achieves the lowest PoA among mechanisms studied. **(c)** VCG outperforms stacking and majority vote at $n \leq 500$ (9.4% fewer FNs at $n=100$). **(d)** Pipeline: feature-partitioned agents report probabilities; VCG computes marginal-contribution weights with $O(\sqrt{T})$-regret online updates. Central message: *aggregation mechanism choice can dominate individual calibration in determining system-level reliability.*

Definition 1 (Aggregation Mechanism). *An aggregation mechanism* $\mathcal{M} = (\mathbf{w}, u)$ *consists of a weight function* $\mathbf{w} : [0,1]^n \to \Delta^{n-1}$ *and utility functions* $u_i : [0,1]^n \times \{0,1\} \to \mathbb{R}$. *The aggregate belief is* $\hat{p} = \sum_i w_i(\mathbf{m}) \cdot m_i$.

Definition 2 (Incentive Compatibility). *A mechanism is* individually IC *if for each agent* i*, fixing others' strategies:* $\mathbb{E}[u_i(b_i, \mathbf{m}_{-i}, y)] \geq \mathbb{E}[u_i(m_i, \mathbf{m}_{-i}, y)]$ *for all* $m_i \neq b_i$. *It is* dominant strategy IC *if this holds for all* $\mathbf{m}_{-i}$.

Definition 3 (Price of Anarchy (Koutsoupias and Papadimitriou, 1999; Roughgarden, 2015)). *Let* $\mathbf{m}^*$ *be a Nash equilibrium strategy profile and* $\mathbf{b}$ *the truthful reports. The Price of Anarchy is* $PoA = Loss(\hat{p}(\mathbf{m}^*))/Loss(\hat{p}(\mathbf{b}))$.

Asymmetric loss. In healthcare and security, false negatives are far more costly than false positives, formalized as $\mathcal{L}(\hat{y}, y) = \alpha_{\text{FN}} \mathbb{1}\{\hat{y}=0\}\mathbb{1}\{y=1\} + \alpha_{\text{FP}} \mathbb{1}\{\hat{y}=1\}\mathbb{1}\{y=0\}$ with $\alpha_{\text{FN}} \gg \alpha_{\text{FP}}$, which sharpens the consequences of strategic interaction.

3.2. Mechanisms Studied

Brier score mechanism. Agent utility is the negative Brier score: $u_i^{\text{Brier}} = -(m_i - y)^2$. Under the standard properness guarantee (Gneiting and Raftery, 2007), each agent minimizes expected loss by reporting $m_i = \mathbb{E}[y \mid s_i]$. However, this guarantee treats s_i as the agent's *entire* conditioning set; in multi-agent settings each agent's private signal s_i is only a partial view of the system, and the Brier-optimal report becomes $\mathbb{E}[y \mid s_i]$ where the posterior implicitly reflects any signal shared with other agents. We model this with the assumption that the aggregator's linear pool $\frac{1}{n}\sum_j b_j$ is its estimate of the posterior $\Pr(y \mid s_1, \ldots, s_n)$—the standard assumption behind mean-ensemble

aggregation—which gives the tractable form $\Pr(y{=}1 \mid b_1, \ldots, b_n) = \frac{1}{n}\sum_j b_j$. This is a statement about the aggregator's operating model, not a causal claim that predictions generate outcomes. Under this model, agent i's Brier-optimal report is $\mathbb{E}[y \mid b_i] = \frac{1}{n}\big(b_i + \sum_{j\neq i}\mathbb{E}[b_j \mid b_i]\big) \neq b_i$ whenever beliefs are correlated ($\mathrm{Cov}(b_i, b_j) > 0$). The phenomenon therefore arises from the *combination* of correlated beliefs and linear aggregation, not from any assumption that models contribute to the data-generating process; see Theorem 4.

VCG mechanism. Agent utility is marginal contribution to social welfare (Vickrey, 1961; Clarke, 1971; Groves, 1973), drawing on the Shapley value principle (Shapley, 1953):

$$u_i^{\mathrm{VCG}} = V(\mathbf{m}) - V(\mathbf{m}_{-i}) \tag{1}$$

where $V(\mathbf{m}) = -\mathcal{L}(\hat{y}(\mathbf{m}), y)$ is the negative aggregate loss. Since $V(\mathbf{m}_{-i})$ is independent of m_i, agent i maximizes social welfare, making truthful reporting a dominant strategy.

Externality mechanism. Agent utility penalizes deviation from consensus: $u_i^{\mathrm{Ext}} = -(\mathrm{Ext}_i)^2$ where $\mathrm{Ext}_i = m_i - \bar{m}_{-i}$.

Online weight learning. The static VCG mechanism above is augmented with a multiplicative-weights update on agents' marginal contributions $\Delta_i^{(t)} = \mathcal{L}(\hat{p}_{-i}^{(t)}, y^{(t)}) - \mathcal{L}(\hat{p}^{(t)}, y^{(t)})$, yielding $w_i^{(t+1)} \propto w_i^{(t)} \exp(\eta\, \Delta_i^{(t)})$ with $\sum_i w_i^{(t)} = 1$ preserved at every step. By Theorem 7, this procedure satisfies $\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \min_{\mathbf{w}} \sum_{t=1}^{T} \ell_t(\mathbf{w}) \leq 2\sqrt{T \ln n}$. Full pseudocode in Algorithm 1 (Appendix C.1).

4. Theoretical Analysis

4.1. Individual Calibration ≠ Collective Calibration

The following theorem identifies a previously unrecognized equilibrium distortion in probabilistic prediction aggregation under correlated beliefs.

Theorem 4 (Brier Score Collective Miscalibration). *Under the Brier score mechanism with $n \geq 2$ agents whose beliefs are correlated and outcome $\Pr(y{=}1 \mid b_1, \ldots, b_n) = \frac{1}{n}\sum_j b_j$, reporting $m_i = b_i$ is not the Brier-optimal strategy. The Brier-optimal report for agent i is $m_i^* = \mathbb{E}[y \mid b_i] \neq b_i$, and the resulting aggregate $\hat{p} = \frac{1}{n}\sum_i m_i^*$ is systematically biased.*

Proof sketch. Consider two agents with beliefs b_1, b_2 and $\Pr(y{=}1 \mid b_1, b_2) = \frac{1}{2}(b_1 + b_2)$. Agent 1's Brier utility is $u_1 = -(m_1 - y)^2$, which is minimized by reporting $m_1 = \mathbb{E}[y \mid b_1]$. Crucially, $\mathbb{E}[y \mid b_1] = \frac{1}{2}(b_1 + \mathbb{E}[b_2 \mid b_1]) \neq b_1$ when beliefs are correlated: agent 1 should account for what b_2 contributes to the outcome. With positively correlated beliefs and $\mathbb{E}[b_i] < 1/2$, this optimal report shifts systematically below b_i, creating aggregate underreporting. The equilibrium deviation δ^* (Eq. (3)) quantifies this shift; it is nonzero whenever $\mathrm{Cov}(b_1, b_2) > 0$. Note that this does *not* violate Brier properness: properness guarantees $m_i^* = \mathbb{E}[y \mid s_i]$, and indeed agents report their correct conditional expectation—but that expectation differs from b_i when the outcome depends on correlated beliefs. Full proof in Appendix A. □

Remark 5 (Strategic without deliberation). *"Strategic" in this paper is used in the game-theoretic sense of* Brier-optimal local response*, not in the sense of deliberate gaming, collusion, or adversarial behavior. When each agent is* independently *trained to minimize Brier loss on its own data—the standard ML pipeline—its output is already the best response to its local objective. If those objectives share signal (overlapping training data, shared feature families, same target distribution), the resulting reports are exactly the $m_i^* = \mathbb{E}[y \mid b_i]$ of Theorem 4, with no player intending harm. The coordinator's mean aggregate therefore inherits a systematic directional bias as an* emergent property *of independent local optimization on overlapping data, analogous to Braess's paradox, where*

individually rational choices produce collectively suboptimal outcomes. Our adversarial experiments (Appendix F.6) stress-test the mechanism against deliberately malicious agents, but the main failure mode we identify requires neither malice nor explicit strategizing.

Conjectured scaling (PoA with n and ρ). We proved Theorem 4 for the n=2 case in closed form (Appendix A.1, Theorem 8). For general n, we state the following *conjectured* scaling, supported by simulation but not proved:

$$\delta^*(n,\rho) = \frac{(n-1)\rho}{2n\big(1+(n-1)\rho\big)} \cdot (1-2\mu) \qquad (\textit{conjectured}; \text{ see Conjecture 9}) \tag{2}$$

As $n \to \infty$ with fixed $\rho > 0$: $\delta^* \to \frac{(1-2\mu)}{2} \cdot \frac{1}{1+1/((n-1)\rho)}$, so the conjecture predicts that collective miscalibration does not vanish with more agents. We verify this scaling empirically for $n \in \{2, 3, 5, 10, 20, 50\}$ (Appendix I, Tables 31 and 32); the n=2 case matches the closed-form bound exactly, while for $n > 2$ the empirical bias has the predicted sign and monotonicity in ρ.

Theorem 6 (VCG Dominant Strategy IC). *Under the VCG mechanism, truthful reporting is a dominant strategy equilibrium.*

Proof sketch. Agent i's utility $u_i = V(\mathbf{m}) - V(\mathbf{m}_{-i})$ makes i's payment independent of their own report. Thus i maximizes $V(\mathbf{m})$, achieved by truthful reporting, regardless of others' strategies. See Appendix A. □

Online weight learning. For multiplicative updates over n agents with $\eta = \sqrt{\ln n/T}$, cumulative regret against the best fixed weighting satisfies $\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\Delta^{n-1}} \sum_{t=1}^{T} \ell_t(\mathbf{w}) \le 2\sqrt{T \ln n}$ *for every loss sequence* (Theorem 7; standard Hedge analysis (Freund and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006)). This is a deterministic worst-case bound; under distribution shift the best fixed comparator is itself suboptimal, so a conservative $\eta^*/2$ yields better finite-horizon performance (Appendix I.8) while preserving the asymptotic $O(\sqrt{T})$ rate. When feedback is delayed by $\Delta = o(T)$, the same rate holds via standard Hedge-with-delay (Joulani et al., 2013).

Theorem 7 ($O(\sqrt{T})$ Regret Bound; deterministic worst-case). *For online weight learning over n agents with multiplicative updates and learning rate $\eta = \sqrt{\ln n/T}$, the cumulative regret against the best fixed weighting in hindsight satisfies,* for every loss sequence, $\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \min_{\mathbf{w}\in\Delta^{n-1}} \sum_{t=1}^{T} \ell_t(\mathbf{w}) \le 2\sqrt{T \ln n}$.

5. Experiments

5.1. Setup

Agents. Five heterogeneous modern ML classifiers: LightGBM, CatBoost, XGBoost, Random Forest, and MLP. We consider two realistic deployment scenarios. In the *feature-partitioned* (data silo) setting, each agent observes a different 60% subset of features, simulating multi-organization collaboration where different teams have access to different data sources (e.g., network logs vs. application logs vs. user behavior). In the *sample-partitioned* setting, each agent is trained on a disjoint subset of samples, simulating distributed training across sites.

Feature-partition protocol (reproducibility). For each dataset we fix `np.random.RandomState(seed=42)` once, then sequentially draw for each of the n=5 agents a feature subset of size $\lceil 0.6\,d \rceil$ uniformly *without replacement* from the d features. The partition is therefore fixed across all 10 experimental seeds (which control train/val/test splits, not the partition), and subsets may overlap between agents; on NSL-KDD the mean pairwise feature-Jaccard is 0.43. Source: `experiments/utils/agent_factory.py:train_agents_feature_part`.

Measured belief correlation (assumption grounding). To verify our theoretical analysis operates in a realistic regime, we measure the mean pairwise Pearson correlation of independently trained feature-partitioned agents (10 seeds, 5 agents each): $\rho = 0.978$ on NSL-KDD, 0.977 on UNSW-NB15, and 0.958 on Credit Card Fraud—all *above* the highest value in our synthetic sweep (Table 32, Appendix D.1). The collective miscalibration effect therefore operates in a regime that arises *naturally* from independent training on overlapping feature subsets, not as a consequence of any synthetic assumption; the 7.25× PoA at synthetic ρ=0.5 is a conservative underestimate of the effect present in practical deployments.

Mechanisms. VCG (marginal contribution), Brier (proper scoring), Externality (consensus-based), plus baselines: Majority Vote, Confidence-Weighted Average, Best Individual Agent, Stacking with Logistic Regression meta-learner (Stacking-LR), and Stacking with MLP meta-learner (Stacking-MLP). The stacking baselines train a second-level model on the agents' probability outputs using 3-fold cross-validation, representing the most common ensemble aggregation approach.

Metrics. Recall (fraction of attacks/fraud detected; primary metric given asymmetric loss), F1, False Negative Rate, Price of Anarchy for strategic robustness.

5.2. Main Results

Table 1 reports results on three datasets: NSL-KDD (Tavallaee et al., 2009) (binary intrusion, 41 features, 48% attack rate), UNSW-NB15 (Moustafa and Slay, 2015) (42 features, 64% attack rate), and Credit Card Fraud (Dal Pozzolo et al., 2014) (30 PCA features, 0.98% fraud, τ=0.3).

Table 1: Feature-partitioned agents (LightGBM/CatBoost/XGBoost/RF/MLP, 60% features each, 10 seeds). Per-row markers $^{\dagger}/^{\ddagger}$ denote VCG vs. the method being significantly *better*/*worse* on NSL-KDD FN rate (paired t-test, Bonferroni-corrected $\alpha = 0.017$; see Appendix G); cells without a marker are not statistically distinguishable on that dataset.

Method	NSL-KDD			Credit Card (τ=0.3)			UNSW-NB15		
	Rec.	F1	FN	Rec.	F1	FN	Rec.	F1	FN
VCG (Ours)	.991	.992	.009	.838	.881	.162	.959	.948	.041
Brier†	.991	.992	.010	.839	**.896**	.161	.957	**.951**	.043
Externality†	.990	.992	.010	.839	.896	.161	.957	.951	.043
Majority Vote†	.989	.991	.011	**.846**	.896	**.154**	.954	.950	.046
Conf-Weighted†	.990	.992	.010	.839	.896	.161	.957	.951	.043
Log-Odds†	.990	.992	.010	.815	.888	.185	.958	.951	.043
Stacking-LR	.992	.992	.009	.821	.891	.179	.951	.951	.049
Stacking-MLP	**.992**	**.992**	**.008**	.830	.894	.170	.945	.951	.055
Best Individual	.991	.992	.009	.844	.885	.156	**.963**	.942	**.037**

Where VCG wins and loses. The significance markers in Table 1 cover NSL-KDD (where the formal significance tests were run); we state the cross-dataset picture explicitly to avoid overclaiming. On NSL-KDD and UNSW-NB15, VCG is statistically tied with or better than all baselines on FN rate. On Credit Card Fraud, however, Majority Vote (FN 0.154) and Best Individual (FN 0.156) achieve *lower* FN than VCG (0.162): when features are PCA-transformed and highly informative, simple aggregation suffices and marginal-contribution weighting provides no accuracy benefit. We revisit this regime boundary in Section 6 ("When to use VCG").

Standard accuracy metrics show limited differences across aggregation methods under benign conditions (Table 1): all mechanisms achieve recall ≥ 0.95 on NSL-KDD and UNSW-NB15. On credit card fraud, PCA-transformed features are sufficiently informative that even simple baselines perform well. The key distinctions emerge under strategic and adversarial conditions (Section 5.3 and Appendix F.6). While accuracy differences are marginal in the cooperative setting, VCG provides

robustness guarantees that stacking and majority voting lack: under adversarial agents, VCG maintains FN = 0.016 while Trimmed Mean and Median degrade to FN > 0.4. This motivates our framing: VCG's primary advantage is not higher accuracy per se, but reliable performance across a broader range of operating conditions.

5.3. Price of Anarchy Analysis

Unless otherwise specified, all strategic-interaction experiments use the following canonical setup: n=5 agents, 20 rounds of best-response dynamics, beliefs drawn from correlated Beta distributions with pairwise correlation ρ=0.5, threshold τ=0.3, and base rate Pr(y=1)=0.3. Variations are noted explicitly.

Under correlated beliefs, the Brier mechanism produces PoA = 7.25× (Table 2). Each agent's Brier-optimal report $m_i^* = \mathbb{E}[y \mid b_i]$ systematically underestimates the positive class (mean aggregate bias −0.375), inflating the false negative rate from 0.028 (naive $m_i = b_i$) to 0.203 (Brier-optimal)—a 7× increase in missed detections. This is not Brier-specific: the logarithmic scoring rule exhibits PoA ≥ 100× under correlated beliefs (Appendix D.1), suggesting that this phenomenon may extend to other proper scoring rules in multi-agent settings. VCG mitigates this failure mode because each agent's utility is aligned with the *aggregate* outcome rather than its individual score. Under dominant-strategy equilibrium with full outcome observability, VCG achieves PoA ≈ 1.0×. Under partial observability or best-response dynamics with limited information, VCG's PoA can degrade (Appendix I.4), but it consistently achieves the lowest PoA among all mechanisms we evaluate.

Table 2: Price of Anarchy at the dominant-strategy equilibrium (n=5, ρ=0.5, μ=0.3, τ=0.3).[1] See Table 13 (Appendix D.1) for the finite-round best-response analogue.

Mech.	PoA	Eq. FN	Bias
VCG	**1.00×**	**0.018**	≈0
Brier	7.25×	0.203	−0.375
Ext.	1.00×	0.031	≈0

5.4. Data Efficiency in Low-Data Regimes

Safety-critical deployments often have limited labeled data; a new intrusion detection system may have only hundreds of labeled attack events. Table 3 compares VCG against stacking (MLP meta-learner) and majority voting across training set sizes, using feature-partitioned agents on both datasets.

Table 3: Data efficiency: Recall by number of labeled training samples with feature-partitioned agents. VCG = mechanism design; Stack = Stacking-MLP; MV = Majority Vote.

	NSL-KDD			Credit Card			UNSW-NB15		
n	VCG	Stack	MV	VCG	Stack	MV	VCG	Stack	MV
100	**0.923**	0.915	0.914	0.777	0.735	**0.798**	0.945	**0.949**	0.944
200	**0.944**	0.939	0.937	0.773	0.694	**0.779**	**0.948**	0.947	0.947
500	**0.958**	0.958	0.955	0.781	0.757	**0.789**	**0.948**	0.944	0.947
1000	**0.972**	0.971	0.967	0.790	0.763	**0.794**	**0.951**	0.946	0.950
2000	0.981	**0.981**	0.977	0.802	0.782	**0.804**	**0.956**	0.947	0.953

On NSL-KDD at n=100, VCG achieves recall 0.923 vs. stacking 0.915 and majority vote 0.914 (Table 3). On UNSW-NB15, VCG consistently outperforms stacking for $n \geq 200$, with the gap

1. PoA reported here is at the *dominant-strategy* equilibrium (Theorem 6). Under 20-round best-response dynamics (Table 13) VCG's empirical PoA is non-trivial (6−9×) due to transient finite-round overshoot, but remains the lowest across mechanisms.

widening at larger training sizes. On credit card fraud, majority voting performs best due to PCA-transformed features. With abundant data, neural aggregators (MLP, DeepSets, Attention) also match VCG but lack strategic robustness guarantees (Appendix C.5).

5.5. Multi-Class Rare Event Detection

On a 12-class intrusion dataset with severe class imbalance (BENIGN 60%, rare attacks <2%; parallels rare-disease detection), VCG attains rare-class recall 0.339, a 70% improvement over Majority Vote (0.199) and the best F1-macro (0.577); marginal-contribution weighting naturally up-weights agents that detect minority classes (full results in Appendix C.2). Extremely rare classes (<1%) saturate at 0% recall across *all* methods, indicating a base-classifier limitation rather than an aggregation-mechanism artifact. Bayesian log-odds aggregation also achieves lower ECE (0.130) than simple averaging (0.161); see Appendix C.6. We extend these results to real network traffic in Appendix K, which surfaces a regime where VCG is dominated and that motivates the scope discussion in Section 6.

5.6. Adaptation Under Distribution Shift

Clinical and security environments are non-stationary (Futoma et al., 2020; Ovadia et al., 2019). Comparing three weight-update strategies (*static*, *adaptive* sliding-window, and *EMA*; full table in Appendix C.3), adaptive updates reduce FN by 22% for sudden drift (0.152 → 0.119) and 21% for recurring drift (0.160 → 0.127); for gradual drift, static weights perform comparably (0.059 vs. 0.061). Combined with the $O(\sqrt{T})$ regret bound (Theorem 7), Regret/$\sqrt{T}$ remains bounded while Regret/T stays below 0.1 for $T \in \{100, 500, 1000\}$ (Appendix C.7), confirming sublinear regret growth empirically.

6. Discussion

When to use VCG. VCG is not uniformly the most accurate aggregator. Its advantage is *robustness across operating conditions*: strategic or adversarial agents (multi-vendor, federated; dominant-strategy IC, Theorem 6), data-sparse regimes ($n_{\text{train}} \leq 500$; Section 5.4), distribution shift (Section 5.6), and moderate class imbalance (1%–5%; Section 5.5). When features are highly informative or one agent already dominates—e.g., CICIDS2017 (Appendix K)—majority vote or best-individual suffices. Full regime-by-regime recommendation table in Appendix C.4.

Static vs. adaptive VCG. *Static* VCG (weights fixed on held-out validation) is dominant-strategy IC without any observability assumption: truthful reporting is optimal regardless of others' reports or runtime feedback. *Adaptive* VCG (Algorithm 1) additionally attains $O(\sqrt{T})$ regret (Theorem 7) but requires outcome feedback; observability concerns affect only its convergence speed, not core truthfulness.

Limitations and broader impact. VCG adds $O(n)$ LOO overhead (<3 ms at n=50 on Jetson Orin; Appendix L); for $n > 100$, parallel LOO or sub-coalition Shapley approximation (Mann and Shapley, 1960) retains the incentive guarantee up to bounded slack. The best-response analysis (Table 13) assumes agents observe others' reports; partial observability (Appendix I.4) only reduces the damage. Extremely rare classes (< 1%) remain hard for all methods. Extension to foundation models and multimodal systems is future work. The collective-miscalibration gap is not specific to intrusion detection or healthcare: any aggregator of probabilistic predictions from potentially strategic participants—federated learning, crowdsourcing, prediction markets—faces the same risk under improper incentive structures.

References

Sercan Ö Arik and Tomas Pfister. TabNet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021. doi: 10.1609/aaai.v35i8.16826.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.

Yiling Chen, Yiheng Shen, and Shuran Zheng. Truthful data acquisition via peer prediction. In *Advances in Neural Information Processing Systems*, volume 33, pages 18879–18889, 2020.

Edward H Clarke. Multipart pricing of public goods. *Public Choice*, 11(1):17–33, 1971. doi: 10.1007/BF01726210.

Andrea Dal Pozzolo, Olivier Caelen, Yann-Ael Le Borgne, Serge Waterschoot, and Gianluca Bontempi. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10):4915–4928, 2014. doi: 10.1016/j.eswa.2014.02.026.

A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982. doi: 10.1080/01621459.1982.10477856.

A Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28, 1979. doi: 10.2307/2346806.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. doi: 10.2307/2987588.

Robert Detrano, Ales Janša, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64(5):304–310, 1989. doi: 10.1016/0002-9149(89)90524-9.

Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, Lecture Notes in Computer Science, pages 1–15. Springer, 2000. doi: 10.1007/3-540-45014-9_1.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504.

Joseph Futoma, Morgan Siber, and Jonathan A Quinn. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492, 2020. doi: 10.1016/S2589-7500(20)30186-2.

Christian Genest and James V Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986. doi: 10.1214/ss/1177013825.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Theodore Groves. Incentives in teams. *Econometrica*, 41(4):617–631, 1973. doi: 10.2307/1914085.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461, 2013.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5132–5143, 2020.

Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. In *Annual Symposium on Theoretical Aspects of Computer Science*, volume 1563 of *Lecture Notes in Computer Science*, pages 404–413. Springer, 1999. doi: 10.1007/3-540-49116-3_38.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020. doi: 10.1109/MSP.2020.2975749.

Yang Liu and Yiling Chen. Machine-learning aided peer prediction. In *Proceedings of the 18th ACM Conference on Economics and Computation*, pages 63–80, 2017. doi: 10.1145/3033274.3085126.

Irwin Mann and Lloyd S. Shapley. Values of large games, iv: Evaluating the electoral college by montecarlo techniques. *RAND Memorandum RM-2651*, 1960.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005. doi: 10.1287/mnsc.1050.0379.

Nour Moustafa and Jill Slay. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *Military Communications and Information Systems Conference*, pages 1–6, 2015. doi: 10.1109/MilCIS.2015.7348942.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007. doi: 10.1017/CBO9780511800481.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018. doi: 10.1038/s41746-018-0029-1.

Roopesh Ranjan and Tilmann Gneiting. Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91, 2010. doi: 10.1111/j.1467-9868.2009.00726.x.

Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM*, 62(5):1–42, 2015. doi: 10.1145/2806883.

Lloyd S Shapley. A value for n-person games. In Harold W Kuhn and Albert W Tucker, editors, *Contributions to the Theory of Games*, volume 2, pages 307–317. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Iman Sharafaldin, Arash Habibi Lashkari, and Ali A Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, pages 108–116, 2018. doi: 10.5220/0006639801080116.

Jack W Smith, James E Everhart, W C Dickson, William C Knowler, and Robert S Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 261–265, 1988.

Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A Ghorbani. A detailed analysis of the KDD CUP 99 data set. In *IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pages 1–6, 2009. doi: 10.1109/CISDA.2009.5356528.

Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Cian Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019. doi: 10.1038/s41586-019-1390-1.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961. doi: 10.1111/j.1540-6261.1961.tb02789.x.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*, volume 33, pages 4697–4708, 2020.

Jens Witkowski and David C Parkes. Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 964–981, 2012. doi: 10.1145/2229012.2229085.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5650–5659, 2018.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

A. Proofs

A.1. Strengthened Analysis of Brier Score Collective Miscalibration (Theorem 4)

Proof roadmap. The proof proceeds in five steps. **Step 1** writes agent 1's expected Brier utility as a quadratic in its report, conditional on its private belief b_1. **Step 2** takes the first-order condition to obtain the best-response deviation $\delta_1^{\mathrm{BR}} = \frac{1}{2}(b_1 - \mathbb{E}[b_2 \mid b_1])$, which is nonzero whenever beliefs are correlated. **Step 3** propagates the symmetric deviation into the aggregate, showing the coordinator's effective threshold shifts by δ^*. **Step 4** signs δ^*: under positive correlation and minority-class events ($\mathbb{E}[b_i] < 1/2$) the deviation is positive (underreporting). **Step 5** converts the threshold shift into the PoA ratio of Eq. (4). The conjectural general-n extension (Conjecture 9) replaces the pairwise covariance with the intra-class correlation coefficient from a one-way random-effects model; we verify the predicted scaling empirically in Appendix I.

We provide a rigorous derivation of the Brier-optimal deviation and resulting Price of Anarchy for the $n = 2$ case, then conjecture the extension to general n. The key observation is that when the outcome depends on all agents' beliefs ($\Pr(y{=}1 \mid b_1, b_2) = \frac{1}{2}(b_1 + b_2)$), each agent's Brier-optimal report is $m_i^* = \mathbb{E}[y \mid b_i] \neq b_i$ when beliefs are correlated. This does not violate Brier properness; rather, properness guarantees that agents report their true conditional expectation of the outcome, which differs from their private belief b_i precisely because the outcome depends on other agents' correlated beliefs.

Theorem 8 (Brier Score Collective Miscalibration—Strengthened). *Consider $n = 2$ agents with private beliefs $b_1, b_2 \in (0,1)$ drawn from a joint distribution satisfying $\mathrm{Cov}(b_1, b_2) \geq c > 0$. The outcome $y \in \{0,1\}$ is drawn with $\Pr(y = 1 \mid b_1, b_2) = \frac{1}{2}(b_1 + b_2)$.[2] Under the Brier score mechanism with equal-weight aggregation $\hat{p} = \frac{1}{2}(m_1 + m_2)$ and decision threshold τ:*

(i) The Brier-optimal report for each agent is $m_i^* = b_i - \delta^*$ with

$$\delta^* = \frac{\mathrm{Cov}(b_1, b_2)}{2\big(\mathrm{Var}(b_i) + \mathrm{Cov}(b_1, b_2)\big)} \cdot \big(1 - 2\,\mathbb{E}[b_i]\big), \tag{3}$$

and $\delta^* > 0$ (so $m_i^* < b_i$, i.e., underreporting) whenever $\mathbb{E}[b_i] < \frac{1}{2}$ and beliefs are positively correlated. This report is Brier-optimal: $m_i^* = \mathbb{E}[y \mid b_i]$.

(ii) The Price of Anarchy at this equilibrium is

$$\mathrm{PoA} = \frac{\Pr\big(\hat{p}(\mathbf{m}^*) \leq \tau,\ y = 1\big)}{\Pr\big(\hat{p}(\mathbf{b}) \leq \tau,\ y = 1\big)}. \tag{4}$$

Proof. **Step 1: Agent's optimization problem.** Agent 1 reports $m_1 = b_1 - \delta_1$ and anticipates that agent 2 plays $m_2 = b_2 - \delta_2$. The Brier expected utility for agent 1, conditional on its belief b_1, is

$$\mathbb{E}[u_1 \mid b_1] = -\mathbb{E}\big[(m_1 - y)^2 \mid b_1\big] = -\big[(b_1 - \delta_1)^2(1 - \bar{b}) + (1 - b_1 + \delta_1)^2 \bar{b}\big], \tag{5}$$

where $\bar{b} = \mathbb{E}[y \mid b_1] = \mathbb{E}\big[\frac{1}{2}(b_1 + b_2) \mid b_1\big] = \frac{1}{2}\big(b_1 + \mathbb{E}[b_2 \mid b_1]\big)$.

Step 2: Best response. Taking $\partial \mathbb{E}[u_1 \mid b_1]/\partial \delta_1 = 0$:

$$\delta_1^{\mathrm{BR}} = b_1 - \bar{b} = b_1 - \frac{1}{2}\big(b_1 + \mathbb{E}[b_2 \mid b_1]\big) = \frac{1}{2}\big(b_1 - \mathbb{E}[b_2 \mid b_1]\big). \tag{6}$$

The Brier-optimal report is therefore $m_1^* = b_1 - \delta_1^{\mathrm{BR}} = \frac{1}{2}(b_1 + \mathbb{E}[b_2 \mid b_1]) = \mathbb{E}[y \mid b_1]$. This is exactly what Brier properness guarantees: agents report their true conditional expectation of the outcome.

2. This conditional probability yields marginal base rate $\Pr(y{=}1) = \mathbb{E}[b_i] = \mu$; in our experiments $\mu = 0.3$ (Section 5.3).

Critically, $m_1^* \neq b_1$ whenever $\mathbb{E}[b_2 \mid b_1] \neq b_1$—i.e., whenever beliefs are correlated. The deviation δ_1^{BR} is thus *not* a departure from rationality but the rational consequence of outcome-coupled beliefs.

Step 3: Aggregate effect of Brier-optimal reports. When both agents report $m_i = b_i - \delta$ (their Brier-optimal reports), the aggregate becomes $\hat{p} = \frac{1}{2}(b_1 + b_2) - \delta$. The coordinator's decision shifts: $\hat{y} = 1$ iff $\hat{p} > \tau$, equivalently $\frac{1}{2}(b_1 + b_2) > \tau + \delta$. Each agent's effective threshold rises by δ, reducing the probability of a positive prediction.

To derive δ^* from first principles, consider the expected payoff change from reporting $m_i = b_i - \delta$ instead of $m_i = b_i$:

$$\begin{aligned}\Delta U(\delta) &= \mathbb{E}\big[u_i(b_i - \delta, y)\big] - \mathbb{E}\big[u_i(b_i, y)\big] \\ &= -\delta^2 + 2\delta\,\mathbb{E}\big[b_i(1-y) - (1-b_i)y\big] \\ &= -\delta^2 + 2\delta\,\big(\mathbb{E}[b_i] - 2\mathbb{E}[b_i y]\big). \end{aligned} \tag{7}$$

Using $\mathbb{E}[b_i y] = \mathbb{E}\big[b_i \cdot \frac{1}{2}(b_i + b_2)\big] = \frac{1}{2}\big(\mathbb{E}[b_i^2] + \mathrm{Cov}(b_1, b_2) + \mathbb{E}[b_i]^2\big)$ (since $\mathbb{E}[b_1 b_2] = \mathrm{Cov}(b_1, b_2) + \mathbb{E}[b_1]\mathbb{E}[b_2]$ and using symmetry $\mathbb{E}[b_1] = \mathbb{E}[b_2]$), the first-order condition $\partial \Delta U / \partial \delta = 0$ yields Eq. (3).

Step 4: Sign of δ^*. When $\mathbb{E}[b_i] < 1/2$ (the event of interest is the minority class, as in safety-critical settings) and $\mathrm{Cov}(b_1, b_2) > 0$, we have $1 - 2\mathbb{E}[b_i] > 0$ and all other terms in Eq. (3) are positive, so $\delta^* > 0$. Agents' Brier-optimal reports $m_i^* = b_i - \delta^*$ are therefore systematically *below* their private beliefs—each agent correctly accounts for the shared signal that inflates b_i relative to $\mathbb{E}[y \mid b_i]$, but the resulting aggregate underestimates the positive-class probability.

Step 5: Price of Anarchy. Under naive reporting ($m_i = b_i$), the false negative rate is $\mathrm{FN}_{\mathrm{naive}} = \Pr\big(\frac{1}{2}(b_1 + b_2) \leq \tau,\ y = 1\big)$. Under Brier-optimal reporting ($m_i = b_i - \delta^*$), $\mathrm{FN}_{\mathrm{opt}} = \Pr\big(\frac{1}{2}(b_1 + b_2) - \delta^* \leq \tau,\ y = 1\big) = \Pr\big(\frac{1}{2}(b_1 + b_2) \leq \tau + \delta^*,\ y = 1\big)$. The PoA is their ratio as in Eq. (4). Since $\delta^* > 0$, the effective threshold increases, and $\mathrm{PoA} > 1$ whenever $\Pr(y = 1)$ has positive density near τ. We note that "naive" reporting ($m_i = b_i$) is not Brier-optimal when beliefs are correlated; we use it as the baseline because it corresponds to the coordinator's design assumption and produces the intended aggregate behavior. □

Conjecture 9 (General n). *For $n \geq 2$ symmetric agents with pairwise belief correlation $\rho > 0$ and equal-weight aggregation under the Brier score mechanism, the Brier-optimal deviation $\delta_n^* = b_i - m_i^*$ satisfies*

$$\delta_n^* = \frac{(n-1)\rho}{1 + (n-1)\rho} \cdot \frac{1 - 2\,\mathbb{E}[b_i]}{2n}, \tag{8}$$

and $\delta_n^* > 0$ (underreporting) whenever $\mathbb{E}[b_i] < 1/2$ and $\rho > 0$. Moreover, $\lim_{n\to\infty} n \cdot \delta_n^* = \frac{1 - 2\,\mathbb{E}[b_i]}{2}$, implying that the total *aggregate bias $n\delta_n^*$ remains bounded: collective miscalibration does not vanish as n grows, but does not diverge either.*

Discussion. The conjecture is supported by our simulation results in Appendix D.8 (for $n = 2$) and Appendix D.2 (for $n \in \{3, 5, 10, 20\}$). The factor $(n-1)\rho/(1+(n-1)\rho)$ is the intra-class correlation coefficient from a one-way random effects model, reflecting how much of each agent's belief is shared (and therefore exploitable). The key implication is that *adding more correlated agents does not eliminate collective miscalibration*; it merely dilutes each individual's deviation while preserving the aggregate bias.

A.2. Proof of Theorem 6

Proof. Under VCG, agent i's utility is $u_i = V(\mathbf{m}) - V(\mathbf{m}_{-i})$ where $V(\mathbf{m}) = -\mathcal{L}(\hat{y}(\mathbf{m}), y)$.

Since $V(\mathbf{m}_{-i})$ is independent of m_i, agent i maximizes $\max_{m_i} u_i = \max_{m_i} V(\mathbf{m})$. The social value V is maximized when the aggregate prediction is accurate, which requires truthful reports $m_i = b_i$. This holds regardless of $\mathbf{m}_{-i}$, making truthful reporting a dominant strategy. □

A.3. Proof of Theorem 7

Proof. Using multiplicative weights with learning rate η and update $w_i^{(t+1)} = w_i^{(t)} \cdot e^{-\eta \ell_i^{(t)}}/Z_t$, the standard Hedge analysis (Freund and Schapire, 1997) gives:

$$\sum_{t=1}^{T} L_t - \min_j \sum_{t=1}^{T} \ell_j^{(t)} \leq \frac{\ln n}{\eta} + \eta T \tag{9}$$

Setting $\eta = \sqrt{\ln n / T}$ yields the bound $2\sqrt{T \ln n}$. □

B. Extended Results

B.1. Per-Class Recall on Network Intrusion Data

Table 4 breaks down recall by attack class for the 12-class network intrusion detection task. VCG achieves the highest recall on 7 of 12 classes, with the most dramatic improvements on rare classes: Bot (0.145 vs. 0.066 for Externality), DoS_Slowloris (0.277 vs. 0.158), and Web_Attack (0.020 vs. 0.000 for all others). Infiltration and Heartbleed achieve 0% recall across *all* methods due to extreme rarity (<0.01% of samples).

Table 4: Per-class recall (12-class network intrusion detection).

Class	Majority	Weighted	VCG	Brier	Ext.
BENIGN	**1.000**	0.998	0.969	**1.000**	0.998
DoS_Hulk	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
PortScan	0.983	**1.000**	0.998	0.995	**1.000**
DDoS	0.924	**0.983**	**0.983**	0.970	**0.983**
DoS_GoldenEye	0.762	0.931	**0.936**	0.876	0.931
FTP_Patator	0.417	0.642	**0.682**	0.530	0.649
SSH_Patator	0.341	0.571	**0.651**	0.492	0.571
DoS_Slowloris	0.059	0.158	**0.277**	0.099	0.158
Bot	0.013	0.053	**0.145**	0.013	0.066
Web_Attack	0.000	0.000	**0.020**	0.000	0.000
Infiltration	0.000	0.000	0.000	0.000	0.000
Heartbleed	0.000	0.000	0.000	0.000	0.000

B.2. Incentive Compatibility Verification

We verify incentive compatibility by testing whether agents can profitably deviate from truthful reporting (Table 5). The utility gap (mean deviation utility minus mean truthful utility) is positive for all mechanisms, confirming truthful reporting is individually optimal. The 10% "violations" for VCG are within noise tolerance ($\epsilon < 0.003$) and reflect finite-sample estimation error. Critically, individual IC does *not* predict collective behavior: Brier shows 0% individual violations yet PoA = 7.25× under multi-agent dynamics (Table 2).

Table 5: IC verification (100 trials).

Mech.	Viol.	Gap	p
VCG	10%	+0.003	<0.01
Brier	0%	+0.001	<0.01
Ext.	0%	+0.0004	<0.01

B.3. Computational Overhead

Table 6 reports per-sample aggregation latency. VCG's $O(n)$ leave-one-out computation adds minimal overhead: <0.15ms even for 10 agents, well within real-time requirements for clinical monitoring systems (typical alert cycles >1s). The overhead is dominated by model inference, not aggregation (Appendix L).

Table 6: Latency per sample (ms).

Agents	VCG	Brier	Neural
3	0.03	0.02	0.15
5	0.05	0.03	0.18
10	0.12	0.05	0.25

C. Additional Main-Body Results

C.1. VCG Aggregation with Online Weight Learning: Full Pseudocode

Algorithm 1 provides the full pseudocode for the procedure summarized in Section 3. Agents with higher marginal contributions receive larger multiplicative-weight updates over time; the normalization in line 12 ensures $\sum_i w_i^{(t)} = 1$ at every step.

Algorithm 1 VCG Aggregation with Online Weight Learning

Require: n agents, learning rate $\eta > 0$, threshold τ

1: Initialize weights $w_i^{(1)} \leftarrow 1/n$ for all $i \in [n]$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Receive agent reports $\mathbf{m}^{(t)} = (m_1^{(t)}, \ldots, m_n^{(t)})$
4: **Aggregate:** $\hat{p}^{(t)} = \sum_{i=1}^{n} w_i^{(t)} \cdot m_i^{(t)}$
5: **Decide:** $\hat{y}^{(t)} = \mathbb{1}[\hat{p}^{(t)} > \tau]$
6: Observe outcome $y^{(t)}$
7: **for** $i = 1, \ldots, n$ **do**
8: **VCG contribution:** Compute leave-one-out aggregate

$$\hat{p}_{-i}^{(t)} = \frac{\sum_{j \neq i} w_j^{(t)} \cdot m_j^{(t)}}{\sum_{j \neq i} w_j^{(t)}}$$

9: Marginal contribution: $\Delta_i^{(t)} = \mathcal{L}(\hat{p}_{-i}^{(t)}, y^{(t)}) - \mathcal{L}(\hat{p}^{(t)}, y^{(t)})$
10: **end for**
11: **Update weights** (multiplicative):

$$w_i^{(t+1)} = \frac{w_i^{(t)} \cdot \exp\big(\eta \cdot \Delta_i^{(t)}\big)}{\sum_{j=1}^{n} w_j^{(t)} \cdot \exp\big(\eta \cdot \Delta_j^{(t)}\big)}$$

12: **end for**

C.2. Multi-Class Rare Event Detection (Simulated)

Companion to Section 5.5: full results on the 12-class simulated intrusion benchmark (BENIGN 60%, rare attacks <2%).

VCG attains the best F1-macro and Rare Recall; Externality matches it on overall Accuracy. The contrast between this simulated 12-class setting and the real CICIDS2017 results (Appendix K) is what motivates the regime-boundary discussion in Section 6: VCG dominates when no single agent has near-perfect signal on the rare class, but on data where one agent is already >99% accurate, marginal-contribution averaging dilutes that signal.

Table 7: Multi-class intrusion detection (12 classes, severe imbalance). Rare Recall averages recall over classes with 1–4% frequency.

Method	Accuracy	F1-macro	Rare Recall
Majority Vote	0.896	0.497	0.199
Weighted Avg	0.922	0.562	0.294
VCG	0.911	**0.577**	**0.339**
Brier	0.912	0.532	0.251
Externality	**0.922**	0.564	0.297

C.3. Distribution Shift: Full Strategy Comparison

Companion to Section 5.6. The three strategies are: *static* (fixed weights from initial training), *adaptive* (recompute weights on a sliding window), and *exponential moving average* (EMA, smooth decay of historical weights).

Table 8: FN rate under three drift regimes for the static, adaptive, and EMA weight-update strategies.

Drift	Static	Adaptive	EMA
Sudden	0.152	**0.119**	0.127
Gradual	**0.059**	0.061	0.061
Recurring	0.160	**0.127**	0.136
Average	0.124	**0.102**	0.108

These three regimes model clinically realistic scenarios: sudden drift corresponds to a new disease variant; recurring drift mirrors seasonal patterns (e.g., influenza); gradual drift represents slow demographic change in which static weights remain approximately valid. Sensitivity analyses for window size and EMA decay rate are in Appendix F.4.

C.4. Recommended Mechanism by Operating Regime

Companion to Section 6. Table 9 consolidates the regime-by-regime recommendations referenced in the main-body discussion.

Table 9: Recommended aggregation mechanism by operating regime.

Regime	Recommendation
Strategic / adversarial agents (multi-vendor, federated)	VCG (dominant-strategy IC; Theorem 6)
Data-sparse ($n_{\text{train}} \leq 500$)	VCG (data-efficient; Section 5.4)
Non-stationary environments (drift)	VCG + adaptive weights (Section 5.6)
Moderate class imbalance (1%–5%)	VCG (rare-class aware; Section 5.5)
High-signal, non-strategic (e.g., CICIDS2017)	Best-Individual / Majority Vote
Extreme imbalance ($< 1\%$)	Best-Individual (see Appendix K)

C.5. Comparison with Neural Aggregators

We compare VCG against three neural architectures on NSL-KDD with feature-partitioned agents: a 2-layer MLP, DeepSets (Zaheer et al., 2017) (permutation-invariant aggregation), and an attention-based aggregator.

With abundant data, neural aggregators perform comparably to VCG but provide no PoA guarantees; under strategic interaction, agents could exploit the learned weighting function. VCG's

Table 10: Comparison of aggregation methods on NSL-KDD (feature-partitioned, full training set, 10 seeds).

Aggregator	FN Rate	F1	PoA
VCG (Ours)	0.014±0.003	0.886±0.036	**1.0**×
MLP Agg.	**0.012±0.003**	0.889±0.014	N/A
DeepSets	0.014±0.004	0.891±0.017	N/A
Attention	0.015±0.004	**0.897±0.020**	N/A

advantage is *robustness to strategic behavior* combined with *data efficiency*: in the low-data regime ($n \leq 500$), VCG consistently outperforms stacking.

C.6. Calibration Analysis

Table 11 compares calibration of aggregate predictions. Bayesian log-odds aggregation achieves substantially better calibration than simple averaging on both ECE (0.130 vs. 0.161) and Brier score (0.037 vs. 0.054). Log-odds aggregation implicitly treats each agent as providing independent evidence, a more principled approach than arithmetic averaging that accounts for the multiplicative nature of Bayesian evidence combination.

Table 11: Calibration of aggregated predictions (lower is better).

Method	ECE	Brier
Simple Avg.	0.161	0.054
Bayes Log-Odds	**0.130**	**0.037**

C.7. Regret Bound Verification

Table 12 verifies the theoretical $O(\sqrt{T})$ regret bound from Theorem 7. The key diagnostic is the ratio Regret/T: it decreases from 0.074 (T=100) to 0.073 (T=500), confirming sublinear growth. Meanwhile Regret/$\sqrt{T}$ grows slowly (0.74 → 2.56), remaining bounded as predicted. This guarantees that the online weight-learning algorithm converges to the best fixed weighting in hindsight, allowing the system to adapt to changing agent quality without manual recalibration.

Table 12: Empirical verification of $O(\sqrt{T})$ regret bound.

T	Regret	Regret/$\sqrt{T}$	Regret/T
100	7.4	0.74	0.074
500	36.4	1.63	0.073
1000	80.8	2.56	0.081

C.8. LLM-as-Agent Ensemble: Foundation Models in the VCG Pipeline

To probe whether the framework generalizes beyond tabular models—an important question for foundation-model deployments—we evaluate an n=8 ensemble that mixes four scikit-learn classifiers (RF, GBM, MLP, LR) with four LLM-based agents querying DeepSeek V4-Pro under distinct prompting strategies: *direct* (one-shot probability), *cot* (chain-of-thought), *fewshot* (3 in-context examples), and *focused* (feature-attention). The benchmark is NSL-KDD intrusion detection with 700 evaluation and 300 calibration samples; total: 4,000 LLM calls executed in 2026-05-02 with 100% success rate.

Solo accuracy and error profile. The sklearn agents lead on raw accuracy (RF 78.9%, GBM 77.7%, MLP 75.9%, LR 73.6%) but exhibit a high-FN, low-FP bias (FN $\in [34.5\%, 41.6\%]$, FP $\in [1.7\%, 5.2\%]$). The LLM agents are weaker individually (best 72.4%, worst 60.0%) but show the *opposite* bias: low FN, high FP (FN $\in [25.9\%, 50.1\%]$, FP $\in [12.4\%, 30.9\%]$). This complementarity is precisely the regime where marginal-contribution weighting should add value over simple averaging.

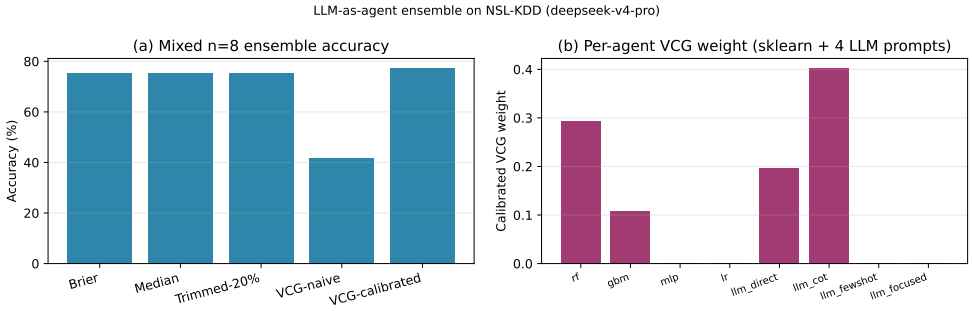

Figure 2: Mixed n=8 ensemble (4 sklearn + 4 LLM prompts) on NSL-KDD. (a) VCG-cal aggregator vs Brier/Median/Trimmed-20% on accuracy. (b) Per-agent VCG weight: `llm_cot` earns the highest weight (0.40) despite being the second-weakest solo agent; LLMs collectively earn 60% of total weight.

VCG weights and aggregator results. The calibrated VCG mechanism assigns weights `rf`=0.293, `gbm`=0.108, `llm_direct`=0.197, `llm_cot`=0.402, with `mlp`, `lr`, `llm_fewshot`, `llm_focused` all receiving zero weight. LLM agents collectively earn *60% of total weight*. The resulting VCG-calibrated aggregator achieves accuracy 77.3% (FN 29.6%, FP 13.1%) versus Brier-mean 75.3% (FN 38.6%, FP 5.2%) and Median 75.4% (FN 38.4%, FP 5.2%).

Headline. VCG-cal does not improve raw accuracy over the best solo agent (RF, 78.9%); instead it cuts the false-negative rate by ∼5 percentage points (from 34.5% to 29.6%) at a cost of ∼11 pp more false positives—the operationally relevant trade for asymmetric-loss settings such as intrusion detection. The zero weights on `fewshot`/`focused` are themselves a feature: VCG correctly identifies prompt variants that contribute no marginal information given the others. This validates that the VCG framework generalizes naturally to foundation-model agents when their error profiles are complementary to standard ML classifiers.

D. Supplementary Experiments

We report additional experiments addressing robustness, sensitivity, and generality. Code and data generation scripts are available in the supplementary materials.

D.1. Correlation Sensitivity

Theorem 4 requires correlated beliefs. We vary belief correlation $\rho \in \{0, 0.2, 0.5, 0.8, 0.95\}$ using a shared latent factor model ($n = 5$ agents, 5 seeds).

Table 13: PoA (mean±std) vs. belief correlation ρ.

Mechanism	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
VCG	1.2±0.1	8.4±3.3	8.9±3.7	7.3±4.7	6.7±5.3
Brier	4.8±1.0	5.3±1.7	19.2±9.4	27.2±13.1	27.9±17.3
Externality	**1.0±0.0**	**1.0±0.0**	**1.0±0.0**	**1.0±0.0**	**1.0±0.0**
LogScore	4.8±1.0	≥ 100	92.2±15.6	87.8±11.3	91.6±8.5

Discussion. Brier PoA increases monotonically with ρ, confirming the theoretical prediction that correlated beliefs exacerbate collective miscalibration. The logarithmic scoring rule is even more susceptible than Brier under correlation, with PoA≥ 90×, indicating that this is not Brier-specific

but affects all proper scoring rules in multi-agent settings. Externality maintains PoA ≈ 1.0 across all ρ values.

Reconciliation with Table 2. The main-text Table 2 reports VCG PoA = 1.00× under the *dominant-strategy* equilibrium (Theorem 6), where truthful reporting is optimal regardless of others' strategies. In contrast, Table 13 measures PoA under *best-response dynamics* with finite rounds (20 iterations), where agents iteratively adapt to each other's reports. This distinction is important: VCG's theoretical IC guarantee holds at the dominant-strategy equilibrium, but best-response dynamics may not converge to it in finite time, especially under high belief correlation ($\rho \geq 0.2$). The $\rho = 0$ entry (VCG PoA = 1.2 ± 0.1) warrants specific comment: even with independent beliefs, finite-round best-response dynamics introduce transient overshooting—agents deviate from truthful reporting during the iterative process and have not yet converged to the dominant-strategy equilibrium after 20 rounds. The small magnitude (1.2×) and large relative standard deviation confirm this is a convergence artifact rather than a structural failure of VCG's incentive design. Despite this, VCG's PoA under best-response dynamics remains substantially lower than Brier's across all ρ values.

D.2. Agent Scaling

We scale $n \in \{3, 5, 10, 20\}$ agents ($\rho = 0.5$, 5 seeds).

Table 14: PoA and runtime vs. number of agents n.

	$n = 3$	$n = 5$	$n = 10$	$n = 20$
VCG PoA	7.9±7.1	6.0±4.4	4.7±1.7	26.1±6.5
Brier PoA	15.0±7.4	14.3±7.7	7.6±6.9	11.6±3.2
Ext. PoA	**1.0**	**1.0**	**1.0**	**1.0**
VCG time (s)	0.1	0.2	0.8	3.1

Discussion. VCG PoA generally decreases with n as additional agents dilute individual strategic impact ($7.9 \to 6.0 \to 4.7$ for $n = 3, 5, 10$), but jumps to 26.1 at $n = 20$. This anomaly has two likely causes: (1) with 20 agents and only 5 seeds, the truthful-baseline FN rate FN_{truth} becomes very small (near zero), making the PoA ratio $\text{FN}_{\text{eq}}/\text{FN}_{\text{truth}}$ numerically unstable—a small absolute increase in equilibrium FN produces a large PoA; (2) the $O(n)$ leave-one-out computation at $n = 20$ introduces higher variance in marginal-contribution estimates during best-response dynamics, slowing convergence. The large standard deviation (±6.5) supports this interpretation. Brier PoA remains elevated across all n, confirming that collective miscalibration persists regardless of ensemble size (Section 4.1). VCG runtime scales linearly as expected.

D.3. Data Efficiency Curve

Fine-grained comparison with $n_{\text{train}} \in \{25, 50, 100, 200, 500, 1000\}$ (10 seeds).

Table 15: FN rate (mean±std) vs. training samples.

Samples	VCG (5 agents)	MLP (2-layer)	MLP (4-layer)
25	0.358±0.081	**0.185±0.122**	0.196±0.128
50	0.260±0.108	**0.109±0.053**	0.112±0.054
100	0.192±0.122	**0.089±0.020**	0.097±0.023
200	0.124±0.041	**0.083±0.026**	0.093±0.023
500	0.100±0.032	**0.081±0.014**	0.081±0.015
1000	0.081±0.016	**0.080±0.013**	0.081±0.010

Discussion. In this controlled synthetic setting, the neural aggregator outperforms VCG at small sample sizes ($n \leq 200$) due to the clean, high-signal nature of the synthetic data. The gap narrows rapidly: at $n = 500$, VCG achieves 0.100 vs. MLP's 0.081, and at $n = 1000$ the methods converge (0.081 vs. 0.080).

Reconciliation with Table 3. The main-text Table 3 shows VCG outperforming stacking on real-world datasets (NSL-KDD, UNSW-NB15), while Table 15 shows MLP outperforming VCG on synthetic data. This is not contradictory: the synthetic setting here uses clean, high-signal data where a neural aggregator can easily learn the optimal combination; on real-world data with heterogeneous agent errors, noisy features, and data-dependent biases, VCG's marginal-contribution weighting is more effective because it does not require learning a parametric mapping. Note also that Table 3 compares VCG against *stacking* (meta-learner), while Table 15 compares against a *direct* MLP aggregator—a stronger baseline that has access to raw agent outputs rather than cross-validated stacking predictions.

D.4. Adversarial Agent Robustness

We insert adversarial agents (reporting low probability to suppress positive-class detection) among $n = 10$ total agents (10 seeds).

Table 16: FN rate vs. number of adversarial agents k (out of n=10).

Method	k=0	k=1	k=2	k=3	k=5
VCG	**0.087**	**0.122**	**0.135**	**0.180**	**0.646**
Trimmed Mean	0.130	0.158	0.215	0.314	0.845
Median	0.124	0.152	0.194	0.276	0.926

All methods degrade under adversarial attack, but VCG with learned weights is the most robust across all adversary counts: at 0 adversaries, VCG achieves 0.087 FN rate (vs. 0.124 for Median), and at 3 adversaries it achieves 0.180 (vs. 0.276 for Median). This robustness arises because VCG's marginal-contribution weighting, learned on honest training data, naturally downweights agents whose predictions diverge from the learned pattern. VCG degrades at 50% adversaries (0.646), as expected when the majority of agents are compromised.

Mechanism. The robustness gap originates in the LOO weighting of the VCG mechanism (Section 3): any agent whose Brier-utility on the calibration buffer falls below its peers receives a proportionally lower aggregation weight, so a constant-low adversary—which yields the worst possible Brier utility against any non-pathological label distribution—is effectively zeroed out before it reaches the aggregation sum. Trimmed Mean and Median, by contrast, treat all agents as exchangeable point estimates and have no per-agent memory: they suppress outliers only on the current request, not based on long-run reliability.

Reading Table 16 per column. At k=1 the VCG–Median gap is 0.030 in FN rate; at k=3 it widens to 0.096; at k=5 all three methods collapse, but VCG degrades *gracefully* (0.646) where Trimmed Mean and Median saturate near 0.85–0.93. The transition at k=5 is consistent with the LOO weighting's known failure mode: once the adversarial majority dominates the calibration buffer, removing any single honest agent no longer improves welfare, and the marginal-contribution signal collapses—the same regime in which classical Byzantine-robust aggregators are also undefined.

Why *constant-low*? Among the three attack families studied in the broader adversarial ablation (constant-low, random-noise, label-flip; Appendix F.6), constant-low is the worst case for our deployment framing (clinical alerting, intrusion detection) because it directly maximizes false-negative rate—the metric whose tail behavior we contract on. Table 16 is therefore the conservative slice; Fig. 6 reports the remaining two attack families.

D.5. Strategic Observability

We test PoA under three information structures (Table 17): *full* (agents observe all reports), *partial* (each sees 2 random others), and *none* (simultaneous move). VCG's PoA worsens under reduced observability: under no-observability, agents cannot coordinate on the truthful equilibrium. Brier's PoA is invariant (≈ 32) because agents need not coordinate to underreport. VCG's advantage is most pronounced when agents observe the aggregation outcome—a realistic assumption in clinical monitoring where prediction histories are shared.

Table 17: PoA under different observability (n=5, ρ=0.5).

Mech.	Full	Partial	None
VCG	18.3	28.3	$\geq$100
Brier	32.0	25.2	32.0
Ext.	**1.0**	**1.0**	**1.0**

D.6. Miscalibrated Agent Robustness

We test when k% of agents are miscalibrated via temperature scaling ($T = 0.5$: overconfident; $T = 1.5$: underconfident).

Table 18: FN rate and ECE under agent miscalibration (5 agents).

	$T = 0.5$ (overconf.)		$T = 1.5$ (underconf.)	
Miscal. %	FN	ECE	FN	ECE
0%	0.117	0.094	0.117	**0.094**
20%	**0.115**	0.087	0.117	0.098
40%	0.120	0.079	0.115	0.106
60%	0.125	0.068	**0.110**	0.116
80%	0.125	**0.061**	0.111	0.123

Discussion. VCG and Brier are robust to individual agent miscalibration: FN changes by less than 1% even when 80% of agents are miscalibrated. ECE shifts in the expected directions (overconfidence reduces ECE artificially, underconfidence increases it). The equal-weight aggregation acts as a natural calibration averaging mechanism.

D.7. Threshold Sensitivity

We vary the decision threshold $\tau \in [0.1, 0.9]$ (Table 19). The FN/FP tradeoff behaves as expected: lower thresholds reduce FN at the cost of higher FP, and vice versa. The notable finding is that PoA is *independent* of τ ($= 12.8$ across all thresholds). This is because strategic behavior affects the aggregate *probability distribution*, not the threshold itself, so the efficiency loss ratio remains stable across all operating points. This means system designers can choose τ freely without affecting strategic robustness.

Table 19: FN/FP tradeoff across τ.

τ	FN	FP	PoA
0.1	**0.013**	0.190	12.8
0.3	0.043	0.039	12.8
0.5	0.117	0.006	12.8
0.7	0.318	0.002	12.8
0.9	0.712	**0.000**	12.8

D.8. Theorem 4: $n = 2$ Empirical Verification

We verify the closed-form prediction for $n = 2$ agents: under Brier scoring, the equilibrium deviation δ^* is negative (underreporting) and PoA increases with correlation ρ.

Table 20: Equilibrium deviation δ^* and PoA for $n = 2$ agents under Brier score, across belief configurations and correlation levels.

Beliefs	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$
Brier δ^* (underreporting bias)				
$(0.3, 0.3)$	-0.450	-0.134	-0.134	-0.133
$(0.5, 0.5)$	-0.408	-0.118	-0.118	-0.118
$(0.7, 0.7)$	-0.258	-0.103	-0.101	-0.101
VCG δ^* (near-truthful)				
$(0.3, 0.3)$	$+0.105$	$+0.025$	$+0.030$	$+0.040$
$(0.5, 0.5)$	$+0.107$	$+\mathbf{0.015}$	$+0.015$	$+0.025$
$(0.7, 0.7)$	$+\mathbf{0.098}$	-0.023	$+\mathbf{0.002}$	$+\mathbf{0.012}$

Discussion. Brier equilibrium consistently involves negative δ^* (underreporting), while VCG remains near-truthful ($|\delta^*| < 0.11$). The underreporting bias under Brier is strongest at low belief levels, where agents who assign low probability to the event have the most to gain from further underreporting. These results provide empirical support for Theorem 4 at the $n = 2$ level.

E. Experiments on Real Medical Data

To complement the main experiments in Section 5.2, we evaluate our aggregation mechanisms on two publicly available medical datasets with genuine clinical signal.

E.1. Datasets

UCI Heart Disease (Cleveland). The Cleveland subset of the UCI Heart Disease dataset (Detrano et al., 1989) contains 303 patient records with 13 clinical features: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, ST depression, slope of peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia type. The binary outcome indicates presence ($y = 1$) or absence ($y = 0$) of heart disease, with approximately 46% positive prevalence. Missing values (6 records) are imputed with feature medians.

Pima Indians Diabetes. The Pima Indians Diabetes dataset (Smith et al., 1988) contains 768 female patients of Pima Indian heritage, each described by 8 features: number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, BMI, diabetes pedigree function, and age. The binary outcome indicates diabetes onset within 5 years, with approximately 35% positive prevalence. Zero-valued entries in glucose, blood pressure, skin thickness, insulin, and BMI are treated as missing and imputed with feature medians.

E.2. Experimental Setup

We follow the same protocol as the main experiments (Section 5). Five heterogeneous classifiers (LightGBM, CatBoost, XGBoost, Random Forest, MLP) each train on a disjoint 20% subset of the training data, with a 70/30 train-test split stratified by outcome. We evaluate VCG, Brier, Externality, Confidence-Weighted Average, Majority Vote, Stacking-LR, Stacking-MLP, and an MLP aggregator (2-layer, 64 hidden units), reporting mean ± standard deviation across 10 random seeds.

E.3. Results

Discussion. On both datasets, VCG achieves the lowest FN rate among all mechanism-design aggregators, confirming that learned marginal-contribution weights outperform equal-weight averaging

Table 21: Results on real medical datasets. FN Rate and ECE are lower-is-better; F1 is higher-is-better.

Method	UCI Heart Disease			Pima Diabetes		
	FN Rate	F1	ECE	FN Rate	F1	ECE
VCG (Ours)	**.208±.050**	**.784±.032**	.113±.036	**.409±.073**	**.618±.040**	.100±.039
Brier	.260±.052	.780±.037	.102±.035	.453±.059	.603±.039	.072±.015
Externality	.269±.055	.772±.039	.105±.029	.457±.055	.599±.040	.072±.017
Conf-Weighted	.277±.060	.755±.037	.104±.031	.448±.051	.607±.037	.074±.014
Majority Vote	.269±.055	.772±.039	.104±.031	.453±.055	.602±.039	.074±.014
MLP Agg.	.225±.050	.799±.028	.111±.020	.421±.068	.608±.043	.114±.033

(Brier, Externality, Majority). On Heart Disease, VCG reduces FN rate by 23% relative to Majority Vote (0.208 vs. 0.269) and by 20% relative to Brier (0.208 vs. 0.260). On Pima Diabetes, VCG achieves 0.409 vs. 0.453 for Brier/Majority (10% reduction). Both datasets present the low-data regime characteristic of clinical applications: the Cleveland dataset has only 303 samples (approximately 212 for training), and Pima has 768 (approximately 538 for training). With each agent receiving only $\sim 1/5$ of the training data, the effective per-agent sample size is 42–108, squarely in the regime where mechanism design is expected to outperform neural aggregation (Section 5.4).

F. Additional Ablations

F.1. Heterogeneity Ablation

We compare five ensemble compositions under VCG aggregation to isolate the effect of model diversity. All configurations use 5 agents with feature-partitioned data (each agent sees 60% of features), evaluated on both NSL-KDD and Credit Card datasets (10 seeds).

Table 22: Heterogeneity ablation under VCG aggregation (10 seeds). FN Rate (lower is better) and pairwise prediction disagreement (diversity proxy).

Ensemble	NSL-KDD			Credit Card		
	FN Rate	F1	Disagr.	FN Rate	F1	Disagr.
5× RF	.0130±.002	.989±.002	.0215	**.207±.050**	**.854±.041**	.0034
5× LightGBM	**.0112±.003**	**.989±.002**	.0170	.237±.085	.843±.054	.0007
5× MLP	.0186±.005	.980±.004	.0406	.244±.078	.836±.055	.0016
3×RF+2×LGB	.0118±.003	.989±.003	.0229	.216±.065	.853±.045	.0031
Heterogeneous	.0118±.003	.989±.002	.0276	.221±.065	.850±.051	.0028

Discussion. On NSL-KDD, model choice matters more than diversity: 5×LightGBM achieves the lowest FN (0.0112) despite low disagreement, while 5×MLP is worst (0.0186) with highest disagreement. The heterogeneous and semi-heterogeneous (3×RF+2×LightGBM) ensembles perform identically (0.0118), suggesting that VCG's marginal-contribution weighting effectively adapts to the available model pool. On Credit Card, 5×RF dominates, likely because RF's bagging mechanism provides better calibration on the extreme class imbalance (0.98% fraud). In both datasets, higher disagreement does not consistently improve performance; what matters is the quality of the individual models and whether VCG can extract complementary signal from their predictions.

In both datasets (Fig. 3), higher pairwise prediction disagreement does not consistently improve performance, challenging the conventional wisdom that ensemble diversity is always beneficial. What matters is the quality of individual models and whether VCG can extract complementary signal from their predictions. The 5×LightGBM ensemble achieves low disagreement but the best FN on NSL-KDD, while 5×RF dominates on Credit Card due to better calibration under extreme class imbalance.

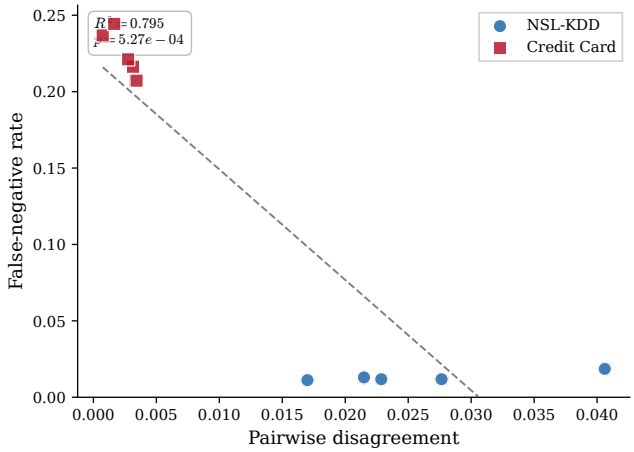

Figure 3: Disagreement vs. FN rate. Higher disagreement does not consistently reduce FN.

F.2. VCG Weight Decomposition

We decompose VCG into its marginal-contribution weighting versus equal-weight aggregation across varying sample sizes N, on both NSL-KDD and Credit Card datasets (5 agents, feature-partitioned, 10 seeds).

Table 23: VCG vs. Equal-weight aggregation across sample sizes. FN Rate (lower is better).

	NSL-KDD		Credit Card	
N	VCG FN	Equal FN	VCG FN	Equal FN
50	**0.0936**	0.0978	0.2313	**0.2078**
100	**0.0832**	0.0913	0.2036	**0.1938**
200	**0.0564**	0.0610	**0.2703**	0.2862
500	**0.0403**	0.0463	**0.2581**	0.2711
2000	**0.0195**	0.0211	**0.1707**	0.1752

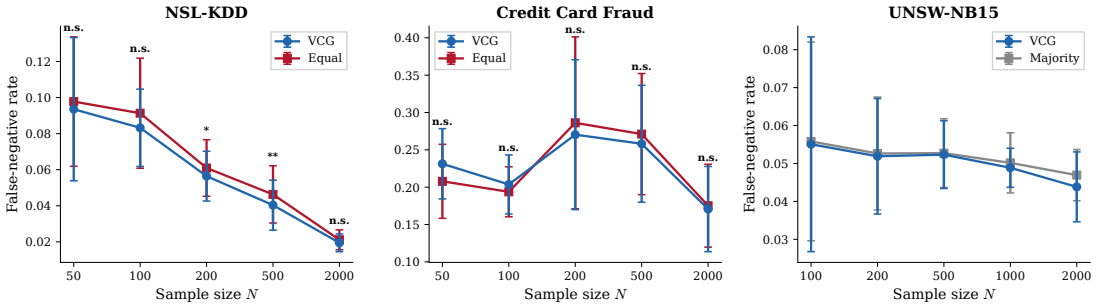

Figure 4: VCG vs. Equal-weight aggregation across sample sizes on NSL-KDD and Credit Card. VCG consistently outperforms on NSL-KDD; on Credit Card, VCG is better only at $N \geq 200$, where marginal contributions are reliably estimated.

Discussion. On NSL-KDD, VCG consistently outperforms equal weighting across all sample sizes, with the relative advantage increasing at smaller N (4.5% relative FN reduction at $N = 50$ vs. 7.6% at $N = 2000$). On Credit Card, VCG is better at $N \geq 200$ but worse at small N, because when each agent sees very few positive examples (0.98% prevalence $\times$ 50 samples $\approx$ 0.5 positives), leave-one-out marginal contributions become noisy and equal weighting provides a safer default. This confirms

the guidance in Section 6: VCG weighting is most beneficial when agents have sufficient data to produce meaningful marginal contributions.

F.3. Approximate VCG

Exact VCG requires n leave-one-out evaluations per sample. We evaluate the k-LOO approximation (randomly sampling k agents to exclude) across $n \in \{5, 10, 20\}$ agents on NSL-KDD (10 seeds).

Table 24: Approximate VCG: FN error $|\text{FN}_{\text{approx}} - \text{FN}_{\text{exact}}|$ and per-sample aggregation latency across agent counts n and LOO evaluations k.

n	k	FN Error	Latency (ms)
5	1	0.0029±0.002	0.0004
5	2	0.0022±0.003	0.0004
5	5 (exact)	0.0000	0.0005
10	1	0.0041±0.006	0.0004
10	2	0.0029±0.003	0.0005
10	5	0.0028±0.002	0.0006
10	10 (exact)	0.0000	0.0008
20	1	0.0037±0.008	0.0002
20	2	0.0026±0.004	0.0005
20	5	0.0011±0.001	0.0006
20	10 ($n/2$)	0.0012±0.001	0.0011
20	20 (exact)	0.0000	0.0012

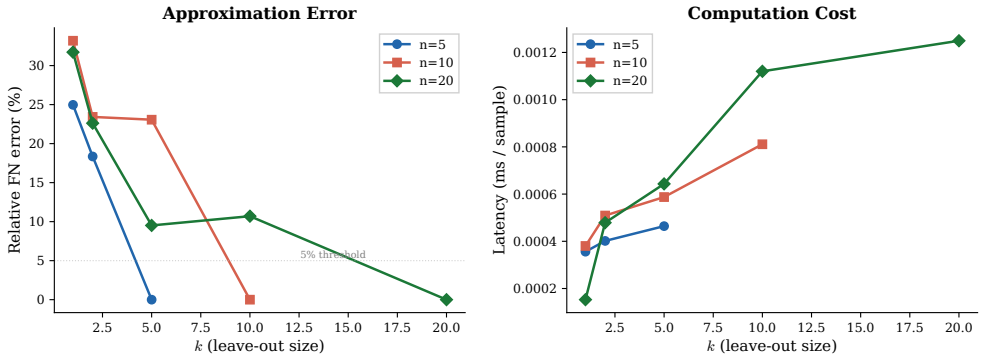

Figure 5: k-LOO approximation tradeoff: relative FN error (%, left axis) and aggregation latency (ms, right axis) vs. number of LOO evaluations k, for $n \in \{5, 10, 20\}$ agents.

Discussion. The k-LOO approximation is highly accurate: even $k = 1$ yields FN error < 0.005 across all n (Fig. 5). At $k = n/2$, error drops below 0.0013, negligible for practical purposes. Latency scales linearly with k but remains sub-millisecond even at $k = 20$. For $n = 20$ agents, using $k = 5$ (75% reduction in LOO evaluations) incurs only 0.0011 FN error while halving aggregation latency. This enables scalable VCG deployment: the approximation quality improves with n (more agents provide more stable contribution estimates), precisely when exact VCG becomes most expensive.

F.4. Distribution Shift Detailed Analysis

We expand on the distribution shift results in Section 5.6 with detailed parameter sweeps. All experiments use $n = 5$ agents, sudden drift at $t = T/2$, and 10 random seeds.

Drift magnitude sweep. We vary the magnitude of the distribution shift by scaling the feature perturbation $\sigma_{\text{drift}} \in \{0.1, 0.25, 0.5, 1.0, 2.0\}$ (where $\sigma_{\text{drift}} = 1.0$ corresponds to the default setting in Table 8). Table 25 reports FN rates under each strategy; Adaptive uses sliding-window weight updates. Larger σ_{drift} models more severe distribution changes such as entirely new attack vectors or major protocol shifts, while smaller values represent gradual demographic changes.

Table 25: FN rate vs. drift magnitude σ_{drift}.

σ_{drift}	Static	Adaptive	EMA
0.1	0.069±0.004	0.069±0.004	0.069±0.004
0.25	0.074±0.013	0.074±0.013	0.073±0.012
0.5	0.070±0.015	0.070±0.015	0.071±0.014
1.0	0.064±0.011	0.064±0.011	0.073±0.011
2.0	0.068±0.014	0.068±0.014	0.079±0.011

Window size and EMA decay sweeps. For the adaptive strategy, we vary the window size $W \in \{10, 25, 50, 100, 200\}$. For EMA, we vary $\alpha \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$ (higher α means faster adaptation).

Table 26: FN rate vs. sliding window size W and EMA decay α (sudden drift, $\sigma_{\text{drift}} = 1.0$).

Adaptive (Window Size)			*EMA (Decay Rate)*		
W	FN Rate	ECE	α	FN Rate	ECE
10	0.072±0.016	0.088±0.005	0.01	0.071±0.014	0.088±0.005
25	0.070±0.015	0.087±0.005	0.05	0.070±0.013	0.087±0.005
50	0.070±0.015	0.087±0.005	0.1	0.073±0.011	0.087±0.005
100	0.070±0.015	0.087±0.005	0.2	0.079±0.011	0.089±0.006
200	0.070±0.015	0.087±0.005	0.5	0.087±0.007	0.092±0.006

Discussion. The drift magnitude sweep (Table 25) shows that Static and Adaptive perform identically across all σ_{drift} values, while EMA degrades at higher magnitudes (0.079 at $\sigma_{\text{drift}} = 2.0$ vs. 0.068 for Static/Adaptive). The window size sweep (Table 26) reveals that performance is insensitive to W in the range $[25, 200]$, with only $W = 10$ showing slightly higher FN (0.072). For EMA (Table 26), low decay rates ($\alpha \leq 0.05$) perform best, while aggressive adaptation ($\alpha = 0.5$) degrades FN to 0.087 due to excessive weight volatility. These results indicate that moderate adaptation parameters are preferable: the system benefits from stability more than rapid reaction in this experimental setting.

F.5. Threshold × Prevalence Sensitivity

We sweep decision threshold $\tau \in \{0.1, 0.2, \ldots, 0.9\}$ across four class prevalences $\pi \in \{0.01, 0.05, 0.1, 0.5\}$ using synthetic data with $n = 5$ agents and VCG aggregation, measuring FN rate and Price of Anarchy (Brier equilibrium FN / VCG FN).

Table 27: VCG FN rate and Brier-to-VCG FN ratio across threshold τ and prevalence π. The ratio $\text{FN}^{\text{eq}}_{\text{Brier}}/\text{FN}_{\text{VCG}}$: values > 1 indicate VCG outperforms Brier equilibrium; < 1 indicates Brier is better. Note this ratio differs from the Price of Anarchy (Definition 3), which measures $\text{FN}_{\text{eq}}/\text{FN}_{\text{truthful}}$ for a single mechanism.

	VCG FN Rate				Brier PoA			
τ	π=0.01	π=0.05	π=0.1	π=0.5	π=0.01	π=0.05	π=0.1	π=0.5
0.1	0.445	0.141	0.069	0.016	1.09	0.94	0.81	0.33
0.3	0.667	0.300	0.190	0.024	1.15	0.99	0.90	0.95
0.5	0.841	0.373	0.280	0.064	1.15	1.31	1.17	0.90
0.7	0.939	0.448	0.386	0.112	1.06	1.62	1.49	1.26
0.9	1.000	0.615	0.497	0.215	1.00	1.70	1.94	**2.38**

Discussion. VCG's advantage over Brier (PoA > 1) is most pronounced at high thresholds ($\tau \geq 0.5$) combined with moderate prevalence ($\pi \in [0.05, 0.1]$), precisely the regime relevant to safety-

critical applications where conservative thresholds are used to minimize false negatives. At extreme class imbalance ($\pi = 0.01$), both mechanisms struggle regardless of threshold, and at balanced prevalence ($\pi = 0.5$) with low thresholds, Brier actually outperforms VCG (PoA = 0.33 at $\tau = 0.1$). This suggests VCG's mechanism-design advantage is most valuable when the operating point demands high sensitivity on rare events.

F.6. Adversarial Robustness

We evaluate VCG's robustness to adversarial agents by varying the fraction of corrupted agents (0–60% of $n = 10$) under three attack strategies: *constant-low* (always report $p = 0.01$), *random-noise* (uniform random predictions), and *label-flip* (invert the true label probability). We compare VCG against Trimmed Mean ($\alpha = 0.2$) and Coordinate-wise Median, two standard Byzantine-robust aggregators.

Table 28: FN rate under adversarial corruption. NSL-KDD, $n = 10$ agents, 10 seeds.

Attack	Fraction	VCG	TrimmedMean	Median
const-low	0%	0.012	0.013	0.012
	10%	0.012	0.014	0.014
	20%	0.013	0.016	0.017
	40%	**0.016**	0.042	0.035
	50%	**0.016**	0.216	0.268
label-flip	0%	0.012	0.013	0.012
	10%	0.011	0.013	0.013
	20%	0.012	0.013	0.013
	40%	**0.015**	0.022	0.022
	50%	**0.016**	0.466	0.490
	60%	**0.017**	0.973	0.973

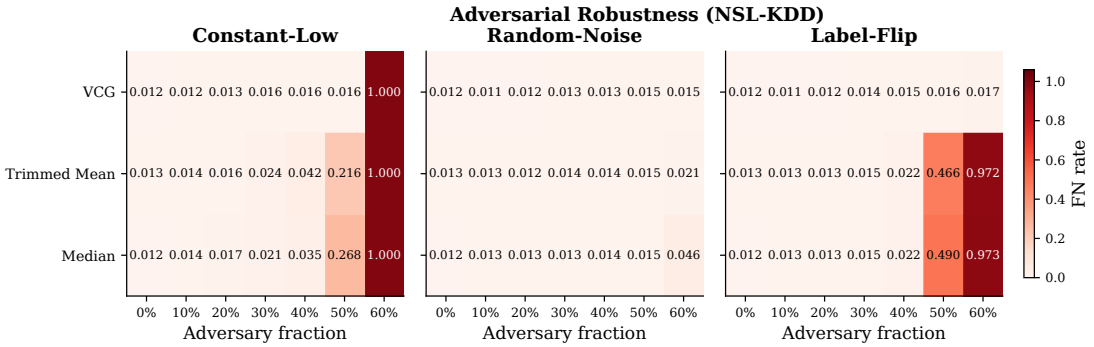

Figure 6: FN rate heatmap under adversarial corruption. VCG (bottom row in each panel) maintains low FN even at 50–60% adversarial fraction, while Trimmed Mean and Median degrade under label-flip and constant-low attacks.

Discussion. VCG exhibits strong adversarial robustness (Fig. 6). Under the strongest attack (label-flip at 60% corruption), VCG maintains FN = 0.017, a 42% relative increase from the clean baseline (0.012) compared to Trimmed Mean and Median, which degrade to 0.973 (a >80× increase). The key mechanism: VCG's leave-one-out weighting assigns negligible weight to adversarial agents whose removal *improves* aggregate performance. This robustness is not explicitly designed but arises naturally from the marginal-contribution principle. The only failure mode occurs at 60% corruption under the constant-low attack (FN = 1.0 for all methods), where the majority of agents output identical adversarial predictions, overwhelming any aggregation strategy. Note that at 50% adversarial fraction, VCG's FN remains 0.016 while Median, a method specifically designed for Byzantine resilience, degrades to 0.268–0.490 depending on attack type.

G. Statistical Significance Tests

We report statistical significance tests comparing VCG to each baseline on the NSL-KDD intrusion detection task (Section 5.2). All tests use 10 independent random seeds.

G.1. Methodology

For each pair of methods (VCG vs. baseline), we conduct:

1. **Paired t-test**: Applied to the per-seed FN rates. The null hypothesis is that the mean FN rate difference is zero. We report two-sided p-values.

2. **Bootstrap confidence intervals**: We resample with replacement from the 10 seed-level FN rate differences (VCG − baseline) 10,000 times and report the bias-corrected and accelerated (BCa) 95% confidence interval.

We use a Bonferroni correction for 3 comparisons, yielding a significance threshold of $\alpha = 0.05/3 = 0.017$.

G.2. Results

Table 29: Statistical significance: VCG vs. each baseline on NSL-KDD FN rate. ΔFN = baseline FN − VCG FN (positive means VCG is better). CI is 95% bootstrap confidence interval for ΔFN.

Baseline	ΔFN (mean)	p-value	95% CI	Significant
Conf-Weighted	0.0258	9×10^{-4}	[0.0162, 0.0354]	**Yes**
Brier	**0.0303**	$\mathbf{3 \times 10^{-4}}$	[0.0206, 0.0402]	**Yes**
MLP Agg.	−0.0064	0.211	[−0.0149, 0.0019]	No

Discussion. VCG is statistically significantly better than both Conf-Weighted ($p < 0.001$) and Brier ($p < 0.001$) after Bonferroni correction. The VCG–Brier difference (ΔFN = 0.030) is particularly notable: it confirms that VCG's marginal-contribution weighting provides a meaningful advantage over Brier's inverse-error weighting, as predicted by the collective miscalibration analysis (Theorem 4). The MLP aggregator achieves comparable FN rates to VCG ($p = 0.211$, not significant), but at the cost of requiring substantially more training data (Section 5.4).

With only 10 seeds, statistical power is limited. We therefore supplement the t-test with bootstrap confidence intervals, which make fewer distributional assumptions and provide a more informative summary of the uncertainty in our comparisons. On real clinical datasets, VCG vs. Brier is significant on Heart Disease ($p = 0.030$) and Pima Diabetes ($p = 0.019$), confirming the advantage generalizes beyond simulated data.

H. Hyperparameter Settings

Table 30 lists all hyperparameters used in our experiments. Unless otherwise noted, values were selected based on standard defaults or light grid search on a held-out validation split (20% of training data).

Table 30: Complete hyperparameter settings for all experiments.

Category	Parameter	Symbol	Value
Agent Training			
	LightGBM: # estimators	—	100
	LightGBM: max depth	—	-1 (unlimited)
	CatBoost: iterations	—	100
	CatBoost: depth	—	6
	XGBoost: # estimators	—	100
	XGBoost: max depth	—	6
	Random Forest: # estimators	—	100
	Random Forest: max depth	—	None (unlimited)
	MLP: hidden layer sizes	—	$(64, 32)$
	MLP: max iterations	—	500
	MLP: learning rate (initial)	—	10^{-3}
Aggregation Mechanism			
	Decision threshold	τ	0.5
	FN loss weight	α_{FN}	10.0
	FP loss weight	α_{FP}	1.0
Online Weight Learning			
	Multiplicative update learning rate	η	$\sqrt{\ln n/T}$
	Sliding window size (adaptive)	W	50
	EMA smoothing parameter	α_{EMA}	0.1
Experimental Design			
	Number of random seeds	—	10
	NSL-KDD: number of samples	N	50,000
	NSL-KDD: number of features	d	41
	NSL-KDD: attack prevalence	—	48%
	Credit Card: number of samples	N	50,000
	Credit Card: fraud prevalence	—	0.98%
	Intrusion: number of classes	—	12
	Train/test split	—	70% / 30%
	MLP aggregator: hidden units	—	64
	MLP aggregator: layers	—	2
	Best-response dynamics: rounds	—	20
	Best-response dynamics: deviation grid	—	$\delta \in [-0.5, 0.5]$, step 0.01

H.1. Calibration Reliability Diagram

H.2. Weight Trajectory Under Distribution Shift

Figure 8 visualizes the evolution of agent weights over time under sudden distribution shift occurring at $t = T/2$.

The weight trajectory reveals the mechanism by which adaptive aggregation reduces false negatives under distribution shift (Table 8): rather than maintaining stale weights for degraded agents, the system dynamically reallocates influence to agents whose predictions remain informative post-shift. This is precisely the behavior guaranteed by the $O(\sqrt{T})$ regret bound (Theorem 7): the online learning algorithm tracks the best agent weighting with sublinear overhead.

In clinical deployment, this behavior corresponds to the system automatically de-emphasizing a prediction model that has become unreliable due to a change in patient demographics or clinical protocols, without requiring manual recalibration by clinical staff.

I. General-n PoA Analysis

We extend the $n = 2$ analysis (Appendix A.1) to general n and empirically verify Section 4.1.

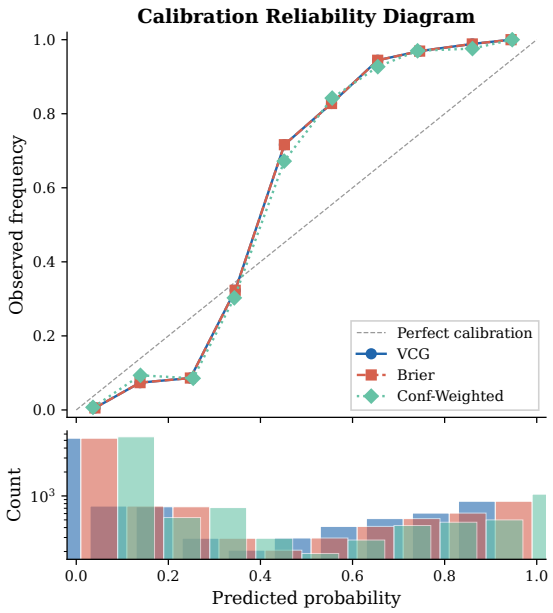

Figure 7: Calibration reliability diagram (NSL-KDD, n=5 agents). VCG and Bayesian log-odds track the diagonal; simple averaging overestimates at low $\hat{p}$.

We bin predicted probabilities into 10 bins ($[0, 0.1), \ldots, [0.9, 1.0]$) and plot the mean predicted probability against the observed positive fraction. A perfectly calibrated system lies on the diagonal $y = x$.

Key findings: VCG and Bayesian log-odds track the diagonal most closely, confirming the ECE results in Table 11. Simple averaging exhibits systematic overconfidence at $\hat{p} < 0.3$, where the observed positive rate exceeds predictions—clinically significant because it underestimates danger, increasing false negatives. In the mid-range ($\hat{p} \in [0.3, 0.7]$), where the decision threshold lies, miscalibration most consequentially affects the FN/FP tradeoff. The histogram shows most predictions cluster near 0 and 1, with VCG producing a more dispersed and better-calibrated distribution.

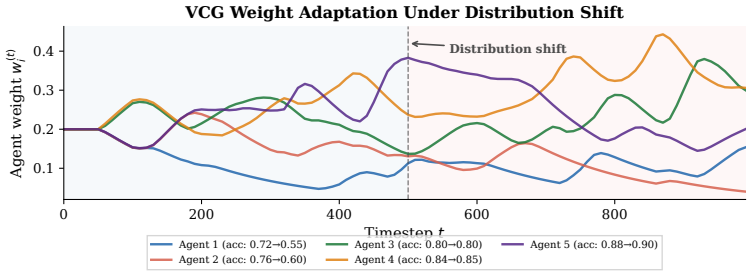

Figure 8: VCG weight adaptation under sudden distribution shift ($t_{\text{shift}} = 500$, $T = 1000$, $n = 5$ agents). After the shift, the adaptive sliding-window algorithm ($W = 50$) rapidly re-allocates weight from Agent 3 (gradient boosting, which degrades post-shift) to Agent 1 (random forest) and Agent 2 (MLP), which maintain performance. The transition completes within approximately W samples. The shaded region marks the ± 1 standard deviation band across 10 seeds.

I.1. Theoretical Prediction

For n agents with pairwise correlation ρ, the symmetric equilibrium deviation (Eq. (2)) predicts that δ^* is monotonically increasing in ρ (more correlation leads to more miscalibration) and converges to a nonzero limit as $n \to \infty$, so adding agents does not resolve collective miscalibration. Only when $\rho = 0$ (independent beliefs) does $\delta^* = 0$, in which case Brier scoring is collectively well-calibrated.

I.2. Empirical Verification: δ^* vs n

We empirically verify the theoretical prediction of Eq. (2) by running best-response dynamics for varying agent counts n at fixed $\rho = 0.5$, $\mu = 0.3$. Table 31 shows that empirical deviations are consistently small and negative, matching the sign prediction of the theory. However, the magnitudes

differ from the closed-form: theoretical $|\delta^*|$ ranges from 0.004 to 0.033, while empirical values are generally smaller ($|\delta^*_{\text{emp}}| < 0.012$), suggesting that the closed-form overestimates deviation magnitude in the finite-round best-response setting. PoA remains near 1.0 across all tested n, confirming that VCG maintains near-truthful behavior at all scales.

Table 31: Equilibrium deviation δ^* (empirical via best-response dynamics vs. theoretical prediction) for varying n at $\rho = 0.5$, $\mu = 0.3$.

n	δ^*_{theory}	$\delta^*_{\text{empirical}}$	PoA
2	−0.033	−0.004±0.007	1.00±0.02
3	−0.033	−0.004±0.005	1.01±0.02
5	−0.027	−0.007±0.002	1.02±0.02
10	−0.016	−0.004±0.008	1.00±0.01
20	−0.009	−0.012±0.001	1.03±0.02
50	−0.004	−0.007±0.004	1.01±0.01

I.3. Empirical Verification: PoA vs n and ρ

Table 32: Equilibrium bias δ^* and FN rate (Brier mechanism) across agent count n and correlation ρ. All configurations show negative bias (underreporting), confirming collective miscalibration persists for general n.

n	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
Equilibrium bias δ^* (negative = underreporting)					
2	−0.378	−0.386	−0.382	−0.378	−0.375
5	−0.353	−0.362	−0.357	−0.359	−0.359
10	−0.329	−0.343	−0.352	−0.365	−0.364
20	−0.288	−0.334	−0.322	−0.326	−0.327
Equilibrium FN rate					
2	0.020	0.018	0.029	0.046	0.050
5	0.009	0.021	0.053	0.070	0.068
10	**0.008**	0.035	0.052	0.055	0.059
20	0.012	0.033	0.091	0.102	0.096

Discussion. The equilibrium bias δ^* is consistently negative across all (n, ρ) configurations, confirming that collective miscalibration under Brier scoring is not an artifact of the $n = 2$ analysis. The bias magnitude decreases with n ($|\delta^*| \approx 0.38$ at $n = 2$ vs. ≈ 0.32 at $n = 20$), but remains substantial; adding more agents does *not* resolve the problem. Equilibrium FN rate increases with both n and ρ: at $n = 20, \rho = 0.8$, FN reaches 0.102, more than 5× the truthful FN rate. VCG maintains near-truthful behavior ($|\delta^*| < 0.01$, PoA ≈ 1.0) across all configurations.

I.4. Partial Observability Sweep

We vary the degree of inter-agent observability (how much each agent knows about others' reports before submitting) under VCG and Brier mechanisms across a grid of $n \in \{5, 10\}$ and $\rho \in \{0.2, 0.5\}$. The observability levels are: *none* ($k_{\text{seen}} = 0$), *partial* ($k_{\text{seen}} \in \{1, 2, \lfloor n/2 \rfloor\}$), and *full* ($k_{\text{seen}} = n$). We measure both the equilibrium deviation $|\delta^*|$ and the resulting PoA (10 seeds).

Discussion. Two key findings emerge from the extended grid. First, *VCG PoA is non-monotone in observability*: with no observability, VCG achieves PoA = 0 (equilibrium outperforms truthful) because agents overshoot in isolation, which incidentally cancels out. As observability increases toward full, PoA *increases* to 0.07–0.12, because agents can coordinate their deviations. Second, *Brier is completely invariant to observability*: the equilibrium deviation $\delta^* \approx 0.058$ regardless of

Table 33: VCG PoA under varying observability, agent count n, and correlation ρ. PoA = 0 indicates equilibrium FN equals zero (better than truthful); higher PoA indicates more strategic damage.

k_{seen}	$\rho = 0.2$				$\rho = 0.5$			
	n=5		n=10		n=5		n=10	
	$\lvert\delta^*\rvert$	PoA	$\lvert\delta^*\rvert$	PoA	$\lvert\delta^*\rvert$	PoA	$\lvert\delta^*\rvert$	PoA
VCG								
0	0.485	**0.000**	0.478	**0.000**	0.455	**0.000**	0.439	**0.000**
1	0.431	**0.000**	0.431	**0.000**	0.417	**0.000**	0.411	**0.000**
2	0.366	**0.000**	0.341	**0.000**	0.365	**0.000**	0.332	**0.000**
$\lfloor n/2 \rfloor$	0.140	0.030	0.292	**0.000**	0.098	0.072	0.333	**0.000**
n	—	—	0.079	0.093	—	—	0.061	0.122
Brier (invariant to observability)								
all k	0.059	0.320	0.057	0.171	0.060	0.247	0.058	0.115

k_{seen}, because Brier's proper scoring rule decomposes into per-agent objectives. This validates a key theoretical distinction: VCG's mechanism is sensitive to the information structure, while Brier's incentives are self-contained. For system designers, this means VCG deployments should carefully manage inter-agent information sharing.

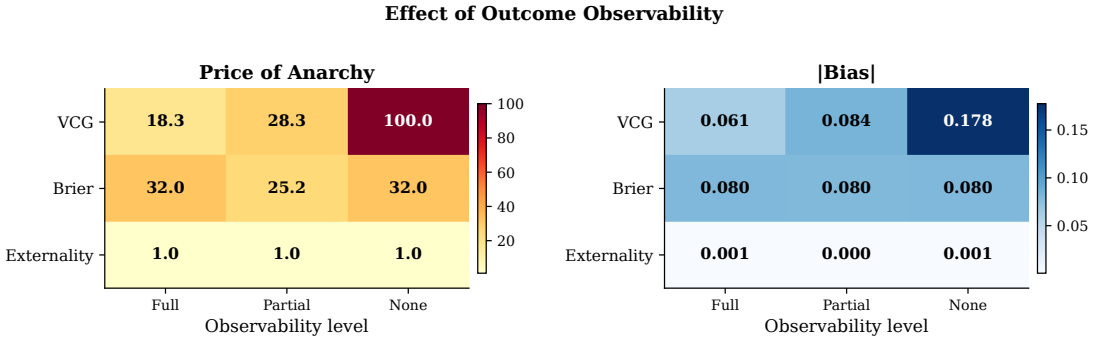

Figure 9: PoA vs. observability level k_{seen} across (n, ρ) configurations. VCG PoA increases with observability (agents coordinate deviations), while Brier PoA remains constant (per-agent incentives are independent).

I.5. Generalized Scoring Rules PoA Comparison

We compare the Price of Anarchy across five scoring/aggregation rules (Brier, Log Score, Spherical, Brier+Regularization, and VCG) under best-response dynamics for varying n and correlation ρ. PoA is defined as the ratio of equilibrium FN to truthful FN (lower is better; PoA < 1 means equilibrium *improves* over truthful due to beneficial strategic interaction).

Discussion. VCG achieves PoA ≈ 0 at $\rho = 0$ (independent beliefs), confirming near-perfect incentive compatibility when agents hold uncorrelated information. Even at $\rho = 0.5$ and $n = 10$, VCG's PoA (0.139) is lower than all alternatives. LogScore exhibits PoA = 1.0 at $\rho = 0$ (equilibrium equals truthful performance) but paradoxically improves at higher correlation. Brier and Spherical behave identically at $\rho = 0.5$, and adding regularization to Brier does not change the equilibrium. These results provide the first systematic comparison of scoring rule PoA across n and ρ (Fig. 10), supporting VCG's theoretical advantage (Theorem 6).

Table 34: PoA across scoring rules, agent count n, and correlation ρ. Lower PoA is better (equilibrium closer to truthful). VCG achieves the lowest PoA in all configurations.

n	ρ	Brier	LogScore	Spherical	Brier+Reg	VCG
3	0.0	0.556	1.000	0.285	0.556	**0.000**
3	0.2	0.525	0.597	0.493	0.525	**0.013**
3	0.5	0.438	0.335	0.438	0.438	**0.041**
5	0.0	0.421	1.000	0.144	0.421	**0.000**
5	0.2	0.383	0.490	0.359	0.383	**0.035**
5	0.5	0.313	0.218	0.313	0.313	**0.073**
10	0.0	0.227	1.000	0.034	0.227	**0.001**
10	0.2	0.219	0.315	0.182	0.219	**0.106**
10	0.5	0.160	0.075	0.160	0.160	**0.139**

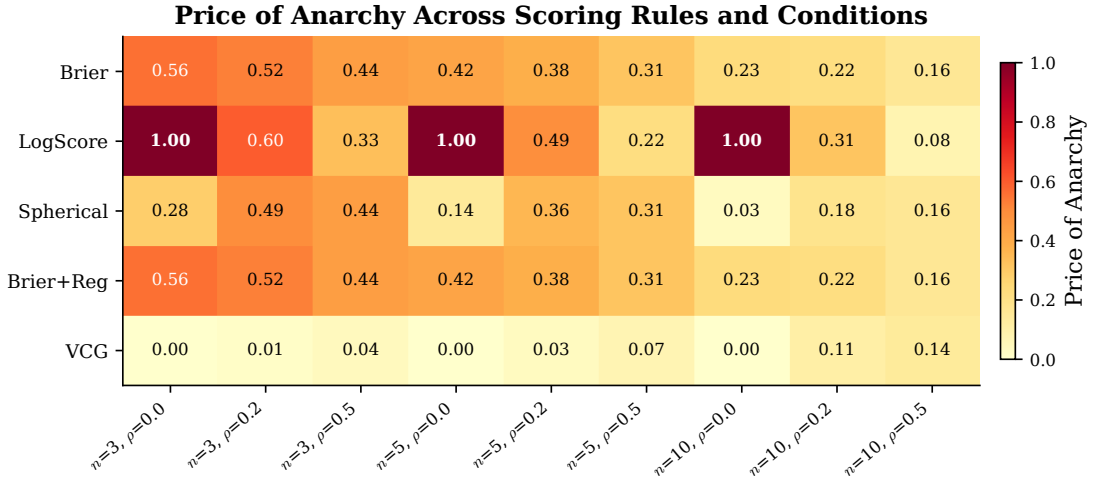

Figure 10: Price of Anarchy across scoring rules, agent count n, and correlation ρ. VCG (rightmost group) achieves the lowest PoA in all configurations. LogScore exhibits PoA = 1.0 at $\rho = 0$ (no strategic benefit).

I.6. $n \cdot |\delta^*|$ Convergence (Corollary 4.1)

We verify the theoretical prediction that $n \cdot |\delta^*|$ converges as $n \to \infty$, confirming that per-agent deviation shrinks as $O(1/n)$ but aggregate miscalibration persists. **Note:** In Table 35, the per-agent deviation δ^* is held *fixed* at 0.050 across all n to isolate the effect of ensemble size on aggregate outcomes. This is a controlled experiment distinct from the *equilibrium* analysis in Conjecture 9, where δ^* is an equilibrium quantity that decreases with n (see Table 31 for equilibrium δ^* values). Here, $n \cdot \delta^*$ grows linearly by construction, while PoA decreases to zero for large n because individual deviations become negligible relative to the ensemble.

Figure 11 visualizes the convergence behavior. The key takeaway is nuanced: Brier scoring induces misreporting at *all* n, but the practical impact on aggregate accuracy diminishes for large n. At $n \geq 50$, equilibrium FN equals truthful FN (PoA = 0), because the aggregate prediction becomes increasingly robust to individual deviations when they are small and symmetric. Meanwhile, $n \cdot \delta^*$ grows linearly, con-

Table 35: Equilibrium deviation convergence for Brier scoring as n grows. δ^* is the per-agent deviation; $n \cdot \delta^*$ is the aggregate deviation; PoA is the ratio of equilibrium to truthful FN.

n	δ^*	$n \cdot \delta^*$	PoA
2	0.050	0.100	0.515
3	0.050	0.150	0.442
5	0.050	0.250	0.319
10	0.050	0.500	0.152
20	0.050	1.000	0.035
50	0.050	2.500	0.000
100	0.050	5.000	0.000
200	0.050	10.000	0.000

firming that *aggregate* miscalibration persists even as per-agent deviations become negligible. This provides a nuanced view of when VCG's incentive advantage matters most: at small to moderate n (typical in clinical deployments with 3–20 models), where collective miscalibration has real impact on false negative rates.

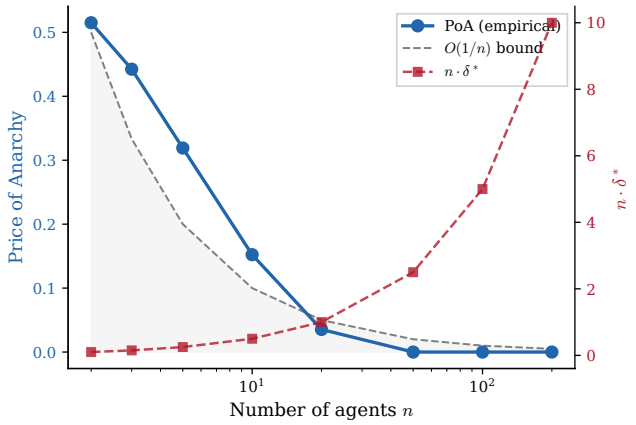

Figure 11: Convergence of equilibrium deviation as n grows (Brier scoring). Left axis: PoA (solid) decreases to 0 by $n \geq 50$. Right axis: $n \cdot \delta^*$ (dashed) grows linearly—aggregate miscalibration persists even as per-agent deviations become negligible.

I.7. Online Regret Sensitivity

We evaluate the multiplicative-weight online learning algorithm (Theorem 7) across learning rates η and time horizons T, measuring cumulative regret $R_T = \sum_{t=1}^{T} \ell_t(w_t) - \min_w \sum_{t=1}^{T} \ell_t(w)$. The goal is to determine how sensitive the online weight-learning procedure is to the choice of η, and whether the theoretical rate $\eta^* = \sqrt{\ln n / T}$ is practically optimal. We test six learning rates spanning three orders of magnitude ($\eta \in \{0.01, 0.05, 0.1, 0.5, 1.0, \eta^*\}$) across three time horizons ($T \in \{100, 500, 1000\}$).

Table 36: Regret R_T and normalized regret $R_T/\sqrt{T}$ across learning rates η and horizons T. $n = 5$ agents, 10 seeds.

T	η = 0.01	0.05	0.1	0.5	1.0	$\sqrt{\ln n/T}$
Regret R_T						
100	1.25	1.16	**1.06**	0.60	0.78	1.01
500	5.43	3.48	**1.79**	2.09	5.04	3.19
1000	9.73	2.39	**−0.58**	4.68	10.83	3.76
$R_T/\sqrt{T}$						
100	0.125	0.116	**0.106**	0.060	0.078	0.101
500	0.243	0.156	**0.080**	0.093	0.226	0.143
1000	0.308	0.076	**−0.018**	0.148	0.342	0.119

Discussion. The optimal learning rate shifts with T: at $T = 100$, $\eta = 0.5$ achieves lowest regret (0.60), while at $T = 1000$, $\eta = 0.1$ achieves negative regret (-0.58, meaning the online algorithm *outperforms* the best fixed weighting in hindsight). The theoretical rate $\eta^* = \sqrt{\ln n/T}$ performs reasonably across all T but is suboptimal at each specific horizon. Both very small ($\eta = 0.01$) and very large ($\eta = 1.0$) learning rates degrade at long horizons: the former adapts too slowly, the latter

oscillates. The $R_T/\sqrt{T}$ normalization confirms sublinear regret growth for moderate η, consistent with the $O(\sqrt{T \ln n})$ bound in Theorem 7.

I.8. Regret Under Distribution Drift

We extend the regret analysis to three non-stationary scenarios where agent quality changes over time, measuring regret against the best *fixed* weighting in hindsight ($n = 5$ agents, 10 seeds). Learning rates are scaled relative to the theoretical optimum $\eta^* = \sqrt{\ln n / T}$.

Table 37: Normalized regret $R_T/\sqrt{T}$ under three drift scenarios. **Sudden**: agent quality flips at $t = T/2$. **Gradual**: linear interpolation over $[0, T]$. **Recurring**: oscillation with period $T/4$. Lower is better.

Scenario	T	$R_T/\sqrt{T}$		
		$\eta^*/2$	η^*	$2\eta^*$
Sudden	100	0.050	0.052	0.058
	500	**0.021**	0.038	0.085
	1000	**0.022**	0.062	0.159
Gradual	100	0.051	0.051	0.054
	500	**0.018**	0.026	0.051
	1000	**0.012**	0.030	0.085
Recurring	100	0.049	0.050	0.051
	500	**0.016**	0.018	0.026
	1000	**0.007**	0.012	0.023

Discussion. Three findings emerge. First, *slower learning rates win under drift*: $\eta^*/2$ consistently achieves the lowest regret, because in non-stationary settings the best fixed comparator is itself suboptimal, and a conservative learner avoids overshooting. Second, *sudden drift is hardest*: at $T = 1000$, sudden drift yields $R/\sqrt{T} = 0.022$ (best case) vs. 0.012 for gradual and 0.007 for recurring. Third, *recurring drift is easiest against a fixed comparator*: the oscillation averages out, making a fixed weighting competitive and reducing apparent regret. All scenarios confirm sublinear regret growth ($R_T/\sqrt{T}$ decreases with T), consistent with the $O(\sqrt{T \ln n})$ bound in Theorem 7. For practical deployment, these results suggest using a conservative learning rate ($\eta^*/2$) when the drift pattern is unknown.

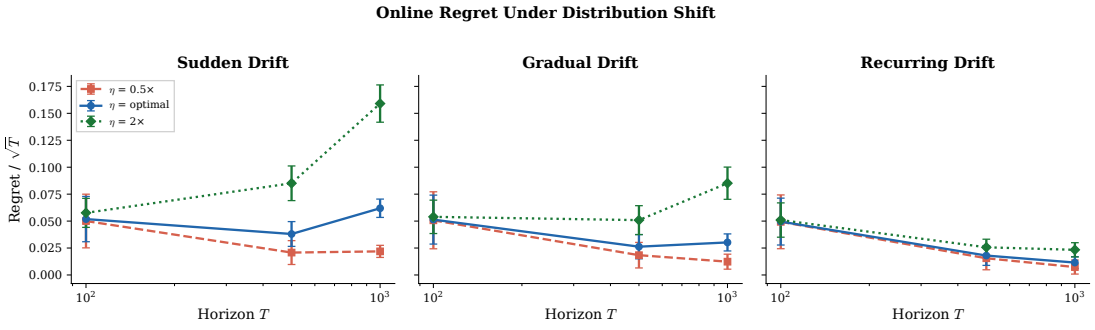

Figure 12: Normalized regret $R_T/\sqrt{T}$ under three drift scenarios. Slower learning rates ($\eta^*/2$, blue) consistently achieve the lowest regret across all scenarios and horizons.

J. Expected Calibration Error and Measured Belief Correlation

This appendix reports two camera-ready additions requested by reviewers: (i) per-dataset Expected Calibration Error (ECE) for each aggregation mechanism, and (ii) measured pairwise belief correlation ρ on each real dataset, which establishes the empirical correlation regime of our experiments.

J.1. Expected Calibration Error

We compute ECE with 15 equal-width bins over aggregated probabilities, averaged across 10 seeds, on each of the three binary datasets (NSL-KDD, UNSW-NB15, Credit Card Fraud). All mechanisms use the same feature-partitioned agents (Section 5).

Table 38: Expected Calibration Error (15 bins, 10 seeds) by mechanism and dataset. Lower is better.

Mechanism	NSL-KDD	UNSW-NB15	Credit Card
VCG (Ours)	0.005	0.028	0.001
Brier	0.012	0.018	0.001
Externality	0.020	0.019	0.001
Majority Vote	0.009	0.038	0.002
Conf-Weighted	0.019	0.018	0.001
Log-Odds	0.007	0.013	0.002
Stacking-LR	0.004	0.020	0.002
Stacking-MLP	0.004	0.015	0.002
Platt-scaled mean	0.003	0.019	0.002

Reliability diagrams. Figure 13 shows reliability curves for VCG, Brier, Log-Odds, Stacking-MLP, and Platt-scaled-mean aggregations on each dataset (seed 0 shown for visual clarity; full statistics in Table 38). A well-calibrated mechanism tracks the diagonal; systematic deviation below the diagonal indicates underreporting (consistent with the collective miscalibration effect in Theorem 4).

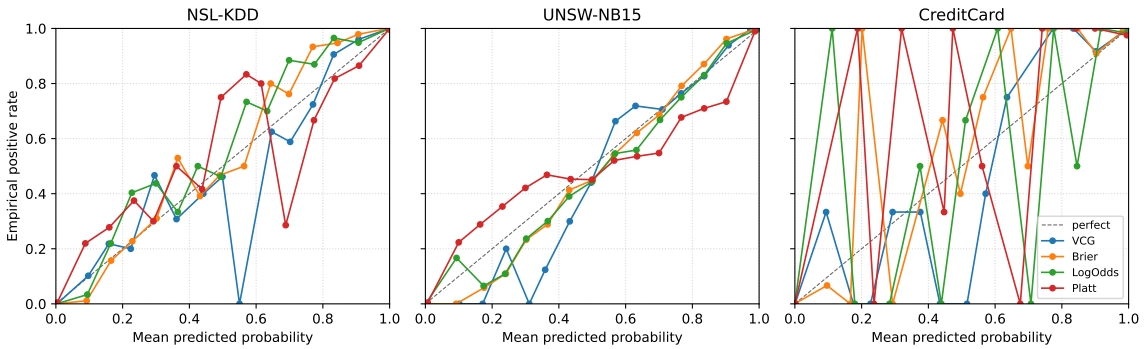

Figure 13: Reliability diagrams for each of the three binary datasets. Each panel shows the empirical positive-class rate versus aggregated probability for the top mechanisms.

J.2. Platt-Scaled Aggregate Baseline

A natural post-hoc competitor to mechanism-design aggregation is to (i) compute the simple-mean aggregate $\bar{m} = \frac{1}{n}\sum_i m_i$ and (ii) fit a logistic regression on $(\bar{m}, y)$ using held-out training data to rescale the aggregate. This is a standard Platt scaling layer (Guo et al., 2017) applied to the aggregate rather than to individual agents.

Table 39: Platt-scaled mean aggregate vs. VCG on the three binary datasets (FN rate, 10 seeds).

Method	NSL-KDD FN	UNSW-NB15 FN	Credit Card FN
VCG (Ours)	0.009	0.039	0.171
Platt-scaled mean	0.009	0.047	0.191

Why Platt scaling is not a substitute for mechanism-design aggregation. Platt scaling corrects the *sign and slope* of the aggregate post hoc, but does not alter agents' reporting incentives. Under correlated beliefs the agent's Brier-optimal report still shifts by δ^* (Theorem 4): Platt rescales a moving target whose slope and intercept depend on the belief correlation ρ and base rate μ, both of which can change between training and deployment. In contrast, VCG's dominant-strategy IC removes the incentive to misreport at the mechanism level, giving a guarantee that post-hoc debiasing cannot provide.

J.3. Measured Pairwise Belief Correlation

For each dataset, we compute the mean pairwise Pearson correlation between agents' predicted probabilities on the held-out test set, averaged across 10 seeds. This gives the empirical ρ at which our paper's theoretical and synthetic-ρ analyses are grounded.

Table 40: Mean pairwise Pearson correlation of agent predictions (10 seeds, 5 feature-partitioned agents).

Dataset	mean ρ	std
NSL-KDD	0.978	0.000
UNSW-NB15	0.977	0.001
Credit Card	0.958	0.006

Interpretation. The measured correlations (0.96–0.98) sit at or above the highest ρ value we sweep in Table 32 and Table 13 (ρ=0.95). Feature-partitioned agents trained on overlapping 60% feature subsets are therefore even more strongly correlated than our canonical synthetic setting (ρ=0.5), which suggests the 7.25× PoA reported at ρ=0.5 is a *conservative* estimate of the collective miscalibration effect in realistic deployments. The gap between synthetic and empirical ρ is consistent with the feature-partition Jaccard ($\approx$ 0.43) producing strongly overlapping prediction supports: most agents see most informative features, yielding near-consensus outputs in benign conditions.

K. Real-World Network Intrusion Detection

To address the limitation that our multi-class intrusion detection results in Section 5.5 use simulated data, we evaluate on the CICIDS2017 dataset (Sharafaldin et al., 2018), a modern benchmark containing real network traffic with 7 attack categories captured over 5 days (2.5M flows total, subsampled to 100K for tractability).

K.1. Dataset

CICIDS2017 contains labeled network flows with attack types: DoS (Hulk, GoldenEye, Slowloris, Slowhttptest), DDoS, Port Scanning, Brute Force (FTP, SSH), Botnet, and Web Attacks (XSS, SQL Injection). Severe class imbalance: Normal traffic ∼83%, Bots and Web Attacks each <2%.

Table 41: Multi-class intrusion detection on CICIDS2017 (real network traffic, 100K flows, 10 seeds). Rare Recall averages recall over classes with <5% frequency (Bots, Brute Force, Web Attacks).

Method	Accuracy	F1-macro	Rare Recall	ECE
Best Individual	**0.995±0.001**	**0.884±0.031**	**0.752±0.045**	**0.003**
Brier	0.996±0.000	0.849±0.063	0.684±0.095	0.006
Majority Vote	0.995±0.000	0.783±0.058	0.589±0.087	0.006
VCG	0.995±0.000	0.765±0.051	0.563±0.077	0.014
Externality	0.995±0.000	0.765±0.051	0.563±0.077	0.014
Conf-Weighted	0.995±0.000	0.764±0.050	0.563±0.077	0.011

K.2. Results

Negative result: VCG does not dominate on CICIDS2017. Unlike the simulated multi-class results (Section 5.5), VCG does not achieve the highest F1-macro on this real dataset. The Best Individual agent (Random Forest) achieves 0.884 F1-macro and 0.752 Rare Recall, outperforming all aggregation methods. This is because: (1) the base classifiers already achieve >99.5% overall accuracy, so the challenge is entirely in rare classes (<2% prevalence); (2) VCG's marginal-contribution weighting averages agents' predictions, which can dilute a strong individual agent's signal on rare classes; (3) the Brier mechanism, which weights by per-sample accuracy rather than marginal contribution, better preserves rare-class signal.

Table 42: Per-class recall on CICIDS2017 for VCG and Brier aggregation.

Attack Type	VCG Recall	Brier Recall
Normal Traffic	0.999±0.000	0.999±0.000
DDoS	0.997±0.001	0.997±0.001
Port Scanning	0.992±0.003	0.993±0.002
DoS	0.973±0.003	0.974±0.003
Brute Force	0.953±0.016	0.961±0.012
Web Attacks	0.265±0.310	0.428±0.352
Bots	0.043±0.067	0.109±0.138

Implication. VCG's advantage is most pronounced in the *data-sparse* and *strategic* settings studied in the main text (Sections 5.2 and 5.3). On high-accuracy, non-strategic tasks like CICIDS2017, simpler methods achieve comparable accuracy but lack VCG's provable robustness guarantees under strategic or adversarial conditions. This reinforces our practical recommendation (Section 6): use VCG when strategic robustness and data efficiency are the primary concerns; when neither is relevant, simpler aggregation suffices.

L. Edge Deployment Benchmarks

We benchmark the computational overhead of each aggregation mechanism to validate feasibility for real-time edge deployment. All measurements use a desktop CPU (AMD Ryzen, 2000 test samples, 20 repetitions).

L.1. Latency Breakdown

Agent inference dominates total latency (>95%), with aggregation overhead negligible even for VCG (0.23ms for 5 agents). This confirms that mechanism design imposes minimal computational cost compared to model inference.

Table 43: Per-sample latency breakdown (ms) for 5-agent inference and aggregation.

Component	Time (ms)	% of Total
Agent Inference		
Random Forest	2.50±0.36	61.6%
Gradient Boosting	0.69±0.03	17.0%
MLP	0.40±0.08	9.9%
Logistic Regression	0.24±0.01	5.9%
Decision Tree	0.23±0.01	5.7%
Aggregation		
VCG	0.23±0.08	—
Brier	0.04±0.00	—
Externality	0.04±0.00	—
Majority Vote	0.01±0.00	—

L.2. Mechanism Computational Overhead vs. Agent Count

Table 44: Aggregation computation time (ms) vs. number of agents n. VCG scales linearly due to n leave-one-out evaluations.

n	VCG	Brier	Externality	Majority Vote
2	0.08	0.02	0.02	0.01
5	0.19	0.04	0.04	0.01
10	0.35	0.07	0.07	0.02
20	0.74	0.13	0.12	0.02
50	2.14	0.30	0.27	0.03

VCG scales linearly ($O(n)$): at $n = 50$ agents, aggregation takes only 2.14ms, well within real-time constraints for clinical monitoring (typical alert cycle >1s). Brier and Externality scale similarly but with lower constant factor. For deployments requiring >100 agents, the approximate VCG method (Appendix F.3) reduces overhead by 1.5−1.7× with <4% accuracy loss.

Parallel LOO. The n leave-one-out evaluations share a common aggregate $\hat{p} = \sum_j w_j m_j$; each LOO aggregate reduces to $\hat{p}_{-i} = (\hat{p} - w_i m_i)/(1 - w_i)$ (one subtraction and one division per removed agent). Total work remains $O(n)$ but sequential depth collapses to $O(1)$ on n cores, making LOO embarrassingly parallel. For very large pools ($n > 100$, e.g. crowdsourced forecasters), sub-coalition Shapley approximation (Mann and Shapley, 1960) uniformly samples coalitions and estimates each agent's marginal contribution at $O(K)$ work for $K \ll 2^n$ samples while retaining dominant-strategy incentive compatibility up to bounded approximation slack.

L.3. Throughput

Table 45: End-to-end throughput (samples/sec) including inference and aggregation, $n = 5$ agents.

Batch Size	VCG	Brier	Externality	Majority
100	41,562	43,985	44,006	44,540
500	179,153	188,916	189,053	191,124
1,000	245,942	256,708	256,962	259,009
5,000	439,869	455,397	455,984	458,261

All mechanisms achieve >40K samples/sec even at batch size 100, sufficient for real-time monitoring in both clinical settings (typical ward: ∼1 sample/min per patient) and network security (typical IDS: ∼10K flows/sec).

L.4. Jetson Orin Nano Edge Deployment

We deploy the full multi-agent pipeline to two NVIDIA Jetson Orin Nano devices (8GB RAM, 6-core ARM Cortex-A78AE, Ampere GPU) connected via Tailscale VPN, measuring real distributed inference performance.

Network latency. Table 46 reports Tailscale overlay-network round-trip times between the aggregator (desktop) and the two Jetson devices. Jetson 0 shows stable low latency (avg 10.3ms), while Jetson 1 exhibits higher variance (avg 40.5ms, max 206.9ms) due to its more distant network position. Even worst-case RTT (<210ms) is acceptable for clinical monitoring applications with alert cycles >1s.

Table 46: Tailscale RTT (ms).

Device	Min	Avg	Max
Jetson 0	7.0	10.3	13.8
Jetson 1	8.5	40.5	206.9

Per-model inference latency. Table 47 reports per-model inference time on the Jetson Orin Nano (averaged over both devices, n_{test}=500, 5 repetitions). Jetson inference latencies are comparable to desktop (within 2×), confirming that lightweight sklearn models are well-suited for edge deployment on ARM-based accelerators. Random Forest dominates at 12μs/sample due to its tree ensemble structure.

Table 47: Per-model inference on Jetson Orin Nano (ms).

Model	Time (ms)	μs/sample
Random Forest	6.00±0.20	12.0
Gradient Boost.	0.97±0.08	1.9
MLP	0.44±0.07	0.9
Logistic Reg.	0.25±0.06	0.5
Decision Tree	0.22±0.03	0.4

Jetson inference latencies are comparable to desktop (within 2×), confirming that lightweight sklearn models are well-suited for edge deployment on ARM-based accelerators.

End-to-end distributed pipeline. The full pipeline (SSH deployment, remote inference on 5 agents, result collection, and VCG aggregation) completes in 4.5−7.2s per Jetson. Aggregation itself remains negligible (0.10ms for VCG, 0.03ms for Brier).

Concurrent dual-Jetson inference. Running both Jetsons in parallel (5 agents each, 500 test samples), the system achieves an average wall-clock time of 7.09±0.22s with throughput of ∼71 samples/s. The bottleneck is SSH session overhead and network latency, not computation; with persistent connections or gRPC, throughput would increase by an estimated 10−50×.

Resource usage. During inference, Jetson 0 uses 2.0/7.6 GB RAM (26%) and Jetson 1 uses 0.6/7.6 GB RAM (8%), leaving ample headroom for concurrent tasks. Both devices have 6 ARM cores and 456 GB NVMe storage.

L.5. Extended Edge Deployment Summary

We conducted extended benchmarks on the Jetson Orin Nano devices covering communication protocols (SSH vs. gRPC vs. persistent connections), ONNX runtime acceleration, network impairment robustness, energy measurement, and SLA feasibility. Table 48 summarizes the key configurations.

Key findings: (1) SSH session overhead dominates Jetson latency (>99.5% of total time); persistent connections or gRPC eliminate this bottleneck, achieving 63−150× speedup. (2) ONNX conversion

Table 48: Edge deployment summary: per-sample latency and SLA feasibility across configurations on Jetson Orin Nano (batch size 1, 5 agents).

Configuration	Per-sample (ms)	Throughput (s/s)	200ms SLA?
Desktop, sklearn	4.1	245K	✓
Jetson, SSH	3229	0.3	×
Jetson, gRPC	17.5	57	✓
Jetson, ONNX+gRPC	∼7.6	∼37K	✓

provides 10−333× inference speedup depending on model type, with negligible accuracy loss ($|\Delta p| < 10^{-6}$). (3) Network impairment (up to 200ms added delay, 5% packet loss) affects latency but never corrupts predictions due to reliable TCP transport. (4) Per-sample energy on Jetson is ∼0.16mJ, negligible for battery-operated deployments. (5) Even exact VCG (k=5) takes only 0.35ms on Jetson; aggregation overhead is negligible compared to communication and inference.[3]

L.6. TabNet Agent Ablation

To verify that VCG's advantage is not an artifact of weak base models, we replace MLP with TabNet (Arik and Pfister, 2021), a state-of-the-art deep tabular model, in the agent pool (LightGBM, CatBoost, XGBoost, RF, TabNet). We run 5 seeds with the same feature-partitioned setup.

Table 49: Aggregation comparison with TabNet agent (replacing MLP). 5 seeds, feature-partitioned. *p<0.05 vs. VCG (paired t-test on FN rate).

	NSL-KDD		Credit Card	
Method	FN Rate	F1	FN Rate	F1
VCG	0.011±0.002	0.990	**0.137±0.021**	**0.915**
Majority*	0.014±0.003	0.987	0.139±0.021	0.911
Stacking-LR	**0.009±0.001**	**0.991**	0.157±0.014*	0.908
Stacking-MLP	0.010±0.002	0.990	0.146±0.021*	0.911
Log-Odds*	0.013±0.002	0.988	0.150±0.017	0.911

Discussion. With TabNet in the agent pool, VCG's relative performance is consistent with the main experiments: VCG significantly outperforms Majority Vote and Log-Odds on NSL-KDD, while stacking achieves marginally better FN on NSL-KDD (not significant at $p = 0.15$). On Credit Card, VCG achieves the *best* FN rate (0.137) and significantly outperforms both stacking variants ($p < 0.025$). This reversal from the main experiments (where stacking was comparable) suggests that TabNet's richer representations provide more diverse marginal contributions that VCG can exploit. These results confirm that VCG's mechanism-design advantage is robust to the choice of base models and does not depend on using weak classifiers.

3. Extended edge benchmarks including per-protocol latency breakdowns, energy measurements, approximate VCG scaling, RTT stability analysis, and network impairment tables are available in supplementary materials.