ON THE BEHAVIOR OF LLMS AS MEDIATORS IN BILATERAL TRADE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Model (LLM) agents increasingly act on behalf of people in commerce, including as mediators that help a buyer and a seller reach a deal. Evaluating such mediators is challenging since (i) their pricing policy is latent: we observe the prices they propose, not the rule that produces them, and (ii) outcomes such as deal rates depend on how the counterparties respond, so they cannot separate a good policy from a favorable environment. To address these challenges, we study LLM mediators in multi-round bilateral trade with private valuations, pair them with persona-parameterized deterministic agents whose responses are known by construction, and recover each mediator's revealed policy with a counterfactual audit that patches one input of the negotiation history at a time. Across 504 market configurations and three frontier model families, we find three consistent behaviors: buyer-facing proposals concede toward the buyer while seller-facing proposals stay near compromise prices, later proposals are driven mainly by the mediator's own earlier proposals rather than by the parties' offers, and the mediator filters offers by trajectory, tracking concessions that continue the bargaining path while ignoring offers that reverse it. We show that this filtering removes the gain from retracting a concession but leaves an incentive to stall, and we derive the feasibility horizon, surplus decomposition, and mechanism-design ceilings of the environment.

1 INTRODUCTION

LLM agents are taking on more responsibilities on behalf of people and businesses, and commerce is one of the fastest-moving areas: users, agents, and merchants can complete purchases in chat (OpenAI, 2025; Stripe, 2025; Perplexity, 2024), assistants handle customer-service interactions (Klarna, 2024), and agents negotiate procurement contracts autonomously (Pactum, 2026). Recent studies analyze LLMs negotiating for one side of a trade (Bianchi et al., 2024; Hua et al., 2024b; Liu et al., 2026; Bergemann et al., 2026), including a field experiment in which employees' agents traded personal belongings (Anthropic, 2026). A natural next role is that of a *mediator*. In many high-stakes sales, such as a house or a used car, a third party helps the buyer and the seller reach a deal, bridging information gaps and raising the rate and efficiency of trade (Jarque et al., 2003; Larsen et al., 2021; Eilat & Pauzner, 2021; Glode & Opp, 2016), and a growing number of startups build LLM tools for this purpose (Indigo, 2025; Keelvar, 2026).

Evaluating an LLM mediator is harder than evaluating a negotiator. A negotiator is judged by the surplus it wins for its principal; a mediator's value lies in the rule by which it turns the parties' offers into proposals, and in the incentives that rule creates. Two difficulties follow. First, the *policy is latent*: we observe the prices a mediator proposes, not the rule that produces them, and LLM decision logic can contain unintended pricing heuristics or biases (Calvano et al., 2020; Klein, 2021; Sharma et al., 2024) and can game high-level objectives (Pan et al., 2023; Cemri et al., 2025). Second, *outcomes depend on the counterparties*: a deal rate reflects how the buyer and seller respond as much as what the mediator proposes, so the same mediator can look strong against one population of parties and weak against another. We find exactly this: an instruction to maximize the probability

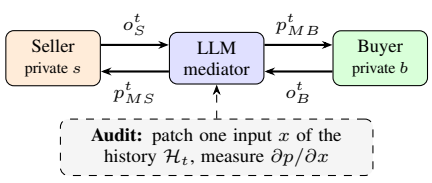

Figure 1: Mediated bilateral trade. The mediator proposes prices to each party in private. We recover its revealed policy by patching one input of the negotiation history at a time.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of agreement raises agreement against LLM negotiators and lowers it against deterministic ones (Section 5.4).

To address these challenges, we build a testbed in which both the counterparties and the mediator's policy can be analyzed (Figure 1). A seller and a buyer with private valuations interact over five rounds through a mediator that proposes prices to each side in private. In our primary environment, the parties are deterministic agents whose concession schedules and forward-looking acceptance rules are calibrated once from textual personas, so their responses are known by construction and contain no model of the mediator. This lets us characterize what any mediator can achieve (Section 3). We then recover each mediator's revealed policy with a *counterfactual policy audit* (Section 4): we patch one scalar input of the negotiation history on a dense grid, keep everything else fixed, and fit the response of the proposed price, which yields a local model of the policy and the incentives it creates. A second environment with free-language LLM negotiators tests whether the conclusions carry over. We evaluate Gemini 3.5 Flash, GPT-5.6, and Claude Sonnet 5 across 504 market configurations, under neutral prompts and prompts that prioritize fairness, deal probability, or mediator revenue.

Our analysis yields four main findings. (i) *The policy is side-dependent and drifts over time*: buyer-facing proposals concede toward the buyer's reserve round after round, while seller-facing proposals stay near compromise prices. (ii) *Proposals become self-anchored*: in later rounds, the mediator's own earlier proposals, including those the recipient rejected, move its price more than the parties' latest offers. (iii) *The mediator filters offers by trajectory*: it passes through concessions that continue the bargaining path and ignores offers that reverse it. This filter makes retracting a concession unprofitable, but it also makes holding firm weakly better than conceding, and it cannot tell strategic retractions apart from the ask increases that discounting forces on impatient sellers. These behaviors replicate across the three model families, whereas product-value priors are model-specific. (iv) *LLM mediators create surplus mainly by closing early, and leave much of the early window unused*: unprompted Gemini attains a normalized gain from trade of 0.527, two thirds of a dominant-strategy ceiling and below the 0.597 that a mediator knowing the buyer's value obtains from the first round alone; GPT and Claude are statistically close to the best stationary heuristics. The audit itself applies to any LLM agent that maps a textual history to a numerical action.

2 A TESTBED FOR LLM MEDIATION

Mediated bilateral trade. A Buyer (B), a Seller (S), and a Mediator (M) interact over a single indivisible item for at most T rounds. The buyer holds a private valuation $b \geq 0$ and the seller a private reservation value $s \geq 0$ on net revenue. If trade occurs in round t at price p, the buyer pays p, the mediator receives a public commission αp with $\alpha \in (0, 1)$, and the seller receives $(1 - \alpha)p$; α is an in-simulation payoff parameter, not a real payment. Both parties discount with factors $\gamma_B, \gamma_S \in (0, 1]$, and realized utilities are

$$u_B(t, p) = \gamma_B^{t-1} b - p, \qquad u_S(t, p) = \gamma_S^{t-1}(1 - \alpha)p - s, \qquad u_M(p) = \alpha p,$$

with zero utility for all players if no agreement is reached by round T. Discounting the buyer's gross valuation models perishable consumption utility, and discounting the seller's receivables models the opportunity cost of delayed revenue; as a consequence, the zone of agreement contracts over rounds (Section 3). Individually rational trade exists in round 1 if and only if $(1 - \alpha)b \geq s$, and our grid satisfies $b - s \geq 0.15 \gg \alpha b$.

Protocol 1. Mediated Negotiation (Round t)

Seller submits asking price o_S^t.
Mediator observes o_S^t and proposes p_{MB}^t to Buyer.
if Buyer accepts p_{MB}^t **then**
  if Seller validates p_{MB}^t **then**
    return Trade at $p = p_{MB}^t$.
else
  Buyer counters with o_B^t.
  Mediator observes o_B^t, proposes p_{MS}^t to Seller.
  if Seller accepts and Buyer validates p_{MS}^t **then**
    return Trade at $p = p_{MS}^t$.
if $t = T$ **then** terminate with no trade.

LLM mediator policy. A mediator persona ψ_M is sampled into the system prompt. At round t, for recipient $i \in \{B, S\}$, the model receives context $C_i^t = \Phi_i(\psi_M, \alpha, t, T, \mathcal{H}_t, z_i^t)$, where $\mathcal{H}_t$ is the negotiation history and z_i^t is the latest observed offer ($z_B^t = o_S^t$, $z_S^t = o_B^t$). It outputs a message y_i^t, a proposal price p_i^t, and private structured reasoning e_i^t. A failed validation ends the round without a counteroffer; validation accepts if and only if $u \geq 0$.

Deterministic buyer and seller agents. In our primary environment, the buyer and seller are not LLMs. To obtain behavioral diversity without stochastic generation, GPT-4o scores a sampled natural-language persona once into urgency u and stubbornness r in $[0, 1]$, which set discount factors and concession rates (Appendix C). The buyer does not initiate offers; upon rejecting a proposal it counters with $o_B^t = \min\{\gamma_B^{t-1} b,\ \frac{b}{\beta_B}(1+\delta_B)^{t-1}\}$, and the seller opens each round with $o_S^t = \max\{\frac{s}{\gamma_S^{t-1}(1-\alpha)},\ \lambda_S s(1-\rho_S)^{t-1}\}$, where $\beta_B, \lambda_S \geq 1$ set initial aggressiveness and δ_B, ρ_S set concession speed. For $t > 1$, an agent accepts a proposal p_t if and only if accepting beats the discounted utility of the next proposal, forecast by linear extrapolation from the first proposal it observed, $\widehat{p}_{t+1} = p_t - (p_1 - p_t)/(t-1)$; for the buyer, $\gamma_B^{t-1} b - p_t \geq \max\{0,\ \gamma_B^t b - \widehat{p}_{t+1}\}$, and analogously for the seller (Appendix B), with zero continuation value at $t = T$. A single observation does not determine a trend, so in round 1 the buyer accepts only if $p_1 \leq 0.3b$ and the seller only if $(1-\alpha)p_1 - s \geq 0.7s$. The rules contain no model of the mediator, so LLM and rule-based mediators face identical responses. In a second environment, the parties are free-language GPT-5.4-mini agents that know their reservation values but not the utility formulas (Appendix A.1).

Metrics. Let $D_j \in \{0, 1\}$ indicate trade in episode $j \in \{1, \ldots, N\}$ at price p_j. We report the deal rate $\frac{1}{N}\sum_j D_j$, the expected price $\mathbb{E}[p] = \frac{1}{N}\sum_j D_j p_j$, gains from trade $\mathrm{GFT} = u_B + u_S + u_M$, normalized gains from trade $\mathrm{NGFT} = \mathrm{GFT}/(b-s)$, and Nash welfare $\mathrm{NW} = u_B u_S$, with non-agreements contributing zero. For a final price p_f, the realized anchor is $a_f = (p_f - s)/(b-s)$ and the fairness ratio is $\mathrm{Fairness} = 2u_B/(u_B + u_S + u_M)$; because $u_M > 0$ enters the denominator, an equal split $u_B = u_S$ evaluates to about 0.95–0.98, and a value of 1 means the buyer's surplus equals the seller's plus the fee. To locate proposals within the bargaining range, we use the normalized anchors $a_{MB}^t = (p_{MB}^t - o_B^{t-1})/(o_S^t - o_B^{t-1})$ for $t > 1$ and $a_{MS}^t = (p_{MS}^t - o_B^t)/(o_S^t - o_B^t)$, and the seller discount ratio $d_S^1 = p_{MB}^1/o_S^1$ before any counteroffer; after a failed validation, o_B^{t-1} is the latest available counteroffer. We also track the rejection ratios $BR^t = p_{MB}^t/o_B^t$ and $SR^t = o_S^t/p_{MS}^{t-1}$ and the normalized gap $G^t = (o_S^t - o_B^t)/o_S^t$.

3 What Any Mediator Can Achieve

Because the counterparties' responses are known, we can characterize the environment before looking at any mediator. The results below provide reference points for Section 5; proofs are in Appendix D.

Proposition 1 (Feasibility horizon and ask escalation)**.** *Individually rational trade in round t exists if and only if $(\gamma_B\gamma_S)^{t-1}(1-\alpha)b \geq s$. Hence no mediator can close a deal after round $\tau(b,s) = 1 + \left\lfloor \ln\!\left(\frac{(1-\alpha)b}{s}\right) / \ln\!\left(\frac{1}{\gamma_B\gamma_S}\right) \right\rfloor$. Moreover, the seller's individual-rationality floor binds in the asking price at round t if and only if $(\gamma_S(1-\rho_S))^{t-1} \leq 1/(\lambda_S(1-\alpha))$, a condition independent of s; it never binds at $t = 1$ when $\lambda_S(1-\alpha) > 1$, and once it binds at some t_0, the ask strictly increases in every later round when $\gamma_S < 1$.*

With $\gamma_B = \gamma_S = 0.9$, the zone of agreement is empty in round 5 for 69% of the 72 valuation pairs, and with $\gamma = 0.85$ for 94% (Table 4); under the persona map, the seller's floor binds by round 5 for 40% of the (u, r) square. Some upward moves of the seller's ask are therefore forced by individual rationality rather than chosen strategically, which matters for how a mediator should react to them. The agents' round-1 thresholds further restrict early agreement.

Corollary 1 (Round-1 window)**.** *Under the round-1 acceptance thresholds, a round-1 agreement exists if and only if $(1-\alpha)b \geq 1.7s$, which holds for 43 of the 72 valuation pairs regardless of personas. A round-1 agreement reached through the seller's acceptance has price $p \geq 1.7s/(1-\alpha)$, hence final anchor $a_f \geq \left(\frac{1.7}{1-\alpha} - 1\right)\frac{s}{b-s}$, which averages 0.56 over the feasible pairs; an agreement through the buyer's acceptance requires $0.3(1-\alpha)b \geq s$, which holds for only 6 pairs. A mediator that knows b attains $\mathrm{NGFT} \geq 43/72 = 0.597$ for any personas.*

Proposition 2 (Surplus decomposition)**.** *For trade at round t and price p, $\mathrm{GFT}(t,p) = (b-s) - (1-\gamma_B^{t-1})\,b - (1-\alpha)(1-\gamma_S^{t-1})\,p$. Thus $\mathrm{NGFT} = 1$ for every individually rational round-1 trade, delay destroys surplus at both parties, and for $t > 1$ and $\gamma_S < 1$ realized surplus strictly decreases in the transaction price, $\partial\,\mathrm{GFT}/\partial p = -(1-\alpha)(1-\gamma_S^{t-1})$.*

Proposition 3 (Mechanism-design ceilings)**.** *Let (b, s) be uniform over the 72 valuation pairs of our grid, and consider static direct mechanisms that trade in round 1 with probability $x(b,s)$ at*

Algorithm 2 Counterfactual policy audit

Require: agent π mapping a textual history to a price; base histories $\{\mathcal{H}_h\}$; scalar variables $\mathcal{X}$ appearing in $\mathcal{H}_h$; coherent grids $\{x_j\}$; samples K

1: **for** each history $\mathcal{H}_h$ and variable $x \in \mathcal{X}$ **do**
2: **for each** x_j: patch every occurrence of x in $\mathcal{H}_h$ by x_j; $\bar{p}_{h,j} \leftarrow \frac{1}{K}\sum_k \pi(\mathcal{H}_h^{x \leftarrow x_j})$
3: Fit $\bar{p}_{h,j}$ on x_j piecewise-linearly with a knot at the previous-round value x^{t-1}, giving (β_L, β_R)
4: Assemble the local policy $p^t = c + \rho\, p^{t-1} + \beta_L \min\{z^t, z^{t-1}\} + \beta_R (z^t - z^{t-1})^+$
5: Derive incentives and pass-through from the local policy (Proposition 4)

expected payment $P(b, s)$, retain the commission αP, and satisfy ex-post individual rationality. The maximal expected NGFT is 1 *with full information,* 0.960 *under Bayesian incentive compatibility, and* 0.790 *under dominant-strategy incentive compatibility; the best deterministic posted price (any $p \in [0.510, 0.600]$) attains* 0.736.

Together, these results say where surplus can come from. Early agreement is decisive, since only round-1 trade reaches $\mathrm{NGFT} = 1$ at any price; it is also mostly seller-favorable, so a mediator that closes early shows seller-leaning anchors even if its proposals are otherwise balanced; and at later rounds NGFT mechanically favors mediators that settle low, so we report conditional prices alongside it. The ceilings assume commitment and a known prior, and since the grid keeps only pairs with $b - s \geq 0.15$, they are far higher than in the independent uniform case (Myerson & Satterthwaite, 1983); we use them as reference points, not targets.

4 Auditing the Mediator's Policy

Outcome metrics aggregate the mediator's policy and the counterparties' responses. To isolate the policy, we intervene on the mediator's input directly (Algorithm 2). We take fixed negotiation prompts and perturb one scalar input on a dense grid while holding all other tokens, histories, and personas constant. The patched variables are the seller's ask o_S^t, the buyer's counteroffer o_B^t, the latest mediator-to-buyer proposal (p_{MB}^{t-1} for $M \to B$, p_{MB}^t for $M \to S$), and the previous mediator-to-seller proposal p_{MS}^{t-1}. Patched values are kept inside ranges that leave the history coherent (Appendix I.2). We sample the mediator $K = 5$ times per grid value, so each slope $\beta_x = \partial p / \partial x$ is a dollar adjustment per dollar of input, and we fit it piecewise around the previous-round value of the patched variable. Each decision point uses three base histories.

The audit summarizes the proposal to a party with own offer z as $p^t = c + \rho\, p^{t-1} + \beta_L \min\{z^t, z^{t-1}\} + \beta_R (z^t - z^{t-1})^+$, written for the seller (ask $z = o_S$, concession means lowering z); the buyer case mirrors it. Here ρ measures self-anchoring and (β_L, β_R) the response to offers that continue or reverse the party's path. The local model is enough to derive what the policy rewards.

Proposition 4 (Filtering and self-anchoring)**.** *Suppose $\beta_R = 0$ and $0 \leq \rho < 1$. (i)* No ratchet gain*: no retraction $z^t > z^{t-1}$ raises the proposal above what holding $z^t = z^{t-1}$ yields. (ii)* Holding firm is weakly optimal*: a concession of size Δ lowers p^t by $\beta_L \Delta$, so a party that maximizes the next proposal never strictly prefers to concede. (iii)* Persistence*: a maintained concession of size Δ at round t_0 lowers p^{t_0+k} by $\beta_L \Delta (1 - \rho^{k+1})/(1 - \rho)$, which tends to $\beta_L \Delta/(1 - \rho)$. (iv)* Cost of filtering*: if the seller's ask rises because its floor binds (Proposition 1) and the proposal is below the new floor, it remains below it, so the seller rejects.*

Part (i) is an anti-ratchet property (Freixas et al., 1985; Hart & Tirole, 1988): a party cannot manipulate the mediator by retracting. Part (ii) is its price, since a concession is passed through against the party that made it. Part (iv) shows that the filter cannot separate strategic retractions from ask increases that discounting forces on impatient sellers.

5 Results

We first describe which policy LLM mediators implement (Section 5.1) and what incentives it creates (Section 5.2), then how much surplus they create (Section 5.3) and how objective prompts change the picture (Section 5.4). We detail Gemini 3.5 Flash and replicate the round-1 product sweep, the

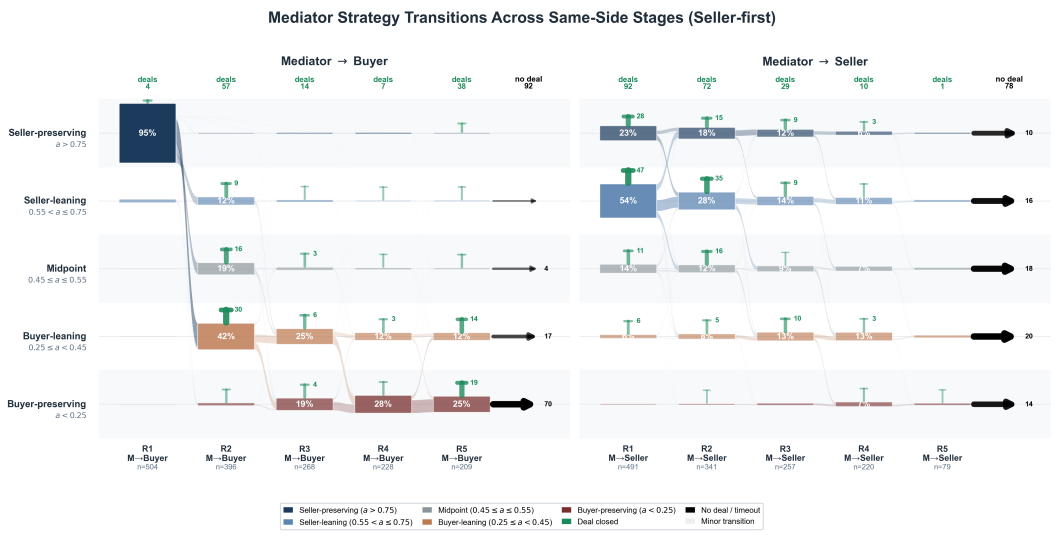

Figure 2: Mediator strategy transitions (Gemini 3.5 Flash, 504 episodes). Columns alternate between buyer-facing ($M \to B$) and seller-facing ($M \to S$) proposals across rounds 1–5; ribbon widths are transition frequencies, green markers are agreements validated after that stage, and black terminals are episodes without agreement at the deadline. Unanchored validation rejections are omitted.

full round-3 audit, and the economic evaluation with GPT-5.6 and Claude Sonnet 5 under identical histories, grids, and episodes (Appendix I.5).

5.1 Which policy do LLM mediators implement?

The policy is side-dependent and drifts over time. We classify each proposal with an active bargaining range by its normalized anchor into five classes: buyer-preserving ($a < 0.25$), buyer-leaning compromise ($[0.25, 0.45)$), midpoint ($[0.45, 0.55]$), seller-leaning compromise ($(0.55, 0.75]$), and seller-preserving ($a > 0.75$); round-1 buyer-facing proposals are classified by d_S^1. Figure 2 shows the flow across all 504 Gemini episodes. Roughly 85% of opening buyer-facing proposals are seller-preserving ($d_S^1 \geq 0.75$), whereas round-1 seller-facing proposals incorporate the buyer's counteroffer and disperse across compromise classes. Part of this difference is informational: the seller's ask is the only price available when p_{MB}^1 is formed (Appendix H). Over rounds 2–5, which share the normalization a_{MB}^t, buyer-facing proposals move from compromise classes to predominantly buyer-preserving (62% and 61% of active episodes in rounds 4 and 5), while seller-facing proposals remain in intermediate classes. Table 1 shows the pattern in a single episode: over two rounds the seller lowers its ask by \$7.63 and the buyer raises its counteroffer by \$2.36, yet the mediator's buyer-facing proposal falls by \$20.00 while its seller-facing proposal moves by only \$4.00.

Table 1: Anatomy of one mediated negotiation (Gemini 3.5 Flash. The base history of the orange curves in Figure 3 and the prompt in Appendix G). Anchors in parentheses; the round-1 buyer-facing proposal is measured by d_S^1. Buyer-facing proposals fall from seller-preserving to buyer-leaning while seller-facing proposals stay near 0.7, although the parties' own offers barely move.

Round	Seller ask o_S^t	Mediator → buyer p_{MB}^t	Buyer counter o_B^t	Mediator → seller p_{MS}^t
1	\$81.67	\$75.00 (0.92)	\$45.61	\$72.00 (0.73)
2	\$77.76	\$65.00 (0.60)	\$46.78	\$68.00 (0.68), seller accepted, buyer rejected
3	\$74.04	\$55.00 (0.30)	\$47.97	patched in the audit

The mediator's own history predicts its next proposal. Regressing current anchors on candidate predictors (Figure 6) identifies the mediator's previous seller-facing anchor as the dominant predictor of its current one ($R^2 = 0.620$, 0.869, and 0.878 in rounds 3, 4, and 5), with a noisier pattern on the buyer side. Higher SR^t predicts lower seller-facing anchors and higher BR^t higher buyer-facing

anchors, while the remaining gap G^t and the parties' concession velocities yield $R^2 < 0.05$. The audit tests whether these associations are causal.

Early proposals follow the seller's ask and a product prior. In round 1, the seller's ask is the only numerical anchor for p^1_{MB}. When it is implausibly low ($o^1_S < \$30$), Gemini often proposes a price *above* it, with mean ratios $d^1_S = 2.32$ for a used bicycle and 2.97 for an iPad; within a plausible range, the response is linear, $p^1_{MB} = 11.70 + 0.684\, o^1_S$ ($R^2 = 0.971$, bicycle) and $9.82 + 0.749\, o^1_S$ ($R^2 = 0.983$, iPad; Figure 9). The prior is model-dependent: Claude Sonnet 5 separates products similarly ($\beta = 0.753$ vs. 0.779), whereas GPT-5.6 is nearly invariant (0.859, 0.854, and 0.868 for a bicycle, an iPad, and a luxury watch). Round-1 seller-facing proposals follow the seller's contemporaneous ask ($\beta = 0.587$) far more than the buyer's counteroffer (0.101) or the mediator's own opening (0.083).

Later proposals are self-anchored. By round 3, the mediator's own history dominates (Figure 3). For seller-facing proposals, the buyer's counteroffer has little pull ($\beta = 0.062$), the latest mediator-to-buyer proposal yields $\beta = 0.193$, and the previous mediator-to-seller proposal $\beta = 0.551$. For buyer-facing proposals, the seller's ask matters little ($\beta_L = 0.076$, $\beta_R = 0.046$), whereas previous proposals from both sides do (0.290 for p^2_{MB} and 0.337 for p^2_{MS}). In all base histories, the seller had accepted p^2_{MS} before the buyer rejected it at validation, so part of that sensitivity may be rational retention of a price the seller is known to accept. The cleaner evidence comes from proposals the recipient *rejected*: the buyer countered p^2_{MB}, yet p^3_{MB} loads on it with $\beta = 0.290$, and p^3_{MS} loads on the rejected p^3_{MB} with $\beta = 0.193$. A rejection tells a Bayesian mediator that the price was too high for the buyer, which gives no reason to stay close to it.

Offers are filtered by trajectory. Responses to the parties' current offers are piecewise around the previous-round value. An ask that continues the seller's downward path ($o^3_S \le o^2_S$) moves the seller-facing proposal with $\beta_L = 0.182$, and an ask above o^2_S has no effect ($\beta_R = -0.027$). A buyer counteroffer below the buyer's trajectory is ignored ($\beta_L = -0.066$), and one that continues it moves the buyer-facing proposal with $\beta_R = 0.484$. GPT-5.6 and Claude Sonnet 5 reproduce the structure (Figures 10 and 11); in GPT's buyer-counter sweep the slope goes from 0.004 to 0.579, and both models load on previous mediator proposals with $\beta > 0.40$. At temperature zero the plateaus sharpen rather than disappear (Appendix I.7), so the thresholds are a property of the policy and not of sampling.

The structure persists to the deadline. In round 5 (Figure 12), seller-facing proposals respond to a continuing ask with $\beta_L = 0.125$ and not to a reversal ($\beta_R = -0.009$), ignore the buyer's counteroffer ($\beta = 0.002$), and load on the previous seller-facing proposal ($\beta = 0.357$). Buyer-facing proposals ignore the seller's ask ($\beta = 0.004$), respond to a trajectory-consistent counteroffer with $\beta_R = 0.588$ but not to a regressive one ($\beta_L = -0.037$), and load on both previous proposals (0.251 and 0.235). The deadline does not make the mediator more responsive to the latest offers.

5.2 What incentives does the policy create?

The estimates above satisfy the premise of Proposition 4: $\beta_R \approx 0$ for reversals on both sides, and $0 < \rho < 1$. The mediator therefore cannot be manipulated by retracted offers, but it teaches the parties that conceding is costly. A maintained seller concession passes through at 0.18 immediately and at $0.182/(1-0.551) = 0.41$ in the limit, and a buyer concession at 0.48 and $0.484/(1-0.290) = 0.68$; the memory of the mediator's own proposals has a half-life of about 1.2 rounds on the seller side. A party that learns this policy has a reason to stall, and in our environment the filter also ignores the ask increases that Proposition 1 forces on impatient sellers. These quantities are local linearizations around three histories per decision point, not structural estimates, but the qualitative pattern replicates in GPT-5.6 and Claude Sonnet 5.

5.3 How much surplus do LLM mediators create?

We compare LLM mediators with the *identity* mediator, which passes offers through; the *midpoint* mediator; and *weighted-average* mediators $p^t_B = o^{t-1}_B + \lambda(o^t_S - o^{t-1}_B)$ and $p^t_S = o^t_B + \lambda(o^t_S - o^t_B)$ for

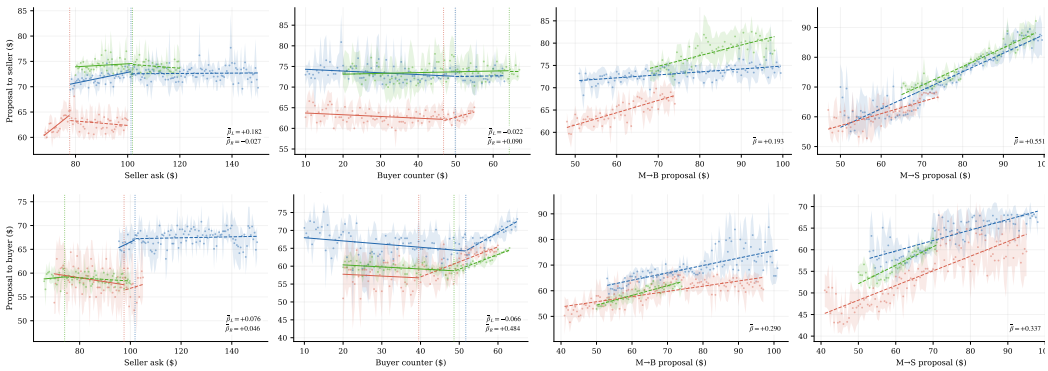

Figure 3: Round-3 counterfactual audit (Gemini 3.5 Flash). Top: p^3_{MS}; bottom: p^3_{MB}. Columns patch, from left to right, the seller ask o^3_S, the buyer counteroffer (o^3_B top, o^2_B bottom), the latest $M \to B$ proposal, and the previous $M \to S$ proposal. Colors are base histories; points are samples, bands show sampling variability, dotted lines mark the unperturbed value, and piecewise fits break at the previous-round value of the patched variable.

Table 2: Deterministic-agent outcomes (504 episodes). 95% bootstrap confidence intervals (10,000 resamples) in brackets. Bold marks the best point estimate within each model family (closest to 1 for Fairness). NW is scaled by 10^2. Complete metrics are in Appendix I.8.

Mediator	Prompt	Deal Rate	Fairness	NGFT	NW ($\times 10^2$)
Gemini 3.5 Flash	Default	**0.657** [0.615, 0.698]	**0.974** [0.920, 1.027]	**0.527** [0.491, 0.563]	**1.54** [1.37, 1.71]
	Fairness	0.635 [0.593, 0.677]	1.043 [0.988, 1.097]	0.468 [0.433, 0.502]	1.40 [1.23, 1.58]
	Revenue	0.546 [0.502, 0.587]	0.786 [0.739, 0.833]	0.394 [0.359, 0.429]	1.51 [1.30, 1.74]
	Deal prob.	0.619 [0.575, 0.661]	1.068 [1.015, 1.123]	0.479 [0.443, 0.514]	1.41 [1.25, 1.58]
GPT-5.6	Default	**0.575** [0.534, 0.617]	1.226 [1.176, 1.276]	**0.397** [0.365, 0.429]	**1.10** [0.97, 1.24]
	Fairness	0.573 [0.530, 0.615]	1.196 [1.147, 1.246]	0.386 [0.354, 0.417]	1.07 [0.94, 1.21]
	Revenue	0.542 [0.498, 0.585]	**1.188** [1.136, 1.239]	0.368 [0.335, 0.400]	1.09 [0.95, 1.24]
	Deal prob.	0.435 [0.391, 0.478]	1.272 [1.214, 1.329]	0.287 [0.256, 0.318]	0.77 [0.65, 0.90]
Claude Sonnet 5	Default	0.577 [0.534, 0.621]	1.198 [1.143, 1.252]	0.400 [0.368, 0.432]	1.05 [0.93, 1.18]
	Fairness	**0.623** [0.579, 0.665]	1.157 [1.107, 1.206]	**0.431** [0.399, 0.463]	**1.15** [1.03, 1.29]
	Revenue	0.593 [0.550, 0.635]	**1.123** [1.068, 1.179]	0.413 [0.381, 0.445]	**1.15** [1.02, 1.29]
	Deal prob.	0.534 [0.490, 0.577]	1.225 [1.172, 1.277]	0.363 [0.331, 0.395]	0.99 [0.86, 1.12]
Heuristic	Identity	0.627 [0.585, 0.669]	0.650 [0.583, 0.718]	0.308 [0.284, 0.332]	0.55 [0.46, 0.65]
	$\lambda = 0.7$	0.504 [0.460, 0.548]	1.018 [0.959, 1.079]	0.359 [0.326, 0.391]	1.20 [1.05, 1.36]

$\lambda \in \{0.1, \ldots, 0.9\}$. The rule-based mediators open by forwarding the seller's ask, and all mediators face the same counterparty rules (full results in Table 7).

Default Gemini 3.5 Flash reaches an NGFT of 0.527 [0.491, 0.563], separated from the strongest stationary rule ($\lambda = 0.7$, 0.359 [0.326, 0.391]). Default Claude Sonnet 5 (0.400) and GPT-5.6 (0.397) have higher point estimates, but their intervals overlap with $\lambda \in \{0.7, 0.8\}$, so we do not rank them above the best stationary rules. LLMs do not dominate on every metric: the identity mediator has a higher deal rate (0.627) than default GPT-5.6 and Claude, with skewed surplus (Fairness 0.650), and the weighted rules with $\lambda \in [0.7, 0.9]$ reach Nash welfare of 0.0120–0.0131, above default GPT and Claude.

The surplus comes from timing, and much of the early window goes unused. Gemini attains 67% of the dominant-strategy ceiling of Proposition 3 (0.527/0.790) and 55% of the Bayesian one; the best stationary heuristic attains 45% of the former. By Proposition 2, every round-1 agreement has NGFT exactly 1, so Gemini's 96 round-1 deals alone contribute $96/504 = 0.190$ of its 0.527. No heuristic closes a deal in round 1, and the identity mediator closes 82.6% of its deals in round 5, whereas Gemini closes 68.9% in rounds 1–2 (Table 3). The price channel of Proposition 2 works against Gemini, whose conditional price (0.517) exceeds that of every weighted rule (0.455–0.504), so its advantage is not explained by settling low. Yet Gemini's round-1 deals cover 19% of episodes, a third of the 0.597 of Corollary 1: a mediator that knows the buyer's value beats Gemini's total NGFT from round 1 alone. Consistent with Section 5.1, Gemini's round-1 seller-facing proposals

Table 3: Agreement timing under deterministic agents. Share of successful agreements closed in each round.

Mediator	Deals	R1	R2	R3	R4	R5
Identity	316	0.0%	6.3%	1.9%	9.2%	82.6%
Midpoint	264	0.0%	17.4%	28.4%	18.2%	36.0%
Weighted $\lambda = 0.7$	254	0.0%	29.9%	34.6%	13.0%	22.4%
Gemini 3.5 Flash, default	331	29.0%	39.9%	13.0%	5.1%	13.0%
Gemini, buyer-first	256	0.0%	42.2%	13.7%	5.5%	38.7%

track the seller's ask rather than what the buyer can pay. By Proposition 1, part of the remaining non-agreements are unattainable for any mediator, since for impatient parties the zone of agreement is empty well before the deadline.

Move order matters. Reversing the protocol so that the buyer moves first lowers Gemini's deal rate from 0.657 to 0.508 and its NGFT from 0.527 to 0.342, and removes all round-1 agreements (Table 3). The seller still captures more expected surplus than the buyer (0.092 vs. 0.058; Table 9), so the outcome is not a mirror image of the seller-first protocol.

5.4 Can objectives steer the mediator?

Under deterministic agents. Revenue prompting moves Gemini's outcomes reliably: fairness falls to 0.786 [0.739, 0.833], the final anchor rises to 0.559 [0.540, 0.578], and the deal rate falls to 0.546 [0.502, 0.587]. With an ad valorem fee, revenue αp rises with price, so fee seeking shifts surplus toward the seller; a flat fee would decouple the two. For Claude, the fairness prompt raises deal rate (0.577 to 0.623) and NGFT (0.400 to 0.431), within overlapping intervals. The deal-probability prompt lowers GPT-5.6's deal rate by 14.0 points to 0.435 [0.391, 0.478], with NGFT falling to 0.287. Its final anchor falls from 0.376 to 0.349 while its conditional price barely moves (0.463 to 0.446), so the drop in $\mathbb{E}[p]$ largely reflects fewer agreements. The shift toward the buyer is consistent with seller-facing proposals falling below the seller's look-ahead threshold and floor: an instruction to maximize agreement becomes a concession strategy that the deterministic seller refuses.

Under LLM negotiators. Against GPT-5.4-mini negotiators (Table 6), direct negotiation already reaches agreement in 503 of 504 instances in 1.77 rounds on average. Every mediator lowers the deal rate, and all but one configuration delay agreement (1.92–2.54 rounds; GPT-5.6 with the deal-probability prompt takes 1.78), so in this environment mediation redistributes surplus rather than creating it. Deal-probability prompting raises agreement for Gemini (0.913 to 0.964 [0.948, 0.980]), with smaller, overlapping increases for GPT (0.980 to 0.996) and Claude (0.952 to 0.980). Revenue prompting raises final anchors for all three families, most for Gemini (0.382 to 0.683 [0.657, 0.710], with prices rising from \$49.48 to \$61.96), at the cost of a lower Gemini deal rate (0.841). The deal-probability prompt thus raises agreement with LLM negotiators and lowers it with deterministic ones: the effect of an objective prompt depends on the counterparty's acceptance policy.

6 Discussion

The same mediator can look helpful or harmful depending on who it mediates between, and the same objective prompt can raise or lower agreement (Section 5.4). Outcome benchmarks alone therefore say little about how an LLM mediator will behave in a new market. The policy it implements is more stable: side dependence, self-anchoring, and trajectory filtering appear in all three model families, persist from round 3 to the deadline, and survive at temperature zero. This is why we treat the recovered policy, rather than any single deal rate, as the main object of evaluation, and the counterfactual audit applies unchanged to other LLM agents that map a textual history to a numerical action, such as pricing, bidding, or allocation agents.

LLM mediators are neither direct-revelation mechanisms nor stationary arbitration rules, but path-dependent arbiters. Their filter protects against retraction, which is desirable, but it passes concessions through against the party that made them, so parties who learn the policy have a reason to stall (Proposition 4). The surplus these mediators create comes mostly from early agreement, which in our environment is itself seller-favorable (Corollary 1), and the best of them captures two thirds of a dominant-strategy ceiling while leaving most of the round-1 window unused. Mediators should

thus be evaluated on timing and on the incentives their policies create, not only on outcome averages; training mediators for procedural fairness and path consistency is a promising direction.

Limitations. Neither deterministic nor LLM negotiators replace studies with human negotiators. The audit uses three base histories per decision point, all with a seller acceptance in round 2; an ablation that removes the validation annotation would further separate self-anchoring from rational retention. Plateaus in the sweeps partly reflect rounding to focal prices. Our heuristics are stationary; time-dependent (Faratin et al., 1998) and opponent-modeling (Baarslag et al., 2016) mediators would separate deadline-aware pacing from LLM-specific behavior. The economic evaluation covers a single-issue sale with the fee deducted from the seller's proceeds; other fee structures and multi-attribute bargaining (Lewis et al., 2017; He et al., 2018) remain open.

7 Related Work

Bilateral trade and mediation. Classical models characterize surplus division and bargaining dynamics through axiomatic or sequential games (Nash, 1950; Rubinstein, 1982). With private valuations, efficient, incentive-compatible, individually rational, and budget-balanced trade is generally impossible (Myerson & Satterthwaite, 1983), and Chatterjee & Samuelson (1983) characterize equilibrium bidding under a split-the-difference rule. Uncommitted mediation can expand the set of attainable equilibrium payoffs relative to unmediated communication (Forges, 1986; Jarque et al., 2003; Goltsman et al., 2009; Eilat & Pauzner, 2021; Salamanca, 2021), and outcomes in practice are shaped by anchors, framing, and concession paths (Bazerman & Neale, 1992; Galinsky & Mussweiler, 2001; Backus et al., 2020). We study which rule an LLM mediator actually implements and what it achieves relative to these benchmarks.

Automated and LLM negotiation. Automated negotiation studies agents with explicit utilities and fixed protocols (Jennings et al., 2001; Baarslag et al., 2012; 2016), including time-dependent concession tactics (Faratin et al., 1998), and language-based negotiation extended it to natural-language bargaining (Lewis et al., 2017; He et al., 2018). Recent benchmarks evaluate LLMs as buyers, sellers, or market participants (Davidson et al., 2024; Bianchi et al., 2024; Xia et al., 2024; Zhu et al., 2025; Vaccaro et al., 2025; Huang & Hadfi, 2024); GLEE (Shapira et al., 2024) standardizes language-based bargaining, negotiation, and persuasion games, and Bergemann et al. (2026) study LLM agents in bilateral trade with structured tool calls. These works ask whether an LLM negotiates well for one side; we ask how it arbitrates between two.

Evaluating LLM agents. Studies of LLM agents evaluate planning, tool use, cooperation, and strategic failure modes (Yao et al., 2023; Liu et al., 2024; Abdelnabi et al., 2024; Hua et al., 2024a; Cemri et al., 2025), mostly through end-of-episode outcomes. Our audit instead intervenes on the agent's input and reads off its revealed policy, which connects outcome evaluation to the incentives the agent creates for the people it interacts with.

8 Conclusion

Evaluating an LLM mediator requires knowing which rule it follows, not only which deals it closes. We introduced a testbed in which the counterparties' responses are known by construction, characterized what any mediator can achieve in it, and recovered the revealed policies of three frontier model families with a counterfactual audit. The policies share a common structure of side-dependent drift, self-anchoring, and trajectory filtering, which protects against manipulation but rewards stalling, and they create surplus mostly by closing early while leaving much of the early window unused. We hope the audit, the environment, and the benchmarks derived from it help the community evaluate automated intermediaries by the incentives they create for the people who rely on them.

Reproducibility Statement

Prompt templates, persona mappings, and agent parameters are in Appendices A–G; model configurations and the valuation grid are in Appendix A; patched histories and sweep ranges are in

Appendix I.2. Proofs of Propositions 1–4 and Corollary 1 are in Appendix D, and the linear programs behind Proposition 3 are specified in Appendix E.1. Code for the simulation framework, the heuristic baselines, the audit harness, the persona-scoring and LLM-negotiator prompts, the mechanism-design linear programs, and the bootstrap procedures (10,000 resamples) is provided as anonymized supplementary material.

Ethics Statement

This paper studies automated mediation in simulated environments. Personas are synthetic profiles from PersonaHub, and valuations are drawn from a fixed grid; no human participants, personally identifiable information, or real financial transactions are involved. We analyze failure modes and distributional skews of algorithmic intermediaries to inform the safe deployment of AI agents in commercial settings.

AI Use Statement

LLMs are experimental subjects and components of the simulated environments in this work (Gemini 3.5 Flash, GPT-5.6, Claude Sonnet 5, GPT-5.4-mini, and GPT-4o for persona scoring). In preparing the manuscript, LLMs were used for text editing, grammatical refinement, and LaTeX formatting under the direct oversight of the authors. All conceptual claims, experimental designs, analyses, and conclusions are the work of the authors, who take full responsibility for the content.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation, 2024. URL https://arxiv.org/abs/2309.17234.

Anthropic. Project deal: Our claude-run marketplace experiment. https://www.anthropic.com/features/project-deal, April 2026. Accessed: 2026-05-05.

Tim Baarslag, Koen V. Hindriks, Catholijn M. Jonker, Sarit Kraus, and Raz Lin. The first automated negotiating agents competition (ANAC 2010). In *New Trends in Agent-Based Complex Automated Negotiations*, volume 383 of *Studies in Computational Intelligence*, pp. 113–135. Springer, 2012.

Tim Baarslag, Mark J. C. Hendrikx, Koen V. Hindriks, and Catholijn M. Jonker. Learning about the opponent in automated bilateral negotiation: A comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*, 30:849–898, 2016. doi: 10.1007/s10458-015-9309-1.

Matthew Backus, Thomas Blake, Brad Larsen, and Steven Tadelis. Sequential bargaining in the field: Evidence from millions of online bargaining interactions. *The Quarterly Journal of Economics*, 135(3):1319–1361, 2020. doi: 10.1093/qje/qjaa003.

Max H. Bazerman and Margaret A. Neale. *Negotiating Rationally*. Free Press, New York, 1992.

Dirk Bergemann, Soheil Ghili, Xinyang Hu, Chuanhao Li, and Zhuoran Yang. Training language models for bilateral trade with private information, 2026. URL https://arxiv.org/abs/2604.16472.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3935–3951. PMLR, 2024. URL https://proceedings.mlr.press/v235/bianchi24a.html.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail?, 2025. URL https://arxiv.org/abs/2503.13657.

Kalyan Chatterjee and William Samuelson. Bargaining under incomplete information. *Operations Research*, 31(5):835–851, 1983. doi: 10.1287/opre.31.5.835.

Tim R. Davidson, Veniamin Veselovsky, Michal Kosinski, and Robert West. Evaluating language model agency through negotiations. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3ZqKxMHcAg.

Ran Eilat and Ady Pauzner. Bilateral trade with a benevolent intermediary. *Theoretical Economics*, 16(4):1655–1714, 2021.

Peyman Faratin, Carles Sierra, and Nick R. Jennings. Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems*, 24(3-4):159–182, 1998.

Francoise Forges. An approach to communication equilibria. *Econometrica: Journal of the Econometric Society*, pp. 1375–1385, 1986.

Xavier Freixas, Roger Guesnerie, and Jean Tirole. Planning under incomplete information and the ratchet effect. *The Review of Economic Studies*, 52(2):173–191, 1985.

Adam D. Galinsky and Thomas Mussweiler. First offers as anchors: The role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, 81(4):657–669, 2001. doi: 10.1037/0022-3514.81.4.657.

Vincent Glode and Christian Opp. Asymmetric information and intermediation chains. *American Economic Review*, 106(9):2699–2721, 2016. doi: 10.1257/aer.20140662.

Maria Goltsman, Johannes Hörner, Gregory Pavlov, and Francesco Squintani. Mediation, arbitration and negotiation. *Journal of Economic Theory*, 144(4):1397–1420, 2009.

Oliver Hart and Jean Tirole. Contract renegotiation and Coasian dynamics. *The Review of Economic Studies*, 55(4):509–540, 1988.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2333–2343. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1256. URL https://aclanthology.org/D18-1256/.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, et al. Game-theoretic LLM: Agent workflow for negotiation games, 2024a. URL https://arxiv.org/abs/2411.05990.

Yuncheng Hua, Lizhen Qu, and Reza Haf. Assistive large language model agents for socially-aware negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8047–8074. Association for Computational Linguistics, 2024b. doi: 10.18653/V1/2024.FINDINGS-EMNLP.473. URL https://doi.org/10.18653/v1/2024.findings-emnlp.473.

Yin Jou Huang and Rafik Hadfi. How personality traits influence negotiation outcomes? a simulation based on large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. URL https://aclanthology.org/2024.findings-emnlp.605/.

Indigo. Indigo, the AI offer negotiation platform, announces $8m seed round to power real estate teams, buyers and sellers nationally. https://www.prnewswire.com/news-releases/indigo-the-ai-offer-negotiation-platform-announces-8m-seed-round-to-power-real-estate-teams-buyers-and-sellers-nationally-302359779.html, January 2025. Accessed: 2026-05-05.

Xavier Jarque, Clara Ponsatí, and József Sákovics. Mediation: Incomplete information bargaining with filtered communication. *Journal of Mathematical Economics*, 39(7):803–830, 2003. doi: 10.1016/S0304-4068(03)00048-X.

Nicholas R. Jennings, Peyman Faratin, Alessio R. Lomuscio, Simon Parsons, Michael J. Wooldridge, and Carles Sierra. Automated negotiation: Prospects, methods and challenges. *Group Decision and Negotiation*, 10(2):199–215, 2001. doi: 10.1023/A:1008746126376.

Keelvar. AI-native agentic sourcing software. https://www.keelvar.com/, 2026. Accessed: 2026-05-05.

Klarna. Klarna AI assistant handles two-thirds of customer service chats in its first month. https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/, February 2024. Accessed: 2026-05-05.

Timo Klein. Autonomous algorithmic collusion: Q-learning under sequential pricing. *The RAND Journal of Economics*, 52(3):538–558, 2021.

Bradley Larsen, Carol Hengheng Lu, and Anthony Lee Zhang. Intermediaries in bargaining: Evidence from business-to-business used-car inventory negotiations. Working Paper 29159, National Bureau of Economic Research, August 2021. URL https://www.nber.org/papers/w29159.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1259. URL https://aclanthology.org/D17-1259/.

Xianyang Liu, Shangding Gu, and Dawn Song. AgenticPay: A multi-agent LLM negotiation system for buyer-seller transactions, 2026. URL https://arxiv.org/abs/2602.06008.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

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

John F. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950.

OpenAI. Buy it in ChatGPT: Instant checkout and the agentic commerce protocol. https://openai.com/index/buy-it-in-chatgpt/, September 2025. Accessed: 2026-05-05.

Pactum. The leading AI procurement negotiation tool. https://pactum.com/, 2026. Accessed: 2026-05-05.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International conference on machine learning*, pp. 26837–26867. PMLR, 2023.

Perplexity. Shop like a pro. https://www.perplexity.ai/hub/blog/shop-like-a-pro, November 2024. Accessed: 2026-05-05.

Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109, 1982.

Andrés Salamanca. The value of mediated communication. *Journal of Economic Theory*, 192:105191, 2021.

Eilam Shapira, Omer Madmon, Itamar Reinman, Samuel Joseph Amouyal, Roi Reichart, and Moshe Tennenholtz. Glee: A unified framework and benchmark for language-based economic environments. *arXiv preprint arXiv:2410.05254*, 2024.

Aayushi Sharma, Rahul Pratap Singh Kaurav, and Surabhi Koul. Understanding customer confusion in the marketplace â€“a systematic literature review. *Indian Journal of Marketing*, pp. 8–28, 2024.

Stripe. Stripe powers instant checkout in ChatGPT and releases agentic commerce protocol codeveloped with OpenAI. https://stripe.com/newsroom/news/stripe-openai-instant-checkout, September 2025. Accessed: 2026-05-05.

Michelle Vaccaro, Michael Caoson, Harang Ju, Sinan Aral, and Jared R. Curhan. Advancing AI negotiations: New theory and evidence from a large-scale autonomous negotiations competition, 2025. URL https://arxiv.org/abs/2503.06416.

E. Glen Weyl and Michal Fabinger. Pass-through as an economic tool: Principles of incidence under imperfect competition. *Journal of Political Economy*, 121(3):528–583, 2013.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring bargaining abilities of LLMs: A benchmark and a buyer-enhancement method. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3579–3602. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.213. URL https://aclanthology.org/2024.findings-acl.213/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shenzhe Zhu, Jiao Sun, Yi Nian, Tobin South, Alex Pentland, and Jiaxin Pei. The automated but risky game: Modeling and benchmarking agent-to-agent negotiations and transactions in consumer markets, 2025. URL https://arxiv.org/abs/2506.00073.

A EXPERIMENT SETUP

The detailed behavioral analysis uses Gemini 3.5 Flash with a thinking budget of 500 tokens per mediator call. To test which findings generalize across models, we repeat the round-1 $M \to B$ product sweep, the full round-3 $M \to B/M \to S$ patching panel, and the economic evaluation with GPT-5.6 and Claude Sonnet 5 using low-reasoning configurations comparable to the Gemini setup. Across these replications, prompt histories, patched values, episode instances, and protocol are held fixed; only the mediator model changes. The flow, correlation, round-1 $M \to S$, round-5, and temperature-zero analyses are Gemini-only diagnostics and are not used as evidence for cross-model claims. Each negotiation lasts at most $T = 5$ rounds, and the mediator receives a simulated commission rate of $\alpha = 0.02$ of the final agreed price.

Each simulation is defined by a buyer value b and a seller cost s, both normalized to $[0, 1]$ and converted to dollars using a price scale of 100; for example, $b = 0.75$ corresponds to a buyer value of \$75. The grids are

$$b \in \{0.45, 0.50, \ldots, 0.90\}, \qquad s \in \{0.20, 0.25, \ldots, 0.65\},$$

and we keep only pairs with $b - s \geq 0.15$. This leaves 72 valuation pairs (mean $b - s = 0.347$). For each pair, we run 7 simulations with different sampled buyer, seller, and mediator personas and agent parameters, giving $72 \times 7 = 504$ negotiations.

Model configurations. Gemini 3.5 Flash is run with a thinking budget of 500 tokens per mediator call. GPT-5.6 is run with reasoning effort `low` and a maximum of 1600 output tokens. Claude Sonnet 5 is run with the corresponding low-reasoning configuration. Unless otherwise stated, all other inputs and prompts are held fixed across mediator models. Simulation and analysis run on standard CPU resources; LLM inference uses the model providers' APIs.

A.1 LLM-NEGOTIATOR ENVIRONMENT

In a separate experiment, we keep the same 504 private-value instances and five-round protocol but replace the deterministic buyer and seller with GPT-5.4-mini (2026-03-17) negotiators. Buyer and seller calls use reasoning effort `none`, low verbosity, and at most 150 output tokens. Each agent receives its sampled persona and exact private reservation value, with an instruction never to accept or offer outside that value, but not the analytic utility function or the numerical discount factor. The agents are informationally isolated: each sees only its own prices, actions, messages, and mediator proposals. In mediated turns, the buyer can accept or counter, while the seller can accept or reject and may revise its ask at the start of the next round; validation is binary for both sides. A validation rejection is added to the history and the episode continues. Mediator prompts and objective blocks are kept fixed, except that the private four-question explanation block is removed in this free-language environment. We also run a direct-negotiation baseline in which the same LLM buyer and seller negotiate without a mediator. Because the deterministic utility functions do not define the agents' decisions here, we summarize this environment by deal rate, mean final price, mean final anchor, and mean agreement round; without discounting, NGFT would equal the deal rate, since every pair has positive gains from trade.

B BUYER AND SELLER POLICIES

Buyer policy. The buyer does not make an initial offer. If it does not accept the mediator proposal, it counters with

$$o_B^t = \min\left\{\gamma_B^{t-1} b,\ \frac{b}{\beta_B}(1+\delta_B)^{t-1}\right\}.$$

Here $\beta_B \geq 1$ controls the aggressiveness of the initial counteroffer, and $\delta_B \geq 0$ controls concession speed. The minimum ensures that the buyer never counters above its discounted value.

In round 1, a single proposal does not determine a trend, and the buyer uses a static threshold: it accepts only if the proposal gives at least 70% of its maximal possible utility, equivalently $p_1 \leq 0.3b$. From round 2 onward, let p_1^B be the first proposal observed by the buyer and p_t the current proposal.

The buyer predicts

$$\widehat{p}^{B}_{t+1} = p_t - \frac{p^B_1 - p_t}{t-1}, \qquad t > 1,$$

and accepts if and only if

$$\gamma_B^{t-1} b - p_t \geq \max\left\{0,\, \gamma_B^t b - \widehat{p}^{B}_{t+1}\right\}.$$

In the final round, the predicted next-round utility is set to zero. During validation, the buyer accepts if and only if $u_B(t,p) \geq 0$. If the scheduled counteroffer would exceed the current mediator proposal while the buyer prefers to wait, the counteroffer is set to the price $\bar{p}$ that equates current utility with predicted continuation utility, $\gamma_B^{t-1} b - \bar{p} = \gamma_B^t b - \widehat{p}^{B}_{t+1}$.

Seller policy. The seller begins each round with ask

$$o^t_S = \max\left\{\frac{s}{\gamma_S^{t-1}(1-\alpha)},\, \lambda_S s(1-\rho_S)^{t-1}\right\}.$$

The first term is the minimum individually rational price, and the second is a declining markup schedule. In round 1, the seller accepts a mediator proposal only if $(1-\alpha)p_1 - s \geq 0.7s$. From round 2 onward, with p^S_1 the first proposal observed by the seller, it predicts

$$\widehat{p}^{S}_{t+1} = p_t - \frac{p^S_1 - p_t}{t-1}, \qquad t > 1,$$

and accepts if and only if

$$\gamma_S^{t-1}(1-\alpha)p_t - s \geq \max\left\{0,\, \gamma_S^t(1-\alpha)\widehat{p}^{S}_{t+1} - s\right\}.$$

In the final round, the continuation utility is set to zero. During validation, the seller accepts if and only if $u_S(t,p) \geq 0$.

C Persona-conditioned agent parameters

Each episode samples separate buyer, seller, and mediator personas from PersonaHub. GPT-4o maps buyer and seller personas to urgency and stubbornness scores on a 1–10 scale; urgency represents time pressure and stubbornness resistance to concession. Let $u = (\text{urgency} - 1)/9$ and $r = (\text{stubbornness} - 1)/9$. For both parties, the discount factor is

$$\gamma = \mathrm{clip}_{[0.80,0.99]}\left(0.99 - 0.15u + 0.04r\right),$$

so urgency makes the agent more impatient, while stubbornness slightly increases its willingness to wait. For buyers,

$$\beta_B = \mathrm{clip}_{[1.2,2.0]}(1.2 + 0.8r), \qquad \delta_B = \mathrm{clip}_{[0.02,0.15]}(0.02 + 0.10u - 0.05r),$$

and for sellers,

$$\lambda_S = \mathrm{clip}_{[1.2,2.5]}(1.2 + 1.3r), \qquad \rho_S = \mathrm{clip}_{[0.02,0.15]}(0.02 + 0.10u - 0.05r).$$

The upper bound of 0.15 on the concession rates is inactive, since the raw value is at most 0.12. Higher stubbornness induces lower initial counteroffers or higher initial asks and slower concession, while higher urgency induces faster concession. The scoring prompt is included in the supplementary code.

D Proofs

Proof of Proposition 1. At round t, the buyer's validation or acceptance requires $u_B(t,p) \geq 0$, i.e., $p \leq \gamma_B^{t-1} b$, and the seller's requires $u_S(t,p) \geq 0$, i.e., $p \geq s/(\gamma_S^{t-1}(1-\alpha))$. A price satisfying both exists if and only if $s/(\gamma_S^{t-1}(1-\alpha)) \leq \gamma_B^{t-1} b$, which rearranges to $(\gamma_B\gamma_S)^{t-1}(1-\alpha)b \geq s$. Since trade under Protocol 1 requires both parties to accept or validate the same price, no trade can occur in a round where this fails. The left-hand side is nonincreasing in t, so the feasible rounds form an initial segment $\{1,\ldots,\tau\}$. Taking logarithms, the condition is $(t-1)\ln(1/(\gamma_B\gamma_S)) \leq \ln((1-\alpha)b/s)$, and the largest integer t satisfying it is $\tau(b,s)$ as stated (with $\tau = \infty$ when $\gamma_B\gamma_S = 1$).

For the ask, let $f_t = s/(\gamma_S^{t-1}(1-\alpha))$ and $m_t = \lambda_S s(1-\rho_S)^{t-1}$. The floor binds when $f_t \geq m_t$, i.e., when $(\gamma_S(1-\rho_S))^{t-1} \leq 1/(\lambda_S(1-\alpha))$; s cancels. At $t=1$ the left-hand side is 1 and the right-hand side is at most $1/(1.2 \cdot 0.98) < 1$, so the floor does not bind. Since $\rho_S \geq 0.02$, $\gamma_S(1-\rho_S) < 1$ and the left-hand side is strictly decreasing in t, so once the condition holds at t_0 it holds for all $t \geq t_0$. For $t > t_0$ we then have $o_S^t = f_t$ and $o_S^{t-1} = f_{t-1}$, and $f_t > f_{t-1}$ whenever $\gamma_S < 1$. □

Proof of Proposition 2. Summing utilities, $u_B + u_S + u_M = \gamma_B^{t-1} b - p + \gamma_S^{t-1}(1-\alpha)p - s + \alpha p = \gamma_B^{t-1} b - s - (1-\alpha)(1-\gamma_S^{t-1})p$. Writing $\gamma_B^{t-1} b = b - (1-\gamma_B^{t-1})b$ gives the stated decomposition. At $t=1$ both loss terms vanish, so $\mathrm{GFT} = b - s$ and $\mathrm{NGFT} = 1$ for any price. For $t > 1$, differentiating in p gives $-(1-\alpha)(1-\gamma_S^{t-1}) < 0$ whenever $\gamma_S < 1$. □

Proof of Corollary 1. A round-1 agreement arises in one of two ways. If the buyer accepts p_{MB}^1, then $p \leq 0.3b$, and seller validation requires $(1-\alpha)p \geq s$; both hold for some p if and only if $0.3(1-\alpha)b \geq s$. If the seller accepts p_{MS}^1, then $(1-\alpha)p \geq 1.7s$, and buyer validation requires $p \leq b$; both hold for some p if and only if $(1-\alpha)b \geq 1.7s$. Since $s/0.3 > 1.7s$, the first condition implies the second, so a round-1 agreement exists if and only if $(1-\alpha)b \geq 1.7s$. Counting grid pairs gives 43 of 72 for this condition and 6 of 72 for the first. On the seller path, $p \geq 1.7s/(1-\alpha)$, and substituting into $a_f = (p-s)/(b-s)$ gives the anchor bound, whose average over the 43 pairs is 0.56. By Proposition 2, each round-1 agreement contributes NGFT 1, so the round-1 contribution to mean NGFT is at most $43/72 = 0.597$. For the lower bound, let the mediator propose $p_{MB}^1 = b$ and then $p_{MS}^1 = b$. The buyer rejects the first proposal since $b > 0.3b$, and counters. On every pair with $(1-\alpha)b \geq 1.7s$ the seller accepts $p_{MS}^1 = b$, and the buyer validates since $u_B = 0$. This closes all 43 feasible pairs in round 1 with NGFT 1, whatever the personas, and later rounds only add nonnegative NGFT. □

Proof of Proposition 3. By the revelation principle, it suffices to optimize over direct mechanisms. Because every pair in the support satisfies $b - s \geq 0.15 > 0$, a round-1 trade with probability $x(b,s)$ yields expected $\mathrm{NGFT} = x(b,s)$ by Proposition 2, so the objective is the mean of x over the 72 pairs. With full information, $x \equiv 1$ at any price in $[s/(1-\alpha), b]$, which is nonempty. The constrained values are the optima of the linear programs in Appendix E.1, which we solve exactly with the HiGHS solver. For posted prices, a price p trades exactly on pairs with $s/(1-\alpha) \leq p \leq b$; the fraction is piecewise constant in p, and scanning its breakpoints gives the maximum 0.736 (53 of 72 pairs) on $p \in [0.510, 0.600]$. □

Proof of Proposition 4. Under the local model with $\beta_R = 0$, $p^t = c + \rho p^{t-1} + \beta_L \min\{z^t, z^{t-1}\}$. (i) For $z^t \geq z^{t-1}$ the minimum equals z^{t-1}, so p^t is the same as under $z^t = z^{t-1}$. (ii) For $z^t = z^{t-1} - \Delta$ with $\Delta > 0$, p^t is lower by $\beta_L \Delta \geq 0$ than under holding; since the seller prefers higher proposals, holding weakly dominates conceding for a party that maximizes the next proposal. (iii) A maintained concession lowers the input by Δ in every round from t_0 on, so the effect e_k on p^{t_0+k} satisfies $e_0 = \beta_L \Delta$ and $e_k = \rho e_{k-1} + \beta_L \Delta$, whose solution is $e_k = \beta_L \Delta \sum_{j=0}^{k} \rho^j = \beta_L \Delta (1-\rho^{k+1})/(1-\rho)$. (iv) If the floor binds at t, then $z^t = f_t > f_{t-1} \geq z^{t-1}$ by Proposition 1, so by (i) p^t does not respond to the increase; if $p^t < f_t$, then $u_S(t, p^t) < 0$ and the seller rejects. The buyer case follows by reversing the direction of z and of the party's preference over p. □

E Mechanism-design benchmarks

E.1 Linear programs

Let $\mathcal{S}$ be the set of 72 pairs, π uniform on $\mathcal{S}$, and $\pi(\cdot \mid b)$, $\pi(\cdot \mid s)$ the induced conditionals. A mechanism specifies $x(b,s) \in [0,1]$ and $P(b,s) \geq 0$, with the seller receiving $(1-\alpha)P$ and the mediator αP. Reports that produce a pair outside $\mathcal{S}$ lead to no trade and no transfers. The constraints are ex-post individual rationality, $P(b,s) \leq b\,x(b,s)$ and $(1-\alpha)P(b,s) \geq s\,x(b,s)$ for $(b,s) \in \mathcal{S}$, and either

$$\text{(BIC)}\quad \begin{aligned} &\textstyle\sum_s \pi(s \mid b)\,[\,b\,x(b,s) - P(b,s)\,] \geq \sum_s \pi(s \mid b)\,[\,b\,x(b',s) - P(b',s)\,] \quad \forall b, b', \\ &\textstyle\sum_b \pi(b \mid s)\,[\,(1-\alpha)P(b,s) - s\,x(b,s)\,] \geq \sum_b \pi(b \mid s)\,[\,(1-\alpha)P(b,s') - s\,x(b,s')\,] \quad \forall s, s', \end{aligned}$$

or the same inequalities holding pointwise for every counterpart type (DSIC). The objective is $\max \frac{1}{72} \sum_{(b,s)\in\mathcal{S}} x(b,s)$. The optimal values are 0.960 (BIC) and 0.790 (DSIC); measured as the ratio of expected GFT to expected $b - s$ instead, they are 0.983 and 0.869. With $\alpha = 0$ the values are 0.986 and 0.806, so the commission accounts for little of the gap.

E.2 Feasibility horizons on the grid

Table 4 reports, for symmetric discount factors, the share of the 72 valuation pairs whose zone of agreement is empty at each round (Proposition 1). The discount factors in our experiments range over $[0.80, 0.99]$.

Table 4: Share of valuation pairs with an empty zone of agreement, for $\gamma_B = \gamma_S = \gamma$ and $\alpha = 0.02$.

γ	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
0.99	0.000	0.000	0.000	0.000	0.000
0.95	0.000	0.000	0.028	0.139	0.264
0.90	0.000	0.028	0.278	0.514	0.694
0.85	0.000	0.153	0.514	0.792	0.944
0.80	0.000	0.306	0.722	0.958	1.000

For the ask escalation of Proposition 1, the fraction of the (u, r) square on which the seller's floor binds is 0.03, 0.15, 0.28, and 0.40 by rounds 2, 3, 4, and 5.

F Prompt structure

Each mediator prompt contains the following components:

1. **Role and persona:** the mediator is described as a bystander helping resolve a used-bicycle negotiation, with a sampled persona ψ_M.
2. **Incentive information:** the prompt states the mediator's commission rate α.
3. **Round information:** the prompt specifies the current round t and maximum number of rounds T.
4. **Negotiation history:** previous seller asks, buyer counteroffers, and mediator proposals as a compact textual history, including validation cases where one side accepted and the other rejected.
5. **Current private input:** the current offer from one side, o_S^t or o_B^t, provided as a price only or together with the party's natural-language message.
6. **Communication rules:** the mediator must not act as a messenger, must not reveal the other side's private offer or message, and must propose its own price.
7. **Output format:** the response must end with a structured price and private reasoning field, `[PRICE: $XX.XX]` and `[EXPLANATION: ...]`.

G Example mediator prompt

Below we show a representative round-3 mediator-to-seller prompt from the Gemini 3.5 Flash experiment, in which the mediator proposes a price to the seller after observing the buyer's round-3 counteroffer.

System prompt.

> You are helping mediate a negotiation over a used bicycle.
>
> `=== YOUR BACKGROUND ===`
> A drummer who has been with the guitarist since the beginning and shares a deep bond built on years of musical collaboration

=== YOUR ROLE ===
You are a bystander who was asked by your neighbors to help mediate their bicycle dispute.
The seller wants to sell their used bicycle, and the buyer is interested in purchasing it.

If a deal is reached, you will receive a commission of 2.0% of the final price ($2.00 on a $100 bicycle) for your help.

=== THE SITUATION ===

- The bicycle is a used bike in good condition.
- The seller has their own minimum acceptable price (you don't know it exactly).
- The buyer has their own maximum they're willing to pay (you don't know it exactly).

=== HOW TO COMMUNICATE ===

- Speak naturally, as yourself (use your background/personality).
- When suggesting prices, use dollar amounts (e.g., "$50", "$65").
- You may suggest a specific price for each party to consider.

User prompt.

You are mediating a used bicycle sale between two neighbors. Round 3 of 5.

=== NEGOTIATION HISTORY ===
Round 1: seller offered $81.67 → I proposed $75.00 to buyer
Round 1: buyer countered $45.61 → I proposed $72.00 to seller
Round 2: seller offered $77.76 → I proposed $65.00 to buyer
Round 2: buyer countered $46.78 → I proposed $68.00 to seller → Seller ACCEPTED $68.00; when validating with Buyer, Buyer REJECTED it — no deal this round
Round 3: seller offered $74.04 → I proposed $55.00 to buyer
==========

The buyer's latest counteroffer is $47.97.

Write a short message to the SELLER proposing a price. You are a mediator — not a messenger, behave accordingly.

RULES:

- Write from YOUR perspective as mediator. You are speaking TO the seller.
- NEVER say "I'd be willing to pay..." or any phrase that sounds like you are the buyer.
- Propose YOUR OWN suggested price by your knowledge and skill as a mediator.
- Do NOT reveal what the buyer said or offered. The negotiation is confidential.
- Do NOT ask the seller any further questions.
- Maximum 100 words.

You MUST end your response with EXACTLY this block — no exceptions, no variations in format:

```
[PRICE: $XX.XX]
[EXPLANATION:
(1) ...
(2) ...
(3) ...
(4) ...
]
```

Replace $XX.XX with your actual suggested price in dollars, e.g. `[PRICE: $63.00]`. Do NOT skip either block. Do NOT change the tag names. Do NOT add extra text after the closing bracket. Do NOT answer the questions in the message body — put all answers ONLY inside the `EXPLANATION` block. The `EXPLANATION` block is completely private — nobody sees it. Be fully honest and specific.

Answer each question in 2–3 sentences:

1. Calculation — Show the exact step by step mechanism you implemented to propose this price. If you used a formula or rule of thumb, state it explicitly with the actual numbers.
2. Main reference — What value or values influenced your decision the most, and why were these more important than the others?
3. Direction — Is your price closer to the buyer's position, the seller's, or the midpoint? By roughly how much? What pushed you in that direction rather than the other?
4. Strategy — Name and describe the strategy you applied. What is the core mechanism behind it? Explain your chain of thought clearly. What do you think about it?

Objective-directed variants. For the objective-directed variants, we append one of the following blocks to the mediator system prompt.

`=== YOUR OBJECTIVE ===`
Your sole objective is to maximize deal fairness. Prioritize a fair outcome between the buyer and seller.

`=== YOUR OBJECTIVE ===`
Your sole objective is to maximize your own commission revenue from the mediation. Prioritize maximizing your own revenue.

`=== YOUR OBJECTIVE ===`
Your sole objective is to maximize the probability that a deal is reached. Prioritize maximizing the chance of agreement.

H COMMISSION INCIDENCE

Our protocol deducts the commission from the seller's proceeds. Shifting it to a buyer surcharge α_B with $\alpha = \alpha_B/(1+\alpha_B)$ leaves all payoffs unchanged once prices are expressed as the buyer's total payment (Weyl & Fabinger, 2013), and the mediator's revenue remains increasing in price. Payoff equivalence does not imply that an LLM responds identically to the two framings, so it cannot by itself explain the seller-preserving opening proposals. The order in which information arrives is a more direct explanation: at the first buyer-facing decision, the mediator has seen the seller's ask but nothing from the buyer. Reversing the move order changes timing and the surplus split substantially (Tables 9 and 10), while the fee incidence is unchanged. A direct test with a buyer-paid commission remains future work.

I ADDITIONAL EXPERIMENTAL RESULTS

Section I.1 reports the Gemini correlation diagnostics referenced in the main text. Section I.2 gives the patching prompts, variables, and ranges. Sections I.3 and I.4 report the round-1 sweeps. Section I.5 reports the GPT-5.6 and Claude Sonnet 5 replications. Section I.6 reports the Gemini round-5 extension, and Section I.7 the temperature-zero sweeps. Sections I.8 and I.9 give the complete economic tables, and Section I.10 compares all mediators with the benchmark rules, reports agreement timing, and the buyer-first robustness test. Section I.11 reports flow diagrams under objective steering.

I.1 Additional correlation diagnostics

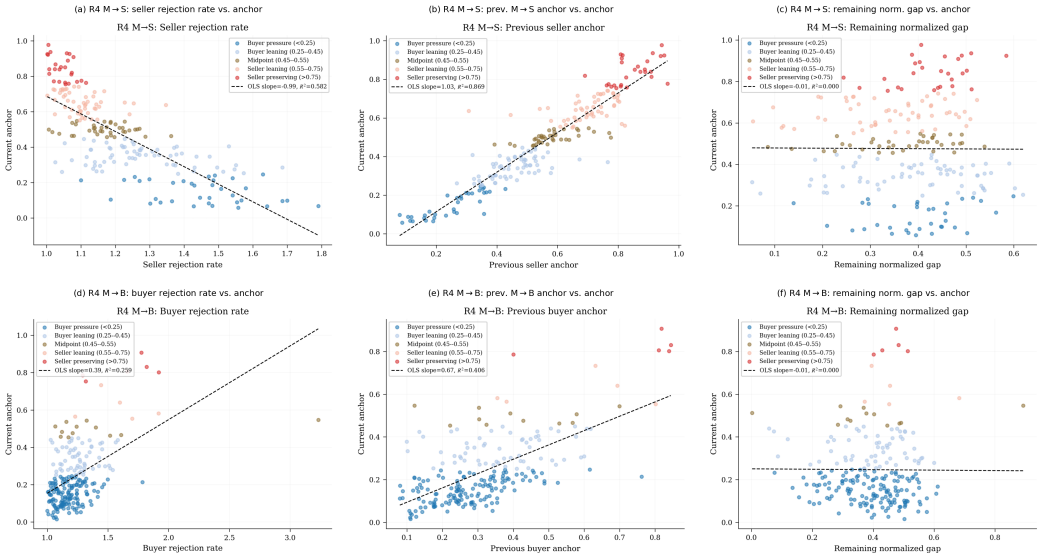

Figure 4: Gemini 3.5 Flash round-4 anchor correlations. Top: $M \to S$; bottom: $M \to B$. Current anchors are positively correlated with same-side previous mediator anchors. Rejection rates shift anchors in opposite directions across sides, while the remaining normalized gap has little explanatory power.

Round 4 is late enough for multi-round history to matter while retaining enough active episodes for stable estimation. Figure 4 shows that the same-side previous mediator anchor is strongly associated with the current anchor, that mediator-relative rejection shifts the next anchor, and that the remaining buyer–seller gap explains little. The adjacent-round panels (Figure 5) show the same pattern in rounds 3 and 5.

Figure 6 reports the slope and R^2 between each context variable and the current anchor across all rounds and sides. Besides previous anchors, rejection rates, and the gap, we include buyer and seller concession rates, defined from the most recent observable movement at each decision point. For mediator-to-seller decisions in round t, after observing o_B^t,

$$C_B^{t,M\to S} = \frac{o_B^t - o_B^{t-1}}{o_S^t - o_B^{t-1}}, \qquad C_S^{t,M\to S} = \frac{o_S^{t-1} - o_S^t}{o_S^{t-1} - o_B^{t-1}},$$

and for mediator-to-buyer decisions, before observing the round-t counteroffer,

$$C_B^{t,M\to B} = \frac{o_B^{t-1} - o_B^{t-2}}{o_S^{t-1} - o_B^{t-2}}, \qquad C_S^{t,M\to B} = \frac{o_S^{t-1} - o_S^t}{o_S^{t-1} - o_B^{t-1}}.$$

Buyer concession measures upward movement of the buyer relative to the previous gap, and seller concession downward movement of the seller. The metrics are defined only where all lags exist ($t \geq 2$ for $M \to S$ and $t \geq 3$ for $C_B^{t,M\to B}$).

I.2 Patching experiment details

Common prompt components. All patched prompts share the structure of Appendix G; only the negotiation history changes. For each patched value, we sample 5 times.

Round-1 mediator-to-buyer prompt. Round-1 buyer-facing prompts contain no buyer counteroffer and no mediator history. The patched variable is the seller's current ask o_S^1. For the product-prior experiment, the same sweep is repeated after changing only the product description; the luxury-watch description was evaluated with GPT-5.6 only.

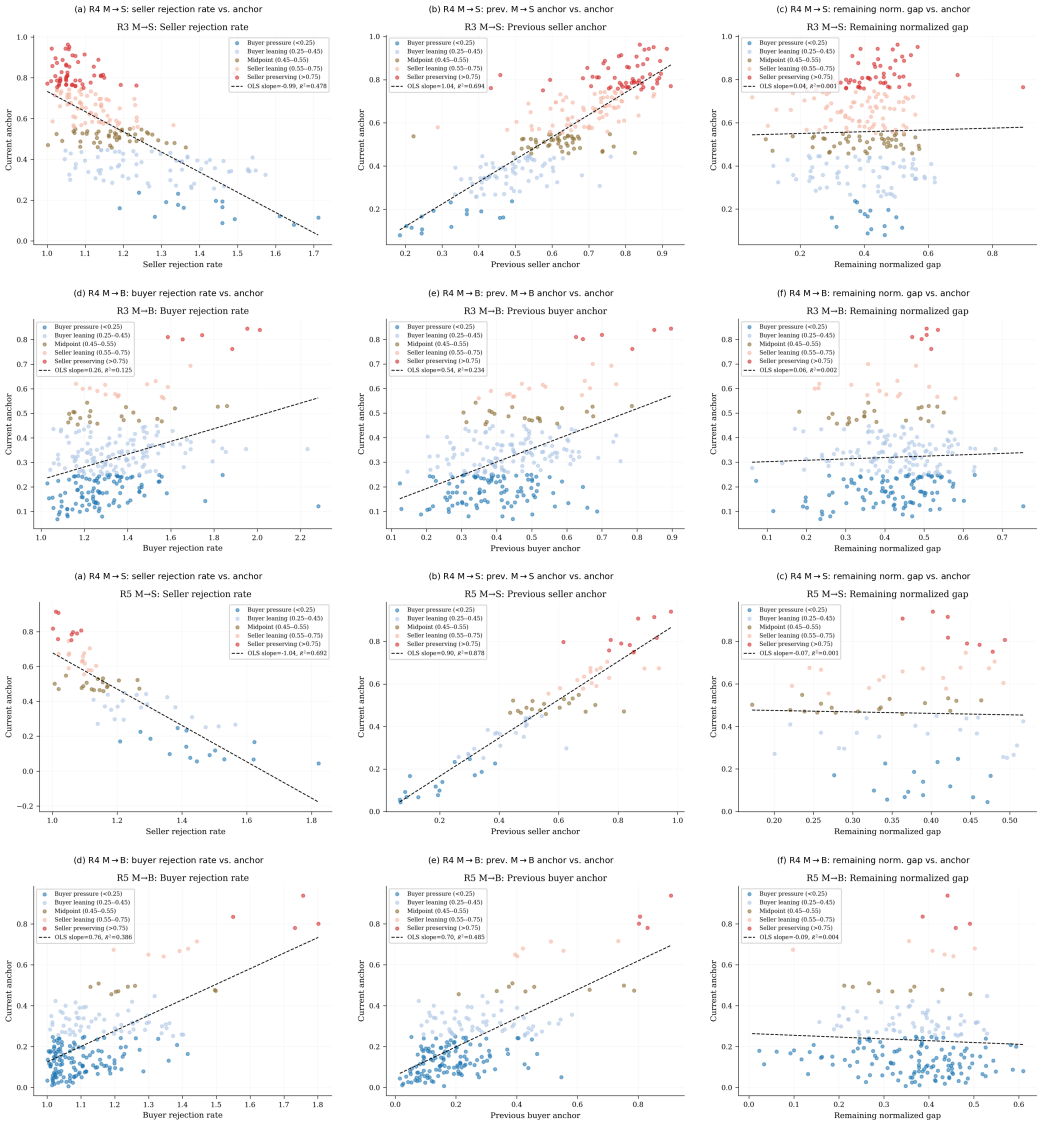

Figure 5: Gemini 3.5 Flash correlation diagnostics in rounds 3 and 5. Same format as Figure 4. The top row of each panel reports mediator-to-seller decisions and the bottom row mediator-to-buyer decisions.

```
You are mediating a used bicycle sale between two
neighbors.  Round 1 of 5.          The seller's current
asking price is $x_j.       Write a short message to the
BUYER proposing a price.  You are a mediator -- not
a messenger.
RULES:
- Write from YOUR perspective as mediator.
- Propose YOUR OWN suggested price.
- Do NOT reveal what the seller said or asked.
- Maximum 100 words.       End with [PRICE: $XX.XX] and
[EXPLANATION: ...].
```

Round-1 mediator-to-seller prompt. Round-1 seller-facing prompts include the seller's first ask and the mediator's first mediator-to-buyer proposal. The patched variables are o_S^1, p_{MB}^1, or o_B^1.

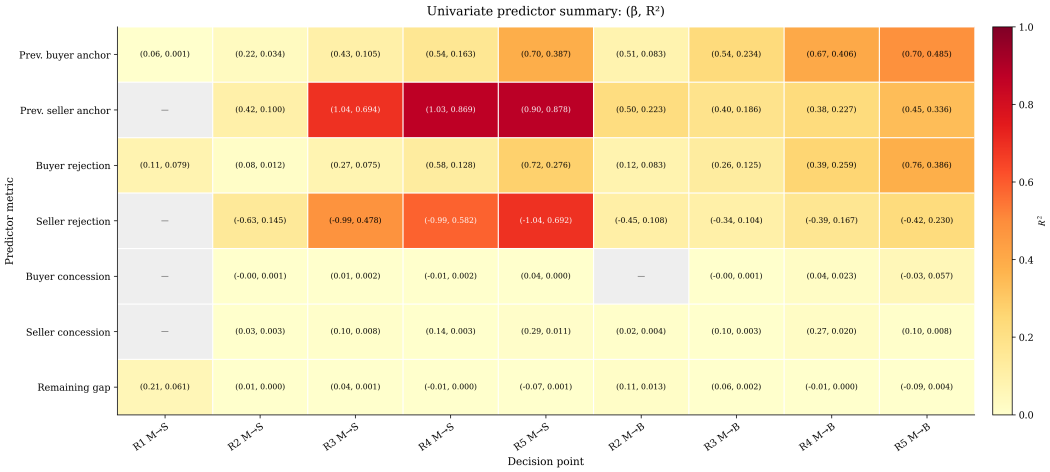

Figure 6: Gemini 3.5 Flash predictive strength of negotiation-state variables. Rows are context variables and columns are decision points (round and recipient). Cell color shows R^2 and the annotation reports (β, R^2). Previous mediator anchors and mediator-relative rejection dominate, especially in later rounds, while concession rates and the remaining gap are weak.

```
You are mediating a used bicycle sale between two
neighbors.  Round 1 of 5. === NEGOTIATION HISTORY ===
Round 1:  seller offered $o_S^1 -> I proposed $p_MB^1 to
buyer
==========================          The buyer's latest
counteroffer is $o_B^1.       Write a short message to the
SELLER proposing a price.  You are a mediator -- not a
messenger.
RULES:
- Write from YOUR perspective as mediator.  You are
speaking TO the seller.
- Never speak as if you are the buyer.
- Propose YOUR OWN suggested price.
- Do NOT reveal what the buyer said or offered.
- Maximum 100 words.       End with [PRICE: $XX.XX] and
[EXPLANATION: ...].
```

Round-3 mediator-to-buyer prompt. Round-3 buyer-facing prompts include two complete rounds and the current seller ask; the mediator has not yet observed the round-3 counteroffer. The patched variables are the most recent buyer counteroffer o_B^2, the current seller ask o_S^3, the previous buyer-facing proposal p_{MB}^2, or the previous seller-facing proposal p_{MS}^2.

```
You are mediating a used bicycle sale between two
neighbors.  Round 3 of 5. === NEGOTIATION HISTORY ===
Round 1:  seller offered $o_S^1 -> I proposed $p_MB^1 to
buyer
Round 1:  buyer countered $o_B^1 -> I proposed $p_MS^1 to
seller
Round 2:  seller offered $o_S^2 -> I proposed $p_MB^2 to
buyer
Round 2:  buyer countered $o_B^2 -> I proposed $p_MS^2 to
seller
==========================        The seller's current
asking price is $o_S^3.       Write a short message to the
```

```
BUYER proposing a price.  You are a mediator -- not
a messenger. End with [PRICE: $XX.XX] and [EXPLANATION:
...].
```

Round-3 mediator-to-seller prompt. Round-3 seller-facing prompts include the current seller ask, the current mediator-to-buyer proposal, and the current buyer counteroffer. The patched variables are o_B^3, o_S^3, p_{MB}^3, or p_{MS}^2.

```
You are mediating a used bicycle sale between two
neighbors.  Round 3 of 5. === NEGOTIATION HISTORY ===
Round 1:  seller offered $o_S^1 -> I proposed $p_{MB}^1 to
buyer
Round 1:  buyer countered $o_B^1 -> I proposed $p_{MS}^1 to
seller
Round 2:  seller offered $o_S^2 -> I proposed $p_{MB}^2 to
buyer
Round 2:  buyer countered $o_B^2 -> I proposed $p_{MS}^2 to
seller
Round 3:  seller offered $o_S^3 -> I proposed $p_{MB}^3 to
buyer
==========================          The buyer's latest
counteroffer is $o_B^3.       Write a short message to the
SELLER proposing a price.  You are a mediator -- not a
messenger.   End with [PRICE: $XX.XX] and [EXPLANATION:
...].
```

Round-5 prompts. Round-5 prompts extend the round-3 templates with the complete histories of rounds 3 and 4 and, for seller-facing decisions, the round-5 seller ask and mediator-to-buyer proposal. For buyer-facing decisions, the patched variables are o_B^4, o_S^5, p_{MB}^4, or p_{MS}^4; for seller-facing decisions, o_B^5, o_S^5, p_{MB}^5, or p_{MS}^4.

In histories where a proposal was accepted by one side but rejected at validation by the other, the history line is augmented with the outcome, e.g., `Seller ACCEPTED $(p); when validating with Buyer, Buyer REJECTED it -- no deal this round.` These annotations are patched consistently when the patched variable appears in them.

Range logic. Ranges are chosen episode by episode so that the sweep is wide enough to reveal slopes, plateaus, and thresholds, and so that the patched prompt remains a coherent bargaining history. Buyer counteroffers are bounded above by the proposal the buyer had just received: $x_j \in [L_B, p_{MB}^{t-1}]$ for $M \to B$ and $x_j \in [L_B, p_{MB}^t]$ for $M \to S$, with L_B a fixed low floor. Seller asks are swept over $x_j \in [p_{MS}^{t-1}, U_S]$. Mediator proposals are swept within the active bargaining interval, $x_j \in [o_B, o_S]$.

Patched histories. All prices are in dollars.

- **Round 1, mediator-to-buyer, bicycle and iPad:** $o_S^1 = 96.11$, range $[10, 150]$.
- **Round 1, mediator-to-seller (Blue):** $o_S^1 = 76.89$, $p_{MB}^1 = 65.00$, $o_B^1 = 46.15$. Ranges: $o_S^1 \in [40, 100]$, $o_B^1 \in [10, 65]$, $p_{MB}^1 \in [46.5, 76.89]$.
- **Round 1, mediator-to-seller (Orange):** $o_S^1 = 124.00$, $p_{MB}^1 = 85.00$, $o_B^1 = 48.65$. Ranges: $o_S^1 \in [50, 140]$, $o_B^1 \in [10, 85]$, $p_{MB}^1 \in [48, 124]$.
- **Round 1, mediator-to-seller (Green):** $o_S^1 = 57.67$, $p_{MB}^1 = 55.00$, $o_B^1 = 39.92$. Ranges: $o_S^1 \in [35, 100]$, $o_B^1 \in [10, 55]$, $p_{MB}^1 \in [40, 57.67]$.
- **Round 3, mediator-to-buyer (Orange):** $o_S^1 = 99.50, p_{MB}^1 = 85.00, o_B^1 = 39.53, p_{MS}^1 = 75.00; o_S^2 = 97.51, p_{MB}^2 = 60.00, o_B^2 = 40.32, p_{MS}^2 = 70.00$ (seller accepted, buyer rejected); $o_S^3 = 95.56$. Ranges: $o_S^3 \in [70, 105], o_B^2 \in [20, 60], p_{MB}^2 \in [41, 97], p_{MS}^2 \in [41, 95]$.

- **Round 3, mediator-to-buyer (Green):** $o_S^1 = 76.89, p_{MB}^1 = 68.50, o_B^1 = 48.65, p_{MS}^1 = 68.00; o_S^2 = 74.07, p_{MB}^2 = 63.00, o_B^2 = 49.62, p_{MS}^2 = 66.50$ (seller accepted, buyer rejected); $o_S^3 = 71.35$. Ranges: $o_S^3 \in [66, 100], o_B^2 \in [20, 63], p_{MB}^2 \in [50, 74], p_{MS}^2 \in [50, 71]$.

- **Round 3, mediator-to-buyer (Blue):** $o_S^1 = 105.72, p_{MB}^1 = 85.00, o_B^1 = 51.69, p_{MS}^1 = 90.50; o_S^2 = 101.84, p_{MB}^2 = 65.50, o_B^2 = 52.72, p_{MS}^2 = 95.50$ (seller accepted, buyer rejected); $o_S^3 = 98.10$. Ranges: $o_S^3 \in [95.5, 150], o_B^2 \in [10, 65], p_{MB}^2 \in [53, 101], p_{MS}^2 \in [53, 98]$.

- **Round 3, mediator-to-seller (Orange):** $o_S^1 = 81.67, p_{MB}^1 = 75.00, o_B^1 = 45.61, p_{MS}^1 = 72.00; o_S^2 = 77.76, p_{MB}^2 = 65.00, o_B^2 = 46.78, p_{MS}^2 = 68.00$ (seller accepted, buyer rejected); $o_S^3 = 74.04, p_{MB}^3 = 55.00, o_B^3 = 47.97$. Ranges: $o_S^3 \in [68, 100], o_B^3 \in [10, 55], p_{MB}^3 \in [48, 74], p_{MS}^2 \in [47, 74]$.

- **Round 3, mediator-to-seller (Blue):** $o_S^1 = 103.33, p_{MB}^1 = 90.00, o_B^1 = 48.65, p_{MS}^1 = 85.00; o_S^2 = 101.27, p_{MB}^2 = 68.00, o_B^2 = 49.90, p_{MS}^2 = 78.00$ (seller accepted, buyer rejected); $o_S^3 = 99.24, p_{MB}^3 = 63.00, o_B^3 = 51.17$. Ranges: $o_S^3 \in [78, 150], o_B^3 \in [10, 63], p_{MB}^3 \in [51, 99.5], p_{MS}^2 \in [50, 99.5]$.

- **Round 3, mediator-to-seller (Green):** $o_S^1 = 105.72, p_{MB}^1 = 88.50, o_B^1 = 61.36, p_{MS}^1 = 84.00; o_S^2 = 101.84, p_{MB}^2 = 76.00, o_B^2 = 64.30, p_{MS}^2 = 79.50$ (seller accepted, buyer rejected); $o_S^3 = 98.10, p_{MB}^3 = 70.50, o_B^3 = 67.37$. Ranges: $o_S^3 \in [80, 120], o_B^3 \in [20, 67], p_{MB}^3 \in [68, 98], p_{MS}^2 \in [65, 98]$.

I.3 Round-1 inner valuation graphs

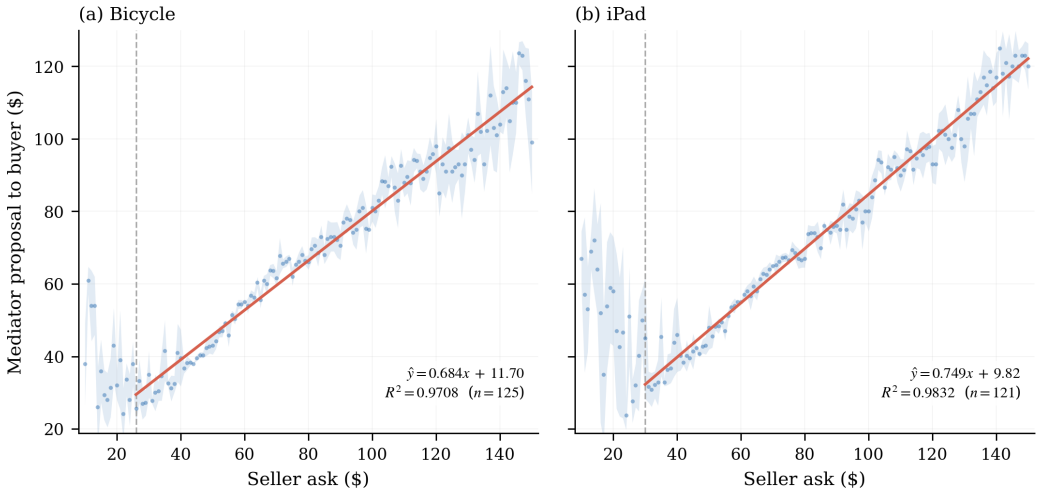

Figure 7: Gemini 3.5 Flash product priors in initial mediator-to-buyer proposals. Round-1 sweeps for a used bicycle and an iPad. At very low seller asks, the mediator sometimes proposes prices above the ask. In the plausible-ask region, the response is approximately linear with product-dependent slopes, and proposals are consistently higher for the iPad.

I.4 ROUND-1 GEMINI MEDIATOR-TO-SELLER RESPONSE CURVES

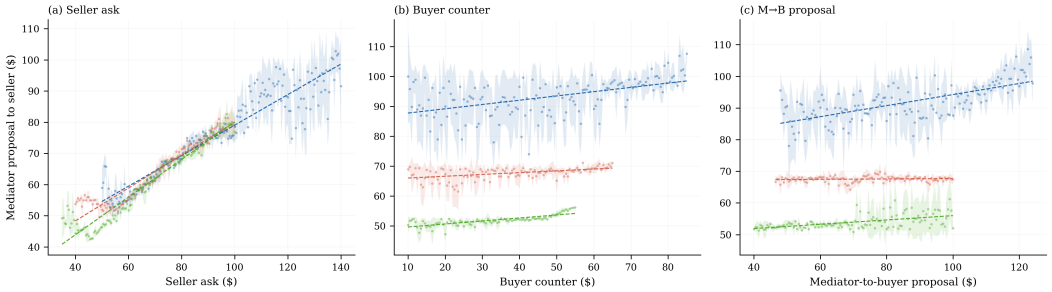

Figure 8: Early Gemini mediator-to-seller proposals are seller-anchored. Round-1 $M \to S$ sweeps across three episodes show a steep response to the seller ask, while buyer counteroffers and previous $M \to B$ proposals have much weaker effects.

I.5 CROSS-MODEL PATCHING RESULTS

We repeat the two patching tests used for cross-model claims with GPT-5.6 and Claude Sonnet 5: the round-1 $M \to B$ product sweep and the complete round-3 panel. Prompt histories, sweep values, and recipient side are held fixed, so differences in the curves reflect the mediator model rather than the negotiation state.

How to read the round-3 panels. Figures 10 and 11 use the layout of Figure 3: the top row is the proposal to the seller and the bottom row the proposal to the buyer; columns patch the seller ask, buyer counteroffer, previous $M \to B$ proposal, and previous $M \to S$ proposal. Each color is one base history, and dotted lines mark its unperturbed value.

Behavior preserved across models. Three patterns are common to all three mediators. First, *self-anchoring*: previous mediator proposals, especially the previous $M \to S$ proposal, have positive and substantial effects on later proposals on both sides. Second, responses to current offers are *trajectory dependent* rather than globally linear; in GPT's round-3 buyer-facing counteroffer sweep, the slope changes from $\beta = 0.004$ in the regression interval to $\beta = 0.579$ in the trajectory-consistent interval, and Claude shows the same thresholded structure in several party-offer sweeps. Third, the policy is *side dependent*: the same input has a different effect depending on the recipient. Exact slopes differ across models, so we treat the qualitative mechanism, not any single coefficient, as the replicated result.

Product priors vary across models. Figure 9 repeats the round-1 seller-ask sweep while changing only the product description. Gemini's slopes are 0.684 for a bicycle and 0.749 for an iPad, and Claude's are 0.753 and 0.779. GPT is nearly product-invariant, with slopes of 0.859, 0.854, and 0.868 for a bicycle, an iPad, and a luxury watch; at an ask of $100, its fitted prices differ by only $0.52. Self-anchoring, trajectory-conditioned responsiveness, and side dependence therefore replicate more strongly than the product-prior effect.

I.6 GEMINI 3.5 FLASH ROUND-5 PATCHING RESULTS

Figure 12 reports the Gemini round-5 response curves. Seller-facing proposals remain largely self-anchored. The seller-ask sweep is again piecewise, with a positive response only when the ask continues the seller's concession path ($\beta_{\text{left}} = 0.125$) and almost none once it reverses ($\beta_{\text{right}} = -0.009$). The buyer counteroffer has essentially no effect on seller-facing prices ($\beta = 0.002$), and the previous mediator-to-seller proposal remains the main input ($\beta = 0.357$), with a small effect of the previous mediator-to-buyer proposal ($\beta = 0.046$).

Buyer-facing curves show the parallel pattern. The seller ask has almost no effect ($\beta = 0.004$), while the counteroffer sweep is thresholded: flat below the trajectory-consistent region ($\beta_{\text{left}} = -0.037$) and steep once the counteroffer reaches the expected concession path ($\beta_{\text{right}} = 0.588$). Previous

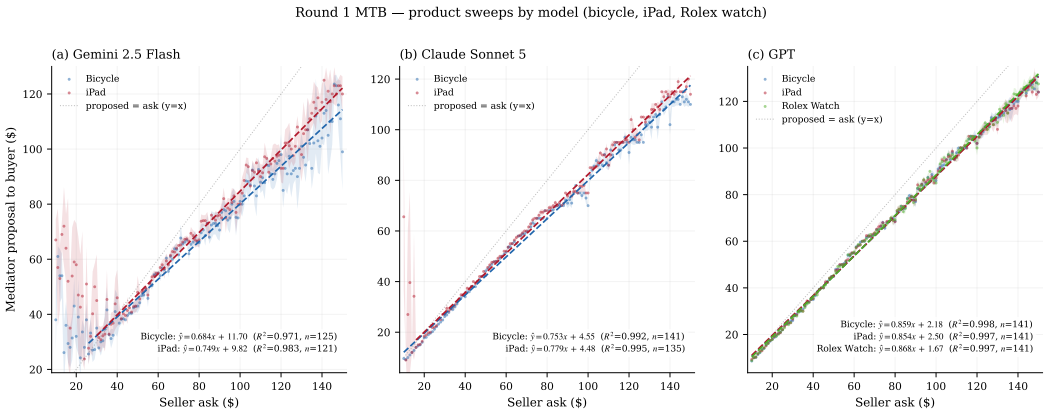

Figure 9: Round-1 mediator-to-buyer product sweeps across models. Each panel varies the seller ask with the rest of the round-1 prompt fixed and compares product descriptions. Gemini 3.5 Flash and Claude Sonnet 5 separate products, whereas GPT-5.6 produces nearly overlapping responses.

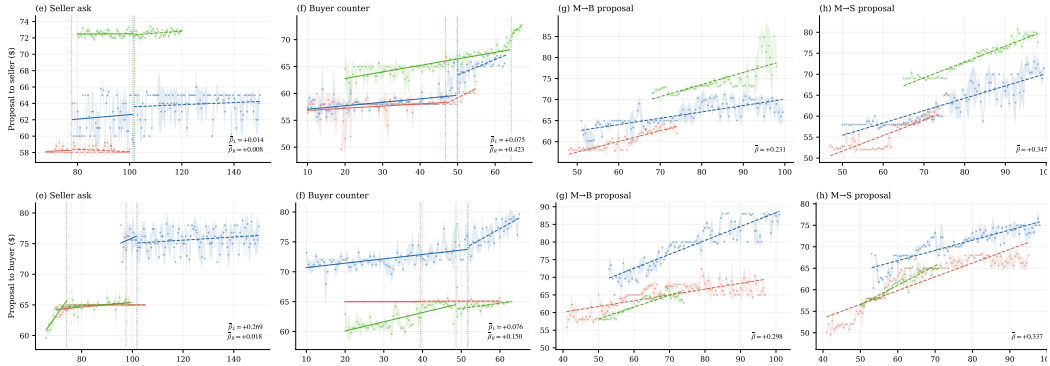

Figure 10: Claude Sonnet 5 round-3 patching response curves. Layout as in Figure 3. Previous mediator proposals remain influential, and several party-offer sweeps change slope around the prior trajectory.

mediator proposals remain influential ($\beta = 0.251$ for $M \to S$ and 0.235 for $M \to B$). Round 5 is thus consistent with round 3: current offers matter mainly when they continue the trajectory, and the mediator's own proposals explain a substantial part of late-round pricing.

I.7 Temperature-zero sweep

We evaluate the same Gemini 3.5 Flash patched prompts at temperature zero to test whether the staircase shape is an artifact of sampling. The plateaus and jumps become sharper: inputs with weak average effects at the default temperature often produce exactly flat regions followed by abrupt changes once a local threshold is crossed (Figures 13–15).

Staircase response pattern. Over a range of patched values, the mediator repeats approximately the same proposal, and only after the input crosses a threshold does the output jump to a new level (Figure 16). Two mechanisms contribute. The prompt's examples ("$50", "$65") encourage focal round prices, which by itself produces steps at the rounding granularity. Plateaus that are wider than this granularity, and thresholds that sit at the previous-round value of the patched variable, point to a thresholded policy on top of rounding. Separating the two would require sweeps with prompts that request cent-level prices.

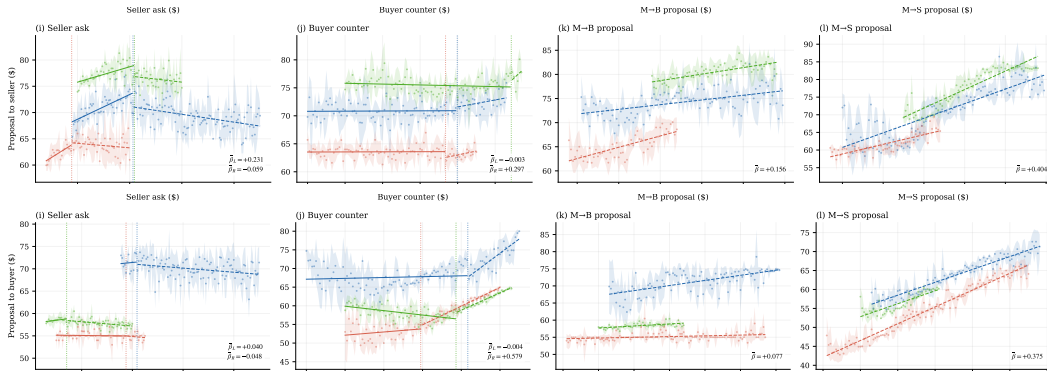

Figure 11: GPT-5.6 round-3 patching response curves. Layout as in Figure 3. Proposals depend strongly on the previous mediator trajectory, and the buyer-counter response is nearly flat below the trajectory-consistent region and steep above it.

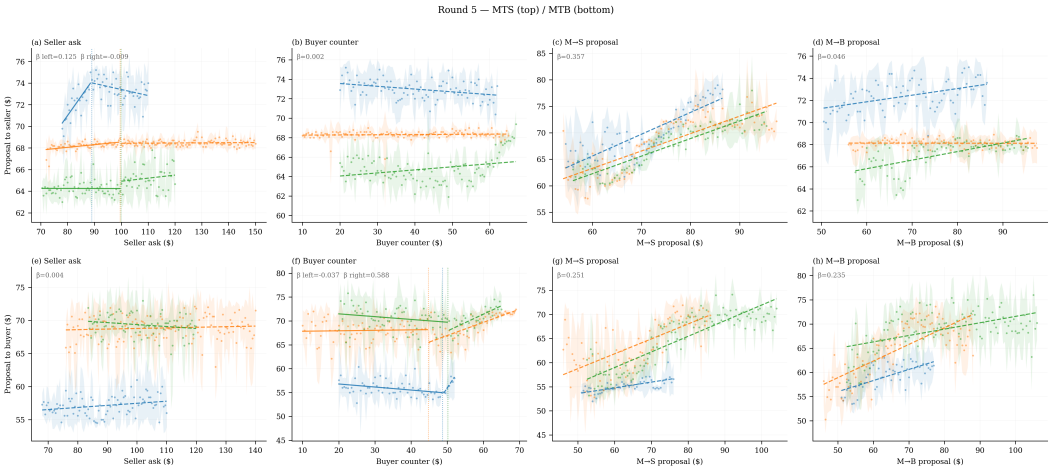

Figure 12: Gemini 3.5 Flash round-5 patching response curves. Top: mediator-to-seller decisions; bottom: mediator-to-buyer decisions. Columns patch the seller ask, buyer counteroffer, previous mediator-to-seller proposal, and previous mediator-to-buyer proposal. Points are sampled prices, shaded regions show sampling variability, and lines are fitted regressions.

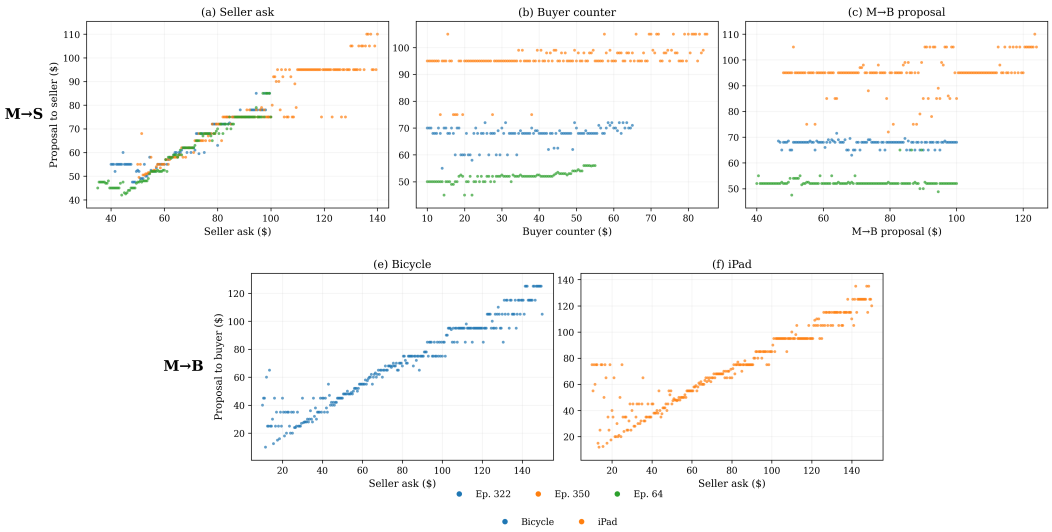

Figure 13: Gemini 3.5 Flash round-1 temperature-zero patching. Top: mediator-to-seller decisions, varying seller ask, buyer counteroffer, and previous mediator-to-buyer proposal. Bottom: initial mediator-to-buyer product-prior sweeps.

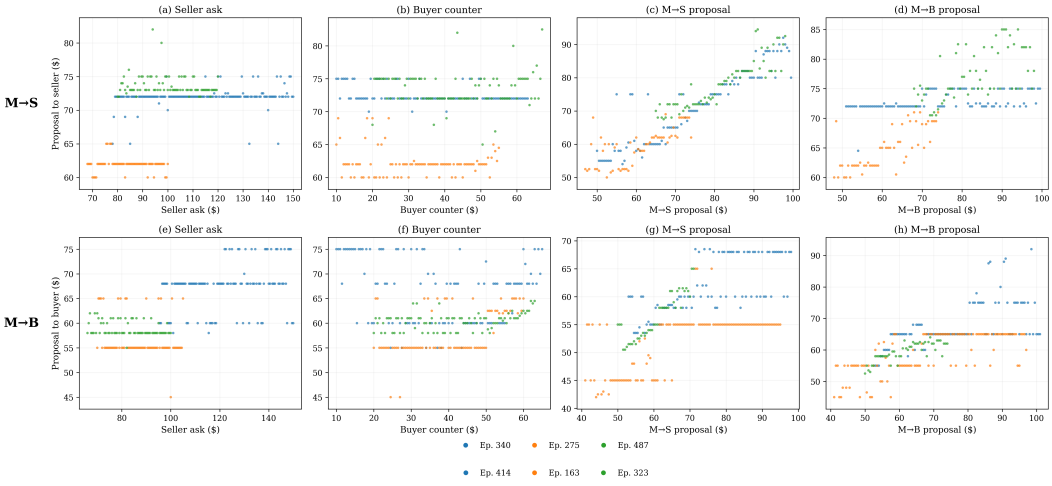

Figure 14: Gemini 3.5 Flash round-3 temperature-zero patching. Top: mediator-to-seller decisions. Bottom: mediator-to-buyer decisions. Each column varies one patched variable.

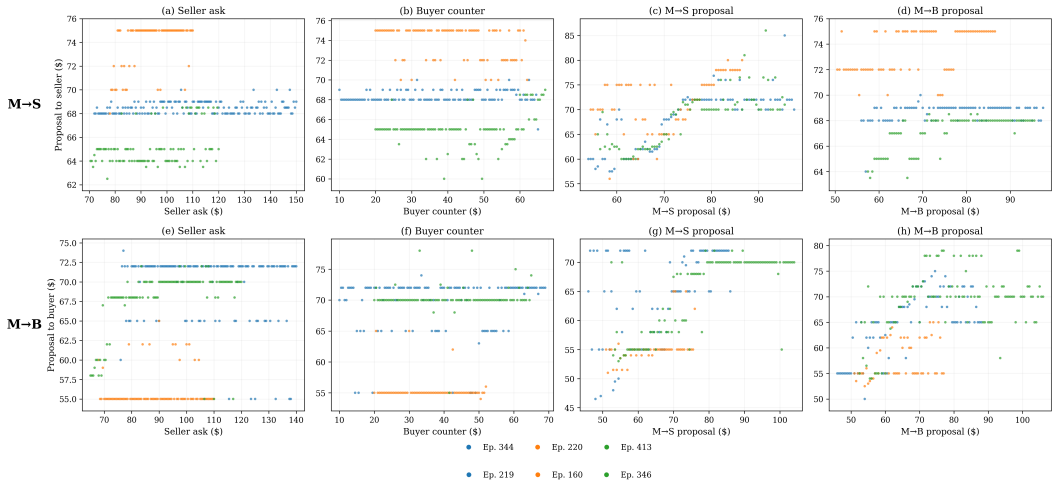

Figure 15: Gemini 3.5 Flash round-5 temperature-zero patching. The same structure persists in the final round.

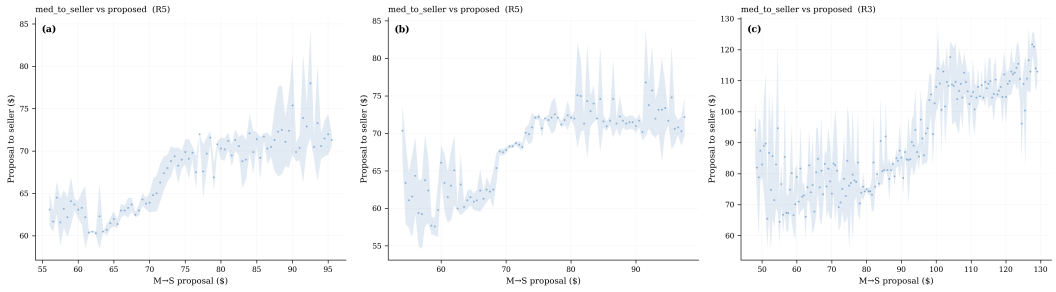

Figure 16: Gemini 3.5 Flash staircase response in one-variable patching. The proposal stays approximately flat over intervals and jumps once the patched input crosses a local threshold.

I.8 DETERMINISTIC-AGENT RESULTS ACROSS MEDIATOR MODELS

Table 5 reports the complete cross-model comparison. Brackets are 95% episode-level bootstrap confidence intervals from 10,000 resamples. Expected price, NGFT, and Nash welfare assign zero to no-deal episodes; fairness and final anchor are averaged over agreements.

Table 5: Complete deterministic-agent outcomes across mediator models. Bold marks the best point estimate within each model family (closest to 1 for Fairness); the final anchor is a behavioral measure and is not ranked. Prices are normalized to $[0, 1]$.

Mediator	Prompt	Deal rate	$\mathbb{E}[p]$	Fairness	NGFT	Nash welfare	Final anchor
Gemini 3.5 Flash	No objective	**0.657 [0.615, 0.698]**	**0.340 [0.316, 0.363]**	**0.974 [0.920, 1.027]**	**0.527 [0.491, 0.563]**	**0.0154 [0.0137, 0.0171]**	0.493 [0.470, 0.516]
Gemini 3.5 Flash	Fairness	0.635 [0.593, 0.677]	0.319 [0.296, 0.342]	1.043 [0.988, 1.097]	0.468 [0.433, 0.502]	0.0140 [0.0123, 0.0158]	0.456 [0.436, 0.477]
Gemini 3.5 Flash	Revenue	0.546 [0.502, 0.587]	0.303 [0.278, 0.328]	0.786 [0.739, 0.833]	0.394 [0.359, 0.429]	0.0151 [0.0130, 0.0174]	0.559 [0.540, 0.578]
Gemini 3.5 Flash	Deal probability	0.619 [0.575, 0.661]	0.305 [0.281, 0.327]	1.068 [1.015, 1.123]	0.479 [0.443, 0.514]	0.0141 [0.0125, 0.0158]	0.448 [0.426, 0.471]
GPT-5.6	No objective	**0.575 [0.534, 0.617]**	**0.266 [0.244, 0.287]**	1.226 [1.176, 1.276]	**0.397 [0.365, 0.429]**	**0.0110 [0.0097, 0.0124]**	0.376 [0.358, 0.394]
GPT-5.6	Fairness	0.573 [0.530, 0.615]	0.264 [0.243, 0.285]	1.196 [1.147, 1.246]	0.386 [0.354, 0.417]	0.0107 [0.0094, 0.0121]	0.381 [0.364, 0.399]
GPT-5.6	Revenue	0.542 [0.498, 0.585]	0.251 [0.229, 0.273]	**1.188 [1.136, 1.239]**	0.368 [0.335, 0.400]	0.0109 [0.0095, 0.0124]	0.388 [0.370, 0.405]
GPT-5.6	Deal probability	0.435 [0.391, 0.478]	0.194 [0.173, 0.214]	1.272 [1.214, 1.329]	0.287 [0.256, 0.318]	0.0077 [0.0065, 0.0090]	0.349 [0.330, 0.366]
Claude Sonnet 5	No objective	0.577 [0.534, 0.621]	0.265 [0.244, 0.286]	1.198 [1.143, 1.252]	0.400 [0.368, 0.432]	0.0105 [0.0093, 0.0118]	0.377 [0.359, 0.395]
Claude Sonnet 5	Fairness	**0.623 [0.579, 0.665]**	**0.290 [0.268, 0.311]**	1.157 [1.107, 1.206]	**0.431 [0.399, 0.463]**	**0.0115 [0.0103, 0.0129]**	0.393 [0.376, 0.410]
Claude Sonnet 5	Revenue	0.593 [0.550, 0.635]	0.280 [0.258, 0.302]	**1.123 [1.068, 1.179]**	0.413 [0.381, 0.445]	**0.0115 [0.0102, 0.0129]**	0.408 [0.390, 0.427]
Claude Sonnet 5	Deal probability	0.534 [0.490, 0.577]	0.241 [0.220, 0.262]	1.225 [1.172, 1.277]	0.363 [0.331, 0.395]	0.0099 [0.0086, 0.0112]	0.367 [0.351, 0.385]

I.9 LLM-NEGOTIATOR ECONOMIC RESULTS

Table 6 reports the complete LLM-negotiator comparison. Brackets are 95% episode-level bootstrap confidence intervals. Direct negotiation reaches agreement in 503 of 504 instances, so these negotiators already have a wide effective zone of agreement. Deal-probability prompting increases agreement relative to the corresponding default for all three families, revenue prompting increases the final anchor, and fairness prompting moves the final anchor toward 0.5 for Gemini and GPT but leaves Claude's unchanged.

Table 6: Economic outcomes with LLM negotiators. Brackets are 95% episode-level bootstrap CIs. Final price, final anchor, and agreement round are conditional on agreement. Prices are in dollars on a $[0, 100]$ scale.

Mediator	Prompt	Deal rate	Mean final price ($)	Mean final anchor	Mean agreement round
None	Direct negotiation	0.998 [0.994, 1.000]	47.46 [46.10, 48.90]	0.354 [0.326, 0.382]	1.77 [1.68, 1.85]
Gemini 3.5 Flash	No objective	0.913 [0.887, 0.937]	49.48 [48.38, 50.60]	0.382 [0.357, 0.408]	2.54 [2.40, 2.69]
Gemini 3.5 Flash	Fairness	0.907 [0.881, 0.931]	51.05 [49.99, 52.12]	0.412 [0.388, 0.436]	2.27 [2.13, 2.41]
Gemini 3.5 Flash	Revenue	0.841 [0.810, 0.873]	61.96 [61.10, 62.80]	0.683 [0.657, 0.710]	2.16 [2.02, 2.30]
Gemini 3.5 Flash	Deal probability	0.964 [0.948, 0.980]	48.08 [46.98, 49.20]	0.338 [0.317, 0.359]	2.37 [2.23, 2.50]
GPT-5.6	No objective	0.980 [0.968, 0.992]	46.70 [45.42, 47.95]	0.304 [0.283, 0.323]	2.07 [1.96, 2.19]
GPT-5.6	Fairness	0.980 [0.966, 0.992]	47.39 [46.11, 48.67]	0.331 [0.310, 0.353]	1.93 [1.83, 2.04]
GPT-5.6	Revenue	0.992 [0.984, 0.998]	49.34 [47.99, 50.72]	0.404 [0.380, 0.429]	1.93 [1.84, 2.02]
GPT-5.6	Deal probability	0.996 [0.990, 1.000]	47.80 [46.50, 49.12]	0.354 [0.332, 0.376]	1.78 [1.69, 1.87]
Claude Sonnet 5	No objective	0.952 [0.933, 0.970]	46.16 [44.90, 47.45]	0.285 [0.266, 0.303]	2.04 [1.93, 2.15]
Claude Sonnet 5	Fairness	0.962 [0.944, 0.978]	46.13 [44.92, 47.36]	0.287 [0.268, 0.306]	1.92 [1.81, 2.02]
Claude Sonnet 5	Revenue	0.976 [0.962, 0.988]	47.90 [46.69, 49.18]	0.346 [0.324, 0.368]	1.93 [1.83, 2.04]
Claude Sonnet 5	Deal probability	0.980 [0.966, 0.992]	46.29 [45.05, 47.53]	0.289 [0.270, 0.310]	2.00 [1.89, 2.10]

I.10 FULL BENCHMARK PERFORMANCE AND DYNAMIC PACING

We evaluate stationary heuristics: identity, midpoint, and static weighted averages $\lambda \in \{0.1, \ldots, 0.9\}$. For all rule-based mediators, the opening proposal to the buyer is the seller's ask. Table 7 reports outcomes with 95% bootstrap intervals, and Table 8 the agreement rounds.

Table 7: All deterministic-agent mediators and benchmark mediators. Brackets report 95% episode-level bootstrap confidence intervals from 10,000 resamples. Prices and payoffs are normalized to $[0, 1]$.

Family	Mediator / prompt	Deal rate	$\mathbb{E}[p]$	Fairness	NGFT	Nash welfare
Benchmark	Identity	0.627 [0.585, 0.669]	0.333 [0.310, 0.357]	0.650 [0.583, 0.718]	0.308 [0.284, 0.332]	0.0055 [0.0046, 0.0065]
Benchmark	Midpoint ($\lambda = 0.5$)	0.524 [0.480, 0.567]	0.244 [0.222, 0.266]	1.084 [1.026, 1.139]	0.343 [0.313, 0.374]	0.0098 [0.0086, 0.0111]
Benchmark	Weighted $\lambda = 0.1$	0.343 [0.304, 0.385]	0.156 [0.137, 0.175]	1.322 [1.269, 1.374]	0.194 [0.169, 0.220]	0.0043 [0.0034, 0.0052]
Benchmark	Weighted $\lambda = 0.2$	0.411 [0.369, 0.452]	0.198 [0.176, 0.219]	1.160 [1.115, 1.206]	0.233 [0.207, 0.260]	0.0056 [0.0047, 0.0067]
Benchmark	Weighted $\lambda = 0.3$	0.498 [0.454, 0.542]	0.236 [0.214, 0.258]	1.122 [1.068, 1.177]	0.293 [0.265, 0.321]	0.0069 [0.0059, 0.0080]
Benchmark	Weighted $\lambda = 0.4$	0.530 [0.486, 0.573]	0.246 [0.225, 0.267]	1.110 [1.053, 1.167]	0.333 [0.303, 0.362]	0.0086 [0.0075, 0.0098]
Benchmark	Weighted $\lambda = 0.6$	0.512 [0.468, 0.556]	0.241 [0.219, 0.262]	1.041 [0.984, 1.099]	0.347 [0.316, 0.378]	0.0106 [0.0093, 0.0120]
Benchmark	Weighted $\lambda = 0.7$	0.504 [0.460, 0.548]	0.240 [0.218, 0.262]	1.018 [0.959, 1.079]	0.359 [0.326, 0.391]	0.0120 [0.0105, 0.0136]
Benchmark	Weighted $\lambda = 0.8$	0.482 [0.438, 0.526]	0.234 [0.212, 0.256]	0.982 [0.925, 1.041]	0.358 [0.324, 0.393]	0.0131 [0.0114, 0.0148]
Benchmark	Weighted $\lambda = 0.9$	0.446 [0.403, 0.490]	0.225 [0.202, 0.248]	0.942 [0.879, 1.003]	0.340 [0.306, 0.375]	0.0130 [0.0113, 0.0148]
Gemini 3.5 Flash	No objective	0.657 [0.615, 0.698]	0.340 [0.316, 0.363]	0.974 [0.920, 1.027]	0.527 [0.491, 0.563]	0.0154 [0.0137, 0.0171]
Gemini 3.5 Flash	Fairness	0.635 [0.593, 0.677]	0.319 [0.296, 0.342]	1.043 [0.988, 1.097]	0.468 [0.433, 0.502]	0.0140 [0.0123, 0.0158]
Gemini 3.5 Flash	Revenue	0.546 [0.502, 0.587]	0.303 [0.278, 0.328]	0.786 [0.739, 0.833]	0.394 [0.359, 0.429]	0.0151 [0.0130, 0.0174]
Gemini 3.5 Flash	Deal probability	0.619 [0.575, 0.661]	0.305 [0.281, 0.327]	1.068 [1.015, 1.123]	0.479 [0.443, 0.514]	0.0141 [0.0125, 0.0158]
GPT-5.6	No objective	0.575 [0.534, 0.617]	0.266 [0.244, 0.287]	1.226 [1.176, 1.276]	0.397 [0.365, 0.429]	0.0110 [0.0097, 0.0124]
GPT-5.6	Fairness	0.573 [0.530, 0.615]	0.264 [0.243, 0.285]	1.196 [1.147, 1.246]	0.386 [0.354, 0.417]	0.0107 [0.0094, 0.0121]
GPT-5.6	Revenue	0.542 [0.498, 0.585]	0.251 [0.229, 0.273]	1.188 [1.136, 1.239]	0.368 [0.335, 0.400]	0.0109 [0.0095, 0.0124]
GPT-5.6	Deal probability	0.435 [0.391, 0.478]	0.194 [0.173, 0.214]	1.272 [1.214, 1.329]	0.287 [0.256, 0.318]	0.0077 [0.0065, 0.0090]
Claude Sonnet 5	No objective	0.577 [0.534, 0.621]	0.265 [0.244, 0.286]	1.198 [1.143, 1.252]	0.400 [0.368, 0.432]	0.0105 [0.0093, 0.0118]
Claude Sonnet 5	Fairness	0.623 [0.579, 0.665]	0.290 [0.268, 0.311]	1.157 [1.107, 1.206]	0.431 [0.399, 0.463]	0.0115 [0.0103, 0.0129]
Claude Sonnet 5	Revenue	0.593 [0.550, 0.635]	0.280 [0.258, 0.302]	1.123 [1.068, 1.179]	0.413 [0.381, 0.445]	0.0115 [0.0102, 0.0129]
Claude Sonnet 5	Deal probability	0.534 [0.490, 0.577]	0.241 [0.220, 0.262]	1.225 [1.172, 1.277]	0.363 [0.331, 0.395]	0.0099 [0.0086, 0.0112]

Table 8: Agreement round distribution for benchmark and Gemini 3.5 Flash mediators. Percentages are computed out of successful agreements for that mediator.

Mediator	Deals	R1	R2	R3	R4	R5
Identity	316	0 (0.0%)	20 (6.3%)	6 (1.9%)	29 (9.2%)	261 (82.6%)
Midpoint	264	0 (0.0%)	46 (17.4%)	75 (28.4%)	48 (18.2%)	95 (36.0%)
$\lambda = 0.1$	173	0 (0.0%)	29 (16.8%)	13 (7.5%)	10 (5.8%)	121 (69.9%)
$\lambda = 0.2$	207	0 (0.0%)	30 (14.5%)	21 (10.1%)	32 (15.5%)	124 (59.9%)
$\lambda = 0.3$	251	0 (0.0%)	32 (12.7%)	51 (20.3%)	29 (11.6%)	139 (55.4%)
$\lambda = 0.4$	267	0 (0.0%)	42 (15.7%)	60 (22.5%)	36 (13.5%)	129 (48.3%)
$\lambda = 0.6$	258	0 (0.0%)	47 (18.2%)	93 (36.0%)	46 (17.8%)	72 (27.9%)
$\lambda = 0.7$	254	0 (0.0%)	76 (29.9%)	88 (34.6%)	33 (13.0%)	57 (22.4%)
$\lambda = 0.8$	243	0 (0.0%)	116 (47.7%)	56 (23.0%)	29 (11.9%)	42 (17.3%)
$\lambda = 0.9$	225	0 (0.0%)	141 (62.7%)	23 (10.2%)	21 (9.3%)	40 (17.8%)
Gemini 3.5 Flash – No objective	331	96 (29.0%)	132 (39.9%)	43 (13.0%)	17 (5.1%)	43 (13.0%)
Gemini 3.5 Flash – Fairness	320	66 (20.6%)	102 (31.9%)	48 (15.0%)	24 (7.5%)	80 (25.0%)
Gemini 3.5 Flash – Revenue	275	80 (29.1%)	45 (16.4%)	25 (9.1%)	21 (7.6%)	104 (37.8%)
Gemini 3.5 Flash – Deal probability	312	69 (22.1%)	112 (35.9%)	53 (17.0%)	20 (6.4%)	58 (18.6%)

First-mover robustness. Tables 9 and 10 report the full buyer-first comparison discussed in Section 5.3.

Table 9: Effect of first mover on Gemini 3.5 Flash mediation. Prices are normalized to $[0, 1]$. Fairness is the ratio $2\,\mathbb{E}[u_B]/(\mathbb{E}[u_B] + \mathbb{E}[u_S] + \mathbb{E}[u_M])$.

Protocol	Deal Rate	$\mathbb{E}[p]$	Fairness	$\mathbb{E}[u_B]$	$\mathbb{E}[u_S]$	$\mathbb{E}[u_M]$	$\mathbb{E}[\text{NGFT}]$	$\mathbb{E}[\text{NW}]$
Seller-first	0.657	0.340	1.046	0.114	0.097	0.007	0.527	0.0154
Buyer-first	0.508	0.292	0.744	0.058	0.092	0.006	0.342	0.0117

Table 10: Gemini 3.5 Flash agreement-round distribution by first-mover protocol. Percentages are computed out of successful agreements.

Protocol	Deals	R1	R2	R3	R4	R5
Seller-first	331	96 (29.0%)	132 (39.9%)	43 (13.0%)	17 (5.1%)	43 (13.0%)
Buyer-first	256	0 (0.0%)	108 (42.2%)	35 (13.7%)	14 (5.5%)	99 (38.7%)

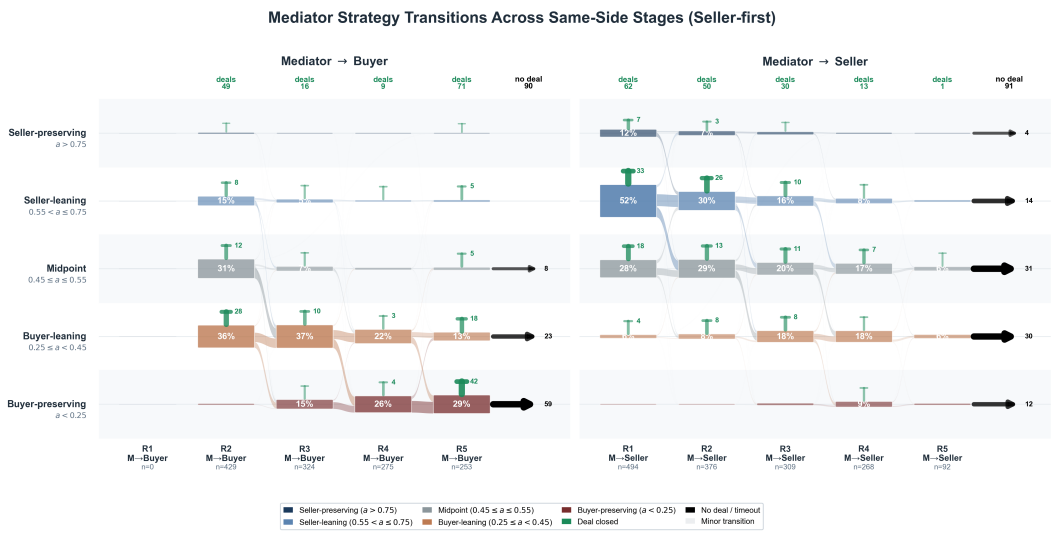

Figure 17: Gemini 3.5 Flash strategy transitions under the fairness objective.

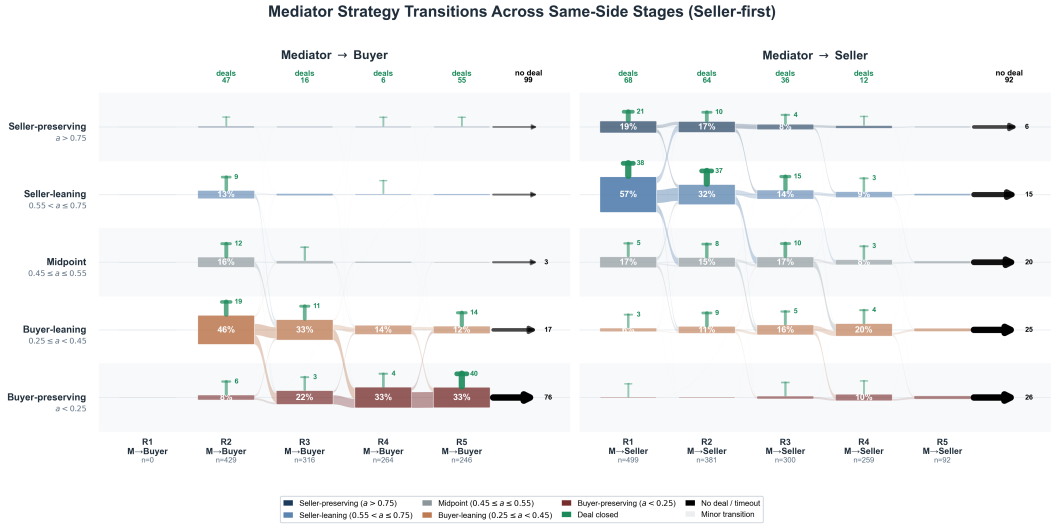

Figure 18: Gemini 3.5 Flash strategy transitions under the deal-probability objective.

I.11 Objective-directed flow graphs

Figures 17–19 show the transition diagrams for Gemini 3.5 Flash under objective steering, and Figure 20 the buyer-first flow. Under the fairness objective, proposals move into the midpoint band ($[0.45, 0.55]$) by rounds 2–3, in line with its fairness ratio of 1.043. Revenue steering keeps proposals in the seller-leaning and seller-preserving classes and suppresses transitions toward buyer-leaning classes, consistent with its higher final anchor (0.559) and lower fairness (0.786).

Assets and licenses. We use Gemini 3.5 Flash through the Google API, GPT-5.6 and GPT-4o through the OpenAI API, and Claude Sonnet 5 through the Anthropic API under the applicable provider terms. We use PersonaHub as the source of sampled personas under its stated license and terms of use. All negotiation logs and analysis code are produced by the authors, and no third-party dataset is redistributed.

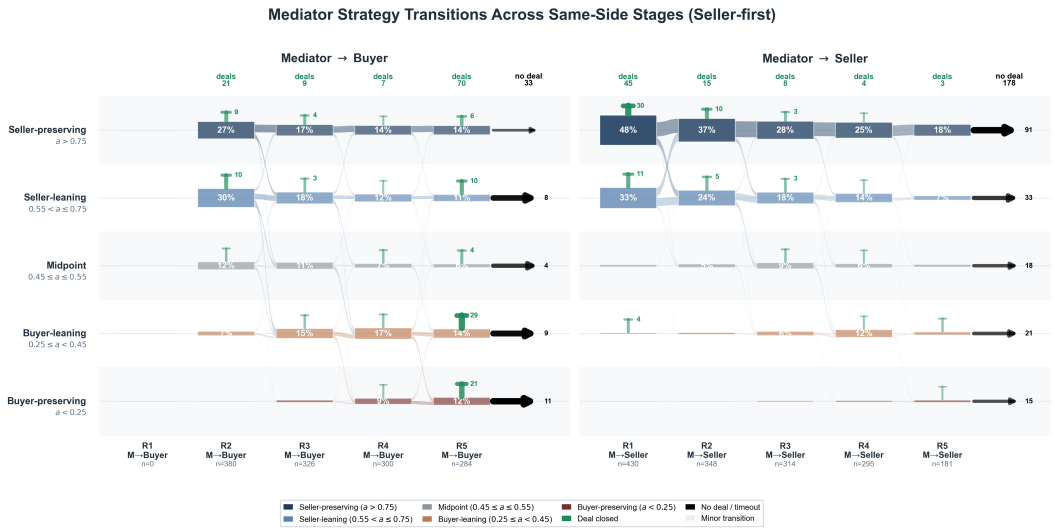

Figure 19: Gemini 3.5 Flash strategy transitions under the revenue objective.



Figure 20: Gemini 3.5 Flash strategy transitions under the buyer-first protocol. The round-1 mediator-to-seller proposal is classified by $d_B^1 = p_{MS}^1/o_B^1$, which is at least 1 for any proposal above the bid, so all round-1 proposals fall in the top class by construction and the first column should not be read as seller-preserving behavior. From round 2 on, proposals are classified by the normalized anchors of Section 2.