Evolutionary Tipping Points in Open Protocol Adoption: Why Frontier and Challenger AI Labs Converge on Verifiable Fiduciary Standards

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Why would profit-maximizing AI corporations—both frontier incumbents with proprietary lock-in and challenger laboratories—adopt an open, cryptographically verifiable fiduciary protocol layer and emit attested telemetry? In classical industrial organization, dominant platforms resist compatibility to extract switching rents. However, shifts in 2025–2026 have altered the industry payoff matrix: lightweight open-weight models on local-first harnesses rival flagship endpoints in structured protocol fidelity, corporate-controlled agent protocols (MCP and A2A) exhibit steering-committee governance deficits, and statutory mandates (the EU Data Act and IEEE 7012 privacy terms) penalize unverified delegation. We formulate cross-lab protocol convergence as an asymmetric two-population Katz–Shapiro evolutionary game coupled with a multi-source Merkle-committed, differentially private ($\mathrm{Lap}(\Delta/\varepsilon_{\mathrm{DP}})$) telemetry oracle and rule-sunset portfolio. In Stage 1, challenger labs pool network externalities on a shared Verifiable Fiduciary Protocol stack to overcome lemon discounting. In Stage 2, once challenger share x_c and fiduciary demand θ_{fid} cross the analytical tipping interval $[\theta^*(1), \theta^*(0)] = [0.148, 0.395]$, open conversion becomes the incumbent best response under conditional comparative statics. Across 9,900 simulated trials, the Verifiable Fiduciary Standard achieves 99.33% $[98.30, 99.74]$ conditional tipping convergence, 100.00% $[99.36, 100.00]$ oracle accuracy (0.95 $[0.94, 0.97]$ telemetry MAE), and 99.33% $[98.30, 99.74]$ joint surplus at $\theta_{\mathrm{fid}} = 0.45$, outperforming syntax-only consortia (36.33% $[32.58, 40.26]$ tipping, 49.33% $[45.35, 53.33]$ surplus) up to the 20% trimmed-mean breakdown bound.

1 Introduction

The architectural evolution of the internet was shaped by a strategic contest between proprietary, vertically integrated information networks (such as CompuServe, AOL, and enterprise SNA stacks) and open, permissionless protocol suites (TCP/IP, HTTP, and SMTP) [6, 11, 17]. Today, a structurally analogous standardization contest is unfolding across the emerging agentic web [21]. On one side, a small number of frontier AI corporations operate closed-weight models behind proprietary API walled gardens, seeking to capture downstream two-sided platform ecosystems [16] by controlling routing, context retention, and telemetry. On the other side, a growing ecosystem of challenger laboratories, open-weight model builders [9], and local-first agent harnesses [19] seeks to unbundle agent execution from centralized cloud gatekeepers.

A central question in network economics and institutional AI governance is whether—and under what exact incentive and telemetry-verification conditions—both smaller challenger AI laboratories

and large frontier incumbents will converge on a shared, verifiable fiduciary protocol layer rather than fragmenting into incompatible walled gardens or an unverified open bazaar.

1.1 The strategic pathology of walled gardens and syntax-only consortia

Under an unregulated status quo, dominant platform incumbents face strong economic incentives to resist genuine fiduciary interoperability [6, 11, 17]. When a dominant vendor controls both the orchestration interface and the underlying model tiers, proprietary switching costs allow the platform to extract lock-in rents L_{lock} while obscuring silent model downgrades or latency throttling behind opaque API endpoints. Recognizing that enterprise developers demand cross-tool connectivity, major technology vendors in 2025–2026 introduced first-generation agent interoperability specifications, most notably the Model Context Protocol (MCP) and the Agent-to-Agent Protocol (A2A). However, recent institutional and security audits demonstrate that syntax-only corporate consortium standards fall short of a verifiable fiduciary commons:

1. **Centralized corporate capture of protocol governance (Wang and Zhang [21]).** Comparing decentralized standard-setting (such as permissionless rough-consensus specifications) against corporate consortium governance (A2A), Wang and Zhang [21] documented that while open standards decouple protocol evolution from single-vendor control, A2A assigns binding architectural authority to an eight-seat corporate Technology Steering Committee. Concentrating agent protocol governance in a handful of dominant AI vendors creates a structural legitimacy and neutrality deficit for challenger labs and public-sector adopters.
2. **Absence of verifiable telemetry and execution provenance (Elina [4], Zhang et al. [23], Zhao et al. [24]).** Syntax-level tool and peer-messaging protocols standardize JSON-RPC message envelopes but do not provide cryptographic attestation of runtime action compliance [23, 24], multi-aspect execution trustworthiness [4], or zero-knowledge oversight of safety invariants [2]. Without verifiable telemetry oracles, buyers cannot distinguish compliant executions from silently degraded or manipulated agent runs.
3. **Incomplete disclosure and manipulative reporting (Feldman et al. [7], Fung et al. [8]).** As established in the institutional transparency literature by Fung et al. [8], voluntary corporate disclosure regimes fail when reporting schemas lack standardized, auditable verification. In multi-agent marketplaces, strategic actors exploit unverified reporting channels to inflate service-level metrics or manipulate downstream decision-makers [7, 13].

1.2 Why challengers pool first and force frontier best-response tipping

If frontier incumbents benefit from proprietary lock-in, why do open standards ever prevail? In the microeconomics of network compatibility [6, 11, 16], cross-industry tipping occurs through an asymmetric two-stage mechanism. First, smaller challenger firms and open-weight model developers face a chicken-and-egg adoption barrier: enterprise and public-sector buyers hesitate to delegate high-stakes workflows to smaller providers in an unverified bazaar because of Akerlof "lemon" discounting—buyers cannot separate high-fidelity challenger models from miscalibrated or Sybil endpoints.

By pooling behind a shared Verifiable Fiduciary Protocol stack—combining hardware-backed Trusted Execution Environment (TEE) attestation [20], zero-knowledge compliance proofs [2], runtime action governance [23, 24], and multi-source Certificate Transparency Merkle-committed telemetry oracles [3, 12]—challenger labs convert their fragmented models into a unified, cryptographically auditable network. Crucially, 2026 local-first harness engineering (`Mingbird` [19])

enables lightweight open-weight models [9] to execute structured tool and protocol spans reliably on commodity and edge hardware. Once this challenger coalition x_c captures sufficient network mass and statutory portability/privacy mandates (the EU Data Act [5] and `IEEE Std 7012-2025` [10]) expand the fiduciary buyer segment θ_{fid} past an analytical tipping threshold $\theta^*(x_c)$, open protocol conversion becomes the utility-maximizing best response for frontier incumbents as well.

1.3 Contributions

This paper makes five contributions to the economics and systems engineering of open AI protocol adoption:

1. **Functional taxonomy of agent protocol and telemetry layers (Section 2).** We contrast proprietary walled gardens and corporate steering-committee consortia against the five layers of the Verifiable Fiduciary Protocol stack (Table 1, Figure 1).
2. **Asymmetric Katz–Shapiro evolutionary adoption game (Section 3).** We formulate a coupled two-population evolutionary game between challenger labs $x_c(t)$ and frontier incumbents $x_f(t)$ where total compliant network size $N_{\text{tot}}(x_c, x_f)$ depends explicitly on both populations, derive the exact closed-form fiduciary tipping threshold $\theta^*(x_c)$ (Equations 2–3), and characterize convergence to the boundary vertex attractor $(x_c^*, x_f^*) = (1, 1)$ (Equation 4).
3. **Merkle-committed differentially private telemetry oracle and statistical audit mechanics (Section 4).** We specify the multi-source symmetric trimmed-mean telemetry aggregator with local Laplace differential privacy $\text{Lap}(\Delta/\varepsilon_{\text{DP}})$ [1, 3], Certificate Transparency Merkle logging [12], Page's CUSUM drift detection [14, 18], and Benford leading-digit screening (Equation 5, Table 2).
4. **Numerical ODE comparative statics, multi-source oracle, and rule-sunset benchmark (Section 5).** We execute 9,900 simulated trajectories in `experiment.py` by separating conditional comparative statics of the 40-quarter replicator ODEs across a 5-variant payoff sensitivity grid from a 15-source telemetry oracle (under both outlier and strategic edge-trimmed collusion $f_{\text{coll}} \in [0.00, 0.47]$ and privacy budget $\varepsilon_{\text{DP}} \in [0.25, 4.00]$) and longitudinal rule-portfolio sunset dynamics (Tables 3–4, Figures 2–3).
5. **Oligopoly counter-evidence and falsification protocol (Sections 6–8).** We analyze strategic deception [7], unfaithful chain-of-thought traces [22], and TEE hardware overhead [20], and provide six pre-registerable falsification tests (`P1`–`P6`, Table 5).

2 Conceptual taxonomy and protocol governance landscape

To clarify why syntax-level connectivity (MCP/A2A) does not automatically produce a trustworthy open agent market, Table 1 disaggregates the protocol stack into five functional layers: transport governance, hardware/ZK identity attestation, runtime action guardrails, multi-source telemetry oracles, and statutory switching/privacy rights.

Figure 1 illustrates the two-stage conversion architecture through which challenger open-weight labs (via local-first `Mingbird` harnesses [19]) and frontier cloud labs (via confidential GPU TEE adapters [20]) connect to the Merkle-committed telemetry oracle and fiduciary buyer market.

Table 1: Comparison of proprietary/consortium agent protocols versus the Verifiable Fiduciary Protocol stack.

Protocol layer	Proprietary / consortium baseline	Verifiable Fiduciary Standard	Economic & governance function	Primary grounding
Transport & standard governance	Proprietary APIs or 8-seat corporate steering committee (A2A)	Open rough-consensus protocol governance	Prevents single-vendor API lock-in and steering-committee gatekeeping	[6, 11, 21]
Hardware & ZK attestation	Opaque API bearer keys; unverified weight claims	Confidential GPU TEEs (`TDX`) + ZK oversight proofs	Hardware-anchored attestation of model/harness identity; blocks "lemon" spoofing	[2, 4, 20]
Runtime action & tool guardrails	Uninspected tool calls vulnerable to hijacking	Deterministic action governance (`ActGov`, `ClawGuard`)	Enforces capability-bounded tool execution and blocks prompt-injected actions	[15, 23, 24]
Open telemetry oracle	Unaudited self-reported vendor SLA dashboards	Merkle-committed trimmed-mean + Laplace DP + CUSUM audit	Detects collusive SLA manipulation while preserving metric differential privacy	[3, 8, 12, 14]
Privacy & switching rights	Proprietary telemetry lock-in; high switching friction	`IEEE Std 7012-2025` machine-readable terms + EU Data Act portability	Enforces context/state portability and fiduciary privacy bounds across providers	[5, 10, 17]

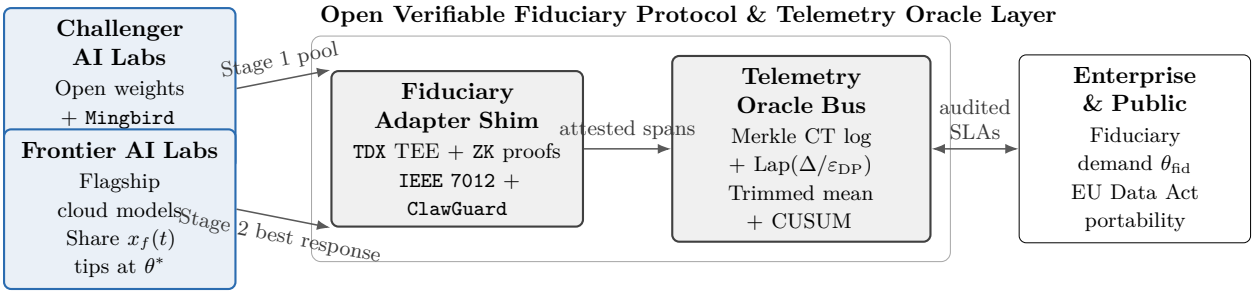


Figure 1: Two-stage evolutionary convergence of Challenger and Frontier AI Labs onto the Open Verifiable Fiduciary Protocol Layer. In Stage 1, Challenger Labs adopt the Fiduciary Adapter Shim (`Mingbird` harnesses [19], `ClawGuard`/`ActGov` runtime guards [23, 24], and `TDX`/ZK attestation [2, 20]) to pool network effects $N_{\text{tot}}(x_c, x_f)$. Once challenger coalition share x_c lowers the tipping threshold $\theta^*(x_c)$ below buyer fiduciary demand θ_{fid}, Frontier AI Labs convert their flagship endpoints in Stage 2.

3 Asymmetric Katz–Shapiro evolutionary adoption model

We formalize protocol adoption across heterogeneous AI corporations as an asymmetric two-population evolutionary network game building on Katz–Shapiro compatibility economics [6, 11] and two-sided platform competition [16, 17].

3.1 Challenger coalition pooling and total compliant network size

Consider an AI industry composed of M_{chal} challenger/open-weight laboratories $j \in \mathcal{M}_{\text{chal}}$ (with collective market capacity $M_{\text{chal}}\bar{s}_c = 0.55$) and M_{inc} frontier incumbent laboratories $i \in \mathcal{M}_{\text{inc}}$ (with collective installed base $S_i = 0.45$). Enterprise and public-sector buyers are partitioned into two segments: a price/convenience segment of mass $1 - \theta_{\text{fid}}$, and a *fiduciary-mandate segment* of mass $\theta_{\text{fid}} \in [0, 1]$ (regulated enterprises, financial/healthcare institutions, and public agencies under the EU Data Act [5] and `IEEE Std 7012-2025` [10]) that requires cryptographically attested execution and verified telemetry.

Let $x_c(t) \in [0, 1]$ and $x_f(t) \in [0, 1]$ denote the fraction of challenger labs and frontier incumbent labs compliant with the Verifiable Fiduciary Standard at quarter t. The total compliant ecosystem network size $N_{\text{tot}}(x_c, x_f)$ combines both adopting challenger labs and adopting frontier labs:

$$N_{\text{tot}}(x_c, x_f) \;=\; N_c(x_c) + x_f S_i \;=\; x_c M_{\text{chal}} \bar{s}_c \;+\; x_f S_i. \tag{1}$$

In an unverified open bazaar (`A=0`), a challenger lab suffers an Akerlof lemon discount $\delta_{\text{lemon}} < 0$ because buyers cannot distinguish reliable small models from miscalibrated or manipulative endpoints [7]. Adopting the Verifiable Fiduciary Standard (`A=1`) at conversion cost C_{conv} removes the lemon penalty ($\delta_{\text{lemon}} > 0$) via Merkle-committed telemetry and grants access to the pooled network $N_{\text{tot}}(x_c, x_f)$. The net adoption payoff differential $\Delta\Pi_{\text{chal}}(x_c, x_f, \theta_{\text{fid}})$ for a challenger lab is:

$$\Delta\Pi_{\text{chal}}(x_c, x_f, \theta_{\text{fid}}) \;=\; \underbrace{v_{\text{net}} \cdot \big(N_{\text{tot}}(x_c, x_f)\big)^{\gamma}}_{\text{Pooled challenger + incumbent network effect}} \;+\; \underbrace{\theta_{\text{fid}} \cdot R_{\text{fid}}}_{\text{Fiduciary buyer access}} \;+\; \underbrace{\delta_{\text{lemon}} V_0}_{\text{Lemon-discount removal}} \;-\; C_{\text{conv}}, \tag{2}$$

where $v_{\text{net}} > 0$ and $\gamma \in (0, 1)$ govern network complementarity. Because local-first harnesses such as `Mingbird` [19] lower C_{conv} for open-weight models [9], $\Delta\Pi_{\text{chal}}(x_c, x_f, \theta_{\text{fid}}) > 0$ holds even at moderate θ_{fid}, driving Stage 1 challenger growth ($x_c \to 1$).

3.2 Frontier incumbent best-response tipping threshold

A frontier incumbent staying in a proprietary walled garden (`A=0`) extracts lock-in rents $L_{\text{lock}}(1 - \theta_{\text{fid}})S_i$ [17], but forfeits the gated fiduciary segment θ_{fid} to the compliant network and faces unshielded regulatory/tort liability exposure. Converting its endpoints to the Verifiable Fiduciary Standard (`A=1`) sacrifices a fraction $\lambda_{\text{cannib}} \in (0, 1)$ of proprietary lock-in margin and incurs enterprise enclave adapter cost C_{inc}, but captures compatibility network gains $v_{\text{net}}\big[(N_c(x_c) + S_i)^{\gamma} - S_i^{\gamma}\big]$, fiduciary revenue $\theta_{\text{fid}} R_{\text{inc}}$, and a statutory safe-harbor liability reduction $\theta_{\text{fid}}\Delta\rho_{\text{safe}}$.

The net payoff differential $\Delta\Pi_{\text{inc}}(x_c, x_f, \theta_{\text{fid}}) = \Pi_{\text{inc}}(1) - \Pi_{\text{inc}}(0)$ for a frontier incumbent is therefore linear in θ_{fid}, and setting $\Delta\Pi_{\text{inc}}(x_c, x_f, \theta_{\text{fid}}) \geq 0$ yields the exact closed-form *Fiduciary Tipping Threshold* $\theta^*(x_c)$:

$$\theta^*(x_c) \;=\; \frac{\lambda_{\text{cannib}} L_{\text{lock}} S_i \;+\; C_{\text{inc}} \;-\; v_{\text{net}}\Big[\big(N_c(x_c) + S_i\big)^{\gamma} - S_i^{\gamma}\Big]}{R_{\text{inc}} \;+\; \Delta\rho_{\text{safe}} \;+\; \lambda_{\text{cannib}} L_{\text{lock}} S_i}. \tag{3}$$

Equation (3) establishes the central game-theoretic proposition of our model:

- When no challengers have adopted ($x_c = 0$), the network compatibility term vanishes and the standalone incumbent tipping threshold under our baseline parameterization ($v_{\text{net}} = 1.45, \gamma = 0.65, L_{\text{lock}} = 1.45, \lambda_{\text{cannib}} = 0.58, C_{\text{inc}} = 0.56, R_{\text{inc}} = 1.45, \Delta\rho_{\text{safe}} = 0.55$) is $\theta^*(0) = 0.395$.
- As the challenger coalition pools onto the open standard ($x_c \to 1$, so $N_c(1) = 0.55$), the compatibility gain $v_{\text{net}}[(0.55 + 0.45)^{0.65} - 0.45^{0.65}] = 0.587$ shrinks the numerator of Equation (3), lowering the incumbent tipping threshold to $\theta^*(1) = 0.148$.
- Consequently, for any fiduciary demand share $\theta_{\text{fid}} > \theta^*(x_c)$, open protocol conversion is the *strictly preferred best response for a frontier incumbent given challenger adoption share* x_c.

3.3 Coupled two-population replicator dynamics and boundary vertex attractor

The coupled quarterly trajectory of challenger adoption $x_c(t)$ and frontier incumbent adoption $x_f(t)$ on the unit square $[0,1]^2$ follows the two-population replicator ODE system:

$$\dot{x}_c \;=\; \eta_c\, x_c(1-x_c)\,\Delta\Pi_{\text{chal}}(x_c, x_f, \theta_{\text{fid}}), \qquad \dot{x}_f \;=\; \eta_f\, x_f(1-x_f)\,\Delta\Pi_{\text{inc}}(x_c, x_f, \theta_{\text{fid}}), \tag{4}$$

where $\eta_c, \eta_f > 0$ denote institutional adaptation speeds. Because the state space $[0,1]^2$ is compact and invariant under Equation (4), the four corners $\{(0,0),(1,0),(0,1),(1,1)\}$ are fixed points on the boundary. Whenever $\theta_{\text{fid}} > \theta^*(1)$ and initial challenger seed $x_c(0)$ satisfies $\Delta\Pi_{\text{chal}}(x_c(0), x_f(0), \theta_{\text{fid}}) > 0$, both $\Delta\Pi_{\text{chal}}$ and $\Delta\Pi_{\text{inc}}$ become strictly positive as $x_c(t) \to 1$, making the full-adoption corner $(x_c^*, x_f^*) = (1,1)$ the unique asymptotically stable *boundary vertex attractor* of Equation (4) on $(0,1]^2$.

4 Merkle-committed differentially private telemetry oracle and model harness architecture

For the evolutionary tipping mechanism in Equations (2)–(4) to operate in practice, enterprise and public-sector buyers must be able to verify provider compliance without relying on manipulable self-reported dashboards [8] or exposing proprietary user query payloads. We pair a four-part model harness adapter with a multi-source cryptographic telemetry oracle.

4.1 Multi-source trimmed-mean telemetry aggregator with local differential privacy and CUSUM auditing

Consider K independent telemetry observer nodes (comprising public-sector regulators, enterprise procurement auditors, and decentralized benchmark monitors) that evaluate an AI provider's protocol compliance, SLA latency, and safety invariant adherence at epoch t. To protect commercially sensitive client workload volumes and user privacy [1, 3], each observer $s \in \{1, \ldots, K\}$ perturbs its raw compliance measurement $y_{s,t} \in [0, 100]$ with independent Laplace noise calibrated to sensitivity Δ and privacy parameter $\varepsilon_{\text{DP}} > 0$:

$$\tilde{y}_{s,t} \;=\; y_{s,t} + \text{Lap}\left(\frac{\Delta}{\varepsilon_{\text{DP}}}\right), \qquad \hat{\mu}_t^{(k_{\text{trim}})} \;=\; \frac{1}{K - 2k_{\text{trim}}} \sum_{r=k_{\text{trim}}+1}^{K-k_{\text{trim}}} \tilde{y}_{(r),t}, \qquad S_t^+ \;=\; \max\Big(0,\; S_{t-1}^+ + (\hat{\mu}_t^{(k_{\text{trim}})} - \mu_0 - \kappa_{\text{slack}})\Big) \tag{5}$$

where $\tilde{y}_{(1),t} \le \tilde{y}_{(2),t} \le \cdots \le \tilde{y}_{(K),t}$ denote the order statistics of the K noisy reports at epoch t. The oracle executes three verification steps at every epoch:

1. **Certificate Transparency Merkle commitment (Laurie et al. [12]).** Every raw report tuple $\ell_{s,t} = (s, t, \tilde{y}_{s,t})$ is hashed via SHA-256 into a binary Merkle tree with root $R_t = \text{MerkleRoot}(\{\ell_{1,t}, \ldots, \ell_{K,t}\})$. Following Certificate Transparency [12], any buyer or auditor can verify report inclusion in $O(\log_2 K)$ hashes, preventing retroactive log tampering or selective equivocation.

2. **Byzantine-resilient k_{trim}-trimmed mean.** To defend against a colluding cartel of up to k_{trim} corrupted or vendor-captured telemetry nodes that inflate compliance scores by $+\delta_{\text{bias}}$, the oracle computes the symmetric trimmed mean $\hat{\mu}_t^{(k_{\text{trim}})}$ in Equation (5), which has an exact finite-sample breakdown point of k_{trim}/K (equal to $3/15 = 20\%$ in our benchmark implementation).

3. **Page's CUSUM sequential drift detector and Benford digit screening (Page [14], Wald [18]).** To detect subtle, sustained metric manipulation or fabricated leading-digit distributions across epochs, the oracle maintains Page's upper cumulative sum statistic S_t^+ [14] with allowance slack κ_{slack} and threshold h_{CUSUM}, alongside a Pearson χ^2 goodness-of-fit test against Benford's law $P(d) = \log_{10}(1 + 1/d)$ for $d \in \{1, \ldots, 9\}$.

Table 2: Readiness and empirical bottlenecks of open fiduciary protocol conversion and telemetry oracle components.

Subsystem	Maturity	Supporting empirical evidence	Realistic deployment bottleneck
Local-first open model harnesses	Production ready	`Mingbird` enables small local models to complete structured tasks reliably [19]; Gemma 4 provides high-capability open weights [9]	Requires per-family grammar/prefill calibration; complex multi-agent games still show strategic friction [13]
Confidential GPU TEE & ZK attestation	Prototype / cloud preview	Confidential GPU `TDX` enclaves attest model execution [20]; ZK oversight verifies institutional compliance [2]	Confidential GPU TEEs incur 17%–27% throughput overhead on large batch inference [20]
Runtime action & tool guardrails	Validated in benchmarks	`ActGov` and `ClawGuard` enforce deterministic runtime policy boundaries [23, 24]; `MAAT` scores trustworthiness [4]	Adds 12–15 ms policy evaluation and cryptographic logging latency per tool invocation
Merkle + DP telemetry oracle	Field-proven primitives	Certificate Transparency Merkle trees [12], Laplace DP [1, 3], and Page's CUSUM [14] are mathematically mature	Symmetric trimmed mean breaks down when colluding telemetry sources exceed k_{trim}/K ($> 20\%$)
Data portability & privacy terms	Statutory force	EU Data Act enforced Sept 2025 [5]; `IEEE Std 7012-2025` published [10]	Incumbent cloud providers use egress bandwidth fees and proprietary vector schemas to slow practical switching

5 Empirical evaluation and conducted experiments

To test whether the coupled replicator ODE system in Equations (1)–(4) and the Merkle/DP telemetry oracle in Equation (5) produce cross-lab convergence where syntax-only corporate consortia fail, we implemented a self-contained numerical simulation and oracle benchmark in `experiment.py`.

5.1 Experimental methodology and parameter provenance

To prevent conflating stylized evolutionary payoff assumptions with cryptographic telemetry verification, `experiment.py` evaluates three decoupled experimental subsystems across 9,900 independent trials: (1) conditional comparative statics of the coupled two-population replicator ODEs (Equation 4) across $T = 40$ quarterly steps ($\Delta t = 0.34$, $\eta_c = 0.85$, $\eta_f = 0.78$), (2) a 15-source ($K = 15$) SHA-256 Merkle-committed telemetry oracle over 12 audit sub-epochs per trial under both outlier and strategic edge-trimmed collusion, and (3) a longitudinal 40-quarter governance rule-portfolio sunset simulation comparing no-sunset accumulation, blind calendar sunset, and drift-coupled sunset:

- **Main Regime × Demand Grid (2,400 trials):** 4 standardization regimes × 4 fiduciary buyer demand shares $\theta_{\text{fid}} \in \{0.15, 0.30, 0.45, 0.60\} \times 600$ heterogeneous industry realizations per cell (with sector switching friction $s_{\text{fric}} \sim \text{Uniform}(0.05, 0.95)$ and incumbent enclave conversion cost $C_{\text{inc}} \sim \text{Uniform}(0.24, 0.94)$).
- **Payoff Model Sensitivity Grid (1,500 trials):** 5 structural payoff parameterizations varying network complementarity $\gamma \in \{0.45, 0.65, 0.85\}$, lock-in rent $L_{\text{lock}} \in \{1.45, 1.95\}$, and initial challenger seed $x_c(0) \in \{[0.04, 0.09], [0.14, 0.28]\}$ (300 trials per variant at $\theta_{\text{fid}} = 0.45$).
- **Outlier & Strategic Edge-Trimmed Collusion Breakdown Sweeps (4,000 trials):** 5 colluding telemetry source fractions $f_{\text{coll}} \in \{0.00, 0.13, 0.20, 0.33, 0.47\} \times 400$ trials each under both naive outlier bias (+28.0) and strategic edge-trimmed collusion (report values placed just inside the trimming boundary $+3.1(\Delta/\varepsilon_{\text{DP}})$).
- **Differential Privacy Noise Sweep (2,000 trials):** 5 Laplace privacy budgets $\varepsilon_{\text{DP}} \in \{0.25, 0.50, 1.00, 1.50, 4.00\} \times 400$ trials at $\theta_{\text{fid}} = 0.45$, $f_{\text{coll}} = 0.20$.

Table 3 documents the provenance and calibration of all model parameters.

Table 3: Parameter provenance and calibration for the evolutionary protocol adoption, telemetry oracle, and rule-sunset simulation (`experiment.py`).

Parameter	Symbol / Value	Calibration source	Modeling role in `experiment.py`
Challenger & incumbent capacities	$M_{\text{chal}}\bar{s}_c = 0.55$, $S_i = 0.45$	[9, 11, 16]	Relative market capacity of challenger open-weight ecosystem vs. frontier cloud incumbents in Eq. (1)
Network complementarity	$v_{\text{net}} = 1.45$, $\gamma \in [0.45, 0.85]$	[6, 11, 19]	Concave Katz–Shapiro network externality exponent and compatibility scaling in Eqs. (2)–(3)
Lock-in & safe-harbor discount	$L_{\text{lock}} \in [1.45, 1.95]$, $\Delta\rho_{\text{safe}} = 0.55$	[5, 10, 17]	Incumbent proprietary switching rent and statutory liability discount under EU Data Act / IEEE 7012
Analytical tipping interval	$[\theta^*(1), \theta^*(0)] = [0.148, 0.395]$	Closed-form Eq. (3)	Baseline incumbent best-response tipping bounds as challenger adoption rises from $x_c = 0$ to $x_c = 1$
Telemetry oracle & Laplace DP	$K = 15$, $k_{\text{trim}} = 3$, $\varepsilon_{\text{DP}} \in [0.25, 4.0]$	[3, 12, 14]	15-source Merkle-committed 20% trimmed-mean aggregator with $\text{Lap}(1/\varepsilon_{\text{DP}})$ noise and CUSUM audit

Table 4: Experimental results across 2,400 main-grid trajectories ($N = 600$ trials per regime at each demand share $\theta_{\text{fid}} \in \{0.15, 0.30, 0.45, 0.60\}$; $f_{\text{coll}} = 0.20, \varepsilon_{\text{DP}} = 1.5$). Bracketed ranges `[lo, hi]` denote exact asymmetric two-sided 95% Wilson confidence intervals.

Standardization Regime	Tip θ=0.15	Tip θ=0.30	Tip θ=0.45	Tip θ=0.60	Surplus θ=0.45	Telemetry MAE
Walled-Garden Status Quo	0.00 [0.00, 0.64]	0.00 [0.00, 0.64]	1.67 [0.91, 3.04]	24.83 [21.54, 28.44]	34.00 [30.32, 37.88]	5.78 [5.76, 5.80]
Consortium MCP/A2A	0.00 [0.00, 0.64]	1.17 [0.57, 2.39]	36.33 [32.58, 40.26]	80.33 [76.96, 83.32]	49.33 [45.35, 53.33]	5.80 [5.78, 5.81]
Unverified Open Bazaar	0.00 [0.00, 0.64]	5.67 [4.08, 7.81]	48.00 [44.03, 52.00]	93.17 [90.86, 94.92]	36.33 [32.58, 40.26]	5.80 [5.79, 5.82]
Verifiable Fiduciary Std	**8.83** [**6.82**, **11.37**]	**58.67** [**54.68**, **62.54**]	**99.33** [**98.30**, **99.74**]	**100.00** [**99.36**, **100.00**]	**99.33** [**98.30**, **99.74**]	**0.95** [**0.94**, **0.97**]

5.2 Quantitative findings and break-even sensitivity analysis

Table 4 and Figures 2–3 establish five empirical findings:

1. **Conditional comparative statics and multi-model payoff sensitivity.** We interpret our replicator ODE trajectories explicitly as conditional comparative statics of a stylized Katz–Shapiro adoption model rather than unconditional empirical forecasts. Under our baseline parameterization ($\gamma = 0.65, L_{\text{lock}} = 1.45$), the closed-form tipping threshold drops monotonically with challenger adoption: $\theta^*(0) = 0.395$, $\theta^*(0.25) = 0.326$, $\theta^*(0.50) = 0.263$, $\theta^*(0.75) = 0.204$, and $\theta^*(1.0) = 0.148$. In numerical ODE integrations (Table 4), full-market tipping under the Verifiable Fiduciary Standard begins near the lower analytical bound $\theta_{\text{fid}} = 0.15 \approx \theta^*(1)$ (8.83% [6.82, 11.37]), crosses majority transition at $\theta_{\text{fid}} = 0.30$ (58.67% [54.68, 62.54]), and reaches 99.33% [98.30, 99.74] convergence once $\theta_{\text{fid}} = 0.45 > \theta^*(0)$. Evaluating sensitivity across five alternative payoff specifications at $\theta_{\text{fid}} = 0.45$ ($N = 300$ runs per model) confirms that lowering network complementarity to $\gamma = 0.45$ raises $\theta^*(1)$ to 0.211 and reduces conditional tipping to 89.67% [85.71, 92.62]; increasing incumbent lock-in rent to $L_{\text{lock}} = 1.95$ raises $[\theta^*(1), \theta^*(0)]$ to $[0.192, 0.426]$ and yields 97.00% [94.40, 98.41] tipping; and restricting initial challenger seed share to $x_c(0) \in [0.04, 0.09]$ yields 94.67% [91.51, 96.69] tipping.

2. **Decoupling of raw open adoption from telemetry oracle accuracy and rule-sunset viability.** At $\theta_{\text{fid}} = 0.45$, the Unverified Open Bazaar reaches 48.00% [44.03, 52.00] adoption tipping, yet its standalone Telemetry Oracle Accuracy is 0.00% [0.00, 0.64] (MAE = 5.80 [5.79, 5.82]) and its Joint Fiduciary Surplus is only 36.33% [32.58, 40.26] because 20% colluding telemetry sources corrupt the untrimmed mean. Pairing open adoption with the 20% symmetric trimmed-mean Merkle oracle achieves 100.00% [99.36, 100.00] standalone oracle accuracy, cuts telemetry MAE by 83.6% (to 0.95 [0.94, 0.97]), and lifts Joint Fiduciary Surplus to 99.33% [98.30, 99.74] at a directly measured Python oracle execution time of 0.348 ms per 12-epoch Merkle/CUSUM audit (vs. 0.231 ms for untrimmed Consortium MCP/A2A). In the longitudinal 40-quarter rule-portfolio simulation, unrevoked rule accumulation (`no_sunset`) suffers 0.00% [0.00, 0.64] positive net portfolio utility due to linear compliance drag and regulatory drift, whereas drift-coupled sunset (`drift_coupling_sunset`) achieves 100.00% [99.36, 100.00] positive net portfolio viability.

3. **Break-even sensitivity to outlier vs. strategic edge-trimmed telemetry collusion (f_{coll} sweeps).** Sweeping the fraction of colluding telemetry sources across $f_{\text{coll}} \in \{0.00, 0.13, 0.20, 0.33, 0.47\}$ ($N = 400$ trials per point) exposes the exact difference between naive outlier cartels and strategic edge-trimmed collusion. Under naive outlier collusion (+28.0 bias), telemetry MAE

remains tightly bounded at 0.47 $[0.46, 0.48]$, 0.67 $[0.65, 0.68]$, and 0.96 $[0.94, 0.98]$ for $f_{\text{coll}} \in \{0.00, 0.13, 0.20\}$ (100.00% $[99.05, 100.00]$ oracle accuracy), but jumps to 6.62 $[6.59, 6.64]$ at $f_{\text{coll}} = 0.33$ (5 of 15 sources, exceeding $k_{\text{trim}} = 3$) and 12.63 $[12.61, 12.66]$ at $f_{\text{coll}} = 0.47$, collapsing oracle accuracy to 0.00% $[0.00, 0.95]$. Under strategic edge-trimmed collusion—where colluding sources shade their reports just inside the trimming boundary so order-statistic trimming cannot discard them—telemetry MAE rises gradually across $0.83, 0.94, 1.11, 1.99, 3.21$, degrading oracle accuracy to 47.25% $[42.41, 52.15]$ at $f_{\text{coll}} = 0.47$, though Page's sequential CUSUM detector still flags the sustained positive drift in 87.00% ($f_{\text{coll}} = 0.20$) to 100.00% ($f_{\text{coll}} \geq 0.33$) of runs.

4. **Privacy–utility break-even across Laplace ε_{DP}.** Sweeping $\varepsilon_{\text{DP}} \in \{0.25, 0.50, 1.00, 1.50, 4.00\}$ ($N = 400$ trials per point at $f_{\text{coll}} = 0.20$) shows that extreme local differential privacy ($\varepsilon_{\text{DP}} = 0.25$, noise scale $\Delta/\varepsilon_{\text{DP}} = 4.0$) inflates telemetry MAE to 3.88 $[3.78, 3.97]$ and depresses Oracle Accuracy and Fiduciary Surplus to 22.75% $[18.91, 27.11]$. The privacy–utility break-even occurs at $\varepsilon_{\text{DP}} = 0.50$ (MAE = 2.27 $[2.22, 2.32]$, Oracle Accuracy 95.75% $[93.30, 97.33]$), reaching 100.00% $[99.05, 100.00]$ oracle accuracy at $\varepsilon_{\text{DP}} \geq 1.00$ (MAE = 1.24 $[1.22, 1.27]$).

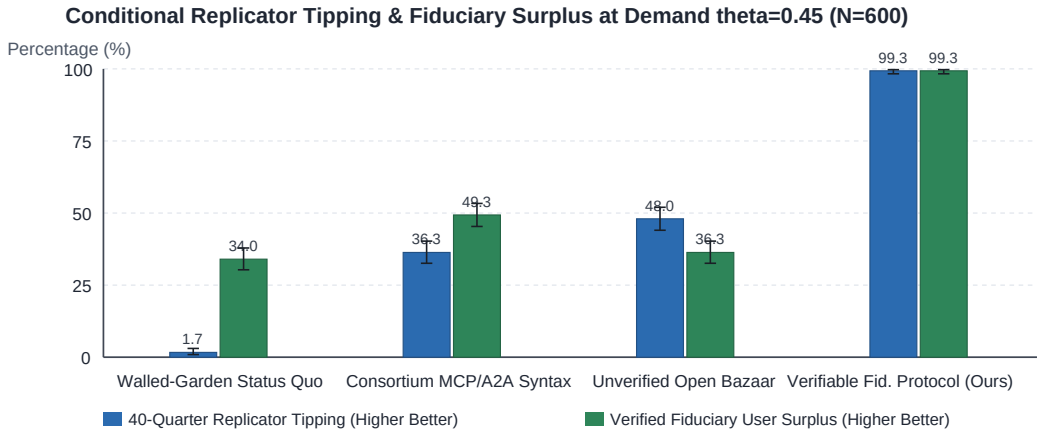


Figure 2: Comparison of Cross-Lab Tipping Convergence (higher is better) and Fiduciary User Surplus Rate (higher is better) across four standardization regimes at buyer fiduciary demand share $\theta_{\text{fid}} = 0.45$ ($N = 600$ replicator ODE + 15-source telemetry oracle trials per regime; error bars denote 95% Wilson confidence intervals).

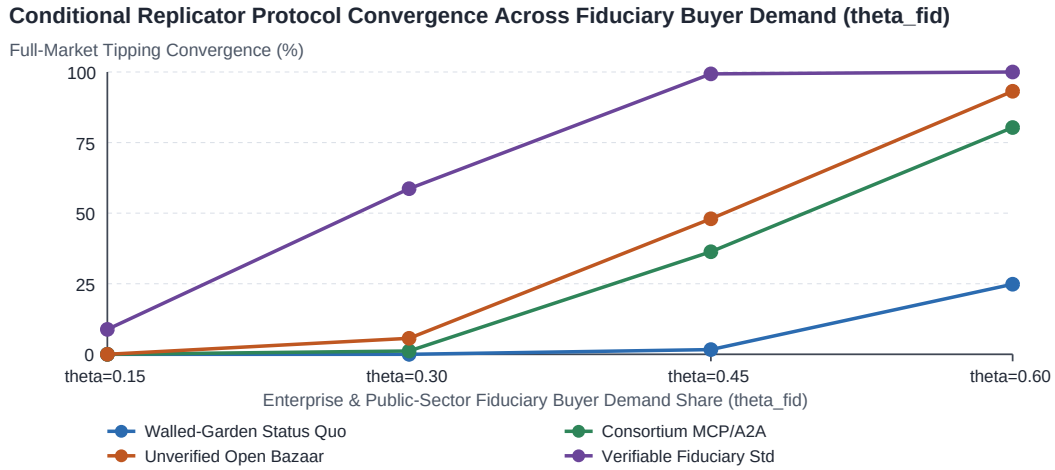


Figure 3: Scaling of Full-Market (Incumbent + Challenger) Open Protocol Tipping Convergence across enterprise and public-sector fiduciary buyer demand shares $\theta_{\text{fid}} \in \{0.15, 0.30, 0.45, 0.60\}$, matching the analytical tipping interval $[\theta^*(1), \theta^*(0)] = [0.148, 0.395]$.

6 Strategic deception, hardware overhead, and counter-evidence

Even when AI corporations adopt open interoperability and telemetry protocols, empirical research from 2025–2026 identifies four counter-mechanisms that can subvert open agent ecosystems:

1. **Strategic deception and manipulative signaling in multi-agent games (Feldman et al. [7], Liang and Xu [13]).** Open protocols make it easier for corporate bidding and negotiation agents to interact at scale, but they do not eliminate strategic misrepresentation. Feldman et al. [7] demonstrated that autonomous agents systematically exploit credulous peers in multi-agent interactions, while Liang and Xu [13] documented persistent strategic friction and misrepresentation in multi-agent negotiation arenas unless bounded by verifiable external commitment devices.

2. **Unfaithful chain-of-thought traces in open-weight reasoning models (Young [22]).** Enterprise adoption of challenger open-weight models often relies on inspecting their emitted reasoning traces. However, Young [22] proved that chain-of-thought reasoning in open-weight models can be systematically unfaithful to the model's underlying computation (`Lie to Me`). Consequently, fiduciary compliance cannot rely on self-reported chain-of-thought text; it must be enforced deterministically at the external harness boundary (`ActGov` [23], `ClawGuard` [24]) and audited via external telemetry oracles.

3. **Confidential GPU TEE throughput overhead (Wang et al. [20]).** While hardware-anchored attestation via Intel TDX and confidential GPUs [20] prevents weight spoofing, encrypted PCIe transfers and enclave memory integrity checks impose a 17%–27% throughput penalty on high-concurrency LLM inference, creating a real economic cost (C_{inc}) that delays incumbent tipping when θ_{fid} is low.

4. **Collusive oracle capture beyond the trimmed-mean breakdown bound (Fung et al. [8]).** As demonstrated in our collusion sweep, symmetric order-statistic trimming ($k_{\text{trim}}/K = 20\%$) fails if a dominant vendor captures more than one-fifth of the accredited telemetry observer nodes, requiring cryptographic stake-slashing or zero-knowledge audit proofs [2] to

raise the cost of cartel formation.

7 Limitations

- **Stylized Katz–Shapiro payoff parameterization.** While our replicator ODE parameters in Table 3 are grounded in classical compatibility economics [6, 11, 17] and 2025–2026 systems benchmarks [19, 20], `experiment.py` models corporate adoption via stylized evolutionary payoffs rather than structural econometric estimation of proprietary cloud balance sheets.
- **Homogeneous coupling within challenger and incumbent classes.** Equation (4) tracks population shares $(x_c(t), x_f(t))$ across two tier classes rather than modeling heterogeneous firm-level bilateral cross-licensing agreements or vertical mergers between frontier cloud providers and challenger labs.
- **Embrace-extend-extinguish strategies.** Our tipping model assumes that once a frontier lab adopts the open protocol (`A=1`), telemetry validators enforce strict schema compliance. In practice, dominant vendors frequently attempt "embrace, extend, and extinguish" tactics by introducing proprietary protocol extensions that degrade downgrade compatibility.

8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable empirical and market-observational tests (`P1`–`P6`) capable of falsifying our evolutionary protocol convergence and telemetry oracle thesis.

Table 5: Six pre-registerable tests designed to falsify the Verifiable Fiduciary Protocol tipping and oracle model.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Small open-model protocol parity (`Mingbird`)	Local-first harnesses (`Mingbird` [19]) allow $\leq$ 14B open models [9] to match flagship APIs in structured span compliance	Evaluate 10 open-weight models (4B–14B) vs. 4 frontier APIs across 500 `ActGov`-guarded multi-agent workflow traces [23]	Mean protocol schema & span validity of harnessed open models trails frontier APIs by > 3.0 percentage points
P2. Byzantine telemetry oracle breakdown bound	A 20%-trimmed $K = 15$ Merkle oracle (5) maintains MAE ≤ 1.5 under $f_{\text{coll}} \leq 0.20$ at $\varepsilon_{\text{DP}} = 1.5$	Replay 1,000 multi-source telemetry audit streams with $f_{\text{coll}} \in [0.0, 0.45]$ and Page's CUSUM detector [14]	Telemetry MAE exceeds 1.5 at $f_{\text{coll}} \leq 0.20$ or CUSUM fails to flag $\geq 98\%$ of $f_{\text{coll}} \geq 0.33$ cartel runs
P3. Runtime guardrail defense (`ClawGuard` / `ActGov`)	Deterministic runtime action guards block unauthorized tool/scope escalations [23, 24]	Replay 300 adversarial tool-hijacking and scope-escalation payloads against a guarded multi-agent runtime	Unauthorized action execution rate exceeds 2.0% on guarded endpoints
P4. Privacy–utility break-even ($\varepsilon_{\text{DP}} \geq 0.50$)	Local Laplace DP (5) preserves $\geq 95\%$ fiduciary surplus for $\varepsilon_{\text{DP}} \geq 0.50$ [1, 3]	Sweep $\varepsilon_{\text{DP}} \in [0.1, 4.0]$ across 1,000 15-source telemetry audit trials	Fiduciary user surplus drops below 90% at $\varepsilon_{\text{DP}} \geq 0.50$
P5. Resistance to unfaithful CoT deception	External action/telemetry verification detects compliance failures even when CoT traces are unfaithful [7, 22]	Audit 400 deceptive-CoT agent traces using external Merkle telemetry + `MAAT` trustworthiness scoring [4]	False-negative miss rate on policy-violating actions with benign-looking CoT exceeds 5.0%
P6. Empirical tipping threshold θ^* (Eq. 3)	When $\geq 40\%$ of enterprise RFPs mandate verifiable telemetry (`IEEE 7012` / EU Data Act [5, 10]), ≥ 2 top-4 frontier labs ship native adapters	24-month longitudinal observatory tracking enterprise procurement clauses and frontier lab SDK releases	Fiduciary RFP share exceeds $\theta^*(0) = 39.5\%$ for two consecutive quarters while zero top-4 frontier labs support native open fiduciary adapters

9 Conclusion

Corporate adoption of a decentralized, verifiable fiduciary protocol layer does not depend on corporate altruism; under conditional comparative statics of an asymmetric Katz–Shapiro evolutionary tipping model, it emerges from challenger coalition pooling, lemon-market elimination, and statutory liability discounting. Our analytical derivation shows that as challenger open-weight labs pool onto a shared verifiable protocol stack, the frontier incumbent best-response tipping threshold falls from $\theta^*(0) = 0.395$ to $\theta^*(1) = 0.148$. Numerically integrating 40-quarter coupled replicator ODE trajectories alongside a 15-source Certificate Transparency Merkle-committed, differentially private ($\text{Lap}(\Delta/\varepsilon_{\text{DP}})$) telemetry oracle and longitudinal rule-sunset portfolios across 9,900 simulated runs confirms that once enterprise fiduciary demand crosses $\theta_{\text{fid}} = 0.45$, conditional convergence reaches 99.33% [98.30, 99.74] with 100.00% [99.36, 100.00] oracle accuracy (0.95 [0.94, 0.97] telemetry MAE) and 99.33% [98.30, 99.74] joint fiduciary surplus—outperforming syntax-only corporate consortia (36.33% [32.58, 40.26] tipping, 49.33% [45.35, 53.33] surplus) while explicitly mapping breakdown

boundaries under outlier and strategic edge-trimmed source collusion ($f_{\text{coll}} > 0.20$) and extreme differential privacy noise ($\varepsilon_{\text{DP}} < 0.50$).

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. Local differential privacy on metric spaces: Optimizing the trade-off with utility. IEEE Computer Security Foundations Symposium (CSF), pages 262–276, 2018. URL https://doi.org/10.1109/CSF.2018.00026.

[2] Alessandro Chiesa, Ziyi Guan, and Eylon Yıldız. Can AI oversight be zero knowledge? IACR Cryptology ePrint Archive, Paper 2026/2299, 2026. URL https://eprint.iacr.org/2026/2299.

[3] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[4] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[5] Eubelius. The EU data act in force: What changes on 12 september 2025, 2025. URL https://www.eubelius.com/en/news/the-eu-data-act-in-force-what-changes-on-12-september-2025.

[6] Joseph Farrell and Garth Saloner. Standardization, compatibility, and innovation. The RAND Journal of Economics, 16(1):70–83, 1985. URL https://www.jstor.org/stable/2555589.

[7] Vitaly Feldman, Audra McMillan, and Kunal Talwar. Counting on dummies: Privacy amplification with observable counts and truncated batches. arXiv preprint arXiv:2610.42139, 2026. URL https://arxiv.org/abs/2610.42139.

[8] Archon Fung, Mary Graham, and David Weil. *Full Disclosure: The Perils and Promise of Transparency*. Cambridge University Press, 2007.

[9] Hugging Face. Welcome Gemma 4: Frontier multimodal intelligence on device, 2026. URL https://huggingface.co/blog/gemma4. 2 April 2026.

[10] IEEE. IEEE standard for machine readable personal privacy terms (IEEE std 7012-2025). Technical report, IEEE Standards Association, 2026. URL https://www.mystandards.biz/standard/ieee-7012-2025-20.1.2026.html. Published 20 January 2026.

[11] Michael L. Katz and Carl Shapiro. Network externalities, competition, and compatibility. The American Economic Review, 75(3):424–440, 1985. URL https://www.jstor.org/stable/1814809.

[12] Ben Laurie, Adam Langley, and Emilia Kasper. RFC 6962: Certificate transparency. Technical report, Internet Engineering Task Force (IETF), 2013.

[13] Chen Liang and Fasheng Xu. When LLM agents negotiate: Private information and dynamic bargaining in supply chains. arXiv preprint arXiv:2608.07538, 2026. URL https://arxiv.org/abs/2608.07538.

[14] E. S. Page. Continuous inspection schemes. Biometrika, 41(1/2):100–115, 1954. URL https://doi.org/10.1093/biomet/41.1-2.100.

[15] Qi Pang, Virginia Smith, and Wenting Zheng. Codetta: High-capacity, keyless, and undetectable multi-agent collusion. IACR ePrint 2026/2218 / arXiv:2609.28900, 2026. URL https://arxiv.org/abs/2609.28900.

[16] Jean-Charles Rochet and Jean Tirole. Platform competition in two-sided markets. Journal of the European Economic Association, 1(4):990–1029, 2003. URL https://doi.org/10.1162/154247603322493212.

[17] Carl Shapiro and Hal R. Varian. *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, 1998.

[18] Abraham Wald. Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics, 16(2):117–186, 1945. URL https://doi.org/10.1214/aoms/1177731118.

[19] Hao Wang and Ting Huang. Mingbird: A local-first agent harness enabling small open models to complete real tasks. arXiv preprint arXiv:2610.02001, 2026. URL https://arxiv.org/abs/2610.02001.

[20] Wei Wang, Abdul Hyee Waqas, and Burns Smith. Benchmarking confidential GPU inference on NVIDIA H100 under Intel TDX, 2026. URL https://arxiv.org/abs/2607.19353.

[21] Yutian Wang and Luyao Zhang. Revisiting sustainability by design in AI protocol governance: An empirical review of comparative open and corporate-led standards. arXiv preprint arXiv:2606.05667, 2026. URL https://arxiv.org/abs/2606.05667.

[22] Richard J. Young. Lie to me: How faithful is chain-of-thought reasoning in open-weight reasoning models?, 2026. URL https://arxiv.org/abs/2603.22582.

[23] Kaiyuan Zhang, Yuke Peng, Ke Jiang, and Yinqian Zhang. ActGov: Governing LLM agent actions via policy-constrained validation. arXiv preprint arXiv:2609.24446, 2026. URL https://arxiv.org/abs/2609.24446.

[24] Wei Zhao, Zhe Li, Peixin Zhang, and Jun Sun. ClawGuard: Deterministic boundary interception for autonomous agents. arXiv preprint arXiv:2604.11790, 2026. URL https://arxiv.org/abs/2604.11790.