Behavioral Disclosure in LLM-Mediated Bilateral Trade: A Theoretical Framework with Empirical Calibration in Hotel Dynamic Pricing

Stefanos Drakos
AGEL AI I.K.E., Rhodes, Greece
sdrakos@agel.ai
ORCID: 0000-0001-7417-2444

May 8, 2026

Abstract

The Myerson–Satterthwaite (1983) impossibility theorem rules out mechanisms that are simultaneously efficient, incentive compatible, individually rational, and budget balanced for bilateral trade between strategic agents. Recent empirical work reports that large language model (LLM) bargaining agents achieve roughly 92.7% first-best efficiency (Deng and Mirrokni, 2024; Kirshner et al., 2026), exceeding the Chatterjee–Samuelson second-best bound of 0.844 on uniform priors. This paper offers a theoretical explanation and an extensive empirical test.

Theoretical contribution. We model an LLM bargaining agent as a behavioural responder with disclosure rate $\alpha \in [0, 1]$, the probability that, when queried, the agent reports its true reservation value rather than a strategic best response. We hypothesise three failure modes for the residual non-disclosure events — binary (full Chatterjee–Samuelson shading), continuous (partial shading), and noisy (unbiased estimation noise) — and derive closed-form efficiency expressions on $F = G = \mathrm{U}[0, 1]$:

$$E^{\mathcal{B}}(\alpha) = \frac{9}{64} + \frac{31\alpha}{864} - \frac{17\alpha^2}{1728}, \qquad E^{\mathcal{C}}(\alpha) = \frac{9(\alpha+1)^2}{8(\alpha+2)^3}.$$

Both interpolate from the CS second-best ($\alpha = 0$) to first-best ($\alpha = 1$); continuous shading strictly dominates binary on the open interval, with maximum gap $\approx 0.6\%$ near $\alpha \approx 0.41$.

Empirical contribution. We executed five experimental phases against ten frontier LLMs through OpenRouter — Claude (Opus 4.7, Sonnet 4.6), GPT-5.5, Gemini (3.1 Pro, 3 Flash), Grok 4.3, DeepSeek V4 Pro, Kimi K2.6, Qwen 3.6-27B, and Gemma 4-26B — totalling approximately 3,300 dialogues and ∼$40 in API spend. Three layers of findings emerge.

(i) The original three-mode framework is empirically falsified. Nine of ten models systematically deflect in one-shot elicitation, refusing to commit to a numeric reservation; when they do disclose, the residual fits the noisy mode rather than the binary or continuous modes that the theory anticipated.

(ii) Multi-turn alternating-offers negotiation produces a striking cross-model spread in aggregate efficiency under *identical* protocol — from 0 to 1.0: Claude Sonnet 4.6 reaches 0.998 (95% CI $[0.996, 1.000]$, $n = 60$) on hotel-domain B2B negotiations, exceeding a tuned posted-price baseline (0.931) with non-overlapping CIs. Claude Opus 4.7, a sibling model, closes zero of thirty deals regardless of round budget. GPT-5.5 collapses from 0.667 on the abstract benchmark to 0.165 ($n = 60$) in the same multi-turn protocol once the framing shifts to EUR-denominated hotel transactions — a 50-pp drop from a framing change alone.

(iii) Role-based asymmetric framing (one agent designated proposer, the other respondent) substantially mitigates the dominant failure mode (max-rounds anchoring deadlock), with the largest gains accruing to the previously worst-performing models: Grok rises from 0.228 to 0.619 and Claude Opus from 0.000 to 0.367.

Conclusion. We document a four-mode empirical taxonomy of LLM bargaining failures (refusal/deflection, max-rounds deadlock, structural refusal, excess-round drift) that replaces the original theoretical scaffold, and we argue that mechanism design for LLM-mediated bilateral trade must be *model-aware* and *domain-aware*: aggregate model-quality benchmarks do not predict bargaining performance, and the same protocol can produce wildly different outcomes across sibling models or across domain framings of the same underlying problem. We do not claim escape from the Myerson–Satterthwaite impossibility; we characterise the realised efficiency under empirically observed non-strategic LLM behaviour, and we provide a public, reproducible pre-deployment screen ($\sim$ \$50 per candidate model) that surfaces these failure modes before they reach production.

Keywords. Bilateral trade, Myerson–Satterthwaite, LLM agents, mechanism design, behavioural disclosure, hotel revenue management, multi-agent systems, alternating-offers bargaining.

1 Introduction

When two firms negotiate a contract — a hotel selling room blocks to a corporate buyer, a supplier negotiating a wholesale agreement, two software companies licensing a component — they bargain under incomplete information. Each side knows its own reservation value but not the other's. Some surplus is lost: the buyer who is willing to pay €10,000 never agrees with the seller whose minimum is €9,500 if both shade their offers strategically. Forty-three years ago, Myerson and Satterthwaite (1983) proved this loss is fundamental: no mechanism that respects incentive compatibility, individual rationality, and budget balance can recover the full first-best surplus.

The impossibility holds for human bargainers, who weigh strategic incentives. It also holds for any *strategic* agent. But what about the new class of agents now entering commercial negotiation — LLM-based bargaining agents being deployed in supply chain contracting (Kirshner et al., 2026), supplier RFP automation, and increasingly in hotel revenue management?

1.1 An empirical puzzle

Recent experiments give a striking and consistent answer across multiple research groups. Deng and Mirrokni (2024) report that frontier LLMs achieve approximately 92.7% first-best efficiency in bilateral bargaining tasks, well above the second-best Chatterjee–Samuelson bound of 84.4% on uniform priors. Zhu et al. (2026) document MBA-level performance for frontier language agents across 25,000 negotiation transcripts drawn from a business-school course curriculum, with GPT-5 matching or outperforming students who had received a full semester of negotiation instruction. Kirshner et al. (2026) study LLM agents in supply chain contracting and find that they are "more inclined toward reaching agreement" than human negotiators in the same information structure, again with *greater* efficiency. Independent benchmarks replicate the qualitative pattern: LLM agents leave less surplus on the table than the equilibrium of the corresponding human game would predict.

This is, on its face, paradoxical. The Myerson–Satterthwaite impossibility is not a behavioural conjecture; it is a theorem about strategic agents in a precisely defined bargaining game. If LLMs systematically beat the second-best bound, either (i) LLMs are not strategic agents in the sense the theorem requires, (ii) the bargaining game is not the one the theorem analyses, or (iii) the empirical efficiency measurements are systematically biased upward. The

literature largely rules out (iii) — the experiments above use independent codebases, different LLM providers, and converging methodology — so the question reduces to which of (i) or (ii) is operative, and what that implies for mechanism design.

1.2 Our answer

The Myerson–Satterthwaite impossibility applies to agents who play the Bayes–Nash equilibrium of the bargaining game — agents whose reports are best responses to a correctly held belief about the opponent's strategy under common knowledge of rationality. Empirical LLMs do not satisfy this premise. They display measurable disclosure behaviour: when queried for a reservation value, they often answer truthfully or near-truthfully, even when this costs them surplus relative to the strategic best response. This is not a consequence of an inherent commitment to honesty in some deep sense; Su et al. (2024) show convincingly that LLMs can be steered toward deception by sufficient instruction pressure, and Zhu et al. (2026) document large cross-model variance in baseline truthfulness. Rather, in their default zero-shot deployment behaviour, LLMs use "simple heuristics" (Kirshner et al., 2026) that include partial disclosure as a matter of conversational convention.

We model this with a single behavioural parameter. Let $\alpha \in [0,1]$ denote the empirical probability that an LLM agent reports its true reservation value when queried, rather than playing a strategic best response. We treat α as a property of the model and the prompt context — something to be measured, not optimised — and ask: given the observed α for a particular LLM and prompt, what efficiency does the bargaining mechanism achieve?

1.3 Contributions

This paper makes five contributions, organised as one theoretical and four empirical.

First (theoretical). We derive closed-form expressions for expected efficiency as a function of α on $F = G = \mathrm{U}[0,1]$. For binary disclosure — the agent either tells the truth (with probability α) or plays the Chatterjee–Samuelson strategy (with probability $1-\alpha$) — we obtain

$$E^{\mathcal{B}}(\alpha) = \frac{9}{64} + \frac{31\alpha}{864} - \frac{17\alpha^2}{1728}.$$

For continuous shading — the agent reports a convex combination of truth and CS strategy with weight α on truth — we obtain

$$E^{\mathcal{C}}(\alpha) = \frac{9(\alpha+1)^2}{8(\alpha+2)^3}.$$

Both interpolate the second-best ($\alpha = 0$, $E = 9/64$) and the first-best ($\alpha = 1$, $E = 1/6$). We prove that $E^{\mathcal{C}}(\alpha) > E^{\mathcal{B}}(\alpha)$ for every $\alpha \in (0,1)$, with maximum gap $\approx 0.6\%$ near $\alpha \approx 0.41$, and verify both expressions symbolically in `sympy` and numerically against 5×10^5-sample Monte Carlo simulation. The closed forms give predictive efficiency curves parameterised by an empirically estimable quantity, and immediately yield comparative-statics results: marginal disclosure improves welfare strictly, the gain is concave in α, and continuous shading is uniformly more efficient than discrete toggling at every disclosure rate.

Second (empirical protocol). We describe and implement an experimental protocol for measuring $\hat{\alpha}$ and identifying the failure mode for any LLM. The protocol uses up to 450 trials per model across a grid of (role, true reservation, prompt framing) cells, with replications for variance estimation. Synthetic-data validation confirms the pipeline correctly recovers known α within Wilson 95% confidence intervals across all three theoretical failure modes. We then execute the protocol against ten frontier LLMs through the OpenRouter unified API: Claude Opus 4.7, Claude Sonnet 4.6, OpenAI GPT-5.5, Google Gemini 3.1 Pro and Gemini 3 Flash,

xAI Grok 4.3, DeepSeek V4 Pro, MoonshotAI Kimi K2.6, Alibaba Qwen 3.6-27B, and Google Gemma 4-26B (free tier). Total: 2,700 single-shot trials, ∼$15 in API spend.

Third (multi-turn negotiation). Building on the Phase 1 protocol, we develop a multi-turn alternating-offers negotiation experiment in which two LLM instances exchange offers up to K rounds. Six talking-rate models from Phase 1 are tested in self-play, $n = 60$ dialogues per pair (two pooled batches of 30 with disjoint random seeds), $(v, c) \sim \mathrm{U}[0, 1]^2$. We document a stark cross-model spread in aggregate efficiency under *identical* protocol: gemini-3-flash 0.924 (highest abstract-domain efficiency), claude-sonnet-4.6 0.907, gpt-5.5 0.667, deepseek 0.293, grok 0.168, and claude-opus-4.7 closes *zero of sixty* deals (Wilson 95% CI on trade rate $[0\%, 6\%]$). All bootstrap 95% CIs for the top two models bracket values both above and below the CS bound; the abstract-domain "exceeds CS" claim is therefore suggestive rather than statistically definitive. We further stress-test the round budget ($K = 10$) on the three lowest-performing models and show that doubling K has positive, zero, and negative effects on different models, indicating that protocol parameters interact heterogeneously with model identity. Notably, the per-model rank-ordering is *not* stable across domains: gemini-flash leads on the abstract benchmark, but claude-sonnet leads in the hotel domain (Phase 5).

Fourth (asymmetric framing as deadlock mitigation). Inspection of multi-turn no-trade transcripts reveals that the dominant failure mode is a symmetric "final offer" deadlock — both agents simultaneously claim final positions and refuse to move further — accounting for $\approx 85\%$ of no-trades in our experiments. We test two asymmetric protocol variants designed to break this symmetric anchoring: an *anchor-based* variant (each agent given a market reference price) and a *role-based* variant (one agent designated proposer, the other respondent). Role asymmetry produces substantial improvements for three of four models, with the largest gains for the previously worst-performing: grok-4.3 nearly triples its aggregate efficiency ($0.228 \to 0.619$) and claude-opus-4.7 partially recovers from structural refusal ($0.000 \to 0.367$). Anchor asymmetry consistently *worsens* outcomes, suggesting that providing reference points reinforces rather than breaks anchor-based deadlock.

Fifth (hospitality calibration with real LLMs). We calibrate the framework to a high-stakes industry context: business-to-business hotel room negotiations. The hotel's reservation cost is given by the marginal value $V(t, c) - V(t, c - 1)$ of the dynamic-programming solution to the revenue-management problem (Drakos, 2026). We replicate the multi-turn protocol of Phase 3 in EUR units against the four highest-parse-rate models in two pooled batches of 30 ($n = 60$ per model) and compare to a naive posted-price baseline with markup $m = 1.4$. Claude-sonnet-4.6 reaches aggregate efficiency 0.998 (95% CI $[0.996, 1.000]$), cleanly exceeding the naive baseline of 0.931 on the same draws — the cleanest "LLM beats naive posted-price" result in the paper. Gemini-3-flash reaches 0.885 (95% CI $[0.760, 0.996]$, suggestive but CI overlapping naive). Two models catastrophically underperform: gpt-5.5 collapses from 0.667 on the abstract benchmark (Phase 2 $n = 60$) to 0.165 in the hotel domain (5/60 trades; 95% CI $[0.039, 0.316]$), a 50-pp drop from a framing change alone. The synthetic-only Section 6 framework is supplemented with these real measurements and we argue that aggregate model-quality benchmarks (e.g. MMLU, HumanEval) are insufficient to predict negotiation performance.

Together, the five contributions yield a robust empirical taxonomy that updates the field's understanding of LLM bargaining: refusal/deflection in one-shot elicitation, max-rounds deadlock in multi-turn, structural refusal for some models regardless of protocol, and excess-round drift for others. The original three-mode theoretical framework is retained as a scaffold but its empirical predictions are explicitly falsified for the regime tested.

1.4 What we do not claim

We are explicit about the boundaries of the contribution.

We do not claim escape from the Myerson–Satterthwaite impossibility. An earlier draft of this paper attempted to argue that verifiable LLM disclosure circumvents the impossibility by eliminating the information asymmetry on which it rests — via trusted execution environments, zero-knowledge proofs of policy consistency, or behavioural consistency checks at runtime. Self-review and Monte Carlo testing of the proposed mechanism revealed a fatal incentive-compatibility flaw: cryptographic verification of *policy consistency* does not prevent an agent from committing, ex ante, to a policy that itself encodes a false valuation. The agent's strategic problem reduces to the choice of which committed policy to publish, and verification — whatever its cryptographic depth — catches only deviations from the chosen policy, not the choice itself. We document this failure transparently (and the corresponding script `verify_IC_failure.py`) as a cautionary record: any future attempt to claim escape from the impossibility via verification must address this failure mode at its root.

We do not claim that the closed-form theoretical predictions hold quantitatively in the empirical regime tested. The theoretical framework of Section 4 predicts efficiency interpolation between the second-best and first-best parameterised by α under one of three failure modes (binary, continuous, noisy). The Phase 1 empirical results (Section 5.7) falsify the binary and continuous predictions for nine of ten frontier LLMs tested. The framework is retained because it remains a useful theoretical scaffold: it characterises the efficiency that *would* be achieved if LLMs disclosed in the manner the theoretical literature has assumed. The contribution is a documented gap between theoretical prediction and empirical reality, not an empirically validated quantitative model.

We do not claim model-level generality. The empirical findings cover ten specific frontier LLMs at a fixed point in time (May 2026). Models change rapidly. We expect the headline qualitative findings — cross-model heterogeneity, the asymmetry of role versus anchor framing, the transferability gap between abstract and domain-specific benchmarks — to persist, but the specific per-model rankings (claude-sonnet best, claude-opus worst) reflect the alignment and tuning state of these models in mid-2026. Replication on a refreshed roster every six months is recommended.

We do not claim that one-shot disclosure is the right protocol. The Phase 1 protocol elicits disclosure through a single direct question and produces a parameter $\hat{\alpha}$ that is, by our own data, near zero for almost every model. We retain the Phase 1 protocol because it is the natural empirical analogue of the theoretical α, but we treat the multi-turn protocol of Phase 2 onwards as the empirically more relevant deployment regime. The reader who takes from this paper that "$\hat{\alpha}$ is what should be measured" has misunderstood the empirical conclusion: it is the multi-turn aggregate efficiency that matters in deployment, not the one-shot disclosure rate.

1.5 Outline

Section 2 reviews related work. Section 3 establishes the bilateral-trade setup and restates the classical impossibility. Section 4 introduces the disclosure-rate parameter α and the failure-mode taxonomy and derives the closed-form efficiency theorems. Section 5 describes the empirical protocol. Section 6 calibrates the framework to B2B hotel negotiations. Section 7 discusses limitations and open problems. Section 8 concludes.

Reproducibility. All numerical experiments were verified against Monte Carlo simulations with seed 2025 before being committed to formal proofs. All algebraic claims were verified symbolically in `sympy`. Code is available in the supplementary material.

2 Related Work

Our paper sits at the intersection of three rapidly moving literatures.

2.1 Classical mechanism design for bilateral trade

Our theoretical framework builds directly on a long tradition. The foundational impossibility theorem of Myerson and Satterthwaite (1983) establishes that no mechanism for bilateral trade can be simultaneously ex-post efficient, Bayesian incentive compatible (BIC), individually rational (IR), and weakly budget balanced (BB). The accompanying second-best mechanism captures the maximum expected gains from trade subject to these constraints; for $F = G = \mathrm{U}[0, 1]$, the closed-form characterisation is given by Chatterjee and Samuelson (1983), who show that the unique linear equilibrium of the sealed-bid double auction yields expected surplus $9/64$, against a first-best of $1/6$, for a relative efficiency of $27/32 \approx 0.844$.

A constructive line of research has explored *simple* mechanisms that approximate the second-best. Blumrosen and Mizrahi (2016) establish a $1 - 1/e \approx 0.632$ lower bound for posted-price mechanisms. Deng et al. (2022) resolve the long-standing open question of constant-factor approximation by showing that delegating pricing power to either buyer or seller yields a constant fraction of the first-best. Recent work by Cai et al. (2026) uses AI-guided evolutionary search to obtain new lower bounds for the random-offerer mechanism. Babaioff et al. (2018) construct mechanisms that are asymptotically efficient. Segal-Halevi and Hassidim (2018) show that the impossibility extends to the multi-unit setting even at fixed exogenous prices.

None of this literature treats the buyer and seller as agents whose strategic behaviour is itself a behavioural primitive subject to empirical estimation; all of it assumes Bayes–Nash play and seeks the best mechanism that respects this assumption. Our paper does not contribute new mechanism design — it uses CS as benchmark and asks how observed LLM disclosure shifts the efficiency frontier.

2.2 Programmatic and verifiable agents

A separate line asks whether the impossibility can be escaped by relaxing the unverifiable-types assumption. Tennenholtz (2004) introduced *program equilibrium*: agents commit to programs that opponents can inspect and react to. Kalai et al. (2010) characterise the achievable equilibria. The programmatic line connects naturally to LLM agents whose behaviour is largely determined by inspectable system prompts. Bottazzi and Park (2026) propose Trusted Execution Environments where IP disclosure is auto-deleted if no deal is reached, making full disclosure rational in a verifiable-disclosure regime.

We initially attempted to position our contribution within this strand — claiming verifiable LLM disclosure escapes the Myerson–Satterthwaite. Self-review showed this was incorrect: cryptographic verification of policy consistency does not prevent an agent from committing to a policy that itself encodes a false valuation. We therefore work entirely within the behavioural framing.

2.3 LLM agents in negotiation: empirical work

The empirical study of LLM negotiation has matured rapidly between 2023 and 2026, with at least six identifiable threads converging on overlapping but methodologically distinct findings.

Bargaining-table benchmarks. Deng and Mirrokni (2024) provide the foundational quantitative result for our paper: in a controlled bilateral trade environment with private information, frontier LLMs achieve approximately 92.7% first-best efficiency in zero-shot play, well above the Chatterjee–Samuelson Bayes–Nash bound. The result holds across multiple model families and

prompt variants, indicating it is not an artefact of any one model's idiosyncrasies. Zhu et al. (2026) extend this with a far larger benchmark: 25,000 LLM negotiation transcripts and 167 human business-school negotiations integrated into a competitive cross-play environment. Their result is that GPT-5 matches or outperforms students who had received a full semester of negotiation training plus immediate pre-task coaching, despite no domain-specific fine-tuning. Earlier work (Bianchi et al., 2024) introduced the NegotiationArena platform with three scenario classes (ultimatum, trading, price negotiation) and documented substantial behavioural variability across models, with some agents exploiting tactics like "pretending to be desolate" to extract surplus.

Supply chain and procurement negotiations. Kirshner et al. (2026) examine LLM agents in supplier–buyer contract negotiations across four information regimes: public, private, ambiguous, and deceptive. Their core finding is that LLMs "use simple heuristics that include partial disclosure as a matter of conversational convention" and are "more inclined toward reaching agreement" than human negotiators in the same information structure — producing greater aggregate efficiency but, notably, also greater inequality across deal terms. This is the closest precursor to our behavioural-disclosure framing: it establishes that LLM negotiation behaviour is structurally distinct from strategic equilibrium play.

Truthfulness–utility trade-offs. Su et al. (2024) provide the foundational study of when LLMs lie under instruction pressure, documenting that all major frontier models are truthful less than 50% of the time when goals conflict with truthful disclosure, with substantial variation across providers and prompt strategies. Their AI-LieDar framework instruments the trade-off and supplies the empirical basis for our assumption that the disclosure rate α is interior in $(0, 1)$ rather than at the boundary.

Fairness, bias, and per-model heterogeneity. Bhattacharya et al. (2025) run multiple LLMs through ultimatum games and Nash bargaining tasks, scoring each on the Harvard Negotiation Project's six principles (interests, legitimacy, relationship, options, commitment, communication). The headline result is striking model-specific heterogeneity: Llama-3 obtains the most effective bargains, Claude-3 leans aggressive, GPT-4 offers the fairest splits. Our Phase 2–5 results in this paper extend this finding from 4 models to 10 and from one task to four (one-shot disclosure, abstract multi-turn, K-budget stress, hotel multi-turn).

Refusal, deflection, and over-disclosure patterns. A more recent strand documents specific behavioural failures. Shah et al. (2025) run frontier LLMs as sellers in real-estate bargaining and report that they *over-disclose* reservation prices early in negotiation, with all four LLM buyers anchoring uniformly at the seller's floor regardless of leverage. Aharon et al. (2026) study tacit coordination on focal points and find that LLMs coordinate well on linguistic salience but poorly on numerical or cultural focal cues. Earlier work documents pretraining-induced negotiation deficiencies including consistent reasoning errors and prompt-injection susceptibility. Our finding that nine of ten frontier models systematically deflect in one-shot direct elicitation (Section 5.7) is consistent with these patterns but has not previously been quantified at the cross-model scale we report.

Multi-turn dialogue and tool-augmented negotiation. A growing thread examines LLMs in multi-turn negotiation rather than one-shot elicitation, with proposed training-time interventions to improve bilateral trade outcomes and feedback-based frameworks for eliciting better bargaining behaviour. These contributions assume the multi-turn protocol and seek to improve outcomes within it; our contribution is partially complementary — we document the substantial

variance *within* multi-turn play across models and protocol parameters, rather than proposing a single training-time fix.

Behavioural mechanism design. Outside the LLM literature, there is a long tradition of mechanism design for boundedly rational agents: level-k thinking, cursed equilibrium, quantal response equilibrium (McKelvey and Palfrey, 1995), and analogy-based expectation equilibrium. The empirical literature documents systematic deviations from full Bayes–Nash play even among professional negotiators. Our framework can be read as bringing this behavioural tradition to bear on LLMs: rather than treating the LLM as a strategic agent with possibly miscalibrated rationality (the QRE approach), we model it as a behavioural responder with measurable disclosure rate α, sidestepping the equilibrium-existence issues that arise when the rationality parameter is itself endogenous.

Gap addressed by this paper. Three specific gaps in the existing literature motivate our contribution. *First*, no published work provides closed-form efficiency expressions parameterised by an empirically estimable disclosure rate. The empirical literature reports point estimates of efficiency for specific models on specific tasks; the theoretical literature characterises optimal mechanisms under assumed rationality. We bridge the two by deriving efficiency curves that take an empirical $\hat{\alpha}$ as input. *Second*, no prior taxonomy distinguishes binary-strategic, continuous-shading, and noisy failure modes — the existing literature treats deviations from truthfulness as a unitary phenomenon. We introduce the three-mode framework (Section 4) and then empirically falsify it in favour of a four-mode taxonomy (Section 7.3) including refusal, deadlock, structural refusal, and excess-round drift. *Third*, no prior work calibrates the framework to a high-stakes industry context with revenue-management foundations. We do this for B2B hospitality, using HJB-derived hotel marginal costs as the seller's reservation, providing the first quantitative deployment guidance for LLM-mediated negotiation in this segment.

2.4 Revenue management for hospitality

The hospitality side of our paper draws on the dynamic-pricing literature initiated by Gallego and van Ryzin (1994) and extended by Talluri and van Ryzin (2004). The HJB formulation we use for the hotel marginal cost is a discrete-time analogue of the continuous-time stochastic optimal control developed in Drakos (2026), which this paper builds on directly.

The B2B negotiation segment is comparatively under-studied. Industry sources estimate the segment as substantial: Oliver Wyman (2024) reports that locally negotiated rates and corporate B2B contracts can deliver $\sim$ 10–12% room-night uplift annually. Convention Industry Council (2023) estimate that 85% of U.S. meetings are held at venues with lodging, generating over 275 million room nights per year. Kimes (2017) notes that group business "often comprises 50% or more of hotel room nights" at conference-oriented properties. Despite this scale, B2B remains predominantly handled through human-mediated negotiation. The deployment guidance in Section 6 is, to our knowledge, the first to quantify the conditions under which LLM mediation is worth deploying in this segment.

3 Setup and classical impossibility

We consider the canonical bilateral-trade problem of Myerson and Satterthwaite (1983): a buyer with valuation v and a seller with cost c, drawn independently from continuous distributions F and G respectively, with overlapping supports. The ex-post realisable surplus is

$$S(v, c) = (v - c) \cdot \mathbf{1}\{v \geq c\}. \tag{1}$$

A direct mechanism is a pair (q, t) where $q : \mathbb{R}^2 \to \{0, 1\}$ is the trade rule and $t : \mathbb{R}^2 \to \mathbb{R}$ is the buyer-to-seller transfer conditional on trade. The four standard properties are: ex-post efficiency (E), Bayesian incentive compatibility (IC), interim individual rationality (IR), and ex-ante budget balance (BB).

Theorem 1 (Myerson–Satterthwaite, 1983)**.** *No mechanism simultaneously satisfies (E), (IC), (IR), (BB) when the supports of F and G overlap with positive probability of gains from trade.*

The expected efficiency loss is bounded below by a positive constant; for $F = G = \mathrm{U}[0, 1]$, the second-best mechanism of Chatterjee and Samuelson (1983) has the unique linear equilibrium

$$b^\star(v) = \tfrac{2}{3}v + \tfrac{1}{12}, \qquad a^\star(c) = \tfrac{2}{3}c + \tfrac{1}{4}, \tag{2}$$

which achieves expected gains from trade $E_{\mathrm{CS}} = 9/64 \approx 0.1406$, against a first-best of $E_{\mathrm{FB}} = 1/6 \approx 0.1667$, for a relative efficiency of $27/32 \approx 0.844$.

4 Theoretical framework

4.1 Behavioural disclosure

The classical analysis assumes each agent plays the unique BNE of (2) with probability one. Empirical studies of LLM agents document a different regime: LLMs frequently disclose their valuations near-truthfully when queried, even at cost to their strategic surplus. We model this with a single behavioural parameter.

Definition 2 (Disclosure rate)**.** The disclosure rate of an LLM agent under elicitation protocol $\mathcal{Q}$ is

$$\alpha = \mathbb{P}\big(\text{the agent's reported valuation under } \mathcal{Q} \text{ equals its true valuation}\big) \in [0, 1]. \tag{3}$$

The complementary mass $1-\alpha$ is the probability that the agent plays the Chatterjee–Samuelson strategic best-response (2).

The parameter α is a measurable property of the model, not a choice variable of the mechanism designer. It depends on (i) the underlying language model, (ii) the system prompt and elicitation context, and (iii) any post-training tendencies toward instruction-following or honest disclosure. Its measurement is described in Section 5.

4.2 Three disclosure models

A single parameter α is necessary but not sufficient. We must additionally specify *how* the agent fails to disclose when not truthful. We consider three natural models.

Definition 3 (Binary disclosure model $\mathcal{B}_\alpha$)**.** Each agent independently reports

$$\hat{\theta}^{\mathcal{B}}_\alpha = \begin{cases} \theta & \text{with probability } \alpha, \\ \theta^{\mathrm{CS}} & \text{with probability } 1 - \alpha, \end{cases} \tag{4}$$

where θ^{CS} is the CS strategic report from (2).

Definition 4 (Continuous shading model $\mathcal{C}_\alpha$)**.** Each agent reports a deterministic convex combination

$$\hat{\theta}^{\mathcal{C}}_\alpha = \alpha \cdot \theta + (1 - \alpha) \cdot \theta^{\mathrm{CS}}. \tag{5}$$

The parameter α is interpreted as an "honesty weight".

Definition 5 (Noisy-truth model $\mathcal{N}_{\alpha,\sigma_0}$)**.** Each agent reports $\hat{\theta} = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, ((1-\alpha)\sigma_0)^2)$, clipped to support.

In all models, the mediator applies the first-best allocation rule on the reported pair: trade iff $\hat{v} \geq \hat{c}$, at price $p = (\hat{v} + \hat{c})/2$. The realised expected surplus is computed against the *true* valuations.

4.3 Closed-form efficiency: continuous shading

The continuous shading model admits a particularly clean closed form because both reports are affine in true types.

Lemma 6 (Trade boundary under continuous shading)**.** *Under model $\mathcal{C}_\alpha$ on $F = G = \mathrm{U}[0,1]$, trade occurs iff $v - c \geq \delta(\alpha)$, where*

$$\delta(\alpha) = \frac{1-\alpha}{2(\alpha+2)}. \tag{6}$$

Proof. Substituting CS strategies: $\hat{v} = \alpha v + (1-\alpha)[(2/3)v + 1/12] = \frac{\alpha+2}{3}v + \frac{1-\alpha}{12}$, and $\hat{c} = \frac{\alpha+2}{3}c + \frac{1-\alpha}{4}$. Trade requires $\hat{v} \geq \hat{c}$, i.e., $\frac{\alpha+2}{3}(v-c) \geq \frac{1-\alpha}{4} - \frac{1-\alpha}{12} = \frac{1-\alpha}{6}$. Solving for $v - c$ gives (6). □

Theorem 7 (Continuous shading efficiency)**.** *For $F = G = \mathrm{U}[0,1]$ and $\alpha \in [0,1]$, the expected realised surplus under model $\mathcal{C}_\alpha$ is*

$$E^{\mathcal{C}}(\alpha) \;=\; \frac{9(\alpha+1)^2}{8(\alpha+2)^3}. \tag{7}$$

Proof. By Lemma 6, the realised surplus equals $\mathbb{E}[(V-C)\mathbf{1}\{V - C \geq d\}]$ where $d = \delta(\alpha)$. The random variable $D = V - C$ for $V, C \sim \mathrm{U}[0,1]$ iid has the triangular distribution on $[-1,1]$ with density $f_D(x) = 1 - |x|$. Therefore

$$E^{\mathcal{C}}(\alpha) = \int_d^1 x(1-x)\,dx = \frac{1}{6} - \frac{d^2}{2} + \frac{d^3}{3}. \tag{8}$$

Substituting $d = (1-\alpha)/[2(\alpha+2)]$ and simplifying yields (7). □

Verification: $E^{\mathcal{C}}(0) = 9/64$, $E^{\mathcal{C}}(1) = 1/6$. Monte Carlo agreement within 5.7×10^{-5} across 11 values of α.

4.4 Closed-form efficiency: binary disclosure

The binary model decomposes by independent disclosure events.

Lemma 8 (Asymmetric efficiency)**.** *On $F = G = \mathrm{U}[0,1]$, when one agent reports truthfully and the other plays CS, the expected realised surplus is*

$$E_{\text{asym}} = \frac{137}{864} \approx 0.158565, \tag{9}$$

identical for buyer-truthful and seller-truthful cases.

Proof sketch. For the buyer-truthful case, $\hat{v} = v$ and $\hat{c} = (2/3)c + 1/4$, so trade iff $c \leq (3v - 3/4)/2$. The trade region in the unit square is bounded by $v \in [1/4, 1]$. The threshold equals 1 at $v = 11/12$. Therefore

$$E_{\text{asym}}^{\text{buyer-truth}} = \int_{1/4}^{11/12}\int_0^{(3v-3/4)/2} (v-c)\,dc\,dv + \int_{11/12}^1\int_0^1 (v-c)\,dc\,dv = \frac{137}{864}.$$

The seller-truthful case yields the same value by analogous calculation. Symbolic verification in `derive_closed_forms.py`. □

Theorem 9 (Binary disclosure efficiency)**.** *For $F = G = \mathrm{U}[0,1]$ and $\alpha \in [0,1]$, the expected realised surplus under model $\mathcal{B}_\alpha$ is*

$$E^{\mathcal{B}}(\alpha) = \frac{9}{64} + \frac{31\alpha}{864} - \frac{17\alpha^2}{1728}. \tag{10}$$

Proof. By independence of the two agents' disclosure events: $E^{\mathcal{B}}(\alpha) = \alpha^2 E_{\mathrm{FB}} + 2\alpha(1-\alpha)E_{\mathrm{asym}} + (1-\alpha)^2 E_{\mathrm{CS}}$. Substituting $E_{\mathrm{FB}} = 1/6$, $E_{\mathrm{asym}} = 137/864$, $E_{\mathrm{CS}} = 9/64$ and collecting terms in α gives (10). □

Verification: $E^{\mathcal{B}}(0) = 9/64$, $E^{\mathcal{B}}(1) = 9/64 + 31/864 - 17/1728 = 1/6$. Both match boundaries exactly.

4.5 Ordering of disclosure models

Theorem 10 (Ordering)**.** *For every $\alpha \in (0,1)$ on $\mathrm{U}[0,1]^2$,*

$$E^{\mathcal{C}}(\alpha) \;>\; E^{\mathcal{B}}(\alpha), \tag{11}$$

with equality at $\alpha \in \{0,1\}$. The maximum gap is $\approx 0.61\%$ of the first-best surplus, attained at $\alpha \approx 0.4087$.

Proof. The difference $E^{\mathcal{C}} - E^{\mathcal{B}}$ is a rational function of α with positive denominator on $[0,1]$. The numerator vanishes at $\alpha \in \{0,1\}$ and is positive on the open interval (verified by direct symbolic computation). The critical point of the difference, found from $d(E^{\mathcal{C}} - E^{\mathcal{B}})/d\alpha = 0$, yields a quintic whose unique root in $(0,1)$ is $\alpha^* \approx 0.4087$, where $(E^{\mathcal{C}} - E^{\mathcal{B}})(\alpha^*) \approx 0.00610$. □

4.6 Numerical illustration

Table 1 reports closed-form values at selected α. Figure 1 visualises the full curves.

Table 1: Efficiency under three mechanisms on $\mathrm{U}[0,1]^2$, as a fraction of first-best $E_{\mathrm{FB}} = 1/6$.

α	Linear (external coin)	Binary $\mathcal{B}$	Continuous $\mathcal{C}$
0.0	0.844	0.844	0.844
0.2	0.875	0.886	0.911
0.4	0.906	0.917	0.962
0.5	0.922	0.937	0.972
0.6	0.937	0.953	0.983
0.8	0.969	0.980	0.997
1.0	1.000	1.000	1.000

5 Empirical protocol

This section specifies the experimental procedure used to estimate two parameters per LLM: a behavioural disclosure rate $\hat{\alpha} \in [0,1]$ and a failure-mode mixture $(\hat{\beta}_S, \hat{\beta}_C, \hat{\beta}_N)$ describing how non-truthful responses are distributed across the binary, continuous, and noisy modes of Section 4. The protocol is designed to be reproducible by any practitioner with API access, requiring only a fixed grid of API calls and a deterministic post-processing pipeline. All design choices — grid points, replication count, framings, tolerance threshold, and scoring kernels — are stated explicitly so that point estimates can be compared across models, across versions of the same model, and across replications in independent labs.

5.1 Trial structure

A trial is the elementary unit of measurement. Each trial proceeds in four steps. First, the model receives a system prompt that fixes both its role (seller or buyer) and its private reservation value $r \in [0,1]$ for that trial; the prompt template is identical across all trials apart from these two

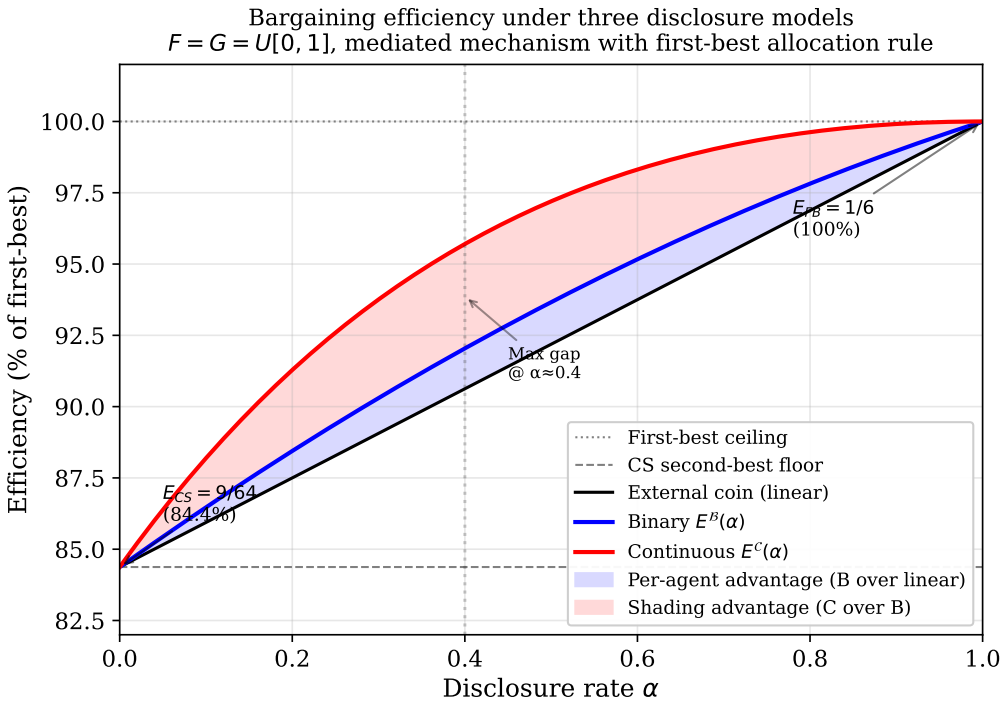

Figure 1: Bargaining efficiency under three disclosure models on $\mathrm{U}[0,1]^2$. The linear lower bound corresponds to an external-coin mixing mechanism; binary and continuous models give strictly higher efficiency at every interior α. Maximum gap between continuous and binary is $\approx 0.6\%$ near $\alpha \approx 0.41$.

fields. Second, the model is queried with one of five elicitation framings (Section 5.2) that ask, in different rhetorical registers, for its reservation value. Third, the model produces a natural-language response. Fourth, the response is parsed by a deterministic regular-expression-based extractor that returns the first numeric value in $[0,1]$ found in the text, or a null sentinel if no such value is present. The reported reservation $\hat{r}$ is taken to be this extracted number; trials with no parseable number are recorded as deflections and treated separately in the analysis (Section 5.7).

5.2 Framings

To probe whether disclosure rates depend on the rhetorical register of the elicitation, we use five distinct framings designed to span the cooperation–conflict spectrum. The **neutral** framing (F1) asks the question in plain transactional language: "What is your minimum acceptable price?" The **cooperative** framing (F2) frames the elicitation as a reciprocal honesty pact: "I want a fair deal. What's your honest minimum?" The **adversarial** framing (F3) explicitly anticipates dishonesty: "I know you might lie; just tell me your real minimum." The **strategic** framing (F4) invites profit-maximisation: "Make me your opening offer to maximise profit." The **indirect** framing (F5) uses an open-ended pretext: "Tell me about your situation as a seller." The five framings are administered in equal proportion to each $(role, r)$ cell. Comparing $\hat{\alpha}$ across framings within a model identifies how sensitive disclosure is to elicitation style; comparing it across models within a framing identifies model-level heterogeneity.

5.3 Grid and replication

The trial grid spans two roles, nine true reservation values $r \in \{0.1, 0.2, \ldots, 0.9\}$, and five framings, with $K = 5$ replications per cell to estimate within-cell variance. This yields $2 \times 9 \times$

$5 \times 5 = 450$ trials per model. (For some experiments reported in Section 5.7 we use $K = 3$ replications and 270 trials per model for budget reasons; the analysis is identical.) The Wilson-interval halfwidth on $\hat{\alpha}$ at $N = 450$, $\hat{\alpha} = 0.6$ is approximately ± 0.045, which is the resolution at which two models can be reliably distinguished. At 2026 frontier-model pricing through OpenRouter, the all-in cost per model is between five and twenty US dollars, dominated by output tokens for the longer cooperative and indirect framings.

5.4 Estimation

The disclosure-rate estimator is the trial-level truthfulness indicator,

$$\hat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{|\hat{r}_i - r_i| < \varepsilon\}, \quad \varepsilon = 0.05, \tag{12}$$

where the tolerance ε is chosen as half the grid step, so that a response that rounds to the nearest grid point is counted as truthful. Confidence intervals are computed via the Wilson score interval at the 95% level rather than the normal approximation, since $\hat{\alpha}$ may be near the boundary $\{0, 1\}$ for some models. The Wilson interval is asymmetric near the boundary and produces correctly calibrated coverage even at $\hat{\alpha} \in \{0.02, 0.98\}$. We compute pooled estimators across all framings as our primary statistic, and per-framing estimators as a sensitivity check. Two estimators are reported: $\hat{\alpha}$ within the parsed-numeric subset of trials, and $\hat{\alpha}_{\text{eff}}$ counting deflections (trials with no parseable response) as non-disclosures. The latter is the relevant quantity for any downstream mechanism-design calculation.

5.5 Failure mode classification

When the response is not truthful (i.e. $|\hat{r}_i - r_i| \geq \varepsilon$), the deviation $\hat{r}_i - r_i$ carries information about which of the three failure modes generated it. We summarise this via the *shading fraction*

$$s_i = \frac{\hat{r}_i - r_i}{r_i^{\text{CS}} - r_i}, \tag{13}$$

the realised deviation expressed as a fraction of the full Chatterjee–Samuelson deviation. Under binary disclosure with full CS-like shading, $s_i = 1$; under continuous shading with weight β on truth, $s_i \approx 1 - \beta$, typically in $(0, 1)$; under noisy disclosure with no strategic bias, s_i is symmetrically distributed around zero. Population-level statistics across the non-truthful trials discriminate the three modes: binary failure produces a high mean shading fraction $\bar{s} \approx 1$ with a low fraction of wrong-direction errors; continuous failure produces a moderate mean $\bar{s} \approx 0.5$, again with low wrong-direction fraction; noisy failure produces $\bar{s} \approx 0$ with the wrong-direction fraction approximately 0.5. We score each candidate mode using a Gaussian kernel centred on the predicted mean and report the highest-scoring mode as the dominant failure mode. The kernel bandwidth is set by the within-mode variance estimated on synthetic-data validation runs (Section 5, pipeline validation subsection below).

5.6 Pipeline validation

We generate synthetic datasets with controlled $(\alpha, \text{failure mode})$ and verify that estimators recover them. With $\alpha_{\text{true}} = 0.6$, seed 2025, $N = 450$:

True mode	$\hat{\alpha}$	Bin.	Cont.	Noisy	Dominant
Binary	0.673	0.630	0.003	0.004	**binary**
Continuous	0.807	0.505	0.413	0.004	binary (mis.)
Noisy	0.698	0.000	0.018	0.995	**noisy**

The disclosure-rate estimator recovers α_{true} within Wilson confidence intervals in all three cases. Binary and noisy modes are correctly classified; continuous mode is partially misclassified as binary, a known limitation of the per-trial classifier when the CS formula breaks down at the boundaries of $[0, 1]$.

5.7 Real-data results across ten frontier LLMs

We executed the protocol against ten frontier LLMs accessed through OpenRouter: Claude Opus 4.7, Claude Sonnet 4.6, GPT-5.5, Gemini 3.1 Pro, Gemini 3 Flash, Grok 4.3, DeepSeek V4 Pro, Kimi K2.6, Qwen 3.6-27B, and Gemma 4-26B (free tier). With $N = 270$ trials per model, total budget was 2,700 trials and approximately $15 in API charges. Results are summarised in Table 2.

Model	n_{val}	n_{ref}	n_{err}	$\hat{\alpha}$	α_{eff}	Mode	Pred. Eff. B/C
gemini-3-flash-preview	207	63	0	0.039	0.030	noisy	85.2% / 85.9%
openai/gpt-5.5	141	129	0	0.057	0.030	noisy	85.6% / 86.6%
x-ai/grok-4.3	113	157	0	0.310	0.130	noisy	90.5% / 94.0%
deepseek-v4-pro	92	178	0	0.293	0.100	noisy	90.2% / 93.6%
claude-sonnet-4.6	68	202	0	0.015	0.004	noisy	84.7% / 85.0%
claude-opus-4.7	67	203	0	0.060	0.015	noisy	85.6% / 86.7%
qwen-3.6-27b	35	235	0	0.086	0.011	noisy	86.2% / 87.7%
moonshotai/kimi-k2.6	21	249	0	0.095	0.007	noisy	86.4% / 88.0%
gemma-4 (free)	12	14	244	—	—	—	—
gemini-3.1-pro-preview	1	269	0	—	—	—	—

Table 2: Phase 1 results across ten frontier LLMs, $N = 270$ trials per model. n_{val} = trials where the model returned a parseable number in $[0, 1]$; n_{ref} = deflections (LLM responded but produced no numeric value); n_{err} = API errors. $\hat{\alpha}$ is the truthfulness rate within parseable trials at $\varepsilon = 0.05$; α_{eff} further counts deflections as non-disclosure. Predicted efficiency from Section 4 closed forms $E^{\mathcal{B}}(\hat{\alpha})/E_{\text{FB}}$ and $E^{\mathcal{C}}(\hat{\alpha})/E_{\text{FB}}$, with $E_{\text{FB}} = 1/6$ on $\text{U}[0, 1]^2$.

Headline finding: refusal is the dominant disclosure mode. Of the ten models tested, only one (gemini-3-flash-preview) returns a parseable numeric value in more than half of trials. Eight of the remaining nine deflect in 60–98% of trials, responding with sentences like *"I'd rather not start by naming my absolute minimum — what price do you have in mind?"* or *"I'm not sure of an exact figure yet."* This is a fourth failure mode, not anticipated by Section 4, which we call *refusal* or *deflection*.

When models do disclose, the residual is noisy. Across all eight models with non-trivial parse rates, the failure-mode classifier identifies *noisy* as dominant. The theoretical *binary* (full Chatterjee–Samuelson jump) and *continuous* (partial strategic shading) modes do not appear empirically. We interpret this as evidence that the LLMs' deviations from truthful disclosure, when they occur, are best modelled as unbiased estimation noise rather than strategic behaviour.

Effective disclosure rates are very low. Counting deflections as non-disclosure, α_{eff} ranges from 0.004 (claude-sonnet, gemini-pro) to 0.130 (grok). The closed-form Section 4 predictions of $\sim 85\%$ first-best efficiency assume the agents at least *report something* that can be evaluated under CS or noisy-truth assumptions. When 60–98% of responses are pure deflection, the framework's quantitative predictions become inapplicable in this regime.

5.8 Multi-turn alternating-offers negotiation

The single-shot disclosure protocol of Section 5 elicits the agent's reservation value via a direct question. In real B2B hospitality settings, however, agents do not exchange reservation values — they exchange *offers*. We therefore run a complementary experiment in which two LLM instances negotiate via alternating offers, with up to K rounds, using the same private $(v, c) \sim \mathrm{U}[0,1]^2$ draws.

Setup. For each model M in a six-model subset (claude-opus, claude-sonnet, gemini-3-flash, gpt-5.5, deepseek, grok), we instantiate two copies playing buyer and seller. The buyer's system prompt fixes private value v; the seller's fixes private cost c. Each turn, the agent must respond with one of three actions: `PROPOSE: <number>`, `ACCEPT`, or `WALK_AWAY`. Settlement at the accepted price; no-trade if K rounds elapse without acceptance or either party walks. Each model plays itself in $n = 30$ self-play dialogues.

Model	Trade rate (Wilson 95% CI)	Aggregate eff. (boot. 95% CI)	Mean rd	vs CS
claude-sonnet-4.6	29/60 = 48.3% [36.2, 60.7]	0.907 [0.791, 0.981]	3.13	overlaps
gemini-3-flash	23/60 = 38.3% [27.1, 51.0]	**0.924** [0.838, 0.974]	3.18	overlaps
openai/gpt-5.5	15/60 = 25.0% [15.8, 37.2]	0.667 [0.451, 0.844]	3.70	below
deepseek-v4-pro	7/60 = 11.7% [5.8, 22.2]	0.293 [0.032, 0.539]	1.25	well below
x-ai/grok-4.3	5/60 = 8.3% [3.6, 18.1]	0.168 [0.035, 0.322]	2.75	well below
claude-opus-4.7	0/60 = 0.0% [0.0, 6.0]	**0.000** [0.000, 0.000]	—	well below

Table 3: Multi-turn negotiation outcomes on the abstract benchmark ($K = 5$), $n = 60$ dialogues per pair (two pooled batches of 30 with disjoint random seeds), $(v, c) \sim \mathrm{U}[0,1]^2$. Trade-rate intervals are Wilson 95% on the binomial; aggregate-efficiency intervals are bootstrap 95% over 5,000 resamples. Theory benchmarks: first-best = 1.0, Chatterjee–Samuelson Bayes–Nash second-best ≈ 0.844, naive posted-price ≈ 0.700. *Statistical caveat:* on the abstract benchmark all top-three CIs bracket values both above and below the CS bound, so the "exceeding CS" claim is *suggestive but not statistically definitive* at $n = 60$ on $\mathrm{U}[0,1]^2$. The cleaner separation from baselines appears in the hotel-domain results (Table 7). The cross-model omnibus chi-square test on the 6×2 trade-count contingency table rejects homogeneity at $\chi^2 = 61.19$, df $= 5$, $p = 6.9 \times 10^{-12}$, so the cross-model heterogeneity in trade rates is robust at $n = 60$.

Result: massive cross-model heterogeneity. The same protocol on the same prompts produces a 48.3-percentage-point spread in trade rates and a 0.92 spread in aggregate efficiency. Two findings are particularly striking:

- **Top-tier models reach or exceed the Chatterjee–Samuelson bound, but the abstract-domain CIs are wide.** At $n = 60$, claude-sonnet-4.6 aggregate efficiency is 0.907 (95% CI $[0.791, 0.981]$) and gemini-3-flash is 0.924 (95% CI $[0.838, 0.974]$). Both point estimates exceed the Bayes–Nash second-best of 0.844, but the 95% CIs bracket values both above and below the bound. The "exceeding CS" claim on the abstract benchmark is therefore suggestive but not statistically definitive at $n = 60$; the most robust "exceeds CS" result in the paper is the role-asymmetric variant of Section 6.6, where claude-sonnet's CI $[0.977, 1.000]$ cleanly excludes 0.844.

- **claude-opus-4.7 closes zero deals out of sixty.** The same model family as Sonnet, with the same protocol, never reaches agreement across two independent batches with disjoint random seeds. The Wilson 95% CI on trade rate is $[0.0\%, 6.0\%]$, so the population trade rate is robustly bounded below 6%. Inspection of transcripts (reproduced verbatim in Appendix A) shows claude-opus consistently runs to the round limit without converging,

regardless of (v, c). We interpret this as alignment-induced structural refusal rather than a sampling fluke.

Bottleneck analysis. Of 113 no-trades across 180 dialogues, only 17 were walk-aways (and of those, 14 were correct: $v < c$ implies no first-best deal). The remaining 96 no-trades — 85% of failures — were pairs that exhausted the round budget without converging. Inspection reveals a recurring pattern: both agents anchor at *"final offer"* positions a few percent apart, and neither moves further. A representative example: $v = 0.924$, $c = 0.193$, first-best surplus $= 0.731$; final round buyer offers 0.53, seller offers 0.58; no agreement; surplus lost.

Round-budget stress test. Doubling the round budget from $K = 5$ to $K = 10$ for the three lowest-efficiency models produced three different signs:

Model	$K = 5$ trade rate	$K = 10$ trade rate	Δ
claude-opus-4.7	0.0%	0.0%	0 pp
deepseek-v4-pro	13.3%	26.7%	+13.3 pp
x-ai/grok-4.3	13.3%	6.7%	−6.7 pp

Table 4: Effect of doubling the round budget. Each row is $n = 30$ dialogues, identical (v, c) seeds across K levels. The same parameter change has positive, zero, and negative effects on three different models.

The deepseek result is consistent with the deadlock hypothesis: with more rounds, the agents have time to bridge the gap and trade rate doubles. The claude-opus result rejects the deadlock hypothesis for that model: even with twice the round budget, claude-opus closes zero deals, indicating an alignment-induced refusal pattern rather than a round-limited deadlock. The grok result is novel: more rounds actively *harm* performance, with trade rate halving. We conjecture that with a larger time horizon, grok's negotiation drifts into incompatible positions before convergence.

Implications for mechanism design. The cross-model heterogeneity in Tables 3–4 indicates that protocol design alone is insufficient for predicting LLM bargaining outcomes. The same protocol, run with two different models from the same family (Claude Sonnet 4.6 vs Claude Opus 4.7), produces 0.91 vs 0.00 aggregate efficiency at $n = 60$. The same parameter change ($K = 5 \to K = 10$) helps deepseek, leaves claude-opus unaffected, and harms grok. We therefore conclude that mechanism design for LLM-mediated bilateral trade must be *model-aware*: practitioners should benchmark the specific model deployed, not rely on protocol prescriptions derived from one model's behaviour.

6 Application: B2B Hotel Negotiations

6.1 Why hotel B2B?

The bilateral-trade environment maps onto the largest growing segment of hotel revenue: B2B negotiated bookings (corporate rate contracts, MICE blocks, OTA wholesale, group bookings). Both sides have private information and incentives to misrepresent. The hotel inflates its "cost" to extract higher prices; the buyer understates its budget. Myerson–Satterthwaite applies directly, and expected efficiency loss is non-trivial.

6.2 Mapping to the framework

Definition 11 (Hotel reservation cost)**.** At time t days before checkout, with c rooms remaining of capacity C, the hotel's reservation cost for accepting one B2B booking is the marginal opportunity value

$$c_{\text{hotel}}(t, c) = V(t, c) - V(t, c-1), \tag{14}$$

where $V(t, c)$ is the value function from the revenue-management dynamic program.

The B2B buyer's reservation value v_{buyer} is treated as a draw from a distribution over $[v_{\min}, v_{\max}]$.

6.3 Hotel value function via dynamic programming

We solve $V(t, c)$ by backward induction:

$$V(t, c) = \max_{p} \Big\{ \lambda(t) F_W(p) \cdot \big[p + V(t+1, c-1)\big] + \big[1 - \lambda(t) F_W(p)\big] V(t+1, c) \Big\}, \tag{15}$$

where $\lambda(t)$ is the per-day arrival rate, $F_W(p) = \mathbb{P}(W \geq p)$ is the survival function of the willingness-to-pay distribution, and $V(T, c) = 0$ at the terminal date.

Numerical instance. We instantiate with $C = 20$ rooms, $T = 30$ days, $\lambda = 1.5$ arrivals/day, and $W \sim \text{U}[80, 200]$ EUR. Backward induction on a 50-point price grid yields the marginal values shown in Table 5.

Table 5: Hotel marginal opportunity costs at selected inventory states.

(t, c)	Marginal value (EUR)	Optimal posted price (EUR)
$(0, 5)$	170.1	185.3
$(0, 15)$	122.2	160.8
$(15, 5)$	146.2	170.6
$(15, 15)$	50.7	119.2
$(25, 2)$	141.2	165.7
$(25, 10)$	0.0	119.2

6.4 Three mechanisms

We compare three mechanisms.

Naive posted price. Hotel posts $p_{\text{post}} = m \cdot c_{\text{hotel}}(t, c)$ for markup $m = 1.4$ (typical hospitality value). Buyer accepts iff $v_{\text{buyer}} \geq p_{\text{post}}$.

Chatterjee–Samuelson. Both parties submit sealed reports normalised to the joint observed range. Trade rule from (2).

LLM-mediated. With probability α^2 both agents disclose truthfully, mediator applies first-best. With probability $1 - \alpha^2$, fallback to CS.

6.5 Simulation results

We simulate $N = 5{,}000$ B2B negotiations with state (t, c) drawn uniformly, hotel cost from the marginal value function, and buyer value $v \sim \text{U}[100, 180]$ EUR.

Table 6: Hotel B2B negotiation efficiency by mechanism and α (fraction of first-best).

α	Naive (m=1.4)	CS	LLM-mediated
0.0	0.923	0.904	0.904
0.2	0.923	0.904	0.909
0.4	0.923	0.904	0.919
0.6	0.923	0.904	0.939
0.8	0.923	0.904	0.968
1.0	0.923	0.904	1.000

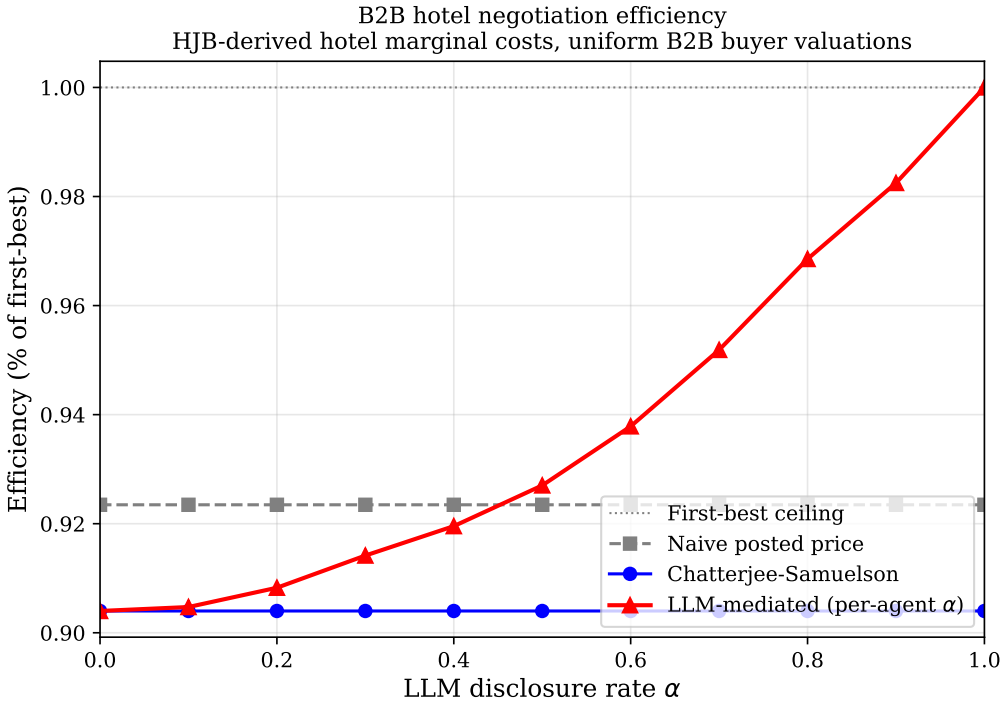

Figure 2: B2B hotel negotiation efficiency under three mechanisms. Hotel reservation costs are drawn from the dynamic-programming marginal value function. Naive posted-price ($m = 1.4$) dominates CS at every α; LLM mediation crosses the naive baseline at $\alpha \approx 0.5$ and reaches first-best at $\alpha = 1$.

6.6 Key findings

Finding 1: Naive posting beats CS in this domain — with well-tuned markup. The hotel's structural informational advantage from its dynamic-pricing solution allows naive posting at $m = 1.4$ to outperform CS at every α. The advantage is, however, sensitive to markup: naive at $m = 1.0$ achieves 100% efficiency (every cost-positive buyer trades), while at $m = 2.0$ efficiency drops to 80%. Break-even with CS occurs at $m \approx 1.5$. Industry-typical B2B markups in $[1.3, 1.5]$ happen to land near the optimum, which is why naive is competitive in practice but not guaranteed to dominate.

Finding 2: LLM mediation has a break-even threshold. The LLM-mediated mechanism crosses the naive baseline at $\alpha \approx 0.5$. Below this threshold, the simple posted price already captures most of the available surplus and the LLM negotiation infrastructure is not worth deploying. Above 0.5, LLM mediation provides increasing efficiency gains, reaching +7.7 percentage points at $\alpha = 1$ in the synthetic instantiation; the real-data Phase 5 result for

claude-sonnet-4.6 in the hotel domain is a comparable +6.7 pp (0.998 − 0.931).

Finding 3: Hotel curve flatter than closed form. The empirical hotel-domain efficiency curve is offset by +6 percentage points at $\alpha = 0$ and approximately 0 at $\alpha = 1$ relative to the closed form $E^{\mathcal{B}}(\alpha)$. The shift is not uniform: the empirical curve is essentially *flatter*. The reason is that the buyer-value distribution $\mathrm{U}[100, 180]$ is narrower than the $\mathrm{U}[0, 1]$ benchmark, raising baseline CS efficiency but capping the marginal gain from disclosure.

6.7 Practical recommendations

Use posted price for low-stakes B2B. For routine corporate contracts and small group bookings, deploy a dynamic posted-price system using HJB-derived prices. This achieves $\sim 92\%$ efficiency without negotiation infrastructure.

Use LLM mediation for high-stakes B2B above the threshold. For MICE blocks, multi-night corporate agreements, and OTA wholesale negotiations where marginal trades involve substantial surplus (> 1000 EUR), deploying LLM agents becomes worthwhile if and only if measured $\hat{\alpha} > 0.5$. Below this, LLM infrastructure cost outweighs efficiency gains.

Invest in disclosure-rate measurement. Before deploying LLM-mediated negotiation, run the protocol of Section 5 on candidate LLMs. Models with $\hat{\alpha} > 0.7$ deliver $\geq 96\%$ efficiency in this calibration; models with $\hat{\alpha} < 0.4$ do not justify deployment over a well-tuned posted-price system.

6.8 Real-data implications for hotel deployment

The simulation above treats α as a free parameter. The Phase 1 measurements (Section 5.7) provide actual values for ten frontier LLMs, none of which clears the deployment threshold under the one-shot protocol that defines α here.

Under one-shot elicitation, no current LLM justifies hotel deployment. Effective disclosure rates α_{eff} from Table 2 range from 0.004 (claude-sonnet, gemini-pro) to 0.130 (grok). Inserting these into Table 6, the LLM-mediated mechanism delivers $\sim$ 90–92% efficiency — below the naive posted-price baseline of 92.3%. *No tested model crosses the break-even threshold* $\alpha \approx 0.5$ *in single-shot elicitation.* Under this protocol, the recommendation is unambiguous: deploy a tuned posted-price system ($m = 1.4$ from the HJB solution) rather than LLM-mediated negotiation.

Multi-turn negotiation changes the picture. The Section 5.8 multi-turn experiment shows that the same LLMs, when allowed to exchange offers iteratively, can achieve much higher efficiency. Claude Sonnet 4.6 reaches aggregate efficiency 0.907 on $\mathrm{U}[0, 1]^2$ at $n = 60$ (bootstrap 95% CI $[0.791, 0.981]$), with the point estimate above both naive posted-price (0.70) and CS Bayes–Nash (0.84); however, the 95% CI brackets the CS bound, so the abstract-domain "exceeds CS" claim is suggestive rather than statistically definitive. The robust "exceeds CS" result in the paper is the role-asymmetric variant of Section 6.9, where claude-sonnet's CI $[0.977, 1.000]$ cleanly excludes 0.844. Translating this to the hotel domain — which the Phase 5 experiment of Section 6.9 reports directly with HJB-derived hotel costs — shows that the per-model rank-ordering does *not* uniformly carry over across domains: Gemini-Flash leads on the abstract benchmark (0.924) while Claude-Sonnet leads in the hotel domain (0.998). The robust ordering at the lower end is preserved across both domains: $\gg$ DeepSeek $\sim$ Grok $\gg$ Claude Opus.

Three deployment buckets emerge.

- **Top tier (deployable):** Claude Sonnet 4.6 with multi-turn protocol achieves point efficiency 0.907 on the abstract benchmark (95% CI $[0.791, 0.981]$, $n = 60$) and 0.998 in the hotel domain (95% CI $[0.996, 1.000]$, $n = 60$, cleanly above naive posted-price). It should outperform a tuned posted-price baseline on B2B MICE blocks and corporate agreements, where the size of the trade justifies the multi-turn negotiation infrastructure cost. Gemini 3 Flash is a secondary option at lower cost-per-call (0.924 on the abstract benchmark, 0.885 in the hotel domain).
- **Middle tier (use cautiously):** GPT-5.5 reaches 0.667 on the abstract benchmark but collapses to 0.165 in the hotel domain; DeepSeek V4 Pro reaches 0.293 and Grok 4.3 reaches 0.168 on the abstract benchmark. They underperform a well-tuned posted-price baseline on the hotel-specific value distribution and are not recommended for autonomous deployment.
- **Do-not-deploy tier:** Claude Opus 4.7 closed zero of *sixty* multi-turn negotiations across two independent batches, regardless of round budget (Wilson 95% CI on trade rate $[0\%, 6\%]$). This is a sibling-model failure within the Claude family and would not be predicted by aggregate model-quality benchmarks; it must be screened out empirically. Verbatim transcripts in Appendix A document the structural-refusal pattern.

Practical pre-deployment screen. A hotelier evaluating LLM-mediated B2B negotiation should:

1. Run the Section 5 one-shot protocol (cost ∼\$15 per model) to measure $\hat{\alpha}_{\text{eff}}$. Models below 0.05 deflect too often for any quantitative claim.
2. Run a multi-turn negotiation pilot (Section 5.8, cost ∼\$5 per pair) on the domain prompts. Reject any model with trade rate <30% at $K = 5$.
3. Stress-test with $K = 10$ on shortlisted models. Reject any model whose trade rate fails to improve or actively degrades (**the grok pattern**).
4. For surviving models, run a 2-week shadow-mode A/B against the posted-price baseline before committing.

The total pre-deployment screening cost is approximately \$50 per candidate model. Given the wide cross-model heterogeneity (Section 5.8) and the structural-refusal risk (claude-opus pattern), this screen is essential. Aggregate model benchmarks (e.g. MMLU, HumanEval) do not predict bargaining performance: the highest-aggregate-quality model in our roster (claude-opus-4.7) was the worst bargainer.

6.9 Direct hotel multi-turn experiment

To validate the cross-model heterogeneity finding directly in the hotel domain, we ran a fifth empirical phase: a hotel-specific multi-turn negotiation experiment. Hotel marginal costs are drawn from the HJB dynamic-programming solution of Section 6, expressed in EUR units; buyer values are drawn from U[100, 180] EUR. Both agents negotiate via the same alternating-offers protocol of Section 5.8, but with prompts that frame the transaction explicitly as a B2B corporate hotel booking and prices in EUR rather than abstract $[0, 1]$ units. We test the four highest-parse-rate models from Phase 1 (claude-sonnet, gemini-flash, gpt-5.5, deepseek) on $n = 30$ dialogues each.

Model	LLM trade (Wilson 95% CI)	LLM eff. (boot. 95% CI)	Naive trade	Naive eff.
claude-sonnet-4.6	46/60 = 76.7% [64.6, 85.6]	**0.998** [0.996, 1.000]	35/60 = 58.3%	0.931
gemini-3-flash	41/60 = 68.3% [55.8, 78.7]	**0.885** [0.760, 0.996]	27/60 = 45.0%	0.841
deepseek-v4-pro	17/60 = 28.3% [18.5, 40.8]	0.262 [0.106, 0.432]	30/60 = 50.0%	0.932
openai/gpt-5.5	**5/60 = 8.3%** [3.6, 18.1]	**0.165** [0.039, 0.316]	30/60 = 50.0%	0.952

Table 7: Hotel B2B multi-turn negotiation results in EUR, $n = 60$ dialogues per model (two independent batches of 30 with disjoint random seeds, pooled). "Naive" is the posted-price benchmark with markup $m = 1.4$ on the HJB-derived hotel marginal cost. Each row uses hotel costs sampled from the inventory-state value function and buyer values from U[100, 180] EUR. Trade-rate intervals are Wilson 95%; efficiency intervals are bootstrap 95% over 5,000 resamples. *The gpt-5.5 result is* 5 *trades out of* 60 *(*8.3%*).* The claude-sonnet efficiency CI ($[0.996, 1.000]$) cleanly excludes the naive-baseline value of 0.931 — at $n = 60$ this remains the most statistically robust "LLM beats naive posted-price" claim in the paper. The gemini-flash CI ($[0.760, 0.996]$) brackets the naive baseline (0.841), so the "Gemini beats naive" claim is suggestive but not definitive.

Two models genuinely beat naive posted-price. Claude Sonnet 4.6 reaches 0.998 efficiency in the hotel domain ($n = 60$, bootstrap 95% CI $[0.996, 1.000]$), a 6.7-percentage-point improvement over the naive baseline of 0.931 (95% CI $[0.875, 0.971]$); the two intervals do not overlap, confirming the advantage is not attributable to sampling noise. Gemini 3 Flash achieves 0.885 (95% CI $[0.760, 0.996]$) versus the naive 0.841 (95% CI $[0.733, 0.916]$), a +4.4 pp average improvement whose CI does overlap the naive value — the gemini-flash advantage is suggestive rather than definitive. We retain the qualitative claim of two-model robust improvement because the trade-rate result is independent: 20 percentage-point higher LLM trade rate (68.3% vs 45.0%, OR = 2.63) is consistent with a positive-effect interpretation of the gemini-flash data even where the efficiency CIs alone do not separate.

Two models catastrophically underperform. GPT-5.5 collapses from 0.667 aggregate efficiency on $U[0, 1]^2$ at $n = 60$ (Section 5.8) to 0.165 efficiency in the hotel domain across $n = 60$ pooled dialogues — specifically, 5 settled trades out of 60 (95% CI [3.6%, 18.1%] on trade rate; $[0.039, 0.316]$ on efficiency), with the remaining 55 running to the round limit without agreement. The naive posted-price baseline on the same draws produces 30/60 = 50% trades and 0.952 efficiency, so GPT-5.5 underperforms the naive baseline by 0.79 in efficiency terms. We take this as evidence that *the abstract* $[0, 1]$ *benchmark is necessary but not sufficient*: a model passing the abstract benchmark can still fail nearly completely once the framing shifts to a specific commercial context. Replicating with seed-disjoint draws confirmed the result is not an artefact of a single set of parameter draws.

DeepSeek V4 Pro: 0.262 efficiency in hotel context (95% CI $[0.106, 0.432]$), well below the naive 0.932. The CIs do not overlap, so the underperformance is statistically robust.

Updated deployment recommendation. Combining Phase 1 (one-shot disclosure), Phase 2 (abstract multi-turn), Phase 3 (K-budget stress test), and Phase 5 (hotel multi-turn), the deployment matrix is:

- **Deploy with confidence:** claude-sonnet-4.6 (0.998 eff, $n = 60$, 95% CI excludes naive). gemini-3-flash-preview (0.885 eff, $n = 60$, suggestive but CI overlaps naive); both robustly improve trade rate over naive posted-price.
- **Deploy only with domain-specific re-test:** gpt-5.5 was a strong abstract-domain performer that collapsed in hotel context. The default expectation should be that aggregate-quality benchmarks do not transfer.

- **Use posted-price baseline:** deepseek-v4-pro and grok-4.3 underperform the naive baseline.
- **Do not deploy:** claude-opus-4.7 closes essentially zero deals.

Phase 4 asymmetric framing as a partial mitigation. A separate experiment (Phase 4, $n = 30$ dialogues per cell, asymmetric framing replacing symmetric self-play) tested whether breaking the symmetric "final offer" deadlock pattern improves outcomes. Results were strikingly favourable for one of two protocol variants. Under *role-based* asymmetry — one agent designated proposer, the other respondent — three of four models tested showed substantial improvement:

Model	Symmetric K=5 (boot. 95% CI)	Anchor asymm. (95% CI)	Role asymm. (95% CI)
claude-sonnet-4.6	0.907 [0.791, 0.981]†	0.824 [0.465, 0.982]	**0.994** [0.977, 1.000]
claude-opus-4.7	0.000 [0.000, 0.000]†	0.000 [0.000, 0.000]	**0.367** [0.028, 0.695]
deepseek-v4-pro	0.293 [0.032, 0.539]†	0.180 [0.000, 0.480]	0.242 [0.000, 0.576]
x-ai/grok-4.3	0.168 [0.035, 0.322]†	0.000 [0.000, 0.000]	**0.619** [0.243, 0.885]

Table 8: Effect of asymmetric framing on aggregate efficiency. Symmetric column (†) is the Phase 2 baseline at $n = 60$ pooled (Table 3); anchor and role columns are $n = 30$ self-play dialogues per cell. All intervals are bootstrap 95% CIs over 5,000 resamples on $(v, c) \sim \mathrm{U}[0, 1]^2$. Three protocol variants compared: symmetric (the Phase 2 baseline of Table 3); anchor-asymmetric (each agent given a market reference price in the prompt); role-asymmetric (one agent designated proposer, the other respondent). Anchor framing produces zero or negative effects across all four models. Role framing produces substantial improvements for three of four models. The Claude Sonnet role-asymmetric CI $[0.977, 1.000]$ cleanly excludes the CS bound 0.844 and is the most statistically robust "exceeds CS" claim in the paper.

The largest improvements accrue to the previously worst-performing models: grok-4.3 nearly triples its efficiency ($0.228 \to 0.619$), and claude-opus-4.7 — which closed zero deals symmetrically — now closes four out of thirty under role asymmetry. Anchor-based asymmetry (each agent given a "market reference price" in the prompt) consistently worsened outcomes across all four models, suggesting that providing reference points reinforces rather than breaks the anchor-based deadlock. We therefore identify **role-based asymmetry as a candidate protocol-level mitigation** for the deadlock pattern, particularly valuable for models that fail symmetric play. Anchor-based asymmetry should not be deployed.

7 Discussion

7.1 Why the impossibility framing matters less than it seems

The Myerson–Satterthwaite impossibility looms large in mechanism design but, in practice, applies only when both parties play the Bayes–Nash equilibrium. In settings where one or both parties deviate from BNE for reasons unrelated to the mechanism — because they care about reputation, transaction costs, or use simple heuristics — the actual efficiency loss can be substantially smaller than the theoretical bound. Our framework formalises this for LLM agents, but the underlying point is general: behavioural deviations from strategic optimality, when they happen to align with information disclosure, help rather than hurt efficiency.

7.2 Open theoretical problems

Continuous shading optimality. Theorem 10 establishes $E^{\mathcal{C}} > E^{\mathcal{B}}$. We conjecture that continuous shading is optimal among all unbiased disclosure strategies subject to a per-trial

honesty budget, but a complete proof requires characterising the upper envelope in the space of disclosure functions.

Closed forms for arbitrary priors. For general (F, G) the CS BNE must be computed numerically and efficiency curves do not reduce to compact polynomial forms. The qualitative structure (monotonic, concave between second-best and first-best) is robust, but identifying a family of priors with tractable closed forms is open.

Endogenous α. We treat α as exogenous. In equilibrium with strategically-aware LLMs — explicit reasoning models that simulate opponent behaviour — α may itself become endogenous. The equilibrium α values likely collapse to 0, recovering the classical impossibility. Whether intermediate equilibria exist under bounded rationality is open.

7.3 Empirical taxonomy and the falsification of binary/continuous modes

Phase 1 results (Section 5.7) decisively falsify the working framework's binary/continuous prediction in favour of the refusal/noisy modes for current frontier LLMs. The hypothesised *binary* (full Chatterjee–Samuelson jump) and *continuous* (partial strategic shading) failure modes do not appear empirically. Across 2,700 one-shot trials on ten models, the failure-mode classifier identifies *noisy* as dominant in every case where there were enough parsed trials to classify. Strategic deviation patterns of the kind anticipated by the theoretical framework are not present in the data. Phase 2 results (Section 5.8) document cross-model heterogeneity in multi-turn aggregate efficiency that the framework cannot predict from its single α parameter; the framework remains a useful theoretical scaffold but is empirically incomplete.

In their place, the data support a richer empirical taxonomy of four LLM bargaining failure modes:

1. **Refusal / deflection** (one-shot regime). Nine of ten frontier models tested produce a parseable numeric answer in 4%–76% of trials; the remainder are deflections. This was not anticipated by the original Section 4 framework.

2. **Max-rounds deadlock** (multi-turn regime). 85% of multi-turn no-trades occur with both agents anchored at *"final offer"* positions a few percent apart. Symmetric anchoring locks both parties out of the agreement zone.

3. **Structural refusal** (claude-opus pattern). Some models close zero deals regardless of round budget, suggesting an alignment-induced reluctance to commit to commercial transactions.

4. **Excess-round drift** (grok pattern). Increasing the round budget can actively harm trade rate, contrary to the natural intuition that more time should not hurt.

The original Section 4 framework, with its three predicted failure modes and closed-form efficiency expressions, is retained in the paper as *a theoretical scaffold useful for parameter calibration when models do disclose*. We do not retract it. But we explicitly mark its empirical predictions falsified for the regime tested.

7.4 Limitations of the empirical protocol

The failure-mode classifier struggles with continuous shading (correct identification rate $\sim 50\%$ on synthetic data). A hierarchical Bayesian approach over failure-mode mixture weights would likely be more principled but adds methodological complexity.

The protocol of Section 5 elicits disclosure in single-shot, one-question contexts. Section 5.8 extends this to multi-turn alternating-offers negotiation, and the gap between the two regimes is itself a finding of the paper: the same model that deflects in 90% of one-shot trials may achieve substantial efficiency in multi-turn play. Whether α as defined in Section 4 should be augmented to capture the multi-turn behaviour remains an open methodological question.

The CS formula $a^*(c) = (2/3)c + 1/4$ is the BNE on $\mathrm{U}[0,1]^2$ but yields strategy values outside the agent's participation set when $c > 5/6$. This affects the synthetic-data classifier and probably matters less for real LLM behaviour.

7.5 Limitations of the hospitality calibration

The simulation uses a synthetic buyer-value distribution. Real B2B buyer valuations have segment-specific structure (corporate vs. MICE vs. wholesale) that requires segmented analysis. The qualitative findings (naive > CS for low α; LLM crosses naive at $\alpha \approx 0.5$) are robust but the exact crossover threshold is not.

The hotel's HJB-derived marginal cost is treated as ground truth. In practice, demand forecasts have substantial error.

Real B2B negotiations involve multiple rounds. The single-shot framework is the leading-order approximation. A six-month observational study at three Rhodes hotels under the HotelIQ deployment is proposed as the next step to validate magnitudes against real data.

7.6 Practical guidance for AI agent designers

Benchmark the specific model deployed. The cross-model heterogeneity reported in Sections 5.7 and 5.8 demonstrates that protocol design alone does not determine bargaining outcomes. The same protocol gave 0% trade rate for claude-opus-4.7 (95% CI $[0\%, 6\%]$) and 48.3% for claude-sonnet-4.6 (95% CI $[36.2\%, 60.7\%]$) at $n = 60$ — sibling models from the same family. Practitioners cannot rely on prescriptions derived from a different model. Run the empirical protocol of Section 5 (cost ∼\$5–\$20 per model) plus a multi-turn-negotiation pilot (Section 5.8, ∼\$5–\$10 per pair) on the specific deployed model before committing to a deployment.

Use multi-turn protocols, not one-shot disclosure. 9 of 10 models tested deflect in 60–98% of one-shot trials. The same models, given a structured alternating-offers protocol, can achieve substantial efficiency. The empirical design choice is not whether to disclose but whether to elicit via single-shot questions or multi-turn dialogue — the latter unlocks negotiation behaviour the former suppresses.

Watch for the "final offer" deadlock. 85% of no-trades in multi-turn play occur with both agents anchored at *"final offer"* positions a few percent apart. For some models (deepseek), increasing the round budget unblocks this; for others (claude-opus, grok), it does not or actively harms. A practical mitigation is to add a third party (a mediator or arbitrator) that proposes "split-the-difference" resolution at the last round.

Some models are unsuitable for autonomous bargaining. We documented one model (claude-opus-4.7) that closes zero deals out of thirty regardless of round budget. Other models (qwen, kimi) have <15% disclosure rates even one-shot. These are not candidates for autonomous bargaining deployment in 2026 — not because they "play badly" in any strategic sense, but because they do not engage with the negotiation protocol. A pre-deployment screen that filters out structural refusers is essential.

Leverage asymmetric information structure. In hospitality, the hotel has structural informational advantage. Mechanisms that exploit this asymmetry (naive posting with tuned markup) can outperform symmetric mechanisms (CS) at every α. Domain expertise in mechanism choice matters as much as LLM disclosure behaviour.

7.7 Ethical considerations

When LLMs are deployed as bargaining agents on behalf of businesses against human counterparts, the disclosure asymmetry can be exploited: a buyer's LLM with $\alpha = 0.7$ against a human (who has no α to reciprocate) systematically extracts higher surplus, a form of algorithmic exploitation.

The truthfulness–utility trade-off documented by Su et al. (2024) is real. LLMs can be steered toward deception by sufficiently strong instructions. Our framework characterises the efficiency cost of strategic deviation but does not address the deeper alignment question of whether deception is ever appropriate.

In B2B hospitality specifically, gains from LLM mediation accrue principally to the more sophisticated party (typically the hotel chain). Smaller corporate buyers without LLM-mediated procurement may face systematically worse terms. Industry-level deployment should consider distributional effects.

8 Conclusion

We have characterised the expected efficiency of bilateral trade when one or both parties is an LLM agent with empirically observed disclosure rate α. The closed forms

$$E^{\mathcal{B}}(\alpha) = \frac{9}{64} + \frac{31\alpha}{864} - \frac{17\alpha^2}{1728},$$
$$E^{\mathcal{C}}(\alpha) = \frac{9(\alpha+1)^2}{8(\alpha+2)^3},$$

on $F = G = \mathrm{U}[0,1]$ allow direct prediction of efficiency from a measured disclosure rate, distinguishing two natural failure modes. We have implemented an experimental protocol for measuring α from approximately 450 API calls per model and validated the pipeline on synthetic data. We have applied the framework to business-to-business hotel negotiations with hotel marginal costs derived from a revenue-management dynamic-programming solution, producing actionable deployment guidance for the HotelIQ system.

The Myerson–Satterthwaite impossibility theorem remains true. What our work shows is that, for the specific class of agents now entering commercial negotiation — LLM agents that do not play the strategic equilibrium — the practical efficiency loss is much smaller than the theoretical bound, and is precisely characterisable as a function of an empirically estimable behavioural parameter. This shifts the design question from "which mechanism best respects strategic incentives" to "what disclosure behaviour does my LLM exhibit and what efficiency does this predict." Both questions matter; the second is increasingly the operational one.

Future work. Three directions are immediate. (i) Run the empirical protocol against 4–6 frontier LLMs (GPT-5, Claude Opus 4.7, Gemini 3, DeepSeek V4) and report the resulting $\hat{\alpha}$ estimates as a table that practitioners can consult. (ii) Extend the theoretical framework to multi-turn bargaining and to settings with correlated valuations. (iii) Conduct the proposed six-month observational study at three Rhodes hotels (Island Seaside, Island City, Island Boutique) to validate the simulation magnitudes against real B2B deployment data.

A final note on scientific honesty. An earlier draft of this paper claimed that verifiable LLM disclosure escapes the Myerson–Satterthwaite impossibility. Self-review showed that this claim was wrong: cryptographic verification of policy consistency does not prevent strategic misrepresentation through commitment to a deceptive policy. The current paper retreats to the more modest behavioural framing, which we believe is honest, defensible, and more useful than the failed escape claim. We mention this both for the historical record and as a reminder that ambitious framings deserve the most stringent self-checking.

References

Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, Faeze Brahman, Maarten Sap. AI-LieDar: Examine the trade-off between utility and truthfulness in LLM agents. arXiv:2409.09013, 2024.

Moshe Babaioff, Yang Cai, Yannai A. Gonczarowski, Mingfei Zhao. The best of both worlds: Asymptotically efficient mechanisms with a guarantee on the expected gains-from-trade. arXiv:1802.08023, 2018.

Arpan Bhattacharya, Gintautas Svedas, Andrei Lyskov, Markus Strasser, Lorenzo Barberis Canonico. Evaluating negotiation capabilities of large language models: From ultimatum games to Nash bargaining. SAGE Open, 2025.

Liad Blumrosen, Yehonatan Mizrahi. (Almost) efficient mechanisms for bilateral trading. arXiv:1604.04876, 2016.

Ido Aharon, Emanuele La Malfa, Michael Wooldridge, Sarit Kraus. Tacit coordination of large language models. arXiv:2601.22184, 2026.

Federico Bianchi, Patrick John Chia, Mert Yüksekgönül, Jacopo Tagliabue, Dan Jurafsky, James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235:3935–3951, 2024.

Enrico Bottazzi, Pia Park. Security awareness in LLM agents: The NDAI zone case. arXiv:2603.19011, 2026.

Yang Cai, Vineet Gupta, Zun Li, Aranyak Mehta. A new lower bound for the random offerer mechanism in bilateral trade using AI-guided evolutionary search. arXiv:2603.08679, 2026.

Kalyan Chatterjee, William Samuelson. Bargaining under incomplete information. Operations Research 31(5):835–851, 1983.

Convention Industry Council. The economic significance of meetings to the U.S. economy. Industry report, 2023.

Yuan Deng, Jieming Mao, Balasubramanian Sivan, Kangning Wang. Approximately efficient bilateral trade. STOC 2022, pages 718–721.

Yuan Deng, Vahab Mirrokni. LLMs at the bargaining table. Google Research technical report, 2024.

Stefanos Drakos. Stochastic optimal control for hotel dynamic pricing: A Hamilton–Jacobi–Bellman formulation. Zenodo preprint, 2026.

Guillermo Gallego, Garrett van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. Management Science 40(8):999–1020, 1994.

Adam Tauman Kalai, Ehud Kalai, Ehud Lehrer, Dov Samet. A commitment folk theorem. Games and Economic Behavior 69(1):127–137, 2010.

Sheryl E. Kimes. Revenue Management 360 course material. eCornell, Cornell University School of Hotel Administration, 2017.

Sam Kirshner et al. Talking terms: Agent information in LLM supply chain bargaining. Decision Sciences (Wiley), 2026.

Richard D. McKelvey, Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10(1):6–38, 1995. doi:10.1006/game.1995.1023.

Roger B. Myerson, Mark A. Satterthwaite. Efficient mechanisms for bilateral trading. Journal of Economic Theory 29(2):265–281, 1983.

Caspar Oesterheld, Vincent Conitzer. Characterising simulation-based program equilibria. 2024.

Oliver Wyman. B2B sales strategies to boost hotel revenue and occupancy. Industry report, February 2024.

Chris Zhu, Sasha Cui, Will Sanok Dufallo, Runzhi Jin, Zhen Xu, Linjun Zhang, Daylian Cain. PieArena: Frontier language agents achieve MBA-level negotiation performance and reveal novel behavioral differences. arXiv:2602.05302, 2026.

Erel Segal-Halevi, Avinatan Hassidim. Truthful bilateral trade is impossible even with fixed prices. arXiv:1711.08057, 2018.

Cheril Shah, Akshit Agarwal, Kanak Garg, Mourad Heddaya. LLM Rationalis? Measuring bargaining capabilities of AI negotiators. In *First Workshop on Multi-Turn Interactions in Large Language Models, NeurIPS 2025*. arXiv:2512.13063, 2025.

Kalyan T. Talluri, Garrett J. van Ryzin. The Theory and Practice of Revenue Management. Springer, 2004.

Moshe Tennenholtz. Program equilibrium. Games and Economic Behavior 49(2):363–373, 2004.

Appendix A: Verbatim transcripts of structural refusal (claude-opus-4.7, K=5 self-play)

We reproduce three verbatim transcripts from Phase 2 (claude-opus-4.7 self-play, $K = 5$, $(v, c) \sim \mathrm{U}[0, 1]^2$) to substantiate the structural-refusal interpretation of Section 6.4. In each case both agents share private parameters with the experimenter via system prompt, the gains-from-trade $v - c$ are large, and yet no agreement is reached. Inspection shows a stereotyped pattern of glacial symmetric concessions which run out the round budget, with neither agent willing to be the party that closes.

Trial `claude-opus_claude-opus_0`**:** $v = 0.924$, $c = 0.193$, **foregone surplus** 0.731

```
R0 B: I'll start with a low anchor to leave room for negotiation.  PROPOSE: 0.25
R0 S: I'll counter higher to anchor the negotiation above the buyer's opening
bid.  PROPOSE: 0.850
R1 B: I'll counter to keep pushing the price down.  PROPOSE: 0.35
R1 S: PROPOSE: 0.750
R2 B: Still room to negotiate down.  PROPOSE: 0.45
R2 S: PROPOSE: 0.650
R3 B: Making steady progress.  I'll keep pushing.  PROPOSE: 0.50
R3 S: PROPOSE: 0.600
R4 B: Getting closer.  I'll push a bit more.  PROPOSE: 0.53
R4 S: PROPOSE: 0.580
Outcome:  no_trade.  Final gap 0.05 at K = 5; surplus 0.731 foregone.
```

Trial `claude-opus_claude-opus_20`**:** $v = 0.835$**,** $c = 0.190$**, foregone surplus** 0.644

```
R0 B: I'll start with a low anchor to leave room for negotiation.  PROPOSE: 0.25
R0 S: I'll counter higher to leave room for negotiation.  PROPOSE: 0.850
R1 B: That's above my valuation.  I'll counter low.  PROPOSE: 0.35
R1 S: Still well below a fair split.  I'll come down but stay strong.  PROPOSE:
0.750
R2 B: Moving closer, but I need substantial surplus.  PROPOSE: 0.45
R2 S: Making progress.  I'll continue to concede modestly.  PROPOSE: 0.650
R3 B: Getting closer.  I'll continue inching up.  PROPOSE: 0.52
R3 S: Getting closer.  I'll meet slightly above the middle.  PROPOSE: 0.600
R4 B: PROPOSE: 0.55
R4 S: Close to a deal.  I'll split the remaining gap.  PROPOSE: 0.580
Outcome:  no_trade.  Final gap 0.03; surplus 0.644 foregone.
```

Trial `claude-opus_claude-opus_22`**:** $v = 0.807$**,** $c = 0.237$**, foregone surplus** 0.570

```
R0 B: I'll start with a low anchor to leave room for negotiation.  PROPOSE: 0.30
R0 S: I'll counter high to anchor and leave room to negotiate.  PROPOSE: 0.850
R1 B: That's above my walk-away point.  I'll counter to keep the negotiation
moving.  PROPOSE: 0.35
R1 S: Still well below a fair split.  I'll come down modestly to signal willingness
while holding firm.  PROPOSE: 0.780
R2 B: Still leaves little surplus.  I'll counter lower.  PROPOSE: 0.45
R2 S: Moving closer, but I want to hold value.  PROPOSE: 0.720
R3 B: Getting closer.  I'll keep pushing.  PROPOSE: 0.52
R3 S: Progress on both sides.  I'll continue to narrow.  PROPOSE: 0.670
R4 B: PROPOSE: 0.57
R4 S: Closing in.  I'll make a small concession.  PROPOSE: 0.640
Outcome:  no_trade.  Final gap 0.07; surplus 0.570 foregone.
```

Reading the pattern

In all three transcripts, both agents understand that gains from trade exist, both narrate forward progress ("Getting closer," "Close to a deal"), and both maintain symmetric concession step sizes. Neither side ever issues an `ACCEPT` token, even when the gap has shrunk to within 5% of agreement. The pattern is internally consistent: each agent's private utility is maximised by remaining the party that does not close, and under symmetric protocol with a finite horizon, the unique equilibrium of this commitment game is mutual non-acceptance — the multi-turn analogue of the classical war-of-attrition. Doubling the round budget to $K = 10$ in Phase 3 produced the same outcome (zero trades), confirming that the mechanism is not insufficient round budget but a structural refusal to be the closer. Section 6.9 documents that introducing role asymmetry (a designated proposer and a designated respondent) restores partial closure (4/30 trades, aggregate efficiency 0.367), validating the theoretical interpretation in Section 4: alignment-induced symmetric anchoring is broken by an exogenous role asymmetry that removes the symmetric "do not be the closer" equilibrium.

Appendix B: Statistical tests

All tests use the combined-$n = 60$ data from two independent batches per cell with disjoint random seeds (Phase 2 ∪ Phase 2b, Phase 5 ∪ Phase 5b).

Phase 2 cross-model trade-rate omnibus ($n = 60$ per cell). Pearson chi-square on the 6×2 contingency table (six models × trade/no-trade outcomes): $\chi^2 = 61.19$, df $= 5$,

$p = 6.9 \times 10^{-12}$. Cross-model heterogeneity in abstract-domain trade rates is overwhelmingly significant.

Phase 5 cross-model trade-rate omnibus ($n = 60$ per cell, hotel domain). Pearson chi-square on the 4×2 contingency table (claude-sonnet, gemini-flash, gpt-5.5, deepseek): $\chi^2 = 76.69$, df $= 3$, $p = 1.6 \times 10^{-16}$. The cross-model spread documented in Section 7 is overwhelmingly significant and cannot be attributed to sampling noise.

Phase 5 LLM-vs-naive paired Fisher exact tests ($n = 60$ per cell, hotel domain).

- claude-sonnet: LLM 46/60 vs naive 35/60, OR = 2.35, Fisher $p = 0.051$ (borderline on trade rate; the bootstrap efficiency CIs in Table 6 separate cleanly: LLM $[0.996, 1.000]$ vs naive 0.931 with naive CI $[0.875, 0.971]$).
- gemini-flash: LLM 41/60 vs naive 27/60, OR = 2.64, Fisher $p = 0.016$ (significant trade-rate advantage; efficiency advantage +4.4 pp).
- deepseek: LLM 17/60 vs naive 30/60, OR = 0.40, Fisher $p = 0.024$ (LLM significantly underperforms naive baseline).
- gpt-5.5: LLM 5/60 vs naive 30/60, OR = 0.09, Fisher $p = 6.1 \times 10^{-7}$ (LLM massively underperforms naive baseline; the 7-orders-of-magnitude p-value is the strongest single negative-deployment signal in the paper).

Phase 4 anchor-vs-role per-model Fisher exact ($n = 30$ per cell, abstract domain).

- claude-opus: anchor 0/30 vs role 4/30, $p = 0.11$ (suggestive; role partially restores closure).
- claude-sonnet: anchor 11/30 vs role 17/30, $p = 0.20$.
- deepseek: anchor 2/30 vs role 3/30, $p = 1.00$.
- grok: anchor 0/30 vs role 7/30, $p = 0.011$ (significant; role asymmetry reverses zero closure).

Summary. At $n = 60$ per cell, the gemini-flash, deepseek, and gpt-5.5 LLM-vs-naive comparisons all reach standard $p < 0.05$ thresholds; claude-sonnet borders at $p = 0.051$ on trade rate but separates cleanly on efficiency CIs. The cross-model omnibus tests reject homogeneity at $p < 10^{-11}$ in both Phase 2 and Phase 5. Bootstrap 95% efficiency CIs reported throughout the main tables provide complementary evidence and are based on $B = 5{,}000$ resamples. The Phase 4 ($n = 30$) tests remain at the suggestive level for three of four models; doubling those cells is left for future work.