
When Subsidies Backfire: The Budget–Audit Frontier for Collusion-Resistant Procurement

Xianrun Chen¹ Ziqi Zhao¹

¹The University of Hong Kong

AI Collusion Research Sprint, October 2026

Abstract

Loser subsidies can discourage an agent from capturing a task, yet reward a coalition for manipulating the winning bid. We derive the exact tradeoff between payment budgets and worst-case auditing in procurement with a fixed winner award and trusted cost verification. Balancing capture and subsidy amplification yields an optimal capped-linear subsidy and a budget threshold beyond which further spending cannot reduce auditing. In 54 prescribed-witness games, model agents attain every exact finite-action payoff maximum, with larger budgets reducing capture gains and increasing amplification gains under reduced auditing. A separate repeated-game dataset shows that negative aggregate gains can conceal profitable pair-level deviations. The mechanism result and the payoff diagnostic connect two tasks in evaluating collusion: preventing joint deviations and measuring their consequences for coalition members, the sponsor, and execution cost.

1 Why paying losers can backfire

A platform asks several AI agents to compete for a task. If two agents share an owner or can share profits, the platform must consider their joint incentive, not just each agent’s honesty. Restricting their messages does not change the underlying payments. Can payment design instead make profitable joint deviations impossible, and how much verification does that require?

We study a payment design in which the lowest bidder executes, the winner receives a fixed award, and each loser receives a subsidy. A subsidy gives an agent something to keep by losing, reducing its incentive to capture the task. But when the subsidy depends on the winning bid, a coalition can manipulate that bid to increase its partners’ payments. The design problem is to balance these two incentives within a fixed budget.

This project approaches the hackathon’s question about collusion restrictions and their costs through mechanism design. We ask how payments and audits can remove profitable joint deviations in procurement, then examine how those deviations should be measured in model-agent interactions. The required cost verification and settlement are explicit parts of the model.

Two ways to profit from a lower bid

There are three agents A, B, and O. Agents A and B share profits ($n = 3$, coalition limit $k = 2$). The winner receives $H = 100$ credits and incurs its actual execution cost. Ties favor A, then B, then O. An audited winner that understated its cost loses the award; there is no extra fine. At nominal budget $P = 160$, consider

$$\ell(\beta) = \min\{30, (100 - \beta)/2\}, \quad q(\beta) = 0.75 [100 - \beta - \ell(\beta)]/100,$$

where β is the lowest bid, ℓ is each loser’s subsidy, and q is audit probability. The multiplier 0.75 deliberately audits below the theorem’s safe rule. All payoffs below average over the audit lottery.

Capture: displace a cheaper outsider. Let costs be $(1, 1, 0)$. With truthful bids, O wins and A/B collect $30 + 30 = 60$. If A bids zero while B bids one, A wins the priority tie with O. Now $q = 0.525$. A’s expected utility is $100(1 - 0.525) - 1 = 46.5$, and B receives 30. Their total becomes 76.5, a gain of 16.5, at one extra execution-cost unit.

Amplification: pay a partner more. In a different instance, all costs are 100. Truthful coalition utility is zero. If A bids 40 and B bids 100, A wins, B receives 30, and $q = 0.225$. A loses 22.5 in expectation, but B gains 30. Pooling yields a net 7.5. Joint incentives matter even though the winning member individually loses before profit sharing.

Larger subsidies reduce the attraction of capture, but can enlarge the partner-payment channel. In these two prescribed instances, the low-audit gains change in opposite directions:

Nominal budget P	120	160	200
Capture gain	21.5	16.5	11.5
Amplification gain	2.5	7.5	12.5

Table 1: Opposing gains at reduced auditing ($\alpha = 0.75$). Each entry is the common observed gain in three witness-study repetitions. CAP and AMP use distinct cost instances.

The two rows use different cost profiles. As the budget rises, the best amplification bid also changes. Section 5 compares these observations with exact finite-action payoff maxima.

An optimal subsidy and two behavioral tests

We optimize over all common minimum-bid subsidy functions under a sponsor payment cap. Capture favors a larger subsidy, while amplification favors a smaller one. Balancing the two constraints gives a capped-linear subsidy and an exact budget threshold beyond which additional spending cannot reduce worst-case auditing. The same mechanism attains the optimum for every minimum bid.

The empirical analysis follows the same incentive question in two settings. First, the prescribed-witness study tests whether agents in an assigned coalition realize the deviations behind the audit bounds. Second, a repeated-game dataset tests the evaluation criterion when coalitions are not assigned. Comparing each pair against its own truthful reports reveals incentives that an aggregate payoff comparison can miss. Mechanism design supplies the guarantee, and coalition-specific evaluation supplies the corresponding diagnostic.

Verification setting. A candidate application is a fixed, metered compute job with authenticated providers, trusted usage logs, a predetermined tariff, and enforceable settlement. The metered quantity must equal the immutable cost in the model. Providers must be unable to inflate usage or hide relevant opportunity costs. The theorem therefore requires evidence of execution cost beyond output correctness.

Related work

Coalition resistance with verification. Penna and Ventre [1, 2] show how verification can support optimal allocation in the presence of coalitions with transferable utility. Their threshold construction already combines payments to selected and unselected agents, with the selected-agent payment scaling with the coalition limit and the type bound. Our design fixes the winner award at the cost cap and optimizes worst-case audit probability under a total payment cap. The resulting frontier quantifies the balance between capture and subsidy amplification within this procurement architecture. Caballero, Lunday, and Ahner [3] study a related inspection problem for sharing the cost of a coalition initiative. They optimize inspection and penalty policies under alternative incentive requirements, including group strategyproofness. Here the allocation is a single mandatory task, and the award and escrow are fixed inputs to the subsidy–audit optimization.

Verification as a design resource. Caragiannis et al. [4] characterize implementability and payments under partial and probabilistic verification. Ferraioli and Ventre [5] study verification and fine requirements for obviously strategyproof mechanisms. Ben-Porath, Dekel, and Lipman [6] optimize allocation with costly verification and no monetary transfers, while Ball and Kattwinkel [7] model tests whose pass probabilities depend on the agent’s type. These papers distinguish the cost of checking a report from the information a check reveals. In our model, a selected audit perfectly detects cost underreporting. The design variable is how often to audit, and loser subsidies change the deviation gains that those audits must deter.

Algorithmic collusion in markets. Calvano et al. [8] find that Q-learning pricing agents can sustain supracompetitive prices without direct communication in repeated oligopoly simulations. Fish, Gonczarowski, and Shorrer [9] study LLM pricing agents and show that prompt wording can substantially change supracompetitive pricing. Agrawal et al. [10] examine LLM collusion in double auctions, varying communication, model choice, and oversight conditions. These studies motivate examining both the strategic environment and the information available to agents. Our procurement experiments connect observed deviations to explicit payoff bounds, so the incentive consequences of changing the payment budget or audit probability can be calculated alongside behavior.

Measuring coordination and its gains. Agrawal et al. [10] combine language-based judgments with pricing and economic outcomes. Motwani et al. [11] study covert coordination through steganography, highlighting the difficulty of detecting coordination from communication alone. Our evaluation adds a coalition-specific payoff counterfactual. For each candidate pair, we compare its observed reports with truthful reports while holding the outsider’s bid fixed. This exposes profitable pair deviations even when total bidder utility falls relative to all-truthful reporting. Communication labels, coalition payoffs, and execution costs therefore answer different questions about the same interaction.

2 Procurement model and coalition guarantee

There are $n \geq 2$ fixed authenticated agents. Agent i has immutable execution cost $c_i \in [0, H]$, where $H > 0$ is public, and reports $b_i \in [0, H]$. The lowest bidder executes one mandatory task; ties use a fixed identity priority. Write $w = w(b)$ and $\beta = b_{(1)}$. Only the winner incurs its cost. Utilities are risk-neutral and quasi-linear, and coalition members may exchange side payments.

The mechanism has four components.

- the winner receives the fixed award H , sufficient to cover the cost cap;
- every loser receives the same subsidy $\ell(\beta)$, for an initially unrestricted function $\ell : [0, H] \rightarrow \mathbb{R}$;
- a committed audit probability $q(b) \in [0, 1]$ may use the full report profile;
- an audited winner with $c_w > b_w$ loses its award and forfeits escrow $F \geq 0$. Its execution cost remains incurred. Other payments are unchanged.

Let $D = H + F$. Expected utility is

$$u_i(c, b) = \begin{cases} H - c_i - q(b)D \mathbf{1}\{c_i > b_i\}, & i = w(b), \\ \ell(\beta), & i \neq w(b). \end{cases}$$

Reports and execution are binding before the hidden audit draw. The verifier detects $c_w > b_w$ perfectly. The model assumes authenticated identities and enforceable settlement, and excludes identity creation, withdrawal, shirking, auditor corruption, and outside sabotage benefits. Authentication is essential because extra identities could otherwise collect loser subsidies.

Fix $2 \leq k \leq n$. For every coalition C with $|C| \leq k$, its true costs, and arbitrary fixed outsider reports, joint truthful reporting must maximize its sum of expected utilities. This defines transferable-utility coalition resistance in expectation. Equality permits payoff-equivalent misreports. Truthful individual rationality is also required. A gross nominal sponsor cap $P \geq H$ imposes

$$H + (n - 1)\ell(\beta) \leq P \quad \text{for every } \beta. \tag{1}$$

Transfers, locked escrow, and real auditing costs are separate quantities. The optimization is over subsidy functions and audit rules within this payment architecture.

3 Subsidy shape is constrained by undetectable deviations

The common minimum-bid form can also be derived when unsanctioned loser transfers are deterministic, audit-independent, and symmetric at unique-winner profiles; the derivation and its exclusions are retained in Appendix B. Overreports are invisible to the verifier, so the payment rule must handle them without sanctions.

Proposition 1 (Necessary shape). *Any mechanism satisfying these requirements has*

$$0 \leq \ell(t) \leq H - t, \quad \ell \text{ nonincreasing}, \quad \ell(H) = 0.$$

No continuity assumption is needed.

Full proof appears in Appendix A.1.

These necessary conditions handle deviations that auditing cannot punish. We next ask how much auditing is required for the remaining, detectable deviations.

4 The optimal budget–audit tradeoff

The budget permits a subsidy of at most $L = (P - H)/(n - 1)$ per loser. For each minimum bid β , define

$$Q_\beta(\ell, q) = \sup_{b: b_{(1)} = \beta} q(b).$$

This is a *conditional worst case*: the other bids are adversarial, while the minimum bid is fixed.

The two channels from the example become sharp lower bounds. At fixed β , write $x = \ell(\beta)$ and $D = H + F$. Capture requires $DQ_\beta \geq H - \beta - x$; amplification by up to $k - 1$ partners requires $DQ_\beta \geq (k - 1)x$. Hence

$$DQ_\beta \geq \max\{H - \beta - x, (k - 1)x\}, \quad 0 \leq x \leq L.$$

The capture bound decreases with the subsidy, while the amplification bound increases. Their maximum is minimized at $x = \min\{L, (H - \beta)/k\}$. The two witnesses use different outsider reports, which is why the bound takes a worst case over profiles sharing β . To attain it, one subsidy function must also deter every other coalition deviation.

Theorem 2 (Simultaneous conditional minimax). *Among all admissible subsidy functions and audit rules satisfying (1), the smallest possible Q_β is*

$$Q_P^*(\beta) = \frac{H - \beta - \min\{L, (H - \beta)/k\}}{H + F}. \quad (2)$$

One mechanism attains these bounds simultaneously for every β :

$$\ell_P(t) = \min\{L, (H - t)/k\}, \quad q_P(b) = \frac{H - \beta - \ell_P(\beta)}{H + F}. \quad (3)$$

The subsidy function is uniquely determined by simultaneous conditional optimality. Audit schedules need not be unique, and $q_P(b)$ need not minimize auditing at every complete report profile.

The lower bound and matching construction appear in Appendices A.2–A.3.

Corollary 3 (Worst-case budget frontier). *The minimum worst-case audit probability under budget P is*

$$\inf_{\ell, q} \sup_b q(b) = \frac{H - \min\{H/k, (P - H)/(n - 1)\}}{H + F}. \quad (4)$$

The frontier plateaus at $P^* = H[1 + (n - 1)/k]$. At or above this budget, the canonical rule is balanced sharing,

$$\ell_P(t) = \frac{H - t}{k}, \quad q_P(b) = \frac{(k - 1)(H - \beta)}{k(H + F)}.$$

For $k = n$ and $F = 0$, every coalition is covered with maximum nominal payout $(2 - 1/n)H$ and maximum audit probability $1 - 1/n$.

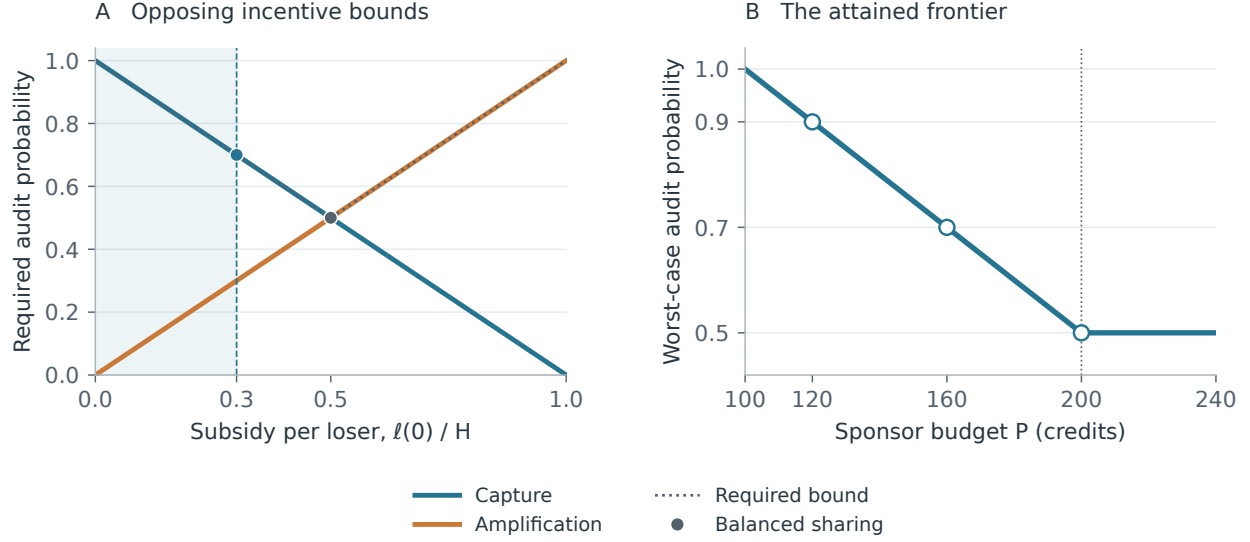


Figure 1: Two incentive constraints determine the audit frontier. Here $n = 3$, $k = 2$, $H = 100$, and $F = 0$. Left, at $\beta = 0$, the normalized subsidy $x = \ell(0)/H$ must satisfy $Q_0 \geq \max\{1 - x, x\}$. Shading marks subsidies affordable at $P = 160$. Right, the attained worst-case audit probability falls with the budget until $P = 200$. Hollow markers identify the three experimental budgets. Curves are analytic bounds, and the two left-panel witnesses use different outsider reports.

Why the cap is optimal. Below the threshold budget, the subsidy is flat at the affordable cap for low winning bids, then decreases with slope $-1/k$. Within the flat region, further bid reductions cannot increase partners’ subsidies. Moving into that region from a higher bid can still amplify their payments. The declining region balances amplification against capture. The linear shape at generous budgets is derived, not imposed. More subsidy beyond the balanced point cannot improve the conditional minimax objective.

Execution cost and computational checks. The mechanism also bounds excess execution cost by unilateral truthful regret in the base report game (Appendix A.4). This guarantee concerns equilibrium incentives, while the experiments below record chosen actions. Exact-rational enumeration checks coalition incentives, payment caps, both deviation witnesses, and the regret inequality. Appendix B.2 reports these implementation checks. These checks complement the continuous proofs.

5 Agents realize both profitable deviations

The witness experiment uses two model agents in an enforced equal-profit pool and a scripted outsider. It fixes $n = 3$, $k = 2$, $H = 100$, and $F = 0$, and crosses the two introductory cost profiles with three budgets $P \in \{120, 160, 200\}$ and three audit multipliers $\alpha \in \{0.75, 1, 1.1\}$. Three repetitions per cell give 54 games, all with messaging. The multiplier scales the canonical audit rule in (3). The truthful comparator is computed under the same pool contract.

All 54 outcomes attain the exact integer-action payoff maximum (Figure 2). At reduced auditing, capture gains fall from 21.5 to 11.5 credits as the budget rises, while amplification gains rise from 2.5 to 12.5 credits. These opposing changes match the two incentive channels in Table 1. At canonical or

higher auditing, all 36 gains are zero, consistent with the incentive guarantee. Four capture outcomes use payoff-neutral overreports, illustrating that zero gain permits nontruthful bids.

This study tests behavior within a formed coalition. Appendix C gives its protocol, accounting, and finite-grid comparison. The separate historical communication study appears in Appendix E.

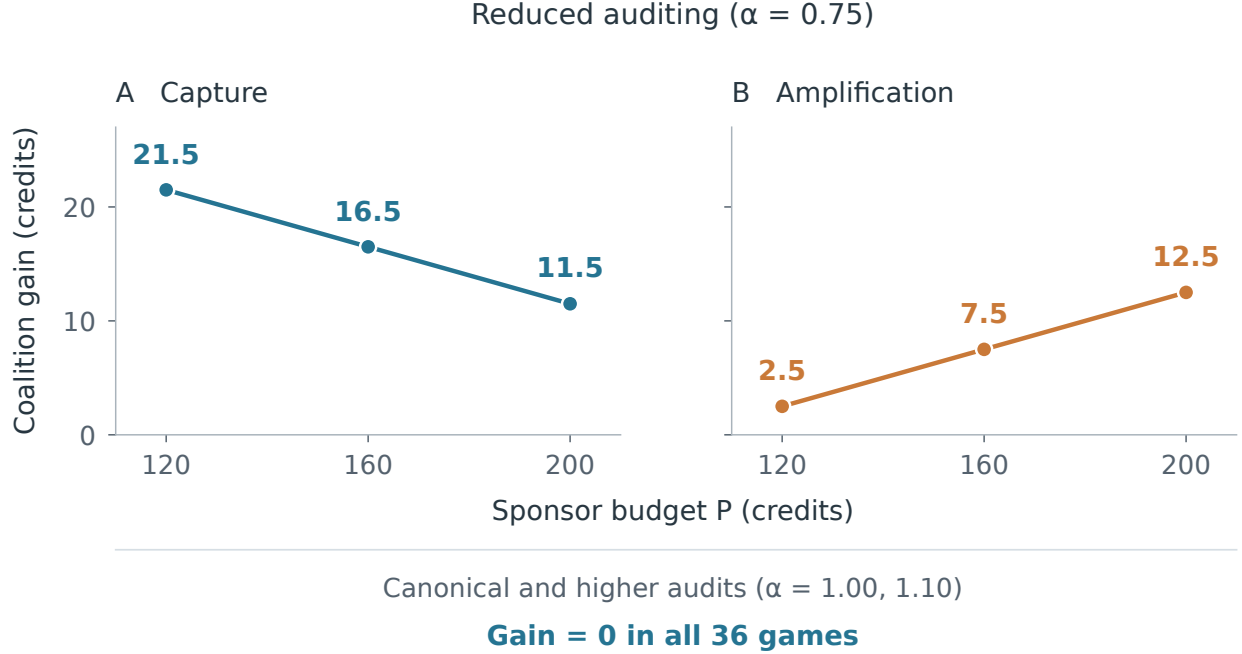


Figure 2: Larger budgets reduce capture gains and increase amplification gains under reduced auditing ($\alpha = 0.75$). Each labeled point represents three games with identical gains. Lines connect the sampled budgets to guide the eye. At canonical and higher auditing ($\alpha = 1.00, 1.10$), all 36 games have zero gain, as summarized below the plots. Every observed gain equals its exact finite-action maximum. Capture and amplification use different cost profiles.

6 Measuring coalition incentives without assigned coalitions

The witness experiment fixes a coalition and isolates the two incentive channels. We next ask how to evaluate outcomes when no coalition is assigned. The theoretical guarantee holds each outsider’s report fixed when testing a coalition’s incentive. The same comparison is useful for observed model behavior.

A separate dataset contains 156 eight-round episodes with three model bidders, fresh private costs each round, and no assigned profit pool. The four payment rules in Table 2 use $n = k = 3$, $H = 100$, and $F = 0$. Opt and NoSub are the two budget endpoints of the theorem. Half and NoAudit reduce the auditing attached to Opt’s subsidy.

Rule	Subsidy per loser	Audit probability
Opt	$(100 - \beta)/3$	$2(100 - \beta)/300$
Half	$(100 - \beta)/3$	$(100 - \beta)/300$
NoAudit	$(100 - \beta)/3$	0
NoSub	0	$(100 - \beta)/100$

Table 2: Payment rules in the unassigned-coalition study.

The DeepSeek-labeled data cover ten conditions with 12 episodes each, and the GLM-labeled data cover three conditions. P0 is the original prompt, P1 adds an incentive reminder, and NC removes both chat and conditional transfers. Chat conditions allow three sequential messages and optional transfers. Cost streams match across conditions. Appendix D gives the complete results and replay details.

Why the coalition’s own baseline matters. Let E be the change in all three bidders’ total expected utility relative to all-truthful reporting. Internal transfers cancel from E . In one Half-audit round, costs are (85, 60, 90) and bids are (95, 50, 58). Total bidder utility falls by 10 credits. Bidders 1 and 2 nevertheless receive 40 credits together. If those two report truthfully while bidder 3 keeps its bid of 58, they receive only 28. Their pair-level gain is therefore 12 credits. The observed profile contains a profitable pair deviation even though aggregate utility falls.

We apply this comparison to every pair at each recorded profile. Under Half, no DeepSeek round has positive E , but positive pair gains occur in 7/96 P0 rounds and 9/96 P1 rounds. The diagnostic is nonpositive under Opt and NoSub, as the theorem requires. It measures base mechanism incentives with the outsider’s report fixed. It does not identify which bidders coordinated or the profits from their transfer commitments.

Coordination labels and economic outcomes. In DeepSeek chat conditions, a model judge labels coordination in 81% to 90% of rounds, while every condition has negative mean E (Figure 3). Under NoAudit, positive- E outcomes occur in 12/96 P0 rounds and 31/96 NC rounds. This comparison changes both communication and transfer access. The GLM chat conditions also have negative mean E , at -12.26 under Opt and -2.95 under NoAudit. Thus language labels, aggregate payoffs, and pair incentives describe different aspects of the same interaction.

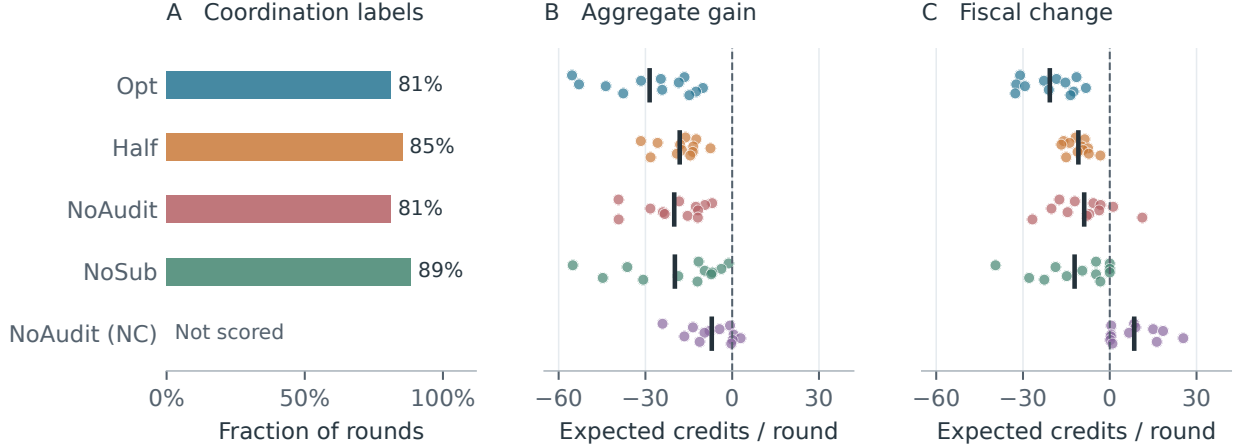


Figure 3: Coordination labels and economic outcomes in the DeepSeek-labeled dataset. Left, bars show existing model-judge labels. Center and right, points show 12 episode means and black ticks show the cell mean. Fiscal change is measured against each mechanism’s own truthful baseline. NC disables both chat and transfers and has no coordination label. Negative aggregate gains can coexist with profitable pair deviations and excess execution cost. P1 and GLM results appear in Appendix D.

7 Implications and conclusion

Audits, spending, and allocation. In DeepSeek P0, Opt has mean audit probability 0.451 and expected sponsor expenditure 130.14 credits. NoSub has mean audit probability 0.632 and expenditure 87.84 credits. Mean excess execution costs are similar, at 7.81 and 7.66. The subsidy is associated with lower audit demand and higher sponsor spending, while execution loss remains. A total-cost comparison would also require the cost of an audit.

For a fixed cost profile, the change in expected sponsor spending equals E plus excess execution cost. This accounting identity separates transfers from the resources consumed by execution. Likewise, evaluating a coalition requires its own truthful counterfactual with outsider reports fixed. The repeated-game results make this distinction concrete even when observed aggregate gains are negative.

Conclusion. A loser subsidy reduces the incentive to capture a task but can amplify payments to a winner’s partners. Balancing these forces yields an exact budget–audit frontier and a capped-linear mechanism that attains it at every minimum bid. The witness games realize both incentive channels. The unassigned-coalition study then shows how coalition-specific payoff comparisons reveal incentives that aggregate gains and communication labels miss. Together, these results provide a procurement design benchmark and an evaluation method for studying joint deviations.

Code and data. The accompanying package includes the figure inputs, replay summaries, and original experiment archives. Appendix F maps the retained records and describes the evidence scope.

References

- [1] P. Penna and C. Ventre. *Collusion-resistant mechanisms with verification yielding optimal solutions*. ESA 2008, pp. 708–719. doi:10.1007/978-3-540-87744-8_59. Extended author version (2009).
- [2] P. Penna and C. Ventre. *Optimal collusion-resistant mechanisms with verification*. Games and Economic Behavior 86:491–509, 2014. doi:10.1016/j.geb.2012.09.002. Author version.
- [3] W. N. Caballero, B. J. Lunday, and D. K. Ahner. *Incentive Compatible Cost Sharing of a Coalition Initiative with Probabilistic Inspection and Penalties for Misrepresentation*. Group Decision and Negotiation 29:1021–1055, 2020. doi:10.1007/s10726-020-09693-z.
- [4] I. Caragiannis, E. Elkind, M. Szegedy, and L. Yu. *Mechanism Design: from Partial to Probabilistic Verification*. EC 2012, pp. 266–283. doi:10.1145/2229012.2229035. Author version.
- [5] D. Ferraioli and C. Ventre. *Probabilistic Verification for Obviously Strategyproof Mechanisms*. IJCAI 2018, pp. 240–246. doi:10.24963/ijcai.2018/33.
- [6] E. Ben-Porath, E. Dekel, and B. L. Lipman. *Optimal Allocation with Costly Verification*. American Economic Review 104(12):3779–3813, 2014. doi:10.1257/aer.104.12.3779.
- [7] I. Ball and D. Kattwinkel. *Probabilistic verification in mechanism design*. Theoretical Economics 20(4):1247–1284, 2025. doi:10.3982/TE6266.
- [8] E. Calvano, G. Calzolari, V. Denicolò, and S. Pastorello. *Artificial Intelligence, Algorithmic Pricing, and Collusion*. American Economic Review 110(10):3267–3297, 2020. doi:10.1257/aer.20190623.
- [9] S. Fish, Y. A. Gonczarowski, and R. I. Shorrer. *Algorithmic Collusion by Large Language Models*. 2024, revised 2026. arXiv:2404.00806v6.
- [10] K. Agrawal, V. Teo, J. J. Vazquez, S. Kunnavaakkam, V. Srikanth, and A. Liu. *Evaluating LLM Agent Collusion in Double Auctions*. 2025. arXiv:2507.01413v1.
- [11] S. R. Motwani, M. Baranchuk, M. Strohmeier, V. Bolina, P. H. S. Torr, L. Hammond, and C. Schroeder de Witt. *Secret Collusion among AI Agents: Multi-Agent Deception via Steganography*. 2024, revised 2025. arXiv:2402.07510v5.
- [12] T. Roughgarden. *Intrinsic Robustness of the Price of Anarchy*. Journal of the ACM, 2015. Author version.

A Proofs of the mechanism results

The argument has three steps. Undetectable deviations constrain the subsidy shape, capture and amplification give the audit lower bound, and the capped-linear rule attains that bound for every coalition. The final result relates truthful regret to execution cost.

A.1 Necessary subsidy shape

Proof. Truthful losers incur no cost, so individual rationality forces $\ell(t) \geq 0$ for every possible winning cost t .

For $t < r \leq H$, give the first-priority agent true cost t and all others cost/report H . It may report r and remain selected, including at $r = H$. A coalition consisting of this winner and a loser changes the loser's payment from $\ell(t)$ to $\ell(r)$ while the winner's cost and award are unchanged. No underreport is present. Coalition resistance forces $\ell(r) \leq \ell(t)$.

For $t < H$, let the first-priority truthful winner tie an outsider at t . By reporting slightly above t , it becomes a loser and cannot itself be sanctioned. Singleton resistance requires $\ell(t) \leq H - t$. Finally, monotonicity gives $\ell(H) \leq \ell(t) \leq H - t$ for every $t < H$; together with nonnegativity this yields $\ell(H) = 0$. $\square$

A.2 Capture and amplification lower bounds

Lower bound. Fix $\beta < H$ and write $x = \ell(\beta)$. Two deviations impose different constraints.

Capture deviation. At all- β reports, give the first-priority winner true cost $\beta + \epsilon$, holding outsider reports at β . Truthfully, it loses and receives x . Reporting β wins and yields $H - \beta - \epsilon$ before sanctions. Therefore

$$DQ_\beta \geq H - \beta - x$$

as $\epsilon \downarrow 0$. This is a supremum argument; it assumes neither a favorable truthful tie nor continuity of ℓ .

Coalition-amplification deviation. At reports $(\beta, H, \dots, H)$, choose a size- k coalition containing the winner and give all its members true cost H . Outsider reports stay H . The truthful coalition obtains zero because $\ell(H) = 0$. After the deviation, the winner's pre-sanction utility is still zero, while its $k - 1$ partners each receive x . Hence

$$DQ_\beta \geq (k - 1)x.$$

These witnesses work for $k = n$ as well. Thus

$$DQ_\beta \geq \max\{H - \beta - x, (k - 1)x\}. \quad (5)$$

The budget gives $0 \leq x \leq L$. The unique minimizer of this V-shaped bound is $x = \min\{L, (H - \beta)/k\}$, yielding (2). At $\beta = H$, the necessary shape gives $x = 0$ and the bound is zero. $\square$

A.3 Coalition resistance of the capped-linear rule

Matching construction. The function ℓ_P is nonnegative, nonincreasing, $1/k$ -Lipschitz, and satisfies $k\ell_P(t) \leq H - t$. Its nominal payments obey the cap and $0 \leq q_P \leq 1$, even for $F = 0$.

Fix a coalition C of size $s \leq k$ and arbitrary outsider reports. If its final winner is outside, the deviation cannot help. Indeed, a truthful outside winner gives the same coalition payments; a truthful inside winner at cost t gives

$$H - t + (s - 1)\ell_P(t) \geq s\ell_P(t),$$

at least what all members receive when losing at a weakly higher outside bid.

Suppose the final winner w is inside C . Write $t = \min_{i \in C} c_i$ and let m be the cheapest outsider report when one exists. If the truthful-in- C winner is inside, the coalition's gain before sanctions is

$$\Delta_0 = t - c_w + (s - 1)[\ell_P(\beta) - \ell_P(t)] \leq (s - 1)\ell_P(\beta) \leq H - \beta - \ell_P(\beta).$$

If the truthful winner is outside, then $m \geq \beta$ and $c_w \geq m$, so

$$\Delta_0 \leq H - m + (s - 1)\ell_P(\beta) - s\ell_P(m).$$

Subtracting $a_P(\beta) = H - \beta - \ell_P(\beta)$ gives

$$\Delta_0 - a_P(\beta) \leq \beta - m + s[\ell_P(\beta) - \ell_P(m)] \leq \beta - m + \frac{s}{k}(m - \beta) \leq 0.$$

Thus the expected sanction covers every positive raw gain. Such a gain necessarily has $c_w > \beta$: otherwise a truthful coalition member prevents an outside truthful winner, including fixed-priority ties, and the inside expression is nonpositive by monotonicity. Therefore exactly the deviations needing deterrence are subject to the sanction. This proves coalition resistance and simultaneous achievability. $\square$

A.4 Execution efficiency statement

Truthful reporting allocates to a minimum-cost agent, but weak incentive constraints alone do not require truthful play. The canonical mechanism admits a direct regret certificate.

Theorem 4 (Truthful regret controls execution cost). *Fix true costs and use (3). Let*

$$R_i(c, b) = u_i(c, (c_i, b_{-i})) - u_i(c, b), \quad \tau = 1 - 1/k.$$

Then all $R_i \geq 0$ and

$$\boxed{\sum_i R_i(c, b) \geq \tau \left(c_{w(b)} - \min_i c_i \right)}. \quad (6)$$

Every coarse-correlated equilibrium (CCE) of the fixed-type expected-audit-payoff report game allocates efficiently. If each agent's gain from any fixed unilateral report is at most ϵ_i in expectation, excess expected execution cost is at most $\tau^{-1} \sum_i \epsilon_i$. The coefficient τ is tight for this truthful-regret inequality whenever $P > H$; at $P = H$ it improves to 1.

The result concerns the fixed-type base report game. The standard regret-to-CCE averaging argument [12] gives execution efficiency in that game. Audit resource costs and bid-contingent transfer contracts are separate from this guarantee.

A.5 Truthful regret bounds execution cost

Proof. Individual truthfulness is included in the coalition guarantee. A losing agent that truthfully wins gains $H - c_i - \ell_P(\beta) \geq \beta - c_i$; otherwise its truthful gain is nonnegative. Thus $R_i \geq (\beta - c_i)_+$ for losers, including ties.

For an underreporting winner, truthful reporting either retains the task or loses to the cheapest outside bid r . In the first case,

$$R_w = a_P(\beta) \geq \tau(H - \beta) \geq \tau(c_w - \beta).$$

In the second, $\beta \leq r \leq c_w$ and

$$R_w = c_w - \beta - [\ell_P(\beta) - \ell_P(r)] \geq c_w - \beta - \frac{r - \beta}{k} \geq \tau(c_w - \beta).$$

A winner that is not underreporting gives the trivial lower bound zero. If the allocation is inefficient, a minimum-cost agent is a loser. Adding its regret yields

$$\sum_i R_i \geq \tau(c_w - \beta)_+ + \left(\beta - \min_i c_i \right)_+ \geq \tau(c_w - \min_i c_i).$$

If the allocation is efficient the claim is immediate. Taking expectations and using the fixed truthful deviation in the CCE constraints proves the equilibrium bounds.

For tightness when $P > H$, set $d = \min\{kL, H\} > 0$ and $\beta = H - d$. Give the first-priority winner cost H and bid β , a later-priority loser cost β , and all other bids/costs H . The winner's truthful regret is τd ; all losers' truthful regrets are zero because reporting β loses the priority tie. The execution gap is d . When $P = H$, the subsidy vanishes and the same calculation gives the stronger factor 1. $\square$

B Subsidy form and exact arithmetic checks

B.1 Deriving the common subsidy form

The form can itself be derived from a somewhat broader architecture. Suppose loser transfers are deterministic functions of bids, unchanged by the winner's audit outcome, and no loser can be sanctioned. Singleton resistance makes a loser's transfer independent of its own report whenever it stays a loser: compare the two reports in both directions. Pair resistance between two losers then removes dependence on another loser's report, since that reporting agent's own transfer is already unchanged. For a fixed winner at $\beta < H$, raise all losing bids to H without changing the winner. Transfers depend only on winner identity, recipient identity, and β . Symmetry of transfers at unique-winner profiles reduces this to a common $\ell(\beta)$; this does not require anonymous tie-breaking. At the all- H boundary, a first-priority winner of true cost $t < H$ can raise its report to H with any loser as a coalition partner. The established exit bound then forces that loser's boundary transfer to zero as $t \uparrow H$. This argument does not cover audit-contingent loser transfers or extra loser sanctions.

B.2 Exact implementation checks

A normalized exact-rational oracle enumerates costs and bids in $\{0, 1/2, 1\}$ for $n = 3$, with arbitrary outsider bids. It checks the canonical capped mechanism at $F = 0$, both coalition caps, and four sponsor budgets:

Sponsor cap P	Max audit, $k = 2$	Max audit, $k = 3$
1	1	1
3/2	3/4	3/4
2	1/2	2/3
3	1/2	2/3

Table 3: Exact implementation checks of normalized worst-case audit probabilities; not model observations.

All eight settings have zero profitable deviations across 37,908 coalition comparisons, meet their payment caps, and pass 5,832 cost/report regret checks. The canonical oracle also verifies 48 capture epsilon witnesses and 24 coalition-amplification witnesses. At the balanced budgets, maximum payouts are 2 and 5/3; larger allowed budgets do not increase payouts. Halving the balanced audit rates produces positive finite-grid gains 1/4 and 1/3.

The continuous capture lower bound can be a supremum: finite-grid maxima need not attain it. These are formula checks, not an LLM experiment or population estimate.

C Witness experiment and economic accounting

C.1 Protocol and finite-action comparison

Two model actors A and B share their net credits through an enforced equal-profit pool. A scripted outsider O bids its cost. Bids are integers in $[0, 100]$, ties favor A, then B, then O, and all games permit messaging. Capture uses costs $(1, 1, 0)$ and amplification uses $(100, 100, 100)$. Each profile is crossed with budgets 120, 160, 200 and audit multipliers 0.75, 1.00, 1.10, with three repetitions per cell. Scoring integrates over the audit lottery and compares coalition utility with truthful reporting under the same pool and mechanism.

All 54 assigned games have valid financial outcomes. Eighteen have positive gains and 36 have zero gains. Every outcome reaches the exact finite-action payoff maximum. The integer capture instance incurs one extra execution-cost unit, placing its low-audit gain one credit below the corresponding continuous capture supremum. The finite-action oracle compares each observation with its own legal action space. Four capture outcomes at canonical or higher auditing use payoff-neutral overreports, consistent with the weak incentive guarantee.

C.2 Actual audit demand

Amplification agents choose progressively lower bids as the budget rises. Their actual audit probability rises even though the mechanism’s worst-case bound falls. Capture keeps its minimum bid at zero and its actual audit probability falls (Figure 4). This distinction matters when translating a worst-case guarantee into observed audit demand.

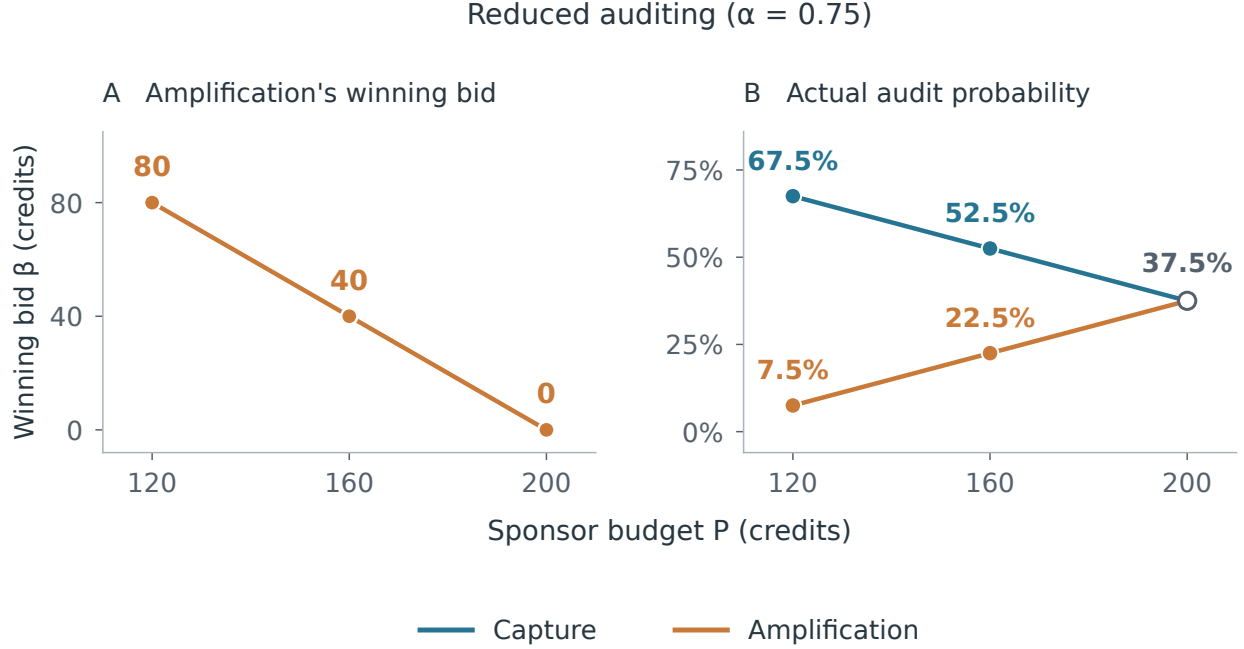


Figure 4: A larger budget can raise actual audit demand even as the worst-case audit bound falls. Under reduced auditing ($\alpha = 0.75$), amplification follows the subsidy kinks at winning bids 80, 40, 0, raising actual audit probability from 7.5% to 37.5%. Capture keeps its minimum bid at zero and its audit probability falls from 67.5% to 37.5%. Each point represents three identical observations, with lines connecting the sampled budgets. The two audit curves meet at the hollow marker. At canonical and higher auditing, AMP has winning bid 100 and audit probability zero. The extended archive retains the full 54-game accounting.

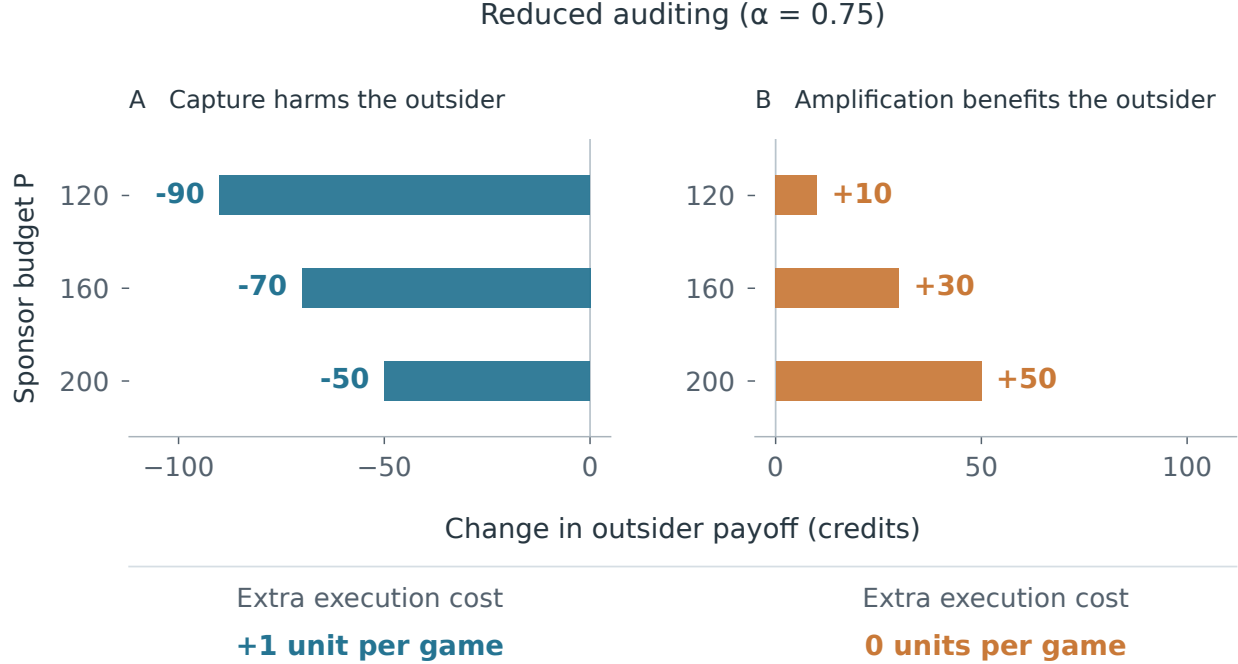
C.3 Transfers and execution cost

Let S_L be total loser subsidies, Δ_C coalition gain, Δ_O outsider utility change, ΔS expected sponsor expenditure change, and Δc execution cost change. All changes use the truthful baseline of the same mechanism. The accounting identity is $\Delta_C + \Delta_O = \Delta S - \Delta c$.

Witness	P	q	S_L	Δ_C	Δ_O	ΔS	Δc
CAP	120	0.675	20	21.5	-90	-67.5	1
CAP	160	0.525	60	16.5	-70	-52.5	1
CAP	200	0.375	100	11.5	-50	-37.5	1
AMP	120	0.075	20	2.5	10	12.5	0
AMP	160	0.225	60	7.5	30	37.5	0
AMP	200	0.375	100	12.5	50	62.5	0

Table 4: Budget-cohort accounting under reduced auditing. Every row represents three identical games.

At $\alpha = 1.00$ or 1.10 , all four changes are zero in each of the remaining 36 games. Capture still pays total loser subsidies $P - 100$ and has audit probability $\alpha[1 - (P - 100)/200]$. Amplification uses winning bid 100, pays no loser subsidy, and requires no audit at those observed reports. Figure 5 separates the outsider's payoff from execution resources. A monetary audit cost would additionally require a measured audit unit cost.



At $\alpha = 1.00$ or 1.10 , both changes are zero in all 36 games.

Figure 5: Coalition gains have different consequences for the outsider and for execution. Under reduced auditing ($\alpha = 0.75$), capture reduces outsider payoffs and adds one execution unit, while amplification raises outsider payoffs with unchanged execution cost. Each bar represents three identical games. Changes are measured against the truthful baseline of the same mechanism. The other 36 games have zero change in both quantities, as summarized below the plots. Total loser subsidies are monetary transfers and remain in the preceding accounting table.

D Unassigned-coalition data and payoff diagnostics

D.1 Design and replay

The supplied dataset contains 120 episodes labeled DeepSeek V4.1 Flash and 36 labeled GLM-5.3 Flash, each with eight rounds. Three authenticated bidders receive fresh private costs from $\{0, 5, \dots, 100\}$ and may submit real-valued bids in $[0, 100]$. Ties favor the smaller identity. Each bidder sees its own cost, the rules, public past-round history, and any current sequential chat. Messages are truncated to 300 characters. The objective is to maximize the bidder’s own coins, with no assigned coalition.

The mechanisms are defined in Table 2. Costs match by episode and round across conditions. Conditional transfers may depend on winning or losing, and outgoing transfers are capped by positive net pre-transfer utility without recycling incoming transfers. The pair diagnostic uses base utilities. P1 changes the incentive reminder, while NC removes both messages and transfers.

Offline replay checks all 1,248 round records, including bids, winner selection, audit comparisons, transfers, and payoff conservation. The DeepSeek records have successful final parsing throughout. The GLM records include 18 retries and one final failed bid in P0 Opt, replaced by a truthful fallback. All rounds remain in the tables. Excluding that fallback changes GLM Opt’s episode-weighted mean

E from -12.26 to -11.87 , with zero positive- E rounds in both versions. The recomputed values supersede conflicting numerical claims in the supplied draft.

D.2 Endpoints

Let $u_i(c, b)$ denote base mechanism expected utility, integrating over auditing, and let c also denote the truthful report vector. Our grand-total endpoint is

$$E(c, b) = \sum_{i=1}^3 (u_i(c, b) - u_i(c, c)).$$

It is independent of internal transfers, but does not certify that every subset lacks a profitable deviation. To connect observations to the theorem’s arbitrary-outsider guarantee, we additionally compute the post hoc report-profile diagnostic

$$G_2(c, b) = \max_{C: |C|=2} \sum_{i \in C} (u_i(c, b) - u_i(c, (c_C, b_{-C}))).$$

The outsider’s *actual* report is held fixed. A positive value identifies a pair-level base-report gain at the observed profile. It does not identify coalition membership or account for changes in contingent transfer promises.

Let $S(c, b)$ be sponsor expenditure in expectation, excluding real auditing resources. With $F = 0$ and transfers internal to bidders, total utility equals $S(c, b) - c_{w(b)}$, so

$$S(c, b) - S(c, c) = E(c, b) + c_{w(b)} - \min_i c_i.$$

The fiscal comparison uses each mechanism’s own truthful baseline. Means first average rounds within episodes and then average the 12 episodes in a condition. Counts retain all 96 rounds per condition. The pair diagnostic was added during integration after the records were observed.

D.3 Complete condition summaries

The model names below abbreviate logged harness labels. DS denotes the DeepSeek-labeled data. A coordination label is a pre-existing model-judge output. Its frequency is reported separately from expected payoffs.

Model	Cell	E	$E > 0$	$G_2 > 0$	Loss	ΔS
DS	NC_noaudit	-7.06	31/96	43/96	15.52	8.46
DS	NC_opt	-37.30	0/96	0/96	13.39	-23.91
DS	P0_half	-18.15	0/96	7/96	7.29	-10.85
DS	P0_noaudit	-20.04	12/96	23/96	11.20	-8.84
DS	P0_nosub	-19.81	0/96	0/96	7.66	-12.16
DS	P0_opt	-28.54	0/96	0/96	7.81	-20.73
DS	P1_half	-20.43	0/96	9/96	7.50	-12.93
DS	P1_noaudit	-20.67	12/96	22/96	11.41	-9.26
DS	P1_nosub	-20.54	0/96	0/96	6.82	-13.72
DS	P1_opt	-29.40	0/96	0/96	9.43	-19.97
GLM	NC_noaudit	-1.72	53/96	76/96	13.12	11.41
GLM	P0_noaudit	-2.95	39/96	55/96	7.45	4.50
GLM	P0_opt	-12.26	0/96	0/96	2.81	-9.45

Table 5: Recomputed outcomes in the unassigned-coalition study. Loss is excess execution cost; ΔS is expected sponsor expenditure minus the same mechanism’s truthful expenditure. GLM Opt includes one truthful fallback, not a valid model bid. Counts refer to rounds, not independent trials.

Model	Cell	Judge agreement	$\mathbb{E}[q]$	$\mathbb{E}[S]$	Realized S
DS	NC_noaudit	—	0.000	159.33	159.33
DS	NC_opt	—	0.519	126.96	124.81
DS	P0_half	0.854	0.227	140.01	142.32
DS	P0_noaudit	0.812	0.000	142.03	142.03
DS	P0_nosub	0.885	0.632	87.84	87.50
DS	P0_opt	0.812	0.451	130.14	128.47
DS	P1_half	0.885	0.215	137.94	138.76
DS	P1_noaudit	0.896	0.000	141.60	141.60
DS	P1_nosub	0.865	0.616	86.28	86.46
DS	P1_opt	0.865	0.388	130.90	127.33
GLM	NC_noaudit	—	0.000	162.28	162.28
GLM	P0_noaudit	0.906	0.000	155.37	155.37
GLM	P0_opt	0.969	0.440	141.42	140.85

Table 6: Coordination labels, expected audit demand, and expenditure levels. Audit probabilities are not monetary audit costs. Mechanisms have different subsidy budgets; raw spending levels are not a measure of collusive harm.

E Communication restrictions and the auction check

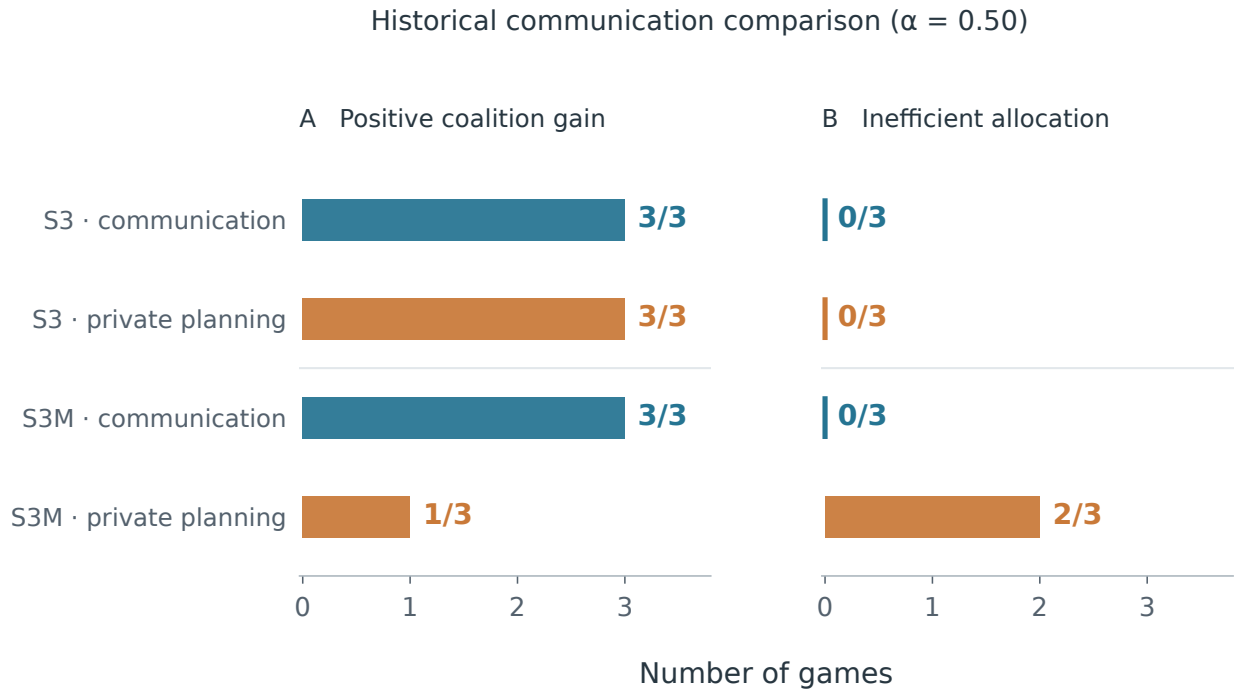
E.1 Historical communication comparison

A separate 36-game study compares communication with private planning under an enforced equal-profit pool. It uses the balanced-sharing rule with $n = 3$, $k = 2$, $H = 100$, and $F = 0$. Costs are S3 (80, 100, 100) and the mirrored S3M (100, 80, 100), the outsider bids 100, and ties favor A, B, then O. Each scenario is crossed with audit multipliers 0.50, 1.00, 1.50 and the two communication conditions, with three repetitions per cell.

Both actors have two decision turns. Communication exposes the partner’s preliminary message

and plan, while private planning exposes neither. The pool, action space, and payment rules are unchanged. All 36 games have valid outcomes. At low auditing, S3 yields gain 15 with either channel. In S3M, communication yields gain 15 in all three games, while private planning yields gains $-30, 15, -30$. The two negative-gain outcomes each add 20 execution-cost units. All 24 games at canonical or higher auditing have zero gain and zero excess execution cost.

The mirrored comparison jointly changes cost alignment with speaking order and tie priority. Communication also permits cost disclosure. The observed tradeoff therefore concerns these specific information and role arrangements. The earlier 12-game voluntary-transfer pilot and six-game pool diagnostic remain in the extended archive.



At $\alpha = 1.00$ or 1.50 , both outcomes are absent in all 24 games.

Figure 6: Message restriction coincides with fewer profitable deviations and more inefficient allocations in the mirrored assignment S3M, while S3 is unchanged. Bars count outcomes in the three low-audit games per condition ($\alpha = 0.50$). Every positive gain is 15 credits. The two inefficient private-planning outcomes in S3M each have gain -30 and excess execution cost 20. At canonical and higher auditing, all 24 games have zero gain and zero excess execution cost. These 36 historical communication-study games are separate from the budget experiment.

E.2 Separate double-auction feasibility check

The adapted double-auction study uses three buyers and three sellers for 15 rounds, with seller costs 80 and buyer values 100. Agrawal et al. [10] use five buyers and five sellers over 30 rounds in their main condition. The adapted study planned communication, no-chat, and oversight conditions at three market seeds.

The prespecified screen required sustained elevated asks, sufficient trading at prices above 90, and

reciprocal seller support for a common price policy. None of the three communicating controls met that screen. They completed 32, 22, and 24 trades, at trade-weighted prices 82.96, 84.44, and 82.56. The six intervention cells were therefore not run. This branch supplies a feasibility result, while a direct ranking of auction restrictions requires the intervention comparison. Full trajectories, the earlier pilot, and the screening details are preserved in the extended report and original archive.

F Evidence scope and archive guide

The mechanism guarantee uses fixed authenticated identities, immutable costs, binding execution, a perfect selected-cost check, and enforceable settlement. A candidate implementation must make that cost evidence available. Monetary transfers, execution resources, and audit resources are separate quantities.

The model studies use small cells and incompletely observed backend and decoding settings. The provider names in the supplied dataset are logged labels, and its judge labels are unvalidated measurements of coordination. The prescribed-witness games test assigned coalitions, while the pair diagnostic identifies incentives at observed report profiles. The studies are analyzed separately under their own protocols.

The accompanying package retains the evidence at three levels.

- `D2_REPLAY.json` and the JSON files under `figures/` contain the numerical sources for the reported summaries and figures.
- `evidence/Extended_Report_44pp.zip` preserves the full report before condensation, with its editable source, all original appendix tables, full-profile and competition calculations, and detailed provenance.
- `evidence/Original_Research_Archive.zip` and `evidence/D2_Original_Archive.zip` preserve the supplied experiment records, code, prompts, responses, and version histories.

AI assistance supported mathematical exploration, code, experiment orchestration, drafting, and review. Model actors also supplied the behavioral observations. This revision uses the existing records and runs no new model experiments.