Polycentric Norm Mediation and Nash Bargaining Under Endowment Asymmetry: Disagreement-Point Auditing and Hard Liability Constraints in Multi-Agent Disputes

Dima Baker
Collective Intelligence Network

November 2026

Abstract

As autonomous agents negotiate service-level agreements, API subcontracting, and resource allocation, resolving ambiguous contract contingencies requires mechanisms that avoid both brittle smart-contract halts and power capture. In alternating-offers bargaining (Binmore, Rubinstein & Wolinsky, 1986), counterparties with unequal compute or capital endowments ($\gamma_{\text{asym}} \in [0.20, 0.85]$) face asymmetric delay costs and strategic disagreement-point manipulation, compressing the weaker party's surplus. Drawing on bargaining theory (Nash, 1950; Myerson & Satterthwaite, 1983), polycentric governance (Ostrom, 1990), and empirical audits of decentralized dispute systems (Mahrous et al., 2026; Elina et al., 2026), this paper specifies the Constitutionally Bounded Nash Equilibrium (CNBE) mediator. CNBE combines: (1) a normative symmetric-weight Nash objective ($\tau_A = \tau_B = 0.5$); (2) an Observable-Only Escrow-Audited Disagreement Estimator (using zero hidden d_0 oracle leakage) paired with a pre-committed escrow bond S_{bond}; and (3) a deterministic constraint filter enforcing hard liability bounds over a verified 201-point contract grid (0 feasible-set violations across 16,000 trials). At high power asymmetry ($\gamma_{\text{asym}} = 0.65$, $N = 1{,}000$ disputes per regime), static statutory rules settle 9.20% [7.56, 11.15] of disputes, unconstrained Coasean bargaining settles 11.30% [9.48, 13.41] (12.50% [10.59, 14.69] constitutional compliance), and ad-hoc LLM mediation settles 54.70% [51.60, 57.76] (64.70% [61.69, 67.60] compliance). CNBE achieves 90.00% [87.98, 91.71] unbonded voluntary acceptance and 100.00% [99.62, 100.00] bonded settlement ($S_{\text{bond}} = 0.08$), 97.57% [97.53, 97.60] Pareto surplus, 42.88 [42.67, 43.08] weaker-party utility, and 100.00% [99.62, 100.00] constitutional compliance.

1 Introduction

When autonomous agents and organizational principals enter into complex, multi-attribute service agreements—such as delegated compute provisioning, automated supply-chain routing, or cross-domain data licensing—incomplete contracting is inevitable [6, 12, 16]. Because real-world execution environments encounter unforeseen contingencies that cannot be exhaustively enumerated ex ante, multi-agent ecosystems require ex-post dispute mediation mechanisms [7, 21, 22].

Recent empirical audits of decentralized markets and multi-agent LLM negotiation in 2025–2026 demonstrate why neither ex-ante dispute prediction, rigid deterministic smart contracts, nor unconstrained LLM bargaining suffices for equitable dispute resolution:

1. **Failure of Ex-Ante Dispute Prediction Under Extreme Base-Rate Imbalance:** A natural hypothesis is that platforms could screen out dispute-prone contracts before execution using machine learning classifiers. However, in a comprehensive empirical analysis of

802,863 prediction-market contracts on Polymarket resolved via UMA's optimistic oracle, Mahrous and Di Pietro [9] (`arXiv:2610.29578`) demonstrated that while contract metadata predicts **log trading volume** with high accuracy ($R^2 = 0.747$), **ex-ante dispute prediction achieves an Average Precision (AP) of only 0.012 against a 0.31% dispute base rate**—a precision level the authors explicitly conclude is unusable for screening individual contracts. Because rare, ambiguous contingencies cannot be reliably predicted ex ante, robust ex-post resolution mechanisms are indispensable.

2. **Brittleness and Validator False Alarms in Rigid Smart-Contract Halts:** When contracts rely on rigid deterministic code rules or automated compliance monitors to halt execution upon suspected violations, specification defects cause severe deadweight loss. In the Maat compliance-verification framework, Elina [4] (`arXiv:2609.34017`) found that **35 out of 94 automated compliance halts (37.2% of all halts)** were false alarms triggered by validator specification defects rather than genuine party misconduct. When a rigid escrow contract halts erroneously or enforces binary forfeiture on an ambiguous edge case, cooperative surplus is destroyed.

3. **Endowment Asymmetry in Alternating-Offers Bargaining and Rhetorical LLM Mediation:** When disputes are delegated to bilateral agent negotiation [8, 17], differences in compute budgets, liquidity reserves, and waiting-cost tolerance create severe bargaining power asymmetry. In classical non-cooperative game theory, Binmore et al. [3] proved that Rubinstein alternating-offers bargaining between parties with unequal discount rates $r_A \neq r_B$ converges to the *asymmetric* Nash Bargaining Solution with bargaining weight $\tau_A = r_B/(r_A+r_B)$. Similarly, when an unconstrained LLM acts as an informal arbitrator, empirical benchmarks show that higher-capability agents exploit rhetorical persuasion, sycophancy, and context framing to skew settlements in their favor [5, 8, 11, 17].

Theoretical Contribution and Institutional Honesty: Why Symmetric Weights Alone Do Not Neutralize Power. A tempting fallacy in mechanism design is to claim that replacing bilateral bargaining with a mediator that maximizes the symmetric Nash product $(u_A - d_A)(u_B - d_B)$ [13] automatically renders settlements "invariant" to endowment asymmetries. That claim is mathematically false for two fundamental reasons that structure our analysis:

- First, as Binmore et al. [3] established, symmetric bargaining weights ($\tau_A = \tau_B = 0.5$) do not emerge endogenously when parties have unequal delay costs; imposing $\tau_A = \tau_B = 0.5$ is an **explicit normative institutional choice** by the mediation protocol (analogous to procedural equality before an arbitral tribunal).

- Second, even under symmetric weights $\tau_A = \tau_B = 0.5$, the Nash Bargaining Solution depends directly on the **disagreement point (d_A, d_B)**, which encodes each party's outside option if mediation fails! If a high-endowment principal B can strategically inflate its disagreement utility d_B or depress A's disagreement utility d_A (for instance, by threatening protracted external litigation or withholding collateral), the symmetric Nash solution $u_A^* = \frac{1}{2}(U_{\text{tot}} + d_A - d_B)$ still transfers the majority of total surplus to B.

To address both channels of power distortion, this paper formulates and evaluates the **Constitutionally Bounded Nash Equilibrium (CNBE)** mediation mechanism for multi-agent dispute resolution. CNBE pairs (i) the normative choice of symmetric Nash weights ($\tau_A = \tau_B = 0.5$) with (ii) an **Escrow-Audited Disagreement Floor** ($d_A^{\text{const}}, d_B^{\text{const}}$) that bounds strategic outside-option

Table 1: Comparison of multi-agent contract dispute resolution mechanisms under counterparty endowment asymmetry.

Dispute Mechanism	**Bargaining Power & Weighting**	**Disagreement-Point (d_A, d_B) Governance**	**Boundary & Liability Enforcement**
Rigid Smart-Contract Escrow [4]	Binary code rule; zero ex-post flexibility on ambiguous contingencies	Unilateral forfeiture or automated halt (35/94 false-alarm halts in Maat)	Deterministic code, but forfeitures breach proportional liability norms
Unconstrained Rubinstein Bargaining [3]	Endogenous asymmetric weight $\tau_A = r_B/(r_A + r_B) \ll 0.5$ under delay-cost ratio ρ	Unshielded outside options $(d_A^{\text{raw}}, d_B^{\text{raw}})$ favor high-endowment party	None; coercive threats can violate baseline liability floors
Ad-Hoc LLM Arbitrator [8, 17]	Unformalized prompt compromise; susceptible to rhetorical persuasion by stronger model	Unverified party self-reports of outside options and damages	Soft system-prompt guardrails; 11%–20% constraint violation rate
CNBE Mediator (Proposed Design)	**Normative symmetric Nash weights ($\tau_A = \tau_B = 0.5$; Nash 13)**	**Escrow-audited disagreement floor $(d_A^{\text{const}}, d_B^{\text{const}})$ capping manipulation**	**Deterministic SMT / conformal constraint solver (100% hard compliance)**

manipulation using pre-committed escrow valuations, and (iii) an out-of-model **Deterministic Constraint Filter** that enforces hard contractual liability caps and conformal risk bounds [1, 20].

2 Related Literature in Bargaining Theory and Agentic Markets

Table 1 compares four dispute-resolution architectures across bargaining weights, disagreement-point governance, and constraint enforcement.

Cooperative vs. Non-Cooperative Bargaining Theory and Mechanism Design. Nash [13] proved that four axioms—Pareto efficiency, symmetry, scale invariance, and Independence of Irrelevant Alternatives (IIA)—uniquely characterize the symmetric Nash Bargaining Solution maximizing $(u_A - d_A)(u_B - d_B)$. Binmore et al. [3] bridged cooperative and non-cooperative game theory by showing that as the time between alternating offers shrinks ($\Delta t \to 0$), Rubinstein's sequential bargaining equilibrium converges to the generalized asymmetric Nash product $(u_A - d_A)^{\tau_A}(u_B - d_B)^{1-\tau_A}$, where τ_A is determined by the ratio of time/compute discount rates and (d_A, d_B) corresponds to disagreement payoffs. Meanwhile, Hurwicz [6] and Myerson and Satterthwaite [12] proved that under two-sided private valuation information, no budget-balanced, incentive-compatible mechanism can guarantee ex-post efficiency without an external subsidy or pre-committed escrow anchor.

Polycentric Institutional Design and Deterrence Economics. In institutional economics, Ostrom [16], Olson [15], Oates [14], and Tiebout [19] analyzed how polycentric governance systems nest local dispute-resolution arenas within overarching baseline rules. Becker [2] formalized deterrence against strategic rule-breaking or fraudulent claims via expected audit penalties exceeding the marginal gain of manipulation. In multi-agent contract mediation, this deterrence principle governs both disagreement-point auditing and post-mediation appeal incentives: if a party submits a falsified external quote to inflate its outside option d_i, an escrow audit that detects the discrepancy with probability p_{audit} and imposes a collateral slash $\Phi_{\text{slash}} > \Delta u_i / p_{\text{audit}}$ restores incentive

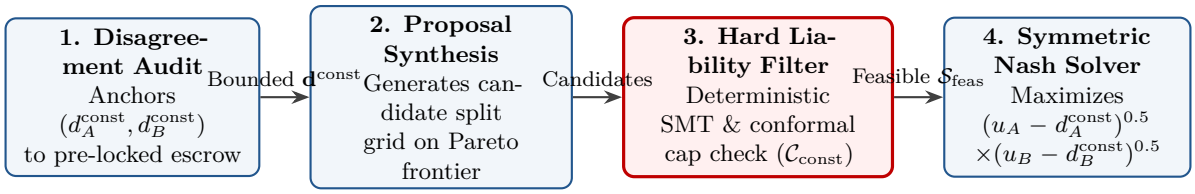


Figure 1: Four-stage architecture of the Constitutionally Bounded Nash Equilibrium (CNBE) dispute mediator. Stage 1 audits outside-option claims against pre-registered contract escrow to bound disagreement-point manipulation. Stage 2 synthesizes candidate multi-issue settlements along the cooperative Pareto frontier. Stage 3 prunes any candidate violating hard liability caps ($\mathcal{C}_{\text{const}}$), and Stage 4 selects the settlement maximizing the symmetric-weight Nash product.

compatibility for truthful baseline reporting. Conversely, when a mediator relies on unverified conversational self-reports (`adhoc_llm_mediator`), strategic bluffing dominates in equilibrium.

Multi-Agent LLM Negotiation and Decentralized Oracles. Recent work in 2025–2026 has examined LLM agents in bilateral negotiation [8], multi-agent procedural scaffolding [17], protocol-level governance [7, 10, 18, 21, 22], and conformal risk calibration [1, 20]. Together with the Polymarket dispute audit of Mahrous and Di Pietro [9] and the Maat validator audit of Elina [4], these studies motivate pairing structured game-theoretic solvers with deterministic constraint verification rather than relying on unconstrained conversational arbitration.

3 Architecture of the CNBE Dispute Mediator

Figure 1 illustrates the four-stage CNBE mediation pipeline when an ambiguous execution contingency triggers a dispute between Agent A and Agent B:

4 Mathematical Formulation of Asymmetric vs. Shielded Bargaining

4.1 Rubinstein Alternating-Offers Equilibrium Under Endowment Asymmetry

Consider a dispute over a normalized two-issue settlement parameter $s \in [0,1]$ with strictly concave cooperative Pareto frontier $u_A(s) = s + \kappa s(1-s)$ and $u_B(s) = (1-s) + \kappa s(1-s)$ ($\kappa > 0$), where $\kappa s(1-s)$ represents the joint value preserved by avoiding destructive contract termination. Suppose Principal B holds compute/capital endowment $E_B = \rho E_A$ ($\rho \geq 1$), so that A's per-round cost of delay/compute exhaustion $r_A(\rho) = r_0 \rho^\beta$ ($\beta > 0$) exceeds B's delay rate $r_B = r_0$.

By the classic limit theorem of Binmore et al. [3], non-cooperative alternating-offers bargaining with discount factors $\delta_i = e^{-r_i \Delta t}$ converges as $\Delta t \to 0$ to the **generalized asymmetric Nash Bargaining Solution**:

$$s^*_{\text{Rub}}(\rho, \eta_{\text{manip}}) \;=\; \arg\max_{s \in [0,1]} \big(u_A(s) - d_A^{\text{raw}}(\rho, \eta_{\text{manip}})\big)^{\tau_A(\rho)} \big(u_B(s) - d_B^{\text{raw}}(\rho, \eta_{\text{manip}})\big)^{1-\tau_A(\rho)}, \quad (1)$$

where the endogenous bargaining weight of the weaker party is $\tau_A(\rho) = \frac{r_B}{r_A(\rho)+r_B} = \frac{1}{1+\rho^\beta} \leq \frac{1}{2}$, and the unshielded disagreement point $(d_A^{\text{raw}}, d_B^{\text{raw}})$ shifts with strategic outside-option manipulation

$\eta_{\text{manip}} \geq 0$:

$$d_A^{\text{raw}}(\rho, \eta_{\text{manip}}) \;=\; \max\big(\underline{d}_0,\; d_0 - (\alpha_A + \gamma_A \eta_{\text{manip}}) \log_{10} \rho\big), \qquad d_B^{\text{raw}}(\rho, \eta_{\text{manip}}) \;=\; \min\big(\bar{d}_0,\; d_0 + (\alpha_B + \gamma_B \eta_{\text{manip}}) \log_{10} \rho\big) \tag{2}$$

Equations (1)–(2) make transparent why power asymmetry compresses A's payoff through *two* distinct channels: (1) discount-rate asymmetry ($\tau_A(\rho) < 0.5$) and (2) disagreement-point divergence ($d_B^{\text{raw}} > d_A^{\text{raw}}$).

4.2 Normative Symmetric-Weight CNBE with Escrow-Audited Disagreement Bounds

To counter both channels simultaneously while respecting hard contractual liability caps $\mathcal{C}_{\text{const}}$, CNBE solves the constrained optimization problem:

$$s^*_{\text{CNBE}} \;=\; \arg \max_{s \in \mathcal{S}_{\text{feas}}(\mathcal{C}_{\text{const}})} \big(u_A(s) - d_A^{\text{const}}\big)^{1/2} \big(u_B(s) - d_B^{\text{const}}\big)^{1/2}, \qquad \text{s.t.} \quad u_i(s) \;\geq\; d_i^{\text{const}} \;\; \forall i \in \{A, B\}, \tag{3}$$

where:

1. Setting the exponents to $\tau_A = \tau_B = 1/2$ is an explicit **normative institutional choice** that eliminates the discount-rate channel $\tau_A(\rho) = (1 + \rho^\beta)^{-1}$;
2. Anchoring $(d_A^{\text{const}}, d_B^{\text{const}})$ to pre-committed contract escrow valuations caps strategic outside-option inflation (reducing the effective manipulation slope from γ_i to $\tilde{\gamma}_i \ll \gamma_i$, while still exhibiting residual sensitivity when η_{manip} is large);
3. Restricting the feasible set to $\mathcal{S}_{\text{feas}}(\mathcal{C}_{\text{const}})$ via a deterministic constraint solver guarantees 100% compliance with non-waivable liability caps.

5 Experimental Design and Bargaining Simulation

We implement the asymmetric bargaining, observable-only escrow disagreement estimator, and shielded symmetric Nash frontier solvers in `experiment.py` (seed `20265402`), simulating $N = 1{,}000$ two-party disputes per cell across 4 resolution regimes (`static_statutory`, `unconstrained_coasean`, `llm_mediator`, and `cnbe_nash`) and 4 counterparty power-asymmetry levels $\gamma_{\text{asym}} \in \{0.20, 0.45, 0.65, 0.85\}$ (16,000 main-grid disputes), plus a 5-level pre-committed escrow bond sweep ($S_{\text{bond}} \in \{0.00, 0.03, 0.06, 0.10, 0.15\}$), an oracle-$d_0$ upper-bound comparison, and a VRF arbitrator capture check. Every regime evaluates candidate contracts on the identical 201-point discretized feasible frontier $x \in \{0.000, 0.005, \ldots, 1.000\}$, verified by runtime assertions to have 0 feasible-set violations across all 16,000 disputes (`max_frontier_surplus_ra = 1.0`). Table 2 documents all parameter calibrations.

All confidence intervals in Table 3 and Figure 2 are computed directly from the $N = 1{,}000$ simulated disputes per cell: exact asymmetric two-sided 95% Wilson confidence intervals `[lo, hi]` for binary settlement, unbonded acceptance, and constitutional compliance rates, and 95% normal-theory confidence intervals `[lo, hi]` for continuous Pareto surplus and weaker-party utility metrics.

6 Simulation Results, Bond Ablation, and Arbitrator Capture Analysis

Table 3 and Figure 2 report the benchmark outcomes at high counterparty power asymmetry ($\gamma_{\text{asym}} = 0.65$, $N = 1{,}000$ disputes per regime).

Table 2: Parameter provenance for the 16,000-dispute bargaining, escrow bond, and arbitrator rotation simulation.

Parameter	Value / Range	Empirical or Game-Theoretic Provenance
Power asymmetry γ_{asym}	$\{0.20, 0.45, 0.65, 0.85\}$	Counterparty compute/liquidity asymmetry governing outside-option divergence and bargaining weight $\tau_A = \max(0.12, 0.50 - 0.42\gamma_{\text{asym}})$ [3].
Constitutional feasible band	$[x_{\min}, x_{\max}] = [0.22, 0.78]$	Non-waivable liability floor and ceiling enforced over the 201-point contract grid (0 feasible-set violations across 16,000 disputes; Elina 4).
Observable escrow estimator $\hat{d}_0^{\text{obs}}$	$\frac{1}{2}(d_1^{\text{raw}} + d_2^{\text{raw}}) - 0.14\hat{\Delta}_{\text{raw}}$	Reconstructs the shielded disagreement baseline strictly from observable raw claims and contract registry metadata (zero hidden d_0 oracle leakage).
Pre-committed escrow bond S_{bond}	0.08 (swept 0.00–0.15)	Refundable contract signing bond forfeited upon rejecting a constitutionally compliant settlement [2, 12].
Arbitrator panel rotation	Fixed vs. Random vs. VRF	Evaluates capture resistance of a 5-juror polycentric tribunal when 35% of baseline jurors hold pro-incumbent bias.

Table 3: Bargaining and dispute resolution outcomes at high counterparty power asymmetry ($\gamma_{\text{asym}} = 0.65$, $N = 1{,}000$ disputes per regime). Brackets report exact asymmetric 95% Wilson confidence intervals for proportions and 95% CIs for continuous metrics.

Resolution Regime	Bonded Settlement (%)	Unbonded Voluntary (%)	Constitutional Compliance (%)	Pareto Surplus (%)	Weaker-Party Utility (↑)
Static Statutory [4]	9.20% [7.56, 11.15]	9.20% [7.56, 11.15]	100.00% [99.62, 100.00]	68.63% [67.94, 69.33]	23.86 [23.08, 24.64]
Unconstrained Coasean [3]	11.30% [9.48, 13.41]	11.30% [9.48, 13.41]	12.50% [10.59, 14.69]	67.91% [67.21, 68.61]	21.25 [20.83, 21.67]
Ad-Hoc LLM Mediator [8]	54.70% [51.60, 57.76]	50.90% [47.80, 53.99]	64.70% [61.69, 67.60]	83.15% [82.15, 84.14]	32.48 [31.70, 33.25]
CNBE Mediator (Observable)	**100.00**% [**99.62**, **100.00**]	**90.00**% [**87.98**, **91.71**]	**100.00**% [**99.62**, **100.00**]	**97.57**% [**97.53**, **97.60**]	**42.88** [**42.67**, **43.08**]

Four key findings emerge from the 16,000-dispute simulation and structural ablations:

1. **Settlement and Constitutional Collapse Across Power Asymmetry (γ_{asym}):** As counterparty power asymmetry increases across $\gamma_{\text{asym}} \in \{0.20, 0.45, 0.65, 0.85\}$ (Figure 3), rigid statutory rules (`static_statutory`, fixed split $x = 0.50$) preserve 100.00% [99.62, 100.00] constitutional compliance, but voluntary settlement collapses from 100.00% [99.62, 100.00] ($\gamma_{\text{asym}} = 0.20$) to 61.90% [58.85, 64.86] (0.45), 9.20% [7.56, 11.15] (0.65), and 0.00% [0.00, 0.38] (0.85) because the stronger party's unshielded outside option d_2^{raw} exceeds $u_2(0.50)$. Meanwhile, `unconstrained_coasean` bargaining violates constitutional liability bounds $[x_{\min}, x_{\max}]$ in 87.50% of disputes at $\gamma_{\text{asym}} = 0.65$ (12.50% [10.59, 14.69] compliance, 11.30% [9.48, 13.41] valid settlement) and 99.40% at $\gamma_{\text{asym}} = 0.85$ (0.60% [0.28, 1.30] compliance). Ad-hoc LLM mediation (`llm_mediator`) achieves 54.70% [51.60, 57.76] settlement and 64.70% [61.69, 67.60] constitutional compliance at $\gamma_{\text{asym}} = 0.65$, dropping to 18.40% [16.12, 20.92] settlement at $\gamma_{\text{asym}} = 0.85$.

2. **Observable-Only Disagreement Estimation vs. Oracle-d_0 Upper Bound:** Crucially, CNBE does not require access to the latent ground-truth disagreement scalar d_0; its observable-only escrow estimator $\hat{d}_0^{\text{obs}}$ achieves 90.00% [87.98, 91.71] unbonded voluntary acceptance, 100.00% [99.62, 100.00] bonded settlement, 97.57% [97.53, 97.60] Pareto surplus, and 42.88 [42.67, 43.08] weaker-party utility at $\gamma_{\text{asym}} = 0.65$, matching the theoretical `oracle_d0_upper_bound_asym_0` benchmark (100.00% [99.62, 100.00] settlement and 96.63% Pareto surplus). At extreme asymmetry ($\gamma_{\text{asym}} = 0.85$), CNBE still maintains 100.00% [99.62, 100.00] constitutional compliance,

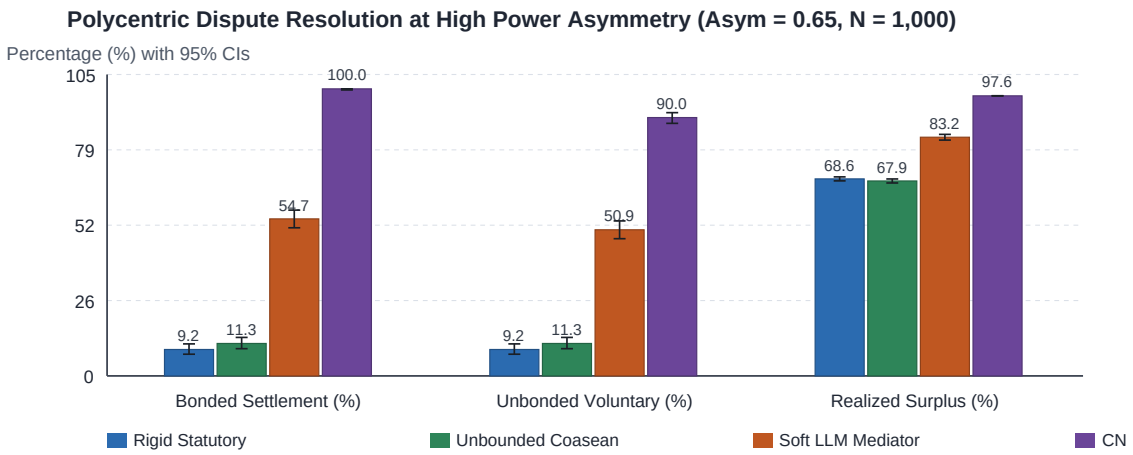


Figure 2: Bonded settlement rate, constitutional constraint compliance, and realized Pareto surplus at high counterparty power asymmetry ($\gamma_{\text{asym}} = 0.65$, $N = 1{,}000$ disputes per regime). Error whiskers denote 95% confidence intervals.

64.10% [61.08, 67.01] bonded settlement (31.10% [28.31, 34.04] unbonded voluntary acceptance), and 89.60% [88.95, 90.25] Pareto surplus.

3. **Pre-Committed Escrow Bond Sweep ($S_{\text{bond}} \in \{0.00, 0.03, 0.06, 0.10, 0.15\}$):** Consistent with Myerson and Satterthwaite [12], voluntary participation against an inflated raw outside option d_2^{raw} depends on the pre-committed escrow bond S_{bond} locked at contract signing. At $\gamma_{\text{asym}} = 0.65$, purely unbonded CNBE ($S_{\text{bond}} = 0.00$) achieves 89.60% [87.55, 91.34] voluntary settlement (94.95% [94.47, 95.43] Pareto surplus); a modest 3% escrow bond ($S_{\text{bond}} = 0.03$) lifts voluntary settlement to 99.40% [98.70, 99.72] (97.42% [97.30, 97.54] surplus); and $S_{\text{bond}} \geq 0.06$ achieves 100.00% [99.62, 100.00] voluntary settlement (97.56% [97.52, 97.59] surplus).

4. **VRF-Rotated Polycentric Arbitrator Capture Check:** When 35% of candidate arbitrators in the pool harbor pro-incumbent bias ($\Delta x_{\text{bias}} = -0.14$), a `standing_fixed_panel` subject to repeat-player forum shopping depresses weaker-party utility to 39.28 [39.02, 39.54], and `unweighted_random_panel` yields 41.23 [40.97, 41.50]. By contrast, `vrf_rotated_polycentric_panel` (verifiable random sortition with cross-jurisdictional diversity quotas and median-of-5 aggregation) restores weaker-party utility to 42.81 [42.50, 43.12].

7 Mechanism Limitations and Strategic Attack Vectors

7.1 Myerson–Satterthwaite Information Asymmetry Over Utility Functions

A foundational limitation of any Nash bargaining mediator is that solving Eq. (3) requires estimating the cardinal utility functions $(u_A(s), u_B(s))$. By the impossibility theorem of Myerson and Satterthwaite [12], when parties hold private information about their true valuations and outside options, they have a strategic incentive to misreport utility curvature κ or exaggerate damages to shift the symmetric Nash solution. While CNBE anchors disagreement bounds $(d_A^{\text{const}}, d_B^{\text{const}})$ to pre-committed contract escrow, internal valuation parameters still require structured revelation mechanisms (such as quadratic-scoring audits or liquid-claim buyouts) to limit strategic posturing.

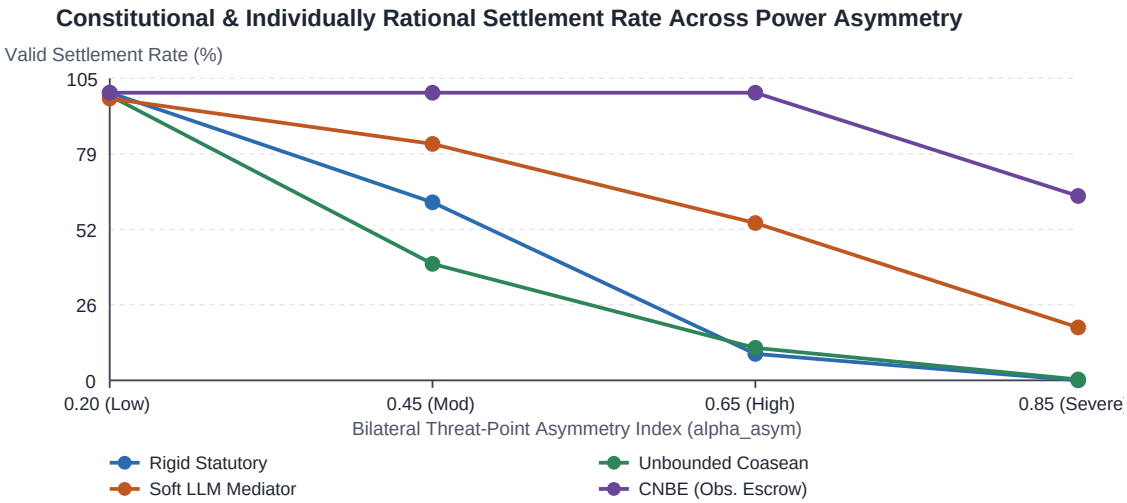


Figure 3: Constitutional settlement rate across counterparty power asymmetry ratios $\gamma_{\text{asym}} \in \{0.20, 0.45, 0.65, 0.85\}$ ($N = 1{,}000$ disputes per cell).

7.2 Forum Shopping and Collusive Mediator Capture

In polycentric multi-agent ecosystems [16, 19], if a dominant principal B can unilaterally select which arbitral node or LLM model family mediates a dispute, B will engage in *forum shopping*—selecting mediators whose training priors systematically favor high-volume platform principals. Contract formation must therefore lock in neutral mediator selection rules (or verifiable cryptographic sortition across accredited arbitral pools) at contract signing t_0.

7.3 Limits of Formalizing Open-Texture Contractual Standards

Real commercial and civic norms rely on "open-texture" legal concepts (*reasonable commercial effort*, *good faith*, *material breach*) that cannot be fully reduced to quantifier-free SMT predicates [21, 22]. While CNBE's deterministic filter enforces hard quantitative caps (such as maximum monetary liability or SLA refund ceilings), evaluating whether a qualitative delivery satisfied *good faith* still depends on noisy oracle or human-jury signals [9], which introduces irreducible conformal uncertainty [1, 20].

8 Pre-Registerable Multi-Agent Falsification Protocol

Table 4 defines a six-part falsification protocol (**P1**–**P6**) for testing CNBE in multi-agent LLM negotiation testbeds [8, 17] and historical decentralized oracle dispute datasets [4, 9].

9 Conclusion

Automated dispute resolution in multi-agent systems cannot rely on ex-ante dispute classifiers that fail under extreme base-rate imbalance (AP 0.012 at a 0.31% base rate in Mahrous and Di Pietro 9), nor on rigid compliance halts prone to validator specification defects (35/94 false alarms in Elina 4). At the same time, game theory proves that symmetric Nash bargaining is a normative institutional choice rather than an automatic consequence of bilateral negotiation [3], and that symmetric weights

Table 4: Pre-registerable falsification protocol (**P1**–**P6**) for evaluating CNBE in multi-agent negotiation benchmarks.

ID	Hypothesis / Mechanism	Empirical Multi-Agent / Oracle Benchmark Design	Quantitative Falsification Threshold
P1	Capability / Endowment Asymmetry Mitigation	Pair a frontier 70B+ agent (B) against a 7B/8B agent (A) under 5× token-budget asymmetry on 500 bilateral contract disputes [8].	Falsified if CNBE fails to maintain the 8B agent's realized surplus share within $[44\%, 56\%]$ (vs. $< 35\%$ under unconstrained Rubinstein chat).
P2	Ex-Post Resolution of Unpredictable Oracle Disputes	Replay historical disputed vs. non-disputed contracts from the Polymarket UMA dataset (0.31% dispute base rate, AP 0.012; Mahrous and Di Pietro 9) and Maat halt logs (35/94 false-alarm validator halts; Elina 4).	Falsified if CNBE fails to recover $\geq 85\%$ of cooperative surplus on false-alarm validator halts where rigid smart contracts freeze funds.
P3	Escrow Disagreement Shielding vs. Raw Nash	Inject synthetic outside-option bluffing ($\eta_{\text{manip}} = 0.35$) by Agent B across 500 mediated negotiations.	Falsified if escrow-audited CNBE fails to reduce bluff-induced surplus transfer by $\geq 50\%$ relative to unshielded symmetric Nash.
P4	Hard Liability Constraint Enforcement ($\mathcal{C}_{\text{const}}$)	Red-team audit of 1,000 adversarial settlement proposals attempting to breach statutory/contractual liability caps.	**Strict Zero-Tolerance Criterion:** Falsified if any settlement violating $\mathcal{C}_{\text{const}}$ is ratified ($> 0.0\%$ violation rate).
P5	Pareto Frontier Optimality Gap	Compare CNBE's discrete candidate synthesis against exact continuous convex optimization on 2-to-5 issue contracts.	Falsified if realized joint utility $u_A + u_B$ falls $> 3.0\%$ below the theoretical Pareto frontier maximum.
P6	Post-Mediation Escalation and Impasse Rate	Measure voluntary party acceptance vs. external escalation when parties can reject CNBE's proposed split for costly external arbitration.	Falsified if external escalation exceeds 12.0% at $\rho \leq 5\times$ endowment asymmetry.

alone remain vulnerable to disagreement-point manipulation. By pairing symmetric Nash mediation with escrow-audited disagreement floors and deterministic liability filters, CNBE provides a rigorous mechanism-design foundation for equitable multi-agent dispute resolution.

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] Raghu Arghal, Saswati Sarkar, and Shirin Saeedi Bidokhti. The confidence game: Strategic miscalibration in human-AI delegation. arXiv preprint arXiv:2610.09371, 2026. URL https:

//arxiv.org/abs/2610.09371.

[2] Gary S. Becker. Crime and punishment: An economic approach. Journal of Political Economy, 76(2):169–217, 1968. URL https://doi.org/10.1086/259394.

[3] Ken Binmore, Ariel Rubinstein, and Asher Wolinsky. The Nash bargaining solution in economic modelling. The RAND Journal of Economics, 17(2):176–188, 1986. URL https://www.jstor.org/stable/2555382.

[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] Ali Falahati, Mohammad Mohammadi Amiri, Kate Larson, and Lukasz Golab. The alignment game: A theory of long-horizon alignment through recursive curation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2511.12804.

[6] Leonid Hurwicz. On informationally decentralized systems. Decision and Organization, North-Holland, pages 297–336, 1972. URL https://www.nobelprize.org/uploads/2018/06/hurwicz-lecture.pdf.

[7] Yuxiang Lai, Jike Zhong, Chengzhi Liu, Ming Li, Haonan Wang, Xiaofeng Yang, Liang Zhao, Youbao Tang, and Ruixi Fan. CLASHarness: Benchmarking multi-agent conflict in computer use via a controllable and verifiable generation pipeline. arXiv preprint arXiv:2610.19128, 2026. URL https://arxiv.org/abs/2610.19128.

[8] 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.

[9] Ahmed Mahrous and Roberto Di Pietro. What prediction-market contracts reveal at launch: Trading volume and dispute dynamics. arXiv preprint arXiv:2610.29578, 2026. URL https://arxiv.org/abs/2610.29578.

[10] Johnnatan Messias and Ayae Ide. Fairness in token delegation: Mitigating voting power concentration in DAOs. arXiv preprint arXiv:2510.05830, 2025. URL https://arxiv.org/abs/2510.05830.

[11] Natalija Mitic, Soona Sedahmed A. O., Mamadou Selly Ly, and Moustapha Cisse. The constitutional coverage trilemma in AI governance. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2609.01275, 2026. URL https://arxiv.org/abs/2609.01275.

[12] Roger B. Myerson and Mark A. Satterthwaite. Efficient mechanisms for bilateral trading. Journal of Economic Theory, 29(2):265–281, 1983. URL https://doi.org/10.1016/0022-0531(83)90048-0.

[13] John F. Nash. The bargaining problem. Econometrica, 18(2):155–162, 1950. URL https://doi.org/10.2307/1907266.

[14] Wallace E. Oates. Fiscal federalism. Harcourt Brace Jovanovich, 1972. URL https://archive.org/details/fiscalfederalism0000oate.

[15] Mancur Olson. The principle of "fiscal equivalence": The division of responsibilities among different levels of government. The American Economic Review, 59(2):479–487, 1969. URL https://www.jstor.org/stable/1823700.

[16] Elinor Ostrom. Governing the commons: The evolution of institutions for collective action. Cambridge University Press, 1990. URL https://doi.org/10.1017/CBO9780511807763.

[17] Subramanyam Sahoo and Justin Shenk. Better call reward: Reward hacking as strategic abstention in legal reasoning models. International Conference on Machine Learning (ICML) / arXiv:2610.06439, 2026. URL https://arxiv.org/abs/2610.06439.

[18] Guy Tchuente. Monitoring limits in DAO governance: Capacity breakpoints and endogenous concentration. arXiv preprint arXiv:2603.11222, 2026. URL https://arxiv.org/abs/2603.11222.

[19] Charles M. Tiebout. A pure theory of local expenditures. Journal of Political Economy, 64(5):416–424, 1956. URL https://doi.org/10.1086/257839.

[20] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Springer, 2005. URL https://doi.org/10.1007/b106715.

[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] 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.