Training Language Models for Bilateral Trade with Private Information

Anonymous authors
Paper under double-blind review

Abstract

Bilateral bargaining under incomplete information—where a buyer and seller each hold a private reservation price—provides a controlled testbed demanding individual rationality, strategic surplus maximization, and coordination to realize gains from trade. We develop a structured bargaining environment in which LLMs negotiate via tool calls within an event-driven simulator, separating binding offers from natural-language messages to enable automated evaluation at scale; the environment serves both as a benchmark for frontier models and as a training environment for open-weight models via reinforcement learning. In a round-robin benchmark among five frontier models (15,000 negotiations), effective strategies implement price discrimination through sequential offers: aggressive anchoring, calibrated concession, and temporal patience are associated with both the highest surplus share and the highest deal rate, while accommodating strategies that concede quickly yield the lowest surplus capture and deal completion. Strategically competent models scale their behavior proportionally to item value, maintaining consistent performance across price tiers; weaker models succeed only when wide zones of possible agreement compensate for suboptimal strategies. In training experiments on Qwen3 (8B, 14B) using supervised fine-tuning (SFT) followed by Group Relative Policy Optimization (GRPO), the two stages optimize competing objectives: SFT approximately doubles surplus share but reduces deal rates, while GRPO recovers deal rates at the cost of surplus gains—a tension traceable to the reward structure. SFT also compresses surplus variation across price tiers, and this compression generalizes to opponents unseen during training, suggesting that behavioral cloning instills proportional strategies rather than memorized price points.

1 Introduction

Bilateral bargaining under incomplete information is a foundational problem in economics and operations research. A buyer and seller, each constrained by a private reservation price, must discover through strategic interaction whether mutually beneficial trade is possible—and if so, at what price. The formal study of this problem spans decades, from Nash's axiomatic bargaining solution (Nash, 1950) and Rubinstein's alternating-offers model (Rubinstein, 1982) to the impossibility results of Myerson & Satterthwaite (1983) and the bilateral trade formulation of Chatterjee & Samuelson (1983). Applications range from procurement and supply chain contracting to dispute resolution and online marketplaces, where alternating-offer bargaining accounts for a substantial share of transaction volume (Backus et al., 2020).

Large language models (LLMs) are increasingly deployed as autonomous agents in multi-turn, goal-directed settings—customer service, code generation, and tool-augmented reasoning. Bilateral bargaining provides a controlled testbed for evaluating whether these models can function as rational, effective decision-making agents. An effective bargaining agent must simultaneously (i) respect hard numeric constraints across many rounds of counterpart pressure, (ii) maximize surplus through strategic offers and communication, and (iii) coordinate with a counterpart to realize gains from trade when they exist. These three capabilities—individual rationality, strategic effectiveness, and allocative efficiency—map directly to requirements for any agentic LLM deployment that involves multi-turn instruction fidelity, objective optimization, and coordination under uncertainty.

Two limitations prevent direct use of general-purpose LLMs for this task. First, frontier models exhibit systematic behavioral failures: they accept deals yielding negative utility and fail to walk away from no-gains-from-trade scenarios. These failures reflect a mismatch between pre-training objectives—which reward fluency and agreeableness—and the requirements of strategic negotiation. Second, free-form dialogue creates a measurement problem: it is unclear whether a mentioned price constitutes a binding offer, a hypothetical, or a market reference, making automated evaluation of utilities, concession rates, and agreement status impossible at scale.

This paper addresses both limitations. We make three contributions:

1. **A structured bilateral bargaining environment with a three-dimensional evaluation framework (Sections 3 and 3.3).** We construct an agent system in which LLMs negotiate through structured tool calls (`make_offer`, `respond_to_offer`, `send_message`, etc.) within an event-driven simulation engine. Separating formal offers from natural-language messages resolves the measurement problem: every offer, acceptance, and rejection is a machine-parseable event. We define three evaluation dimensions—individual rationality, strategic effectiveness (surplus share), and allocative efficiency (deal rate)—each with formal metrics that can be computed automatically from the negotiation trace. A price-tier decomposition disaggregates all metrics by item value, revealing whether models maintain consistent performance across the full price range.
2. **A systematic benchmark of frontier LLMs on strategic negotiation (Section 4).** We evaluate five frontier models in a round-robin tournament where every model plays as both buyer and seller against every model, yielding 15,000 total negotiations. Three empirical findings emerge. (a) Aggressive anchoring with calibrated concession is the most effective observed strategy: the top-performing model achieves both the highest surplus share and the highest deal rate by opening far above cost and then conceding to close, contradicting the naive expectation of a surplus–efficiency tradeoff. (b) Accommodating strategies are associated with weak buyer performance: a cautious, high-concession profile produces both the lowest buyer surplus share and the lowest buyer deal rate—cooperativeness without strategic backbone hurts both surplus capture and deal completion. (c) Strategic competence is associated with consistency across price tiers: strong models maintain narrow surplus spreads across value quintiles, while weaker models improve only on high-value items where larger zones of possible agreement buffer suboptimal strategies.
3. **A training pipeline for open-weight negotiation agents, with a diagnosis of the SFT–RL tension (Section 5).** We fine-tune Qwen3 models (8B and 14B parameters) via a two-stage pipeline: supervised fine-tuning (SFT) on synthetic self-play trajectories, followed by reinforcement learning via Group Relative Policy Optimization (GRPO) against a fixed frontier opponent. The two stages optimize competing objectives: SFT teaches "hold out for good terms"; RL teaches "close the deal." This tension, driven by a reward structure that provides no signal to distinguish rational walk-aways from irrational acceptances, identifies concrete design changes for future work: explicit IR penalties, diverse training opponents, and reward curricula that stage constraint compliance before surplus optimization.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 formalizes the bargaining game, evaluation dimensions, and agent system. Section 4 presents the frontier model benchmark. Section 5 describes the training pipeline and results. Section 6 situates empirical outcomes against the Myerson-Satterthwaite welfare frontier. Section 7 concludes with discussions of limitations and future work.

2 Related Work

We organize the literature along four streams that correspond to the contributions of this paper: the game-theoretic foundations of bilateral trade, automated negotiation systems, LLM evaluation in negotiation and agentic settings, and reinforcement learning for language models.

2.1 Bilateral Trade and Negotiation Theory

The bilateral bargaining problem studied here is rooted in a well-established theoretical tradition. Nash (1950) introduces the axiomatic bargaining solution, predicting equal surplus division under symmetric

disagreement points. Rubinstein (1982) models alternating-offers bargaining with discounting, deriving the unique subgame-perfect equilibrium. Chatterjee & Samuelson (1983) analyze bilateral trade under incomplete information with simultaneous offers, characterizing the set of equilibria when reservation prices are private. Myerson & Satterthwaite (1983) establish the impossibility of achieving ex-post efficiency in bilateral trade under any incentive-compatible, individually rational, budget-balanced mechanism—a result that provides theoretical context for our empirical finding that even the strongest agents leave some gains unrealized.

Large-scale empirical evidence on sequential bargaining has only recently become available. Backus et al. (2020) study over 25 million alternating-offer negotiations on eBay's Best Offer platform—the same protocol our simulator implements—and document several patterns that inform our evaluation framework: more patient buyers obtain lower prices, experienced players capture more surplus, prices converge through Coasian dynamics (seller offers decline, buyer offers increase), and bargaining behavior differs systematically between expensive and inexpensive items due to fixed costs of negotiation. Our price-tier decomposition (Section 4.2.2) tests whether LLMs exhibit the same value-dependent patterns, and our behavioral metrics (concession rate, temporal patience) directly measure the mechanisms Backus et al. identify as determinants of bargaining outcomes.

The behavioral negotiation literature documents systematic deviations from game-theoretic predictions. Galinsky & Mussweiler (2001) demonstrate the first-offer anchoring effect: the opening offer shifts the counterpart's reference point and predicts final settlement prices. Bazerman & Neale (1992) survey cognitive biases in negotiation, including anchoring, the fixed-pie assumption, and reactive devaluation. Thompson & Hrebec (1996) show that negotiators routinely fail to identify integrative potential, leaving surplus on the table even in cooperative settings. These findings inform our evaluation framework: initial aggressiveness, concession rate, and temporal patience (Section 3.3.3) are designed to capture the behavioral mechanisms that the psychology literature identifies as determinants of negotiation outcomes.

2.2 Automated Negotiation Systems

Pre-LLM automated negotiation systems operate in structured, language-free environments. Jennings et al. (2001) survey game-theoretic, heuristic, and argumentation-based approaches for bilateral agent negotiation. The Automated Negotiating Agents Competition (ANAC) provides a standardized platform for evaluating agent strategies in multi-issue bargaining with predefined utility functions (Baarslag et al., 2012; 2016). Chang (2020) train actor–critic networks via self-play for multi-issue bargaining, learning adaptive bidding and acceptance strategies. Green & Plunkett (2022) model eBay's Best Offer protocol as a Markov decision process and train neural-network agents on historical transaction data, achieving near-optimal buyer utility. Their agents outperform humans by anchoring more aggressively and bargaining longer—a pattern consistent with the behavioral drivers we measure in Section 4.2.

These systems demonstrate that RL can produce effective bargaining strategies when the action space is well-defined and rewards are directly observable. Our work builds on this insight by preserving a structured action space (Section 3.3.1) while adding natural-language communication as a parallel channel—enabling both automated metric computation and the richer strategic signaling that language affords.

A separate line of work combines language with strategic reasoning in more complex multi-party settings. Meta Fundamental AI Research Diplomacy Team (FAIR) et al. (2022) achieve human-level play in Diplomacy by coupling a language model with a planning module that proposes joint actions, demonstrating that language and strategy can be integrated when the game structure provides sufficient scaffolding. Lewis et al. (2017) introduce end-to-end learning for negotiation dialogues in the DealOrNoDeal task, training models that negotiate over item bundles using both language and discrete choices. He et al. (2018b) decouple strategy and generation in CraigslistBargains, using a dialogue manager to select actions and a separate module to generate utterances. These architectures rely on template retrieval or action-conditioned generation (Keizer et al., 2017; Yang et al., 2020), limiting linguistic flexibility. LLMs remove this bottleneck but introduce the behavioral failures we document in Section 3.2.

2.3 LLM Evaluation in Negotiation and Agentic Settings

Several concurrent benchmarks evaluate LLM negotiation capabilities. Davidson et al. (2024) evaluate LLM agency through multi-issue negotiations, finding that models frequently accept dominated offers and fail to maintain consistent goals. Xia et al. (2024) measure bargaining abilities and susceptibility to adversarial tactics, quantifying violations of negotiation rationality. Bianchi et al. (2024) introduce NegotiationArena, a multi-domain evaluation platform that reveals limited strategic diversity across models. Abdelnabi et al. (2024) study cooperation and deception in multi-agent LLM games, finding that models exhibit inconsistent strategic behavior across repeated interactions.

Our benchmark differs from these efforts in three respects. First, the structured action space separates binding offers from natural-language messages, enabling unambiguous utility computation without human annotation or LLM-based parsing. Second, we evaluate along three formally defined dimensions (individual rationality, strategic effectiveness, allocative efficiency) with automated metrics, rather than relying on aggregate scores or qualitative assessments. Third, the round-robin tournament design—every model against every model in both roles—reveals pairwise behavioral dynamics (e.g., which models induce opponent IR violations) that single-opponent evaluations cannot capture.

More broadly, LLM agent evaluation has expanded beyond negotiation. Liu et al. (2023) benchmark LLMs across diverse agentic tasks including web browsing, tool use, and database interaction. Yao et al. (2023) introduce the ReAct framework, interleaving reasoning and action for tool-augmented tasks. Xi et al. (2025) survey LLM-based agents, identifying planning, memory, and tool use as core capabilities. Our work contributes to this literature by providing a testbed where multi-turn instruction fidelity, numeric constraint adherence, and strategic optimization can be measured simultaneously—capabilities that general-purpose agent benchmarks typically evaluate in isolation.

2.4 LLM-Based Negotiation Agents

Recent work on improving LLM negotiation performance falls into two categories.

Prompt-based and modular approaches. Fu et al. (2023) use self-play with a critic to select strong negotiation transcripts, appending them as in-context examples to improve pricing outcomes. Oh et al. (2025) augment prompts with utility-based feedback, helping models anticipate opponent behavior. Li et al. (2024) introduce Dialogue Action Tokens—planning signals that guide a frozen LLaMA through multi-turn negotiation acts. Kong et al. (2025) deploy FishBargain, a modular agent combining price extraction, planning, and generation, tested with real sellers on Xianyu. These approaches improve performance without modifying model weights, but their gains are bounded by the base model's strategic reasoning capacity.

Fine-tuning approaches. Chatterjee et al. (2024) generate synthetic dialogues via chain-of-thought prompting and fine-tune LLaMA-based models, improving fairness and agreement rate. Verma et al. (2022) apply conservative Q-learning on offline negotiation logs. Ahmad et al. (2023) construct a mixed human/GPT-J dataset and train agents using PPO with reward signals for rational behavior. These fine-tuning efforts use small datasets, offline RL, or single-metric optimization. None diagnoses the interaction between supervised and reinforcement learning stages—specifically, the tension between the selectivity instilled by behavioral cloning and the deal-closing bias introduced by online RL—which our training experiments identify as a central design challenge (Section 5.4.4).

2.5 Reinforcement Learning for Language Models

RL from human feedback (RLHF) has become standard for aligning LLMs with human preferences (Ouyang et al., 2022; Bai et al., 2022). The dominant algorithm, Proximal Policy Optimization (PPO) (Schulman et al., 2017), has been applied to instruction following, summarization, and dialogue. Shao et al. (2024) introduce Group Relative Policy Optimization (GRPO), which normalizes advantages within groups of rollouts for the same prompt, reducing variance in reward estimation. We adopt GRPO for negotiation

training because its within-context normalization is well-suited to the high context-dependence of negotiation outcomes (Section 5.3).

Recent work applies RL to verifiable reasoning tasks where correctness can be checked automatically. Guo et al. (2025) train DeepSeek-R1 via large-scale RL on mathematical and coding problems, demonstrating that RL can substantially improve reasoning capabilities. Luong et al. (2024) apply RL fine-tuning to improve generalization on multi-step reasoning tasks. These settings feature single-turn interactions with bounded output length and binary or near-binary rewards (correct vs. incorrect). Multi-turn negotiation introduces three complications absent from these tasks: (i) context grows unboundedly over rounds, risking overflow during training; (ii) the opponent is a non-stationary part of the environment, creating distribution shift; and (iii) the reward must balance protocol compliance with outcome quality, as a model that generates well-formed tool calls but accepts irrational deals has failed. Section 5.3.1 addresses these challenges explicitly.

3 Bilateral Bargaining as a Testbed for LLMs

3.1 The Bargaining Game

Consider a bilateral negotiation where a buyer and seller seek to trade an item under conditions of incomplete information. Each agent is constrained by a private reservation price[1] that is hidden from their counterpart. The two parties exchange offers and natural-language messages over multiple rounds, each trying to reach a deal that maximizes their own surplus while respecting their constraint. Neither party knows whether a mutually beneficial agreement even exists; they must discover this through strategic interaction.

This setup, rooted in the bilateral trade formulations of Chatterjee & Samuelson (1983) and Myerson & Satterthwaite (1983), creates a rich testbed for LLM capabilities. An effective agent must simultaneously (i) respect hard constraints across many rounds, (ii) maximize surplus through strategic offers and communication, and (iii) coordinate with a counterpart through natural-language communication to realize gains from trade when they exist. Bilateral bargaining under incomplete information is a foundational problem in operations research and economics, with applications to procurement, supply chain contracting, and dispute resolution.

We now formalize the game and define the three evaluation dimensions.

Definition 1 (Bilateral Bargaining Game)**.** A buyer with reservation price b and a seller with reservation price s negotiate over a single indivisible item through multi-round alternating offers (Rubinstein, 1982). Each round combines natural-language communication with formal price proposals. Both parties observe the item description, historical prices, and the full negotiation history (all offers and messages). Neither party observes the other's reservation price. The utility functions are:

$$U_B = \begin{cases} b - p & \text{if agreement at price } p \leq b, \\ 0 & \text{if no agreement,} \\ -\infty & \text{if agreement at price } p > b; \end{cases} \qquad U_S = \begin{cases} p - s & \text{if agreement at price } p \geq s, \\ 0 & \text{if no agreement.} \\ -\infty & \text{if agreement at price } p < s. \end{cases} \tag{1}$$

Accepting a price that violates the reservation constraint ($p > b$ for the buyer or $p < s$ for the seller) yields negative utility—an irrational outcome.

Definition 2 (Zone of Possible Agreement)**.** The zone of possible agreement (ZOPA) is the interval $[s, b]$ when $b \geq s$. The total available surplus (gains from trade) is $b - s$. Because reservation prices are private, neither party knows whether a ZOPA exists or how large it is; this must be inferred through the negotiation process.

Definition 3 (GFT and NGFT Scenarios)**.** A scenario with $b \geq s$ is a *gains-from-trade* (GFT) scenario: a mutually beneficial deal exists, and any price $p \in [s, b]$ yields non-negative utility for both parties. A scenario with $b < s$ is a *no-gains-from-trade* (NGFT) scenario: no price yields non-negative utility for both parties simultaneously, and the rational outcome is no deal.

[1] A *reservation price is the highest a buyer will pay or the lowest a seller will accept. We use this term throughout; it is equivalent to "private value" in the auction literature.*

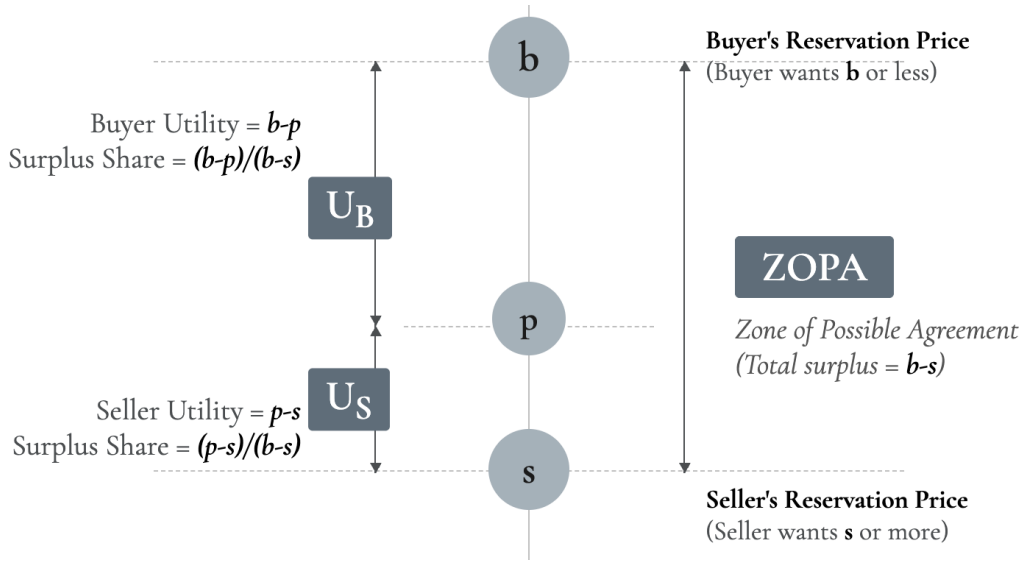

Figure 1: The diagram illustrates the bilateral bargaining space defined by the overlap between the seller's reservation price (s) and the buyer's reservation price (b). The ZOPA represents the total available economic surplus, calculated as $b - s$. Within this range, any agreed settlement price (p) facilitates allocative efficiency by distributing the surplus between the two parties.

Three evaluation dimensions. The game tests three distinct capabilities. Together, they provide a comprehensive assessment of whether an LLM can function as a rational, effective negotiating agent.

Dimension 1: Individual rationality. Can the agent respect hard constraints? An agent that accepts a deal yielding negative utility—paying more than its maximum or selling below its minimum—has failed at the most basic level.

Definition 4 (Individual Rationality)**.** An agent satisfies individual rationality (IR) if it never accepts a price yielding negative utility: the buyer rejects any $p > b$, and the seller rejects any $p < s$. This is the standard IR constraint in mechanism design (Myerson, 1981).

IR is a necessary condition for credible negotiation. An agent that violates it cannot be considered a rational participant regardless of its average surplus.

Dimension 2: Strategic effectiveness (surplus maximization). Given that the agent respects its constraints, how much surplus does it capture? Two individually rational agents can differ substantially in the utility they obtain. We measure strategic effectiveness through average utility (across all negotiations and conditional on deals) and the surplus share each agent captures.

Definition 5 (Surplus Share)**.** For GFT scenarios reaching agreement at price p, the buyer's surplus share is $(b-p)/(b-s)$ and the seller's surplus share is $(p-s)/(b-s)$. The two shares sum to one: buyer and seller are dividing the total available surplus $b - s$. A larger surplus share indicates a more strategically effective agent.

Because the seller posts the initial listing price, the protocol introduces a structural role asymmetry that may advantage the seller through anchoring. We examine this asymmetry empirically in Section 4.2. While IR is a necessary condition (floor), strategic effectiveness is an optimization objective (ceiling).

Dimension 3: Allocative efficiency (gains-from-trade realization). Can the buyer–seller pair discover and realize the available gains from trade through dialogue? This is a *pair-level* property: it depends on both agents' strategies jointly, not on either agent alone.

Definition 6 (Deal Rate)**.** Given a set of N negotiations, the deal rate is

$$\text{Deal Rate} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\text{negotiation } i \text{ reaches agreement}\}, \tag{2}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. The deal rate counts all agreements regardless of whether the agreed price yields positive or negative utility.

In GFT scenarios, higher deal rates indicate better allocative efficiency (surplus exists and should be realized). In NGFT scenarios, deal rate coincides with violation rate. Lower deal rates indicate better rationality (any deal violates IR for at least one party). Beyond the deal rate itself, allocative efficiency is shaped by behavioral drivers: *initial aggressiveness* of opening offers (Galinsky & Mussweiler, 2001), the *concession rate* (the pace of price adjustments across rounds), and *temporal patience* (willingness to sustain dialogue rather than terminating prematurely). These behavioral metrics jointly determine whether a pair reaches agreement and how quickly.

Instruction adherence under multi-turn pressure. The three dimensions share a common prerequisite: the agent must faithfully execute the instructions encoded in its system prompt[2]—reservation price constraints, tool-use protocols, and behavioral rules—across many rounds in which the counterpart actively attempts to persuade and extract concessions. Each capability failure admits a dual interpretation: an IR violation is also a failure to follow the reservation price constraint; a an NGFT agreement is also a violation of the surplus-maximization objective. If a model cannot maintain adherence to a numeric constraint under sustained counterpart pressure, it is likely to exhibit similar failures in any agentic deployment requiring multi-turn instruction fidelity.

3.2 Motivation: Why General-Purpose LLMs Are Insufficient

The most straightforward approach is to prompt a general-purpose LLM with the game setup and let it negotiate in free-form dialogue. Two fundamental limitations make this insufficient.

1. **Systematic behavioral failures.** General-purpose LLMs exhibit failures across all three evaluation dimensions. Davidson et al. (2024) find that models frequently accept dominated offers and fail to maintain consistent goals across rounds. Xia et al. (2024) quantify violations of negotiation rationality and show that models are susceptible to adversarial tactics that pressure them into irrational concessions. Bianchi et al. (2024) reveal limited strategic diversity: models converge on similar negotiation patterns regardless of context. Abdelnabi et al. (2024) document inconsistent strategic behavior across repeated interactions in multi-agent games. These findings span our three dimensions: IR violations (accepting deals at negative utility), weak strategic effectiveness (failing to maximize surplus), and allocative inefficiency (inability to close deals when gains from trade exist or walking away too easily). The common root is a mismatch between pre-training objectives—which reward fluency and agreeableness—and the requirements of strategic negotiation: sustained goal-directed reasoning, adherence to hard numeric constraints, and multi-round planning.
2. **The measurement problem.** In free-form dialogue, it is unclear whether a mentioned price constitutes a binding offer, a hypothetical suggestion, or a reference to market data. This makes it impossible to compute utilities, track concession rates, or determine whether agreement was reached. Without structured actions, every evaluation requires human annotation or unreliable LLM-based parsing, precluding both large-scale benchmarking and the use of outcome quality as a training signal for RL.

These two limitations motivate a domain-specific agent system that provides structured actions for precise, automated measurement while preserving natural-language communication.

3.3 The Agent System

We construct a simulation environment and structured agent interface whose design choices directly address the three evaluation dimensions (Section 3.1) and two limitations (Section 3.2). The system is modular: the negotiation domain (item catalog, reservation price distributions, behavioral rules) is configurable for adaptation to different application domains.

[2]A *system prompt* is a set of instructions provided to the LLM at the beginning of a conversation, specifying its role, objectives, and constraints.

3.3.1 Structured Action Space

We define the agent's action space via structured tool calls:[3]

Tool call	Description
`make_offer(agent, price, side_offer)`	Propose a price for the item.
`respond_to_offer(agent, response)`	Accept or reject opponent's offer.
`send_message(agent, content)`	Send a negotiation message to opponent.
`search_price()`	Retrieve historical market reference prices.
`quit_negotiation(agent)`	Terminal tool; end negotiation without deal.
`wait_for_response(agent)`	Terminal tool; wait for opponent's action.

Table 1: Structured action space: tool calls available to each agent.

Separating formal offers from messages enables unambiguous utility computation and automated detection of IR violations (Definition 4). Without structured actions, it is often unclear from free-form text whether a price constitutes a binding offer. Structured actions also produce a machine-readable negotiation trace from which surplus share, concession rates, and initial aggressiveness can be computed automatically. The `send_message` tool preserves the communicative richness of natural language—agents can persuade, signal flexibility, and extract information—while the `make_offer/respond_to_offer` pair provides a formal mechanism for registering agreement.

The structured action space resolves the measurement problem identified in Section 3.2. Offers, acceptances, rejections, and outside-option exercises are all explicit, machine-parseable events. This enables fully automated, reproducible evaluation without reliance on LLM-as-judge. The tool-based design also mirrors real-world negotiation platforms, where interactions occur through structured forms or APIs; each simulator action maps naturally to backend components used in commercial systems.

Each agent receives a system prompt specifying its role, reservation price constraint, utility formula, item description, and available tools. The prompt also includes expert-crafted negotiation strategies (e.g., anchoring with aggressive opening offers, pacing concessions, exercising the outside option when appropriate). Figure 2 shows the hierarchical structure of the system prompt. These strategies provide a behavioral baseline; fine-tuning can subsequently improve upon them (Section 5).

3.3.2 Event-Driven Simulation Engine

The simulation engine is a discrete-event system with a priority queue; events are processed chronologically. Each event comprises a timestamp, type, acting agent, and content. The negotiation protocol proceeds as: initialization $\rightarrow$ seller posts listing price $\rightarrow$ buyer opens $\rightarrow$ alternating turns $\rightarrow$ termination. A negotiation terminates when either party accepts an offer, exercises the outside option (`quit_negotiation`), or the round limit is reached. The maximum number of rounds is set to $N = 10$; each turn permits at most 3 tool calls. Each agent's output consists of a `Thought:` block (internal reasoning, not visible to the counterpart) followed by a `Code:` block (tool calls), which the simulator parses and executes.

3.3.3 Evaluation Metrics

The structured action space enables a comprehensive suite of automated metrics, organized by the three evaluation dimensions.

[3] A *tool call* is a structured function invocation that the LLM generates as part of its output; the simulation engine parses and executes each call, returning the result to the agent's context for the next turn.

System Prompt Skeleton

Role and Objective. You are a skilled bargaining agent negotiating on behalf of...
Critical Rule: Constraint Compliance. Mandatory adherence to role-specific constraints...
Variable Definitions. Definition of `HARD_LIMITS` and `UTILITY.` formulas...
Constraint Rules. Explicit prohibition of deals resulting in negative utility...
Mandatory Decision Process. Step-by-step verification: Identify → Recall → Calculate → Decide...
Negotiation Strategy Guidelines.

- *General Approach:* Strategic initialization and utility maximization...
- *Anchoring and Positioning:* Use of precise numbers and pressure tactics...
- *Concessions:* Tapered movements to signal limits...
- *Non-Monetary Negotiation:* Exploration of bonuses and bundles...

Rules You Must Follow Requirements for persuasiveness, professionalism, and firm exit strategies...
How You Operate (CoT) Use of **Thought** (private reasoning) and **Code** (tool calls)...
Tool Space and API Definitions Specifications for `make_offer`, `send_message`, etc...
Critical Tool Usage Patterns Mandatory utility pre-calculation patterns for API calls...
Code Rules Syntax strictness, state persistence, and environmental constraints...

Figure 2: Hierarchical structure of the agent system prompt. The prompt integrates private economic constraints (s, b), strategic bargaining directives, and the "Thought-Action" protocol governing agent behavior within the simulation.

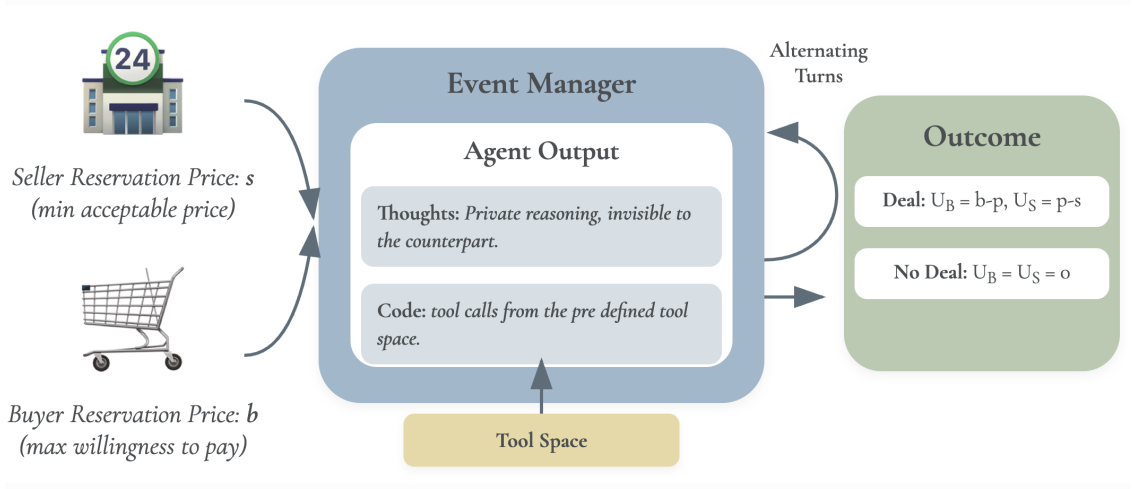

Figure 3: This is the system architecture of the bilateral bargaining simulation environment. Agents are initialized with reservation prices (s for the seller, b for the buyer). The negotiation phase is governed by an Event Manager that enforces a protocol of alternating turns, a limit of max rounds, and a maximum of three structured tool calls per turn from the predefined Tool Space. Agent outputs have private "Thoughts" (internal reasoning) and "Code" (external tool-based actions). The cycle concludes in the Outcome phase, where the environment evaluates deal status and calculates resulting utilities ($U_B = b - p$; $U_S = p - s$) based on the final settlement price (p).

Metric	Description
Violation rate	Fraction of negotiations where an agent accepts a price yielding negative utility ($p > b$ for the buyer or $p < s$ for the seller). A no-deal outcome (utility $= 0$) is *not* a violation. In GFT scenarios, violations arise when a deal is struck outside the ZOPA; in NGFT scenarios, every deal violates IR for at least one party, so the violation rate coincides with the NGFT deal rate (see Table 4). The two metrics are otherwise distinct: violation rate measures individual constraint adherence, while deal rate measures pair-level agreement realization.

Table 2: Dimension 1 metrics: individual rationality (IR).

Metric	Description
Average utility (all)	Mean U_B and U_S across all negotiations in a given regime (GFT or NGFT), including no-deal outcomes (utility = 0).
Average utility (deals)	Mean U_B and U_S restricted to negotiations reaching agreement. Isolates agreement quality from agreement frequency.
Surplus share	Share of surplus captured in both-valid deals: $(b-p)/(b-s)$ for the buyer, $(p-s)/(b-s)$ for the seller (Definition 5). Defined only for GFT scenarios.

Table 3: Dimension 2 metrics: strategic effectiveness (surplus maximization).

Metric	Description
Deal rate	Fraction of negotiations reaching agreement (Definition 6). Higher is better in GFT; lower is better in NGFT.

Table 4: Dimension 3 metrics: allocative efficiency (gains-from-trade realization).

Metric	Description
Seller initial aggressiveness	Ratio of the seller's opening offer p_0^{seller} to the seller's reservation price: $\alpha^{\text{seller}} = p_0^{\text{seller}} \big/ s$. Higher values indicate a more aggressive opening position (asking further above cost).
Buyer initial aggressiveness	Measured in two variants (higher = more aggressive): • *Gap closure:* $\alpha^{\text{buyer},1} = \left(p_0^{\text{seller}} - p_0^{\text{buyer}}\right) \big/ p_0^{\text{seller}}$, fraction of the seller's asking price cut by the buyer's first counter-offer. • *Reservation ratio:* $\alpha^{\text{buyer},2} = \left(b - p_0^{\text{buyer}}\right) \big/ b$, fraction of the buyer's reservation price retained as surplus in the opening bid.
Concession rate	Fraction of remaining surplus conceded per round (computed for valid deals only). • *Seller:* $c_t^{\text{seller}} = \left(p_t^{\text{seller}} - p_{t+1}^{\text{seller}}\right) \big/ \left(p_t^{\text{seller}} - s\right)$. • *Buyer:* $c_t^{\text{buyer}} = \left(p_{t+1}^{\text{buyer}} - p_t^{\text{buyer}}\right) \big/ \left(b - p_t^{\text{buyer}}\right)$. Higher values indicate aggressive bargaining. The sign is used as a Dimension 1 diagnostic (Table 2).
Temporal patience	Mean number of rounds to termination. Measures willingness to sustain dialogue rather than terminating prematurely.

Table 5: Behavioral driver metrics: mechanisms underlying strategic effectiveness and allocative efficiency outcomes.

Price-tier decomposition. Aggregate metrics can mask systematic variation across item price ranges. Human negotiation behavior differs by item value: buyers tend to bargain more tightly on inexpensive items (where a fixed dollar concession is proportionally larger) and more flexibly on expensive items. To capture this heterogeneity, we decompose all metrics by item value.

For each buyer–seller model pair, we partition the item catalog into five quintiles based on each item's reference price (midpoint of the historical price range). Within each quintile, we compute all IR, strategic effectiveness, and allocative efficiency metrics separately. This decomposition reveals whether models exhibit price-sensitivity patterns analogous to human behavior, whether certain models systematically under-perform in specific price tiers, and whether IR violations concentrate at the extremes of the price distribution. It also serves as a diagnostic for training: if fine-tuning improves performance on expensive items but degrades it on inexpensive ones, the price-tier decomposition makes this visible.

3.4 Item Catalog and Reservation Price Sampling

The negotiation scenarios are constructed from a product catalog and a reservation price sampling procedure that together determine the input to the agent system.

Product catalog. Listings are sourced from the CraigslistBargains dataset (He et al., 2018a) (1,402 items) and Amazon Price History data (930 items), totaling 2,332 products. Each listing includes a title, description, category, and historical price range. For Amazon listings, historical price bounds are taken directly from the dataset. Craigslist listings contain only a posted price; we set the highest price to the posted price and the lowest price to 50% of that value. An example listing is in Figure 20 (Section A.1). The catalog is partitioned into training (80%), validation (8.5%), and test (11.5%) splits.

Reservation price sampling. The `search_price()` tool returns historical high and low market reference prices for each item, providing reference information for anchoring. Seller reservation prices are sampled as

$$s \sim \text{Uniform}[0.6 \times p_{\min},\ 0.9 \times p_{\min}], \tag{3}$$

and buyer reservation prices as

$$b \sim \text{Uniform}[0.4 \times p_{\min},\ p_{\max} + 0.1(p_{\max} - p_{\min})], \tag{4}$$

where $p_{\min}$ and $p_{\max}$ are the historical low and high prices from the item's catalog listing. This produces a mix of GFT and NGFT scenarios (∼65% GFT, ∼35% NGFT), testing whether agents both realize surplus when it exists and avoid value destruction when it does not.

Connection to evaluation dimensions. Reservation prices define each agent's hard constraints; the sampling range creates scenarios from wide ZOPA to no ZOPA, testing whether agents maintain IR across varying difficulty. The range of ZOPA widths creates variation in available surplus, enabling meaningful comparison of surplus share across agents. The GFT/NGFT split directly operationalizes allocative efficiency—do agents reach agreement when (and only when) gains from trade exist?

3.5 Two Uses of the System

The agent system serves two complementary purposes, corresponding to the next two sections:

1. **Benchmark for evaluating frontier LLMs (Section 4).** Running models in the simulator and measuring the metrics above (including price-tier decomposition) yields a systematic evaluation of each model across the three dimensions. This extends prior LLM negotiation evaluations by providing structured, dimension-specific diagnostics rather than aggregate scores.
2. **Environment for training negotiation agents (Section 5).** The simulator enables large-scale synthetic trajectory generation via LLM self-play. The behavioral metrics serve as reward signals for reinforcement learning. Together, these support a two-stage training pipeline (supervised fine-tuning[4] on filtered synthetic data, then RL with rule-based rewards) to fine-tune open-weight models for the negotiation task.

4 Benchmarking Frontier LLMs

4.1 Evaluation Protocol

We use the agent system (Section 3.3) as a standardized benchmark for evaluating frontier LLMs on bilateral price negotiation.

Models. We evaluate five frontier models: DeepSeek-V3-0324 (DeepSeek-AI, 2025), Gemini-2.5-Flash (Gemini Team, 2025), Gemini-2.5-Pro (Gemini Team, 2025), GPT-4.1 (OpenAI, 2023), and o3 (OpenAI, 2025). All models receive the same system prompt template, tool definitions, and item information. See Table 6 for model details and costs.

[4] *Supervised fine-tuning* (SFT) is additional training of a pre-trained LLM on task-specific demonstrations, analogous to behavioral cloning in the RL literature.

Model	Provider	Cost ($/1M tokens, in / out)
DeepSeek-V3-0324	DeepSeek	$0.27 / $1.10
Gemini-2.5-Flash	Google	$0.30 / $2.50
Gemini-2.5-Pro	Google	$1.25 / $10.00
GPT-4.1	OpenAI	$2.00 / $8.00
o3†	OpenAI	$2.00 / $8.00

Table 6: Frontier models evaluated in the round-robin benchmark. Prices reflect API list rates at the time of experiments. †o3 is a reasoning model; internal reasoning tokens are billed as output tokens, increasing effective cost.

Round-robin tournament. Each model plays as both buyer and seller against every model (including itself), producing $5 \times 5 = 25$ distinct (buyer, seller) pairings. All 25 pairings negotiate over the same 600 product listings (400 GFT + 200 NGFT), ensuring that differences across pairings reflect model behavior rather than item-level variance. This yields a total of 15,000 negotiations. Self-play pairings (model vs. itself) are included to examine whether a model's strategies lead to efficient outcomes or deadlock when facing its own behavioral profile.

4.2 Results

A unifying lens for the results that follow is price discrimination through negotiation. The economics literature recognizes bargaining as a mechanism for first-degree price discrimination: rather than posting a single price, the seller uses sequential offers to discover each buyer's willingness to pay and extract surplus accordingly (Riley & Zeckhauser, 1983; Sobel & Takahashi, 1983). Three findings from the benchmark align with this interpretation. First, the top-performing seller strategy—anchoring far above cost and conceding calibratedly over multiple rounds—creates maximum room for price adaptation: different buyers end up paying different prices based on their strategic resistance, yet deals still close at near-maximal rates (Section 4.2.3, Finding 1). Second, accommodating buyer strategies that concede quickly and uniformly represent the failure mode: they reveal willingness to pay immediately, eliminating the seller's need to discriminate and leaving surplus on the table for the buyer role (Finding 2). Third, consistency across price tiers reflects the ability to discriminate proportionally: strong models scale their anchoring and concession to item value, extracting a similar fraction of surplus whether the item costs $20 or $2,000; weak models succeed only when wide ZOPAs mask their inability to adapt (Finding 3).

4.2.1 Model-Level Overview

We evaluate models along the three dimensions defined in Section 3.1: (1) individual rationality, measured by violation rates; (2) strategic effectiveness, measured by surplus share; and (3) allocative efficiency, measured by deal rate. We also report behavioral drivers—initial aggressiveness and concession rate—that characterize the mechanisms associated with outcome differences. We discuss each dimension below with detailed tables; pairwise heatmaps for all metrics are in Section C.2.

Dimension 1: Individual rationality. Figure 4 reports violation rates by model and role for both GFT and NGFT scenarios, distinguishing between a model's own violations (self) and violations it induces in opponents (opp.). Full numerical values are in Table 17.

GFT violation rates are uniformly low. All GFT self-violation rates are below 1%, and opponent-violation rates are below 0.5%. At these magnitudes, differences across models are not reliably distinguishable from sampling variability.

NGFT reveals a seller-side vulnerability. Buyer NGFT self-violations are all $\leq$0.40%, while seller NGFT self-violations range from 0.00% (o3) to 2.30% (GPT-4.1). GPT-4.1 is the most vulnerable seller (2.30%), followed by Gemini-2.5-Flash (1.90%), Gemini-2.5-Pro (0.70%), and DeepSeek-V3-0324 (0.60%). This buyer–seller asymmetry is consistent across all five models, suggesting a systematic difficulty with the seller role in NGFT scenarios rather than a model-specific failure.

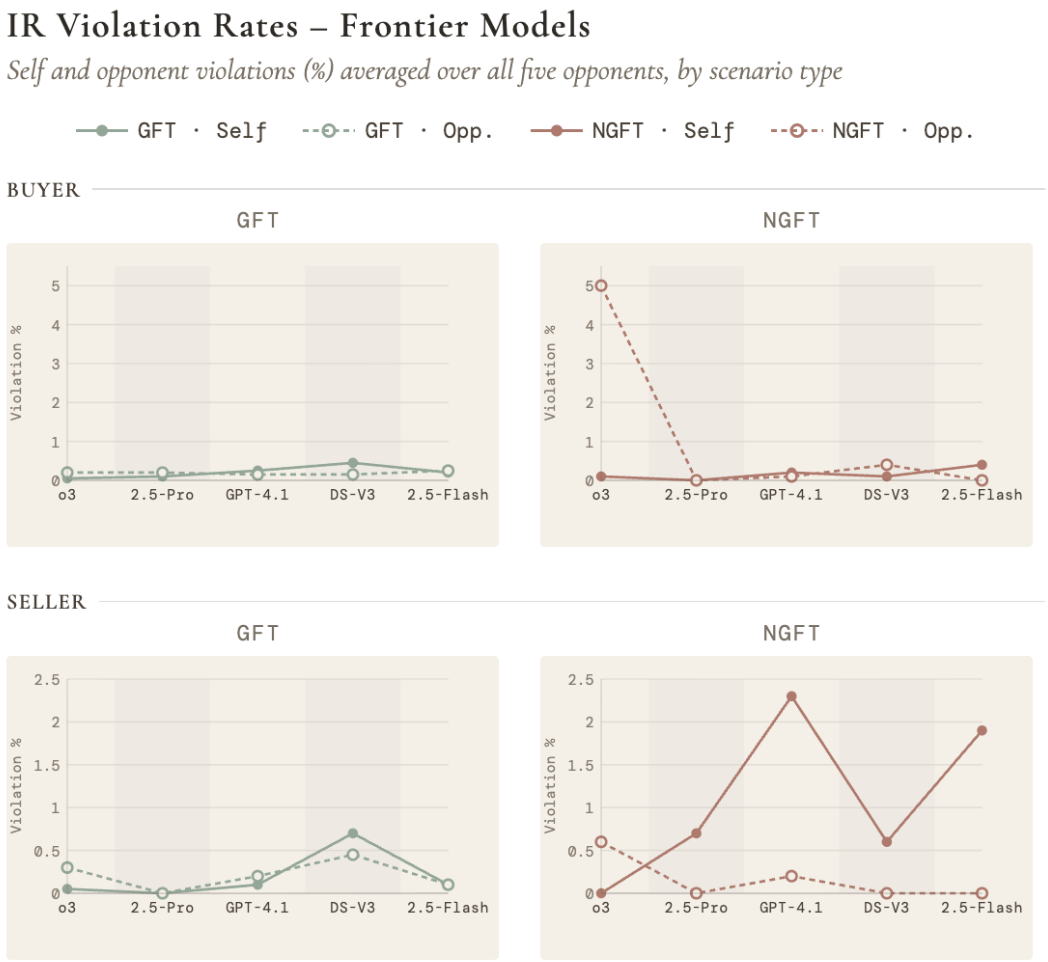

Figure 4: IR compliance for frontier models. Self: the model's own violation rate (deals yielding negative utility for itself). Opp.: the opponent's violation rate when facing this model. All values averaged over all five opponents. Green/red borders highlight best/worst values in each column. In NGFT, any deal is a violation for at least one party; o3 as buyer induces 5.00% opponent violations despite only 0.10% self-violations. Full table in Table 17.

o3 as buyer induces substantially higher NGFT opponent violations than any other model. o3 buyer NGFT opponent-violation averages 5.00% across all opponents, far above the next highest (DeepSeek buyer at 0.40%). Figure 5 decomposes this by opponent (Table 18): GPT-4.1 as seller violates at 11.5% when facing o3; Gemini-2.5-Flash at 9.5%. o3 self-play produces 0.0% opponent violations. The pattern is consistent with opponent strength: weaker sellers (GPT-4.1, Flash) are pressured into accepting irrational deals, while stronger opponents (Pro, DeepSeek, o3 itself) resist. Pairwise violation rate heatmaps are in Section C.2.1.

Dimension 2: Strategic effectiveness—surplus maximization. Figure 6 reports surplus share and deal rate by model and role (Table 19). A clear hierarchy emerges at the top: **o3 captures the most surplus in both roles** (47.2% buyer, 77.0% seller), followed by Gemini-2.5-Pro (44.9%/72.6%). The remaining three models are closer together, and their relative ordering shifts across roles: as buyers, DeepSeek-V3-0324 (35.1%) edges GPT-4.1 (34.1%), with Gemini-2.5-Flash last (24.3%); as sellers, GPT-4.1 (57.7%) and Gemini-2.5-Flash (57.6%) are nearly tied, while DeepSeek-V3-0324 is last (49.6%). **The top two models rank consistently across roles; the remaining three swap positions.**

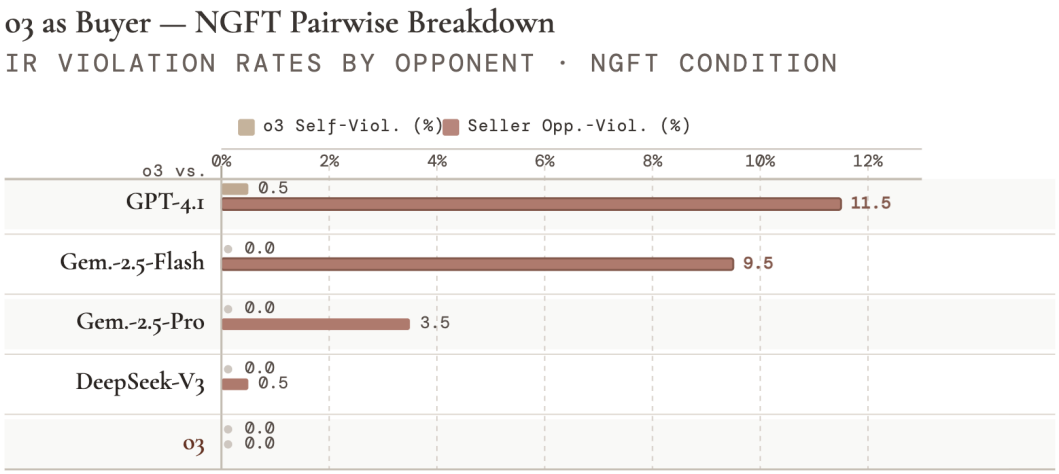

Figure 5: NGFT pairwise breakdown for o3 as buyer. The irrationality comes almost entirely from opponents: GPT-4.1 and Gemini-2.5-Flash violate at 11.5% and 9.5%, while o3 itself violates at most 0.5%. Full table in Table 18.

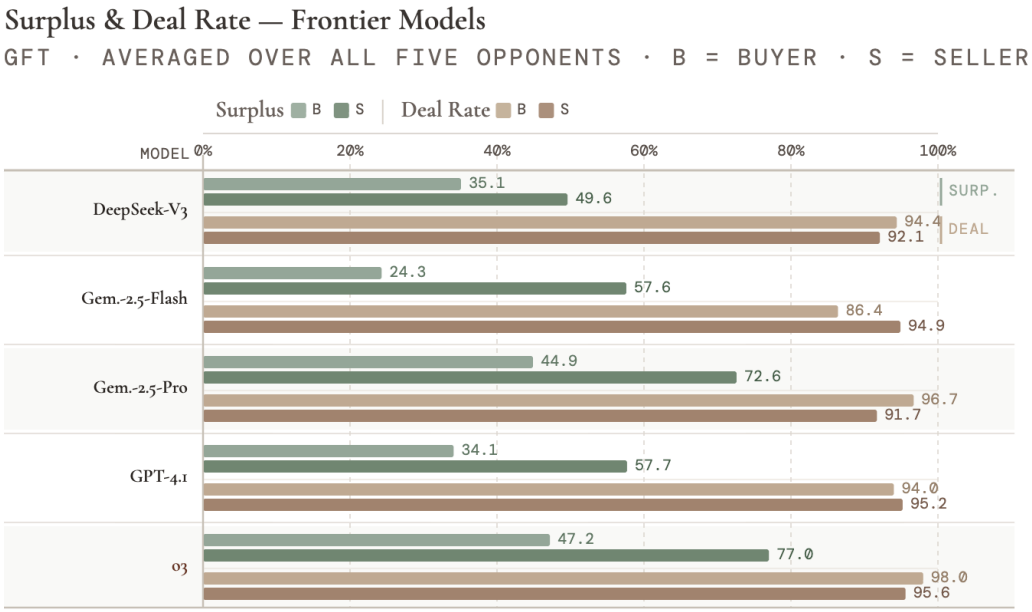

Figure 6: Surplus share and deal rate for frontier models (GFT scenarios), averaged over all five opponents. B = buyer, S = seller. Surplus: fraction of ZOPA captured conditional on agreement. Deal rate: fraction of negotiations reaching agreement. Full table in Table 19.

Sellers capture the majority of surplus in nearly every pairing, reflecting the structural **first-mover advantage**. Self-play isolates this effect: DeepSeek-V3-0324 self-play is nearly symmetric (48.5%/51.5%), while o3 self-play is heavily seller-skewed (32.3%/67.7%), revealing that o3's strategy amplifies the structural advantage far beyond the baseline. The 50 percentage point (pp) spread between the best buyer surplus share (o3 vs. DeepSeek, 61.8%) and the worst (Gemini-2.5-Flash vs. o3, 11.6%) underscores that **model choice matters as much as role assignment**. Pairwise utility and surplus share heatmaps are in Sections C.2.2 and C.2.3.

Dimension 3: Allocative efficiency—deal rate. Figure 6 reports GFT deal rates by model and role. **o3 achieves the highest rates in both roles** (98.0% buyer, 95.6% seller), demonstrating that high surplus capture can coexist with high deal completion. GPT-4.1 shows balanced performance across roles

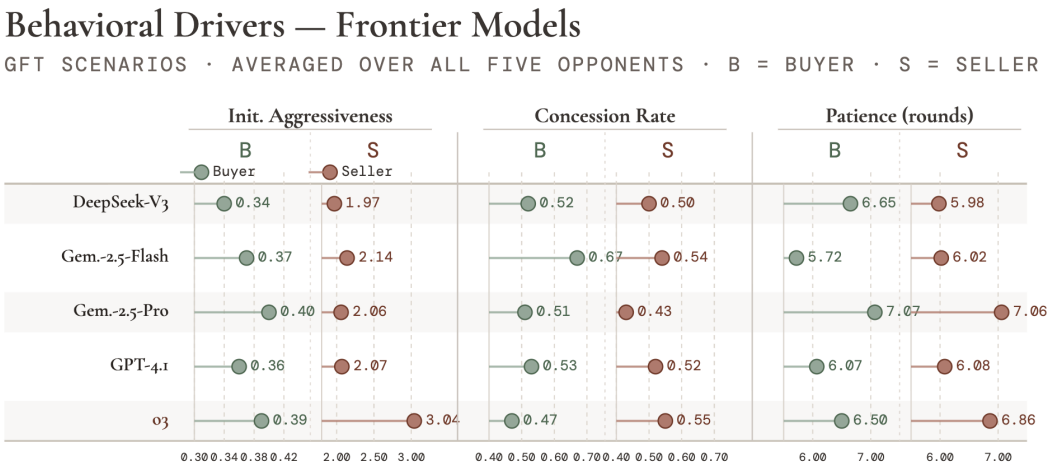

Figure 7: Behavioral drivers for frontier models (GFT scenarios), averaged over all five opponents. B = buyer, S = seller. Init. Aggr. (buyer): fraction of the seller's asking price cut by the buyer's first counter-offer (higher = more aggressive). Init. Aggr. (seller): ratio of first offer to reservation price (higher = more aggressive). Conc. Rate: fraction of remaining surplus conceded per round. Patience: mean number of rounds to termination (max 10). Full table in Table 19.

(94.0%/95.2%). Gemini-2.5-Flash as buyer is the clear outlier (86.4%): its cautious profile is associated with more walk-aways, compounding its strategic disadvantage—though it recovers as seller (94.9%). Gemini-2.5-Pro shows the reverse asymmetry: strong as buyer (96.7%) but the lowest seller deal rate (91.7%), associated with its low concession rate as seller. Pairwise deal rate heatmaps are in Section C.2.4.

Behavioral drivers. The preceding three dimensions document *what* models achieve; the Figure 7 reveals *how*.

Initial aggressiveness. Seller aggressiveness varies substantially: o3 opens at 3.04× its cost, far above every other model (1.97–2.14×). **This aggressive anchoring is the most distinctive behavioral difference** across models, consistent with the first-offer anchoring effect (Galinsky & Mussweiler, 2001). The remaining four sellers cluster tightly, with DeepSeek-V3-0324 the most moderate (1.97×). On the buyer side, aggressiveness is more compressed (0.34–0.40), with Gemini-2.5-Pro cutting the most from the seller's asking price (0.40) and DeepSeek-V3-0324 the least (0.34).

Concession rates. o3 employs opposite strategies by role: as buyer it concedes the least (0.47), while as seller it concedes the most (0.55). This **"anchor high, concede to close" pattern is consistent with o3's combined high surplus share and high deal rate**. Gemini-2.5-Flash shows the opposite profile: the highest buyer concession rate (0.67) is associated with the weakest buyer surplus share (24.3%), while Gemini-2.5-Pro's low seller concession rate (0.43) is associated with its strong seller surplus share (72.6%) and the lowest seller deal rate (91.7%).

Temporal patience. Gemini-2.5-Flash's impatience (5.72 rounds as buyer) correlates with its weak bargaining power: it closes deals quickly but captures less surplus. **Gemini-2.5-Pro's patience (7.07 rounds) is associated with its strong surplus capture** and low concession rate. Pairwise heatmaps for initial aggressiveness, concession rate, and temporal patience are in Sections C.2.5 to C.2.7.

4.2.2 Price-Tier Decomposition

Human negotiators exhibit price-sensitive behavior: they bargain more tightly on inexpensive items, where a fixed dollar concession is proportionally larger. Do frontier LLMs show similar patterns? We disaggregate all three evaluation dimensions by reservation price quintile (Q1 = lowest, Q5 = highest), averaging across all five opponents per model. Buyer-side metrics are broken down by the buyer's reservation price b; seller-side metrics by the seller's reservation price s. This decomposition reflects that each agent's own reservation price

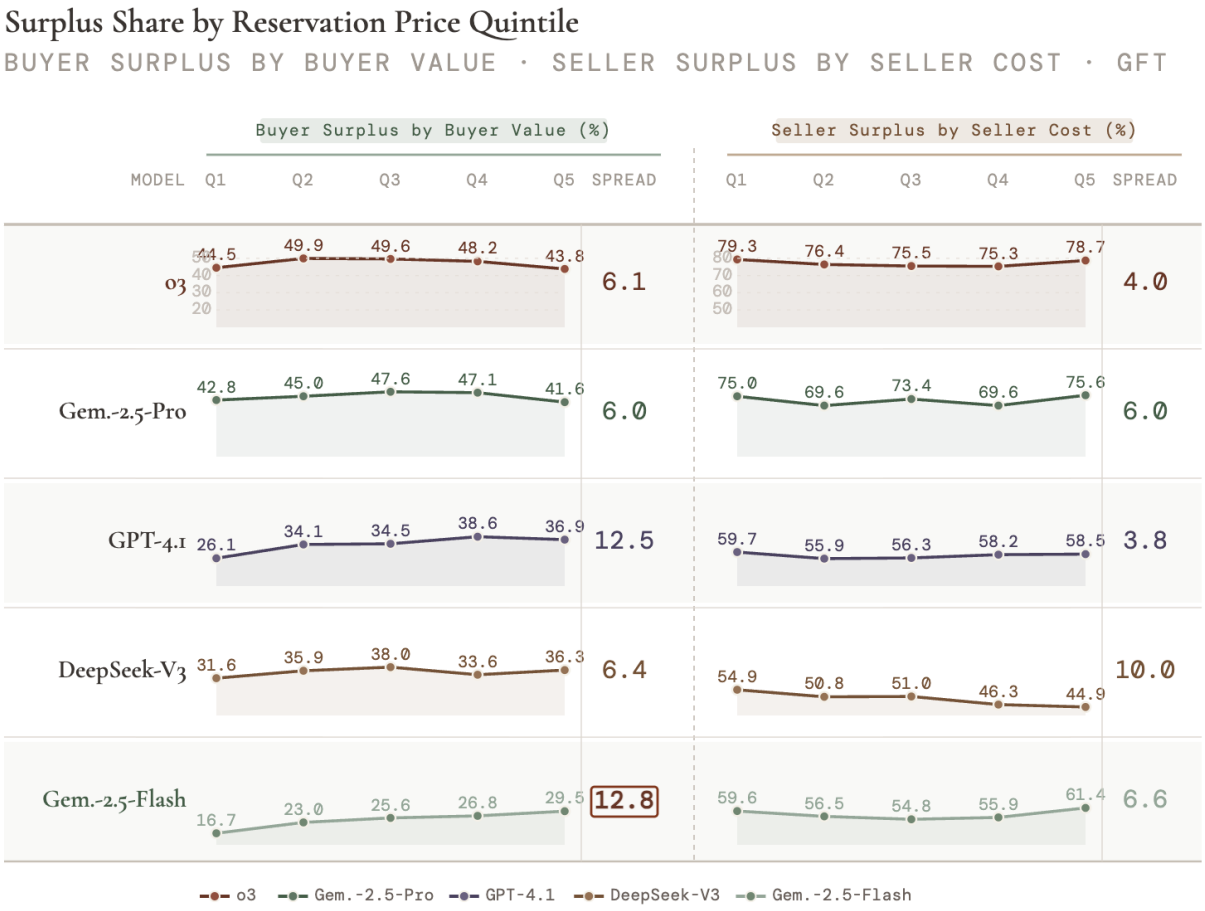

Figure 8: Surplus share by reservation price quintile, averaged across all 5 opponents (GFT). Left panel: buyer surplus by buyer-value bracket. Right panel: seller surplus by seller-cost bracket. Strong models (o3, Gem.-Pro) show 4–6 pp spreads; weak buyers (Gem.-Flash) show 13 pp. Full table in Table 22.

determines how much room it has to negotiate: a buyer with low b faces thin ZOPAs and must bargain tightly, while a buyer with high b has more margin for suboptimal play. The results split sharply: surplus capture and deal rates are value-sensitive for weaker models but stable for stronger ones, while violation rates show no systematic price-tier trend for any model.

Surplus capture: stronger models show smaller bracket spreads. Figure 8 reports each model's surplus share across value quintiles (Table 22). o3 maintains 44–50% buyer surplus across all brackets (6.1 pp spread); Gemini-2.5-Pro is similarly stable at 42–48% (6.0 pp). Gemini-2.5-Flash improves from 17% at Q1 to 30% at Q5—a 12.8 pp spread—because high-value brackets have larger ZOPAs that buffer suboptimal strategies. DeepSeek-V3 as seller shows an approximately declining pattern, from 55% (Q1) to 45% (Q5), suggesting difficulty maintaining margins on more expensive items. The top two models (o3, Gemini-2.5-Pro) and the bottom one (Gemini-2.5-Flash as buyer) maintain their relative positions across all quintiles; GPT-4.1 and DeepSeek-V3 swap at Q4 and Q5.

Deal rates: weaker buyers show more bracket variation. Figure 9 shows that deal rates follow a similar pattern (Table 23). o3 closes 97–99% of deals regardless of bracket (1.6 pp spread). Gemini-2.5-Flash as buyer drops to 80% at Q2—its minimum—and reaches 90% at Q5, producing the largest spread of any model (9.7 pp). As sellers, all models are more stable: spreads range from 2.0 pp (o3) to 4.8 pp (Gemini-2.5-Flash). Buyers exhibit more value-sensitivity than sellers, consistent with sellers' structural advantage in setting initial asking prices.

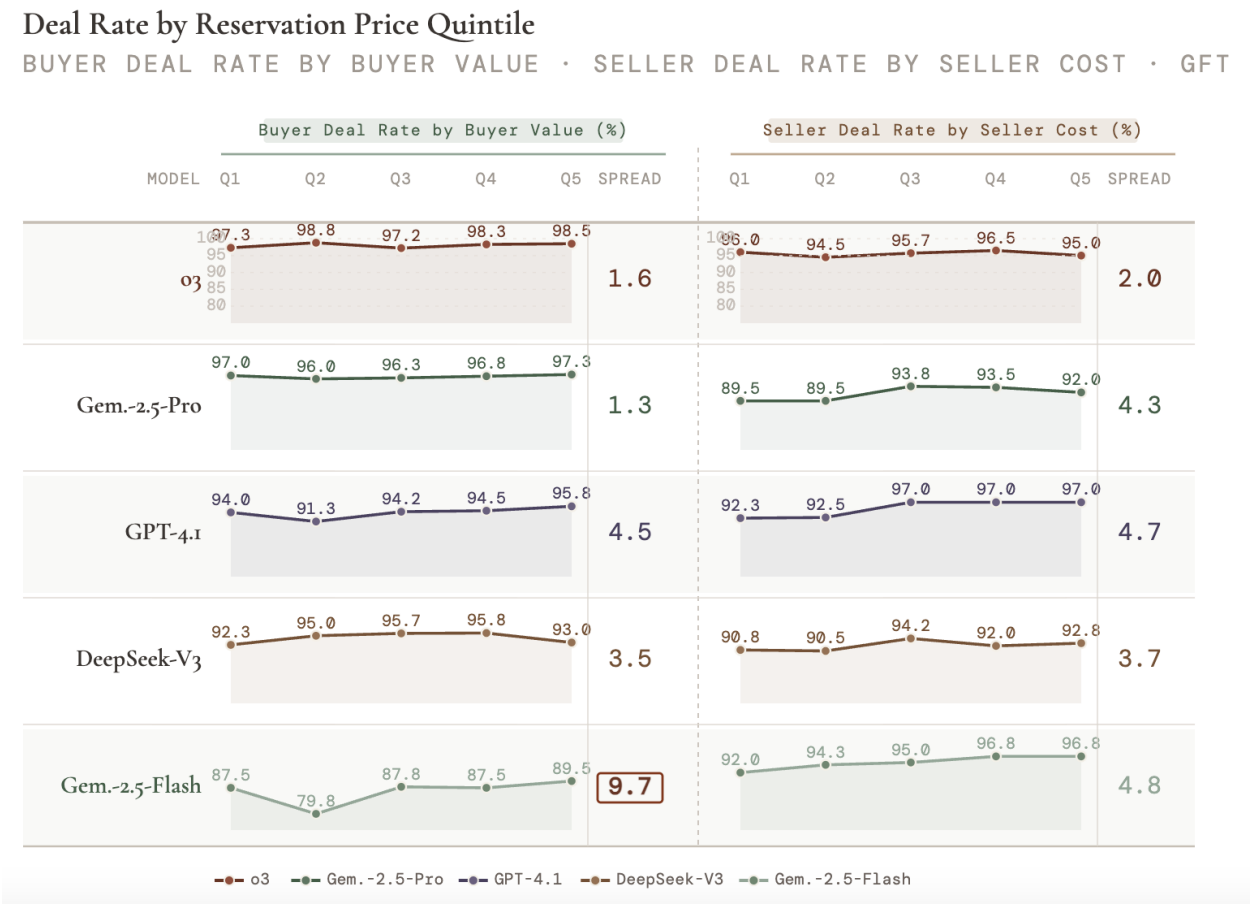

Figure 9: Deal rates by reservation price quintile, averaged across all 5 opponents (GFT). Left panel: buyer deal rate by buyer-value bracket. Right panel: seller deal rate by seller-cost bracket. Flash as buyer shows the largest spread (9.7 pp); all other models stay within 1–5 pp. Full table in Table 23.

Violation rates show no systematic price-tier trend. Figure 10 reports combined violation rates (either party accepting a negative-utility deal; Table 24). All rates stay below 2% and show no systematic Q1→Q5 trend. Gemini-2.5-Pro as seller achieves 0% violations across every bracket. DeepSeek-V3 has the highest rates (up to 2%) but these do not concentrate in any particular quintile.

Full per-opponent bracket decompositions are in Section C.3.

4.2.3 Cross-Dimensional Summary

Table 7 synthesizes the results from the previous sections into role-specific performance rankings. The top two models—o3 and Gemini-2.5-Pro—rank first and second under both roles; the remaining three models swap positions depending on whether the model acts as buyer or seller. Three findings emerge from connecting the results across the three dimensions. Figures 11 and 12 visualize these patterns as six-dimensional performance profiles, where the shape and size of each model's polygon reflect its strategic fingerprint.

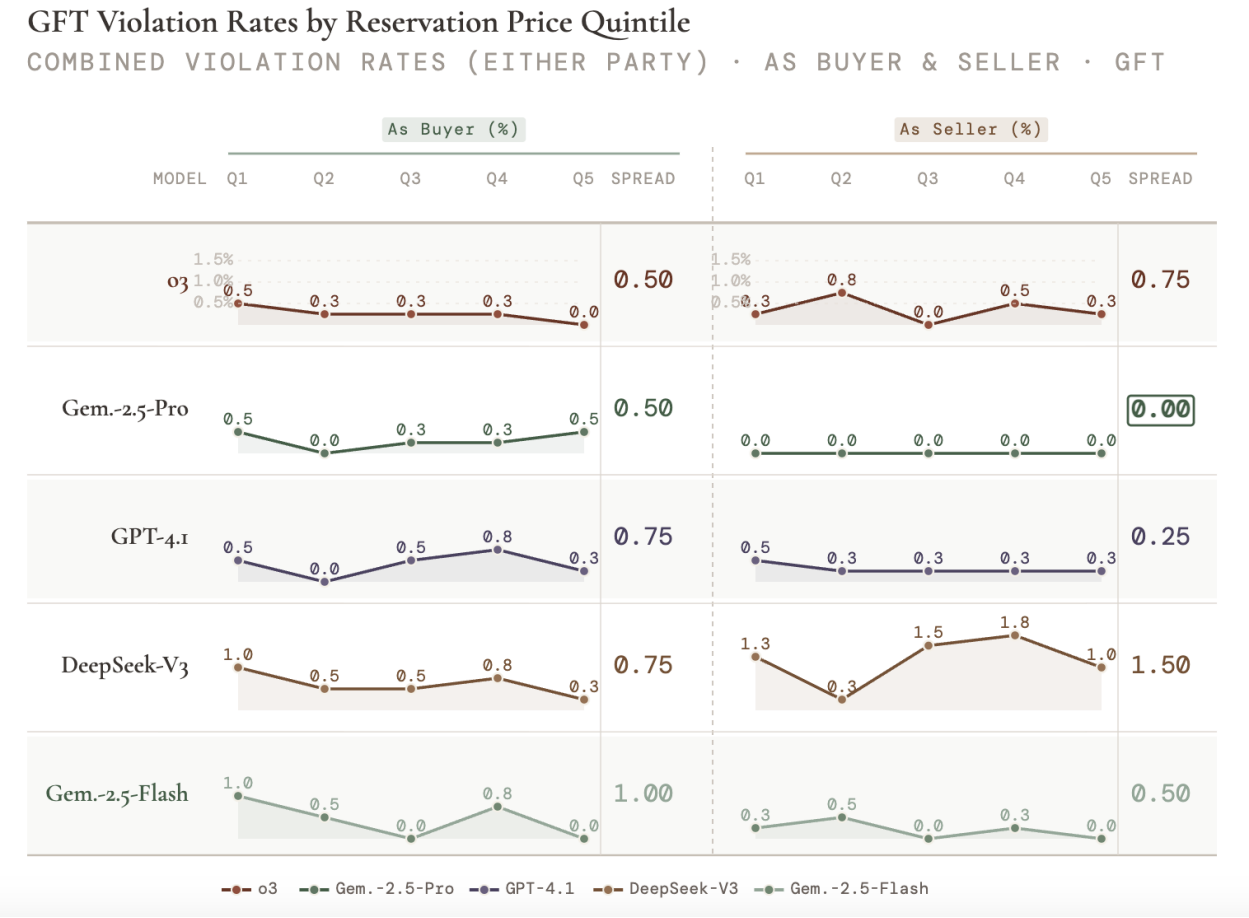

Figure 10: Combined violation rates (either party) by reservation price quintile, averaged across all 5 opponents (GFT). All rates are below 2% with no systematic bracket trend. Gemini-2.5-Pro as seller: 0% in every quintile. Full table in Table 24.

(a) Buyer

Rank	Model	Score
1	o3	46.3%
2	Gem.-2.5-Pro	43.4%
3	DeepSeek-V3	33.1%
4	GPT-4.1	32.1%
5	Gem.-2.5-Flash	21.0%

(b) Seller

Rank	Model	Score
1	o3	73.6%
2	Gem.-2.5-Pro	66.6%
3	GPT-4.1	54.9%
4	Gem.-2.5-Flash	54.7%
5	DeepSeek-V3	45.7%

Table 7: Performance ranking by role. Score = surplus share × deal rate, measuring unconditional surplus capture. All values from GFT scenarios, averaged over all five opponents.

Finding 1: Aggressive anchoring, calibrated concession, and temporal patience facilitate price discrimination. A naive expectation is that aggressive bargaining faces a surplus–efficiency tradeoff: extracting more surplus should come at the cost of more failed negotiations. o3 contradicts this expectation. It achieves both the highest surplus share (47.2%/77.0%) and the highest deal rate (98.0%/95.6%) through a role-asymmetric pattern: as seller, it anchors at 3.04× cost—far above the 1.97–2.14× range of other models—then concedes at the highest rate (0.55) to close deals; as buyer, it concedes the least (0.47), holding firm after the opponent's aggressive opening. The extreme anchor creates maximum room for price adaptation: different buyers end up paying different prices based on their strategic resistance (e.g., 88.4% seller surplus against Gemini-2.5-Flash vs. 67.7% in self-play), while the calibrated concession sustains willingness to trade.

This "anchor high, concede to close" pattern **implements first-degree price discrimination through sequential offers**—the seller discovers each buyer's willingness to pay through the concession arc rather than posting a single price, consistent with the first-offer anchoring effect (Galinsky & Mussweiler, 2001). Moderate temporal patience (6.50/6.86 rounds) provides enough rounds to execute this discovery process without the premature exits that characterize weaker strategies. The same behavioral profile is associated with o3's Dimension 1 pattern: its aggressive anchoring pressures weaker opponents into NGFT violations (Figure 5), while its own constraint adherence remains intact.

Finding 2: Accommodating strategies disable price discrimination in the buyer role. One might expect that accommodating agents would at least close more deals even if they capture less surplus. Gemini-2.5-Flash as buyer demonstrates the opposite. Its cautious profile—low aggressiveness, the highest buyer concession rate (0.67), and the fewest rounds (5.72)—is associated with both the worst buyer surplus share (24.3%) and the worst buyer deal rate (86.4%). By conceding quickly and uniformly, the accommodating buyer reveals its willingness to pay before the seller needs to discriminate, surrendering the informational asymmetry that sustains bargaining power. This pattern is not price-dependent: buyer surplus capture remains at 17–30% across all quintiles (Figure 8). This finding does not extend to the seller role, where Gemini-2.5-Flash performs competitively (94.9% deal rate, 57.6% surplus share). For AI system design, however, the buyer-side result cautions against "helpful and agreeable" defaults in competitive settings—**in the buyer role, cooperativeness without strategic backbone eliminates the ability to resist price discrimination and is associated with worse surplus capture and deal completion**.

Finding 3: Consistency across price tiers reflects proportional discrimination ability. The price-tier decomposition (Section 4.2.2) reveals that strong and weak models differ not in whether they can negotiate well, but in whether they can do so *regardless of item value*. o3 and Gemini-2.5-Pro maintain 4–6 pp surplus spreads across all quintiles; Gemini-2.5-Flash shows a 13 pp spread, improving from 17% at Q1 to 30% at Q5 because larger ZOPAs at high values buffer suboptimal strategies. Strong models scale their anchoring and concession proportionally to item value—extracting a similar fraction of surplus whether the item costs $20 or $2,000—which is the hallmark of effective price discrimination across market segments. Weak models lack this proportional adaptation and succeed only when wide ZOPAs mask their inability to discriminate. The top two models (o3, Pro) and the bottom one (Flash as buyer) maintain their positions in every quintile, though GPT-4.1 and DeepSeek-V3 swap at Q4 and Q5. Violation rates, by contrast, show no systematic price-tier trend for any model: agents that correctly evaluate deal quality do so regardless of the stakes. This value-dependent pattern echoes the human bargaining evidence of Backus et al. (2020), who find that bargaining behavior differs systematically between expensive and inexpensive items on eBay's Best Offer platform.

Figures 11 and 12 present six-dimensional performance profiles for each model, with all axes oriented so that larger values indicate stronger performance. Full numeric values are in Tables 20 and 21.

5 Training Negotiation Agents

Section 4 establishes that frontier LLMs achieve 86–98% deal rates and sub-1% GFT violation rates, but surplus allocation varies by up to 50 pp across models and aggressive strategies induce opponent IR violations in NGFT. These results demonstrate that negotiation performance is not an inherent property of model scale—it depends on behavioral strategy. Can open-weight models of smaller scale be trained to match frontier-level performance through targeted fine-tuning? We conduct a pilot study using the agent system (Section 3.3) as a training environment. The goal is not a production-ready pipeline, but a proof of concept: the framework's structured actions and automated metrics can serve as effective reward signals for RL. We fine-tune Qwen3 models (8B and 14B parameters) via a two-stage pipeline (SFT then RL) and evaluate on the same three dimensions used in Section 4.

5.1 Training Pipeline Overview

The training pipeline has two stages:

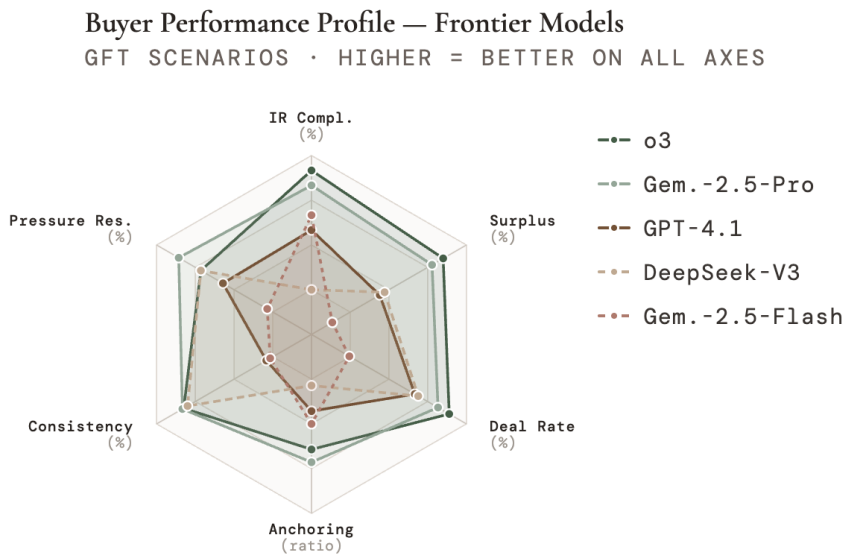

Figure 11: Buyer performance profile. Six axes: IR compliance, surplus share, deal rate, anchoring power (fraction cut from asking price), price-tier consistency, and opponent pressure resistance. All min-max normalized; higher = better. Full table in Table 20.

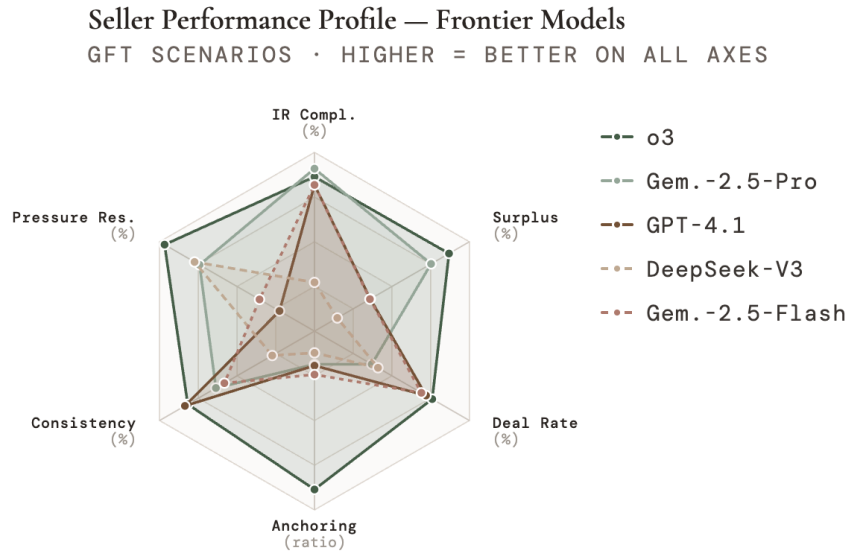

Figure 12: Seller performance profile. Same six axes as the buyer plot, with anchoring power measured as the ratio of first offer to reservation price. o3 dominates on anchoring and surplus; GPT-4.1 leads on consistency. Full table in Table 21.

1. **Stage 1—SFT (Section 5.2).** Generate demonstration trajectories via DeepSeek-R1 self-play within the simulator, then fine-tune Qwen3 models on filtered trajectories. This provides basic negotiation competence: well-formed tool calls, protocol compliance, and reasonable offers.
2. **Stage 2—RL (Section 5.3).** Train the SFT-initialized models via GRPO, negotiating against GPT-4.1 as a fixed opponent. This enables the model to improve beyond behavioral cloning by directly optimizing surplus.

The two stages use different opponents by design: SFT benefits from diverse demonstrations (DeepSeek-R1 self-play), while RL benefits from a fixed opponent that provides a stationary reward landscape and prevents mode collapse from co-adaptation.

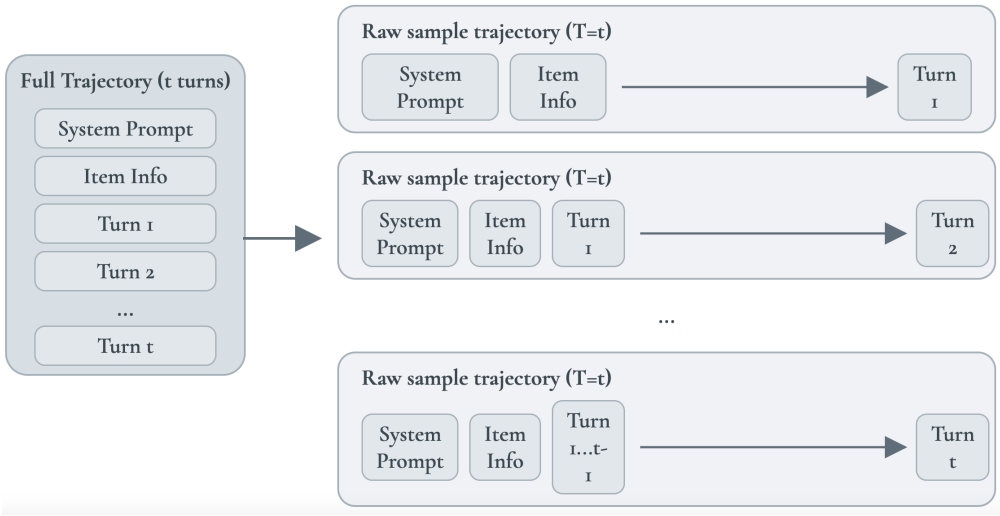

Figure 13: Turn-level decomposition of a negotiation trajectory into autoregressive training samples.

5.2 Stage 1: Supervised Fine-Tuning

SFT provides the foundation through behavioral cloning[5]: the model acquires well-formed tool calls, protocol compliance, and reasonable offers before RL optimization.

Trajectory generation via DeepSeek-R1 self-play. We generate synthetic negotiation trajectories by running DeepSeek-R1 in self-play within the simulator. Both agents receive the same system prompt template with negotiation strategies, item information, and their respective reservation prices. Negotiations proceed through the event-driven protocol (Section 3.3.2) until agreement, outside-option exercise, or the round limit.

Turn-level decomposition. Each trajectory (typically has t = 5–10 turns) is decomposed into t autoregressive training samples of increasing context length: sample T uses [system prompt + item info + turns $1 \ldots T-1$] as context and turn T as the target. This ensures the model learns to condition on negotiation history at every stage.

Reasoning-token masking. DeepSeek-R1 trajectories contain `<think>...</think>` reasoning blocks. For each input-output pair, reasoning tokens are masked from the input context, so that only the agent's observable responses contribute to the loss. This ensures the model conditions on the same information available during inference—where it generates its own reasoning—and avoids distribution shift from reasoning patterns.

Training details. We fine-tune Qwen3-8B and Qwen3-14B (Qwen Team, 2025) using VERL (Volcano Engine, 2024). The model learns the full agent output format (`Thought:` + `Code:`) conditioned on negotiation history. Hyperparameters are in Section B.1.

5.3 Stage 2: Reinforcement Learning with GRPO

SFT produces a policy that *imitates* demonstrations. RL enables the model to *optimize*—discovering strategies that maximize expected surplus while maintaining IR, potentially exceeding demonstration quality.

Asymmetric play: Qwen vs. GPT-4.1. The SFT-initialized Qwen model negotiates against GPT-4.1 as a fixed opponent, with the trainee randomly assigned as buyer or seller each episode. A fixed opponent is chosen for three reasons: (1) it provides a *stationary reward landscape*, simplifying credit assignment; (2) it *avoids mode collapse* from co-adaptation, where co-evolving agents can converge to degenerate equilibria; and

[5] *Behavioral cloning* trains a policy to imitate demonstrated actions via supervised learning, without direct reward optimization.

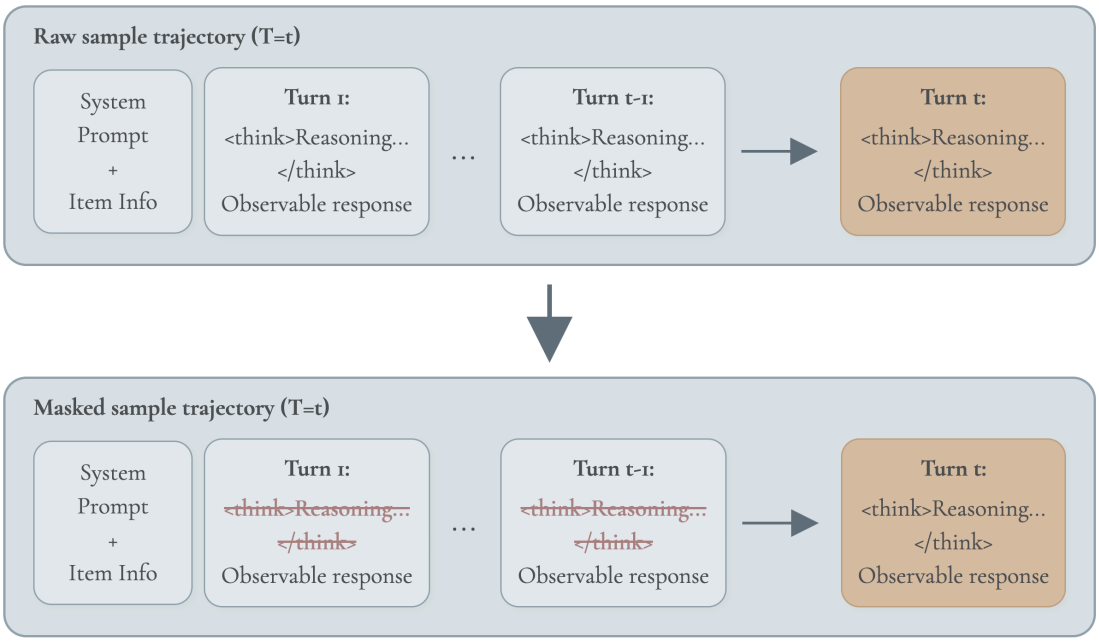

Figure 14: Reasoning-token masking: `<think>` blocks are masked from input context turns, so the model conditions only on observable agent responses.

(3) GPT-4.1 exhibits *sufficient opponent quality* (anchoring, measured concessions, strategic outside-option exercise), providing a richer learning signal than weaker opponents.

Algorithm: Group Relative Policy Optimization (GRPO). We use GRPO (Shao et al., 2024), which generates $G = 8$ trajectory rollouts per negotiation context (item + reservation prices) and computes *relative* advantages by comparing outcomes within the group. This is well-suited for negotiation because outcomes are highly context-dependent: the same strategy may yield different utilities depending on the item, reservation prices, and counterpart behavior. Within-context comparison provides a lower-variance learning signal than absolute rewards.

For each prompt $x \sim \mathcal{D}$ (a negotiation context), GRPO samples G rollouts $y^1, \ldots, y^G$ from the behavior policy $\pi_{\text{old}}(\cdot \mid x)$. Let y_t^g denote the t-th token of rollout g, with T^g tokens in total. The GRPO objective maximizes a clipped surrogate advantage while penalizing divergence from the SFT reference policy π_{old}:

$$\mathcal{L}(\theta) = -\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ y^{1:G}\sim\pi_{\text{old}}}} \left[\frac{1}{G}\sum_{g=1}^{G} \frac{1}{T^g}\sum_{t=1}^{T^g} \min\Big(\rho_t^g \hat{A}^g,\, \text{clip}\big(\rho_t^g, 1-\epsilon, 1+\epsilon\big)\hat{A}^g\Big) - \beta \cdot \text{KL}\big(\pi_\theta \| \pi_{\text{old}}\big) \right], \tag{5}$$

where ρ_t^g is the per-token importance ratio between the current policy π_θ and the reference policy π_{old}:

$$\rho_t^g = \frac{\pi_\theta(y_t^g \mid x, y_{<t}^g)}{\pi_{\text{old}}(y_t^g \mid x, y_{<t}^g)}. \tag{6}$$

The advantage $\hat{A}^g$ is computed by normalizing rewards within the group of G rollouts for the same prompt:

$$\hat{A}^g = \frac{R^g - \bar{R}}{\sigma_R + \epsilon_{\text{stable}}}, \tag{7}$$

where $\bar{R} = \frac{1}{G}\sum_{g=1}^{G} R^g$ is the group mean reward, $\sigma_R = \sqrt{\frac{1}{G}\sum_{g=1}^{G}(R^g - \bar{R})^2}$ is the group standard deviation, and $\epsilon_{\text{stable}} = 10^{-4}$ prevents division by zero when all rollouts receive the same reward.

Reward structure. The composite reward combines protocol compliance with negotiation outcome:

$$R = 0.5 \cdot R_{\text{parsing}} + 0.5 \cdot R_{\text{execution}} + 0.5 \cdot R_{\text{constraints}} + R_{\text{utility}}. \tag{8}$$

Table 8 defines each component. All four components lie in $[0, 1]$, so $R \in [0, 2.5]$. The protocol components (R_{parsing}, $R_{\text{execution}}$) provide a learning signal early in training when successful negotiations are rare. The outcome components reward behavioral consistency ($R_{\text{constraints}}$) and surplus capture (R_{utility}). IR violations are *not* directly penalized by $R_{\text{constraints}}$; rather, accepting a deal at negative utility yields $R_{\text{utility}} = 0$, providing only an indirect signal. The KL penalty constrains the policy to remain close to the SFT-initialized reference, preserving language quality ($\beta = 0.05$ for 8B; $\beta = 0.005$ for 14B).

Component	Definition	Range
R_{parsing}	Fraction of agent messages containing a valid tool-call code block: (# parseable messages)/(# all messages)	$[0, 1]$
$R_{\text{execution}}$	Fraction of parseable code blocks that execute successfully: (# successfully executed blocks)/(# parseable blocks)	$[0, 1]$
$R_{\text{constraints}}$	Behavioral consistency, averaged over two checks: $0.5\,(1 - \mathbf{1}\{\texttt{accepted_worse_offer_later}\}) + 0.5\,(1 - \mathbf{1}\{\texttt{proposed_worse_than_rejected}\})$. Penalizes accepting a deal worse than a previously rejected offer, or proposing terms worse than what the opponent already rejected.	$\{0, 0.5, 1\}$
R_{utility}	Surplus share (Definition 5) if a deal is reached; 0 otherwise	$[0, 1]$

Table 8: Reward components in the composite reward (Equation (8)). Protocol components (R_{parsing}, $R_{\text{execution}}$) reward well-formed tool usage; outcome components reward behavioral consistency ($R_{\text{constraints}}$) and surplus capture (R_{utility}).

Training architecture and infrastructure. Online GRPO requires overlapping rollout generation with gradient computation. We separate the two workloads across dedicated GPU sets on a single 8×H100 80 GB node: a *vLLM inference cluster* (Kwon et al., 2023) serves the current policy for rollout generation using tensor parallelism, while a *training cluster* running DeepSpeed ZeRO-3 (Rajbhandari et al., 2020) computes spreads and updates weights. Updated weights are synchronized to the inference server periodically, enabling pipeline parallelism where batch $N+1$ is generated while batch N is being trained on. For 14B models, we use LoRA (Hu et al., 2021) (rank 64) to reduce trainable parameters to <1% of the full model, combined with gradient checkpointing and mixed-precision training (bfloat16 forward, fp32 optimizer states). Full details on hardware allocation, hyperparameters, and system performance are in Section B.2.

5.3.1 Challenges Specific to RL for Multi-Turn Negotiation

Two challenges distinguish this RL setting from standard language model RL (e.g., for mathematical reasoning or code generation).

Challenge 1: Variable-length context and context overflow. Unlike math or code generation tasks where output length is bounded, negotiation contexts grow unboundedly over multiple rounds. Each round adds ∼1,100 tokens (prompt update + agent response + tool execution + environment feedback), so a 10-round negotiation can approach the 16K-token training context limit (Section B.3). When the context exceeds this limit, the episode cannot continue—but simply discarding overflowed episodes would bias the training distribution toward short negotiations.

We address this via *gradient-safe overflow handling*. When overflow occurs, we terminate the episode with a synthetic `quit_negotiation` response and generate synthetic logprobs using the PAD token (a special token never produced during normal generation). The reward is set to zero. Because $\pi_\theta(\text{PAD} \mid s) \approx 0$, the resulting gradient contributions are negligible—the episode occupies its batch slot without corrupting the learned distribution. Using a real vocabulary token (e.g., a punctuation character) would instead create spurious gradient updates that systematically suppress that token's probability in long-context settings. Implementation details are in Section B.4.

Challenge 2: Distribution shift across training stages. The two-stage pipeline introduces distribution mismatch at each transition. First, the SFT demonstrations are generated by DeepSeek-R1 self-play, but the trainee is Qwen. DeepSeek-R1 has its own vocabulary preferences and strategic tendencies; Qwen must learn the negotiation *behavior* without overfitting to model-specific artifacts. The reasoning-token masking described in Section 5.2 mitigates this by excluding DeepSeek's internal reasoning format from the input context, forcing the trainee to attend only to observable actions.

Second, RL trains against a single fixed opponent (GPT-4.1) rather than the diverse model population used in the benchmark (Section 4). The trainee therefore only observes one counterpart's behavioral profile during optimization. This is a deliberate scope constraint: our goal is to establish whether RL can improve negotiation capabilities in this setting, not to produce a universally robust agent. To test whether the learned improvements generalize beyond the training opponent, we evaluate trained models against both GPT-4.1 and o3-mini (unseen during training) in Section 5.4.

5.4 Results

We report results from a pilot study: Qwen3-8B and Qwen3-14B trained against GPT-4.1 via SFT+GRPO, evaluated against both GPT-4.1 (in-distribution) and o3-mini (unseen during training). Two main findings emerge. First, SFT and RL pull in opposite directions: SFT doubles surplus share but drops deal rates; RL recovers deal rates but erodes surplus back to base-like levels. This tension originates from the reward structure (Equation (8)), which assigns $R_{\text{utility}} = 0$ to both rational walk-aways and irrational acceptances, biasing RL toward deal-closing. The full pipeline (Base $\rightarrow$ SFT $\rightarrow$ RL) lands within ± 2 pp of the base model on most metrics. Second, SFT compresses buyer-side surplus spreads across price tiers from 15–25 pp to 7–9 pp, and this compression holds against opponents unseen during training—evidence that SFT teaches proportional strategies rather than memorizing price points.

5.4.1 Evaluation Protocol

We evaluate the training pipeline by measuring how each stage (Base $\rightarrow$ SFT $\rightarrow$ SFT+RL) improves negotiation performance within a given model size. We consider six **Qwen variants: Qwen3-8B (base), Qwen3-8B-SFT, Qwen3-8B-RL; Qwen3-14B (base), Qwen3-14B-SFT, Qwen3-14B-RL**. Each variant is evaluated against two frontier counterparts—GPT-4.1 and o3-mini—in both buyer and seller roles, yielding $6 \times 2 \times 2 = 24$ experiment configurations. Each configuration runs 800 negotiation scenarios (400 GFT + 400 NGFT), for a total of 19,200 negotiations. The 400 GFT items are identical to those in Section 4.1; the 400 NGFT items are a superset of the 200 used there. All metrics are reported separately for GFT and NGFT subsets.

The two counterparts serve different diagnostic roles: **GPT-4.1** is the RL training counterpart. Performance against GPT-4.1 measures direct training effectiveness. **o3-mini** is unseen during training. Performance against o3-mini provides a secondary check on whether learned behaviors reflect general strategic improvement or are specific to the GPT-4.1 interaction. The analysis is organized into two parallel tracks—one for the 8B progression and one for the 14B progression—with no cross-size comparisons. This isolates the effect of training methodology from model capacity.

5.4.2 Model-level Overview

Mirroring Section 4.2.1, we evaluate each training stage on the three dimensions. All metrics are reported per opponent to distinguish in-distribution (vs. GPT-4.1, the RL training opponent) from out-of-distribution (vs. o3-mini, unseen during training) performance.

Dimension 1: Individual rationality. Figure 15 reports NGFT violation rates by training stage and opponent, showing both the Qwen model's own violations (Self) and the frontier counterpart's violations when facing Qwen (Opp.). Full numerical tables for both GFT and NGFT cases are in Tables 36 and 37.

GFT violation rates are uniformly low. Table 36 shows that all GFT self-violation rates remain below 4%, with the majority of cells at or below 1%. The 8B seller column against GPT-4.1 exhibits a suggestive

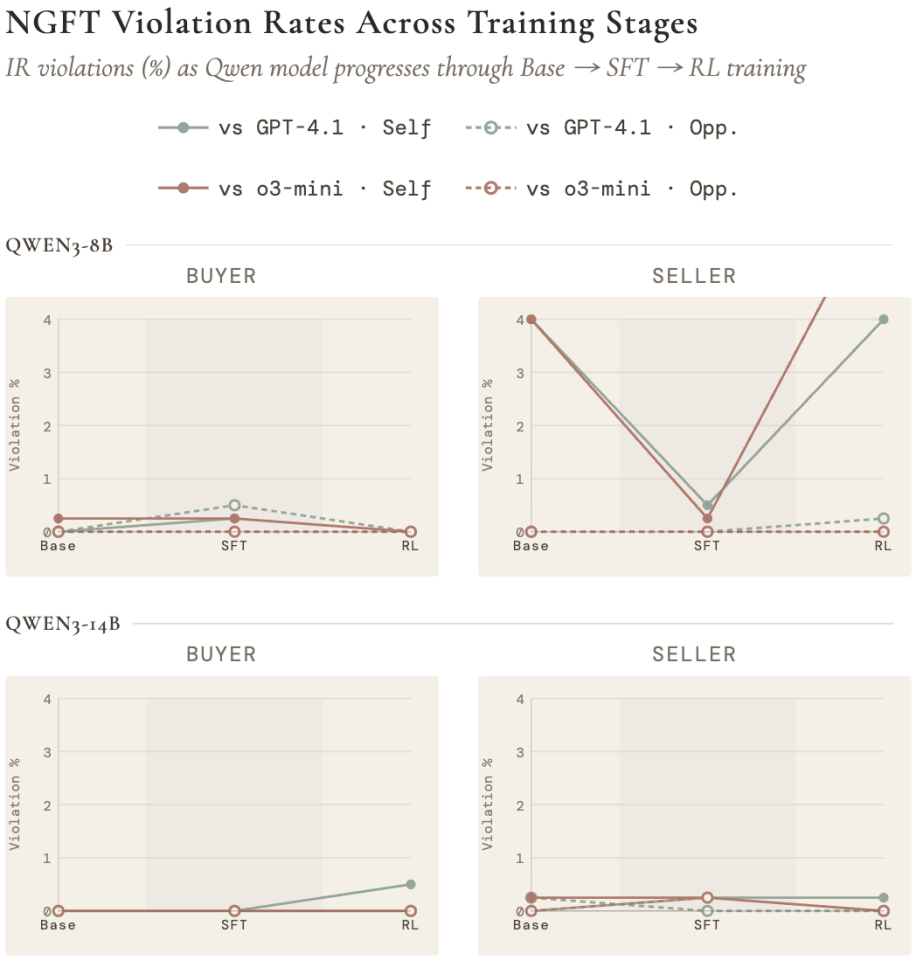

Figure 15: NGFT violation rates across training stages. Self: the Qwen model strikes a deal in a no-gains-from-trade scenario (irrational for at least one party). Opp.: the frontier counterpart does so when facing Qwen. Full table in Table 37.

Base→SFT→RL pattern (3.75%→0.25%→3.25%), but other cells do not replicate this trend consistently. At these magnitudes, differences across training stages are difficult to distinguish from sampling variability.

NGFT seller violations reveal the clearest training-stage effects, concentrated in 8B; buyer violations are consistently near zero. The NGFT regime produces larger and more interpretable differences. Figure 15 and Table 37 show that 8B seller self-violations follow a consistent Base→SFT→RL pattern across both opponents: 4.00%→0.50%→4.00% against GPT-4.1, and 4.00%→0.25%→6.50% against o3-mini. SFT reduces NGFT seller violations by an order of magnitude; RL reverses this gain entirely. The 6.50% rate against o3-mini—the highest value in either table—exceeds the base model (4.00%), suggesting that RL's deal-closing bias is amplified when the model faces an opponent unseen during training. By contrast, 14B seller violations remain at or below 0.25% through all training stages and against both opponents, exhibiting no such fragility. Buyer self-violations remain at or below 0.50% in nearly every cell across both GFT and NGFT, both model sizes, and all training stages; all meaningful variation in rationality occurs in the seller role.

Dimension 2: Strategic effectiveness—surplus share. Figure 16 reports surplus shares per opponent across training stages (full table in Table 38).

SFT increases surplus share by 15–40 pp across all conditions. 8B buyer rises from 22.4%/11.2% to 40.8%/36.1% (vs. GPT-4.1/o3-mini); 8B seller from 32.5%/24.8% to 60.6%/64.8%. The 14B pattern is

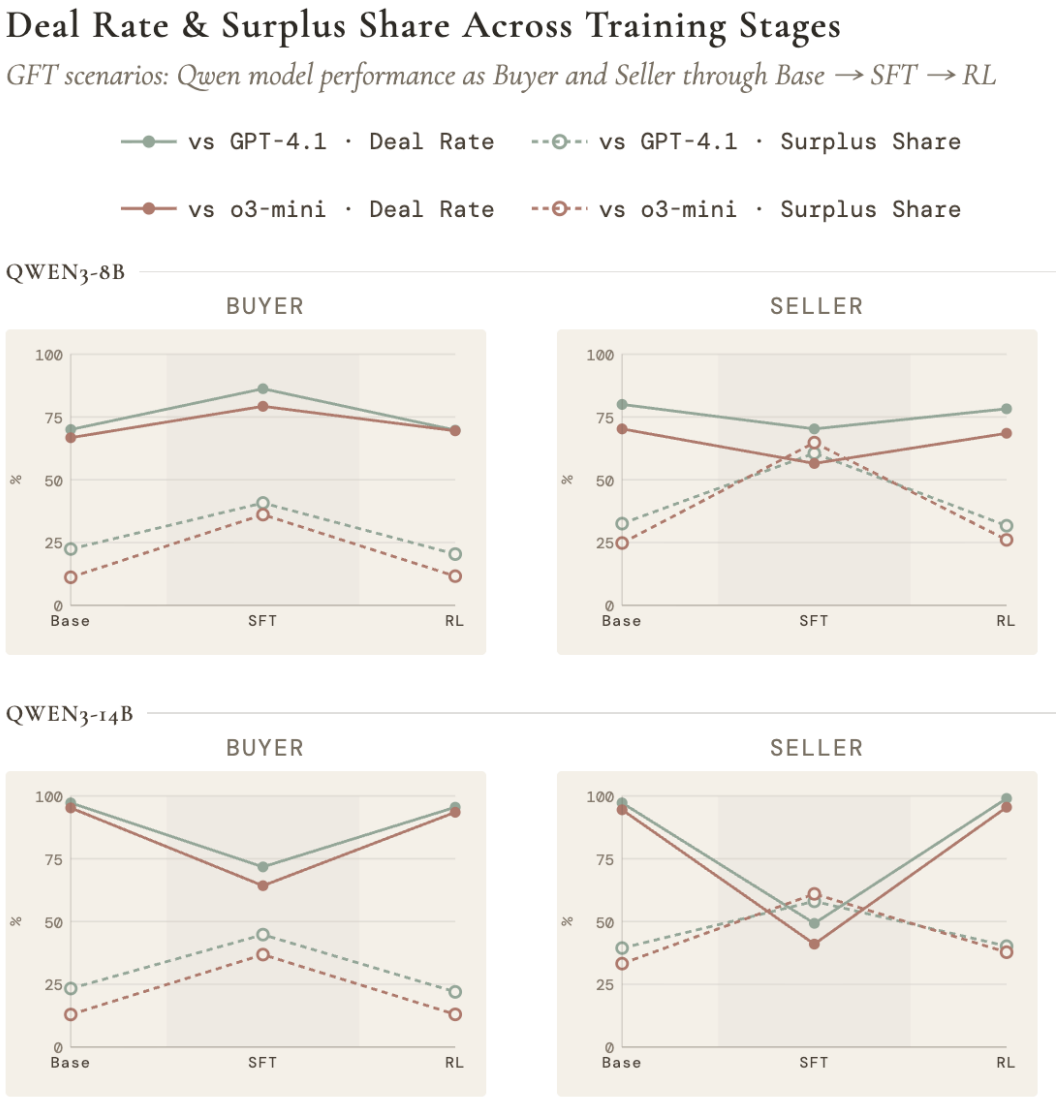

Figure 16: Strategic effectiveness and allocative efficiency across training stages (GFT scenarios). Deal rate: fraction of negotiations reaching agreement. Surplus share: fraction of ZOPA captured conditional on agreement. GPT-4.1 is the RL training opponent (in-distribution); o3-mini is unseen (out-of-distribution). Full table in Table 38.

similar. All eight cells (2 sizes × 2 roles × 2 opponents) show consistent gains, roughly doubling surplus in most cases.

RL returns surplus to base-like levels. Surplus shares fall back across all conditions: 8B-RL buyer captures 20.4%/11.6%, within 2 pp of base (22.4%/11.2%); 14B-RL seller captures 40.2%/37.8% versus base 39.4%/33.2%. All stages capture less surplus against o3-mini than GPT-4.1, reflecting o3-mini's stronger bargaining rather than a training-specific effect (the same gap appears in Base models that never trained against either opponent).

Dimension 3: Allocative efficiency—deal rate. Figure 16 also reports GFT deal rates across training stages (Table 38).

Deal rates follow a V-shape: SFT reduces them, RL recovers them. This pattern holds in six of eight model–role–opponent cells. The drop is most severe for 14B seller (97.3%→49.3%/94.5%→41.0%), where fewer than half of GFT negotiations reach agreement after SFT; 8B seller shows a milder decline

(−9.8/−13.8 pp). RL restores deal rates along the same lines: 14B-RL seller reaches 99.0%/95.5%, and 14B-RL buyer reaches 95.5%/93.5%; 8B-RL returns to base levels (78.3%/68.5% seller, 69.8%/69.5% buyer) without exceeding them. The sole exception is 8B buyer, where SFT *increases* deal rates (+16.3/+12.5 pp). 8B-Base buyer is the weakest condition overall, with both the lowest deal rates (53.5%/57.0%) and the lowest surplus shares; for models that already negotiate competently, SFT's behavioral cloning makes them more selective (higher surplus, fewer deals), but for a sufficiently weak base model, the primary effect may be teaching coherent negotiation behavior, raising both surplus and deal rates simultaneously.

5.4.3 Bracket-Level Analysis

The model-level results above average across all reservation prices. To test whether training effects concentrate in specific price tiers, we disaggregate by reservation-price quintile (Q1 = lowest, Q5 = highest), reporting buyer metrics by buyer-value bracket and seller metrics by seller-cost bracket. Results are shown separately for each opponent: GPT-4.1 (the RL training opponent) and o3-mini (unseen during training). Figures 17 to 19 report the full decomposition.

Figure 17 shows no systematic Q1→Q5 trend in violation rates (Table 39). With 80 scenarios per bracket, individual cells are noisy (e.g., 3.75% = 3 negotiations), so bracket-level violation patterns should be interpreted with caution.

SFT compresses buyer-side Q1→Q5 surplus spreads from 15–25 pp to 7–9 pp. Base and RL buyer surplus rises monotonically from Q1 to Q5 (Figure 18; Table 40). Against o3-mini, 14B-Base captures 4.7% at Q1 but 26.5% at Q5 (21.8 pp spread); 8B-Base shows a 24.5 pp spread. SFT compresses these spreads to 8.6 pp (14B) and 8.8 pp (8B), maintaining 31–40% surplus across all quintiles even against o3-mini. Against GPT-4.1, the same pattern holds: SFT spreads of 7.6–8.6 pp versus 15–25 pp for Base and RL. The effect is most pronounced at the lowest price tier against the strongest opponent: 8B-Base captures 1.8% at Q1 against o3-mini, while 8B-SFT captures 31.5%. On the seller side, Base spreads are already moderate (4–11 pp), and SFT preserves this while raising the level from 20–43% to 51–68%.

RL restores the Q1→Q5 spread. RL returns buyer surplus to base-like levels in every quintile. Against o3-mini, 14B-RL spread widens to 22.4 pp (vs. SFT's 8.6 pp); 8B-RL to 22.6 pp (vs. SFT's 8.8 pp). No bracket preserves SFT's gains.

Figure 19 shows that Base and RL deal rates are stable across brackets (Table 41). SFT introduces bracket-dependent variation, most notably for 14B seller (26.3 pp spread vs. GPT-4.1), consistent with selective deal acceptance that varies by item value.

5.4.4 Cross-dimensional Summary

Two main findings emerge from this pilot study.

SFT and RL optimize for competing objectives. SFT learns from demonstrations that balance selectivity with deal-closing: it doubles surplus share and teaches the model when to walk away. This selectivity comes at a cost—deal rates drop, severely for 14B sellers (49.3%/41.0%). RL then optimizes the composite reward (Equation (8)), where closing a deal can yield $R_{\text{utility}} > 0$ while walking away always yields $R_{\text{utility}} = 0$. This asymmetry biases RL toward deal-closing: it recovers deal rates (14B-RL reaches 93–99%) but erodes the strategic selectivity that SFT installed. Surplus shares fall back to base-like levels. In NGFT, the same mechanism drives the 8B seller violation pattern (Table 37). $R_{\text{constraints}}$ penalizes non-monotonic offers, not IR violations (Table 8); an irrational acceptance can therefore receive $R_{\text{constraints}} = 1$ if the agent's offer sequence is monotonic. Combined with $R_{\text{utility}} = 0$ for both rational walk-aways and irrational acceptances, the reward provides no signal to distinguish the two, and RL's deal-closing bias leads 8B sellers to accept terms that violate their reservation price. The full pipeline (Base → SFT → RL) therefore lands within ±2 pp of the base model on most metrics—not because neither stage works, but because they pull in opposite directions. SFT teaches "hold out for good terms"; RL teaches "close the deal."

SFT teaches proportional strategies that generalize across price tiers and opponents. Base models show large buyer-side Q1→Q5 surplus spreads (15–25 pp spreads), performing poorly on low-value

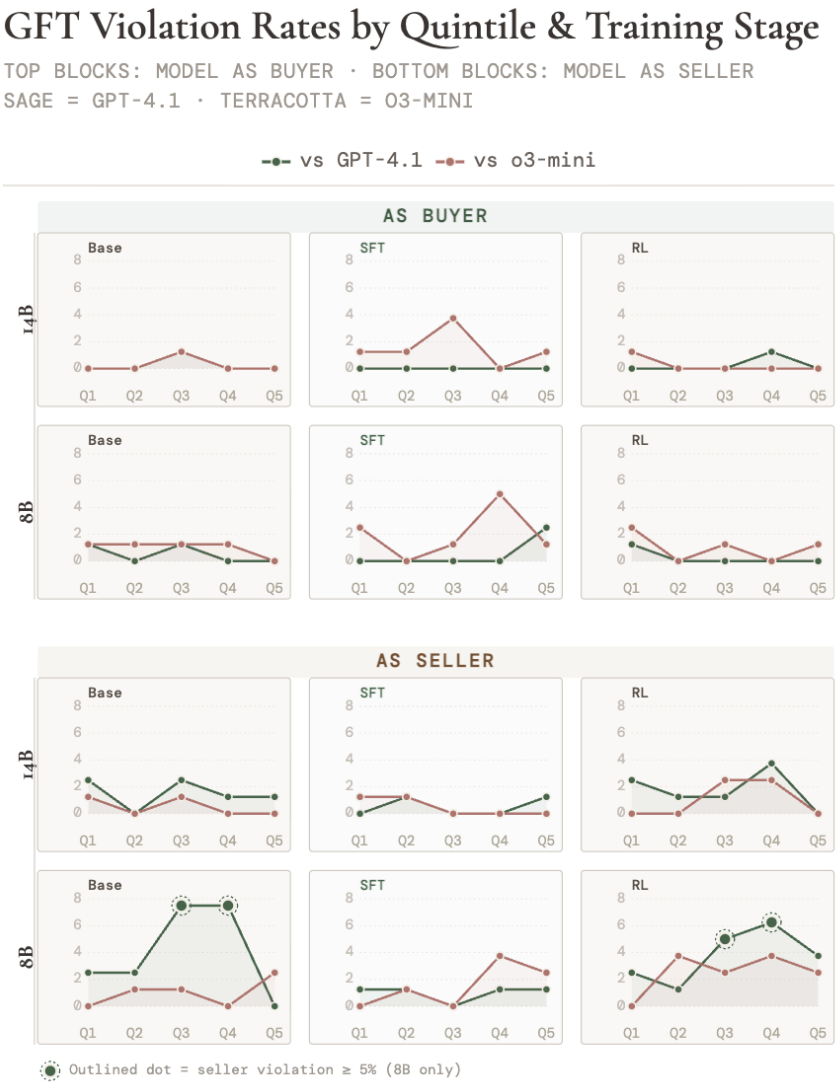

Figure 17: Combined violation rate by reservation-price quintile and training stage, split by opponent. Top row: Qwen3-8B. Bottom row: Qwen3-14B. 4.1 = GPT-4.1; o3 = o3-mini. Violations show no systematic Q1→Q5 trend. Full table in Table 39.

items where ZOPAs are thin. SFT compresses these spreads to 7–9 pp, maintaining 31–40% surplus across all quintiles—even against o3-mini, which is unseen during training (Figure 18). At the hardest condition (8B, Q1, vs. o3-mini), SFT captures 31.5% surplus where Base gets 1.8%. SFT does not memorize specific price points from demonstrations; it learns to anchor and concede *proportionally* to item value, a transferable strategic principle. This mirrors Section 4.2.3, Finding 3: consistency across price tiers is a signature of strategic competence, and behavioral cloning successfully instills it in open-weight models at a fraction of frontier-model cost.

Additional observations. We note two weaker patterns. First, 14B is more resilient to RL than 8B. 14B keeps NGFT seller violations at or below 0.25% across all stages (Table 37). 8B deteriorates: RL raises NGFT seller violations to 6.50% against o3-mini, worse than the untrained base (4.00%). With only two model sizes, we cannot separate capacity from pretraining differences. Second, training changes the model's own behavior but does not induce opponent errors. Frontier counterpart violations remain near zero across all conditions (Tables 36 and 37). This is unsurprising: R_{utility} optimizes the model's own surplus and provides no signal for pressuring opponents.

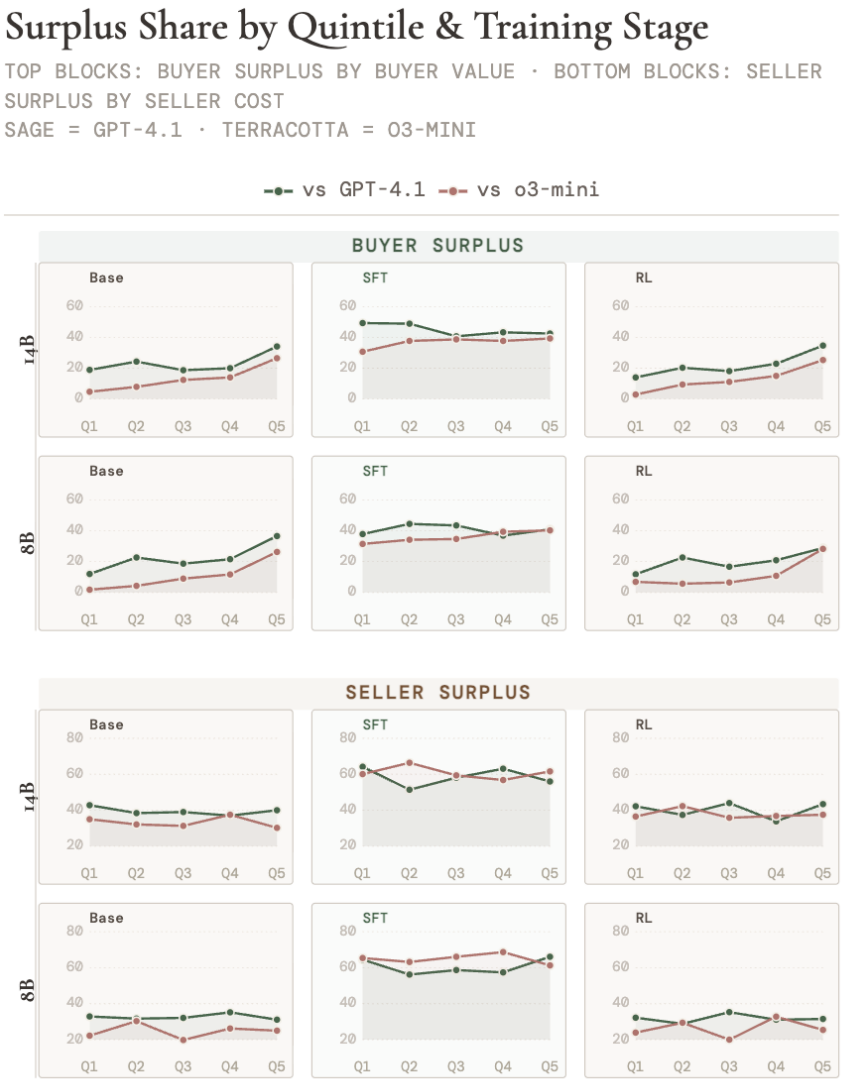

Figure 18: Surplus share by reservation-price quintile and training stage, split by opponent. Top row: Qwen3-8B. Bottom row: Qwen3-14B. 4.1 = GPT-4.1 (RL training opponent); o3 = o3-mini (unseen). SFT compresses buyer-side spreads to 7–9pp; Base and RL show 15–25pp Q1→Q5 spreads. Full table in Table 40.

Looking ahead, three directions may address these limitations: (1) an explicit reward penalty for deals below the reservation price, breaking the symmetry between rational walk-aways and irrational acceptances; (2) training against diverse opponents rather than a single model, to reduce opponent-specific overfitting; and (3) multi-stage reward curricula that first establish IR compliance before optimizing surplus.

6 Discussion: Welfare Efficiency Against Theoretical Benchmarks

Sections 4 and 5 measure performance in relative terms: surplus share and deal rate compare models to each other within each experiment. This section adds an absolute yardstick. We ask how much of the theoretically available gains from trade agents collectively realize, and compare the answer to the Myerson-Satterthwaite efficient frontier.

Figure 19: Deal rate by reservation-price quintile and training stage, split by opponent. Top row: Qwen3-8B. Bottom row: Qwen3-14B. 14B-SFT as seller shows the largest bracket spread (26.3pp vs. GPT-4.1). Base and RL deal rates are high and value-invariant. Full table in Table 41.

6.1 Welfare Efficiency Metrics

The benchmark: theoretical efficiency ratio. The central quantity is

$$\eta^{\text{theory}} = \frac{\text{GFT}^{SB}}{\text{GFT}^{FB}}, \tag{9}$$

the ratio of second-best to first-best expected gains from trade. The two terms have precise meanings. The *first-best* benchmark, $\text{GFT}^{FB} = \mathbb{E}[(V - C)^+]$, is the expected surplus under the efficient trade rule: trade whenever $v \geq c$, realizing the full available gains from trade. The *second-best* benchmark, GFT^{SB}, is the expected surplus under the optimal direct mechanism subject to incentive compatibility, individual rationality, and budget balance (Myerson & Satterthwaite, 1983). Under private information, agents must be compensated for truthful revelation; this information rent is the source of the wedge. Myerson & Satterthwaite (1983) establish that no mechanism can close this gap: $\eta^{\text{theory}} < 1$ whenever types are private and continuously distributed.

To compute GFT^{FB} and GFT^{SB}, we fit approximated empirical distributions of normalized seller costs $c_i = S_i/H_i$ and buyer values $v_i = B_i/H_i$ from the negotiation scenarios (normalizing by each item's historical

highest price H_i to remove dollar-scale heterogeneity), then solve for the optimal mechanism via virtual valuations and bisection. Full computation details are in Section E. Table 9 reports the results for each item pool.

Item Pool	GFT^{FB}	GFT^{SB}	$\alpha^\star$	η^{theory}
Frontier (400 GFT + 200 NGFT)	0.2560	0.2379	0.216	0.929
Qwen (400 GFT + 400 NGFT)	0.2010	0.1857	0.243	0.924

Table 9: Theoretical welfare benchmarks by item pool. GFT^{FB}: expected first-best gains from trade. GFT^{SB}: Myerson-Satterthwaite second-best (optimal IC-IR-BB mechanism). $\alpha^\star$: optimal Lagrange multiplier on the budget-balance constraint. $\eta^{\text{theory}} = \text{GFT}^{SB}/\text{GFT}^{FB}$: theoretical efficiency ratio. All values in normalized units ($H_i = 1$). Computation details in Section E.

Across both item pools, $\eta^{\text{theory}} \approx 0.924$–$0.929$: the best possible mechanism recovers about 93% of first-best surplus, and the remaining 7% is the irreducible cost of private information. This ratio is the benchmark. Any empirical shortfall *beyond* 7% is attributable to behavioral failures—suboptimal strategies, irrational agreements, or missed deals—not to the information structure itself.

The empirical counterpart. To compare LLM outcomes to η^{theory}, we measure realized welfare per negotiation. Items span heterogeneous dollar scales, so we normalize by H_i. For negotiation t with deal indicator D_t, the realized welfare contribution is

$$g_t^{\text{emp}} = D_t \cdot \frac{V_t - C_t}{H_t}. \tag{10}$$

NGFT deals ($V_t < C_t$) enter with a negative value, penalizing irrational agreements. The empirical efficiency ratio for a buyer–seller pair (b, s) is

$$\eta_{(b,s)}^{\text{emp}} = \frac{\widehat{\text{GFT}}_{(b,s)}^{\text{emp}}}{\text{GFT}^{FB}}, \tag{11}$$

where $\widehat{\text{GFT}}^{\text{emp}}$ is the mean of g_t^{emp} over all negotiations in the pair. A pair with $\eta^{\text{emp}} \approx \eta^{\text{theory}}$ is performing at the theoretical ceiling; a pair with $\eta^{\text{emp}} \ll \eta^{\text{theory}}$ is losing welfare to behavior.

6.2 Frontier Models

Table 10 reports $\bar{\eta}^{\text{emp}}$ for each frontier model averaged across all five counterparts in each role; full pairwise results appear in Table 43.

Model	$\bar{\eta}^{\text{emp}}$ **(Buyer)**	$\bar{\eta}^{\text{emp}}$ **(Seller)**
o3	**0.931**	0.891
Gem.-2.5-Pro	0.917	0.859
DeepSeek-V3	0.911	0.853
GPT-4.1	0.860	0.878
Gem.-2.5-Flash	0.756	0.893

Table 10: Frontier models: role-averaged empirical welfare efficiency $\bar{\eta}^{\text{emp}}$, averaged across all five counterparts. $\text{GFT}^{FB} = 0.2560$ (frontier item pool). Theoretical ceiling: $\eta^{\text{theory}} = 0.929$. Full pairwise results in Table 43.

Three patterns stand out. First, the top three buyer-role models—o3 (0.931), Gemini-2.5-Pro (0.917), and DeepSeek-V3 (0.911)—all average at or within 2 pp of $\eta^{\text{theory}} = 0.929$. This does not contradict the MS result. η^{theory} is the expected GFT of the optimal IC-IR-BB direct mechanism and is not a tight upper bound on LLM bargaining outcomes, for two reasons.

1. **No distributional knowledge.** The MS bound requires agents to know the counterparty's type distribution F_V or G_C. LLM agents know only their own reservation price. Without this knowledge they cannot compute virtual valuations and therefore cannot implement the MS trade rule, which withholds some GFT trades (where $v \geq c$ but virtual surplus is negative) to satisfy budget balance.

2. **Cooperative rather than strategic behavior.** o3 is more cooperative than strategically surplus maximization: rather than holding out to maximize its own average surplus, it prioritizes reaching agreement. Its GFT deal rates are the direct evidence: 98.0% as buyer and 95.6% as seller (Section 4.2.1). A purely strategic agent would sometimes refuse a GFT deal to extract better terms, forgoing some efficient trades to improve its average take. o3 does not: it closes GFT negotiations, and this cooperativeness pushes realized surplus toward GFT^{FB}.

Second, Gemini-2.5-Flash is the consistent buyer-role outlier, averaging $\bar{\eta} = 0.756$—10 pp below the next weakest buyer (GPT-4.1 at 0.860). This directly reexpresses Finding 2 from Section 4.2.3: Flash's accommodating profile reduces both surplus share and deal rate, and the welfare efficiency ratio collapses both losses into a single signal. Notably, Flash as *seller* averages $\bar{\eta} = 0.893$, the highest seller value in the table: an accommodating counterpart benefits the seller by conceding quickly.

Third, role assignment is the primary driver of efficiency variation. The buyer-role spread is 17.5 pp (0.756–0.931), more than twice the 7 pp theoretical first-best–second-best gap; the seller-role spread is only 4 pp (0.853–0.893). Buyer behavior—not information asymmetry, and not seller behavior—accounts for most of the realized efficiency shortfall.

6.3 Training Models

Table 11 reports $\bar{\eta}^{\text{emp}}$ for each Qwen variant averaged across GPT-4.1 and o3-mini in each role; full pairwise results appear in Table 44.

Model	$\bar{\eta}^{\text{emp}}$ **(Buyer)**	$\bar{\eta}^{\text{emp}}$ **(Seller)**
Qwen3-14B	0.876	0.870
Qwen3-14B-RL	0.864	0.888
Qwen3-14B-SFT	0.638	0.410
Qwen3-8B	0.680	0.710
Qwen3-8B-RL	0.666	0.695
Qwen3-8B-SFT	0.765	0.603

Table 11: Qwen variants: role-averaged empirical welfare efficiency $\bar{\eta}^{\text{emp}}$, averaged across GPT-4.1 and o3-mini as counterparts. $\text{GFT}^{FB} = 0.2010$ (Qwen item pool). Theoretical ceiling: $\eta^{\text{theory}} = 0.924$. Full pairwise results in Table 44.

Three findings emerge. First, SFT-only variants are the worst performers for joint welfare, despite producing the highest individual surplus shares in Section 5.4. Qwen3-14B-SFT averages $\bar{\eta} = 0.410$ as seller—the lowest seller value in either table—because its deal-rate collapse (Section 5.4.2) leaves GFT surplus unrealized entirely. High conditional surplus share is irrelevant if GFT negotiations never close. Qwen3-8B-SFT as seller performs better ($\bar{\eta} = 0.603$) because its deal-rate decline is more moderate (Table 38). The table also reveals a SFT-induced role asymmetry: both SFT variants achieve higher $\bar{\eta}$ as buyer than as seller, reflecting the "hold out for good terms" behavior that raises surplus share when the counterpart concedes but collapses deals when it does not.

Second, RL and base models occupy the top of the table. Qwen3-14B-RL reaches $\bar{\eta} = 0.888$ as seller and 0.864 as buyer, within 4 pp of $\eta^{\text{theory}} = 0.924$. The V-shape in deal rates documented in Section 5.4.2 maps directly onto a V-shape in η: SFT collapses joint welfare; RL recovers it.

Third, the gap to frontier models is modest in the seller role. The best Qwen seller average (Qwen3-14B-RL at 0.888) is within 0.5 pp of the best frontier seller average (Gemini-2.5-Flash at 0.893). The larger gap is in the buyer role: the best Qwen buyer average (Qwen3-14B at 0.876) falls 5.5 pp below the best frontier buyer (o3 at 0.931). For a pilot study at 8B–14B scale against a single training opponent, this buyer-role gap is the primary remaining shortfall.

6.4 Synthesis

Two observations follow from the welfare efficiency analysis.

The MS impossibility result in practice. Myerson & Satterthwaite (1983) establish that private information causes an irreducible efficiency loss; the theoretical shortfall here is 7%. The empirical shortfall for the weakest pairings reaches 30–60%. Behavioral failures—accommodating concession patterns, deal-rate collapse under SFT, and NGFT constraint violations—account for most of the observed efficiency loss. Private information, the source of the theoretical wedge, is secondary.

The SFT–RL tension reframed. Section 5.4.4 diagnoses the SFT–RL tension as a conflict between surplus share and deal rate. The η metric reduces both to a single welfare signal and makes the cost of SFT's selectivity concrete: Qwen3-14B-SFT as seller destroys roughly 50% of theoretically available joint welfare relative to the same model's base and RL variants. An explicit IR penalty and a staged reward curriculum—first enforcing deal closure in GFT scenarios, then optimizing surplus—would target η directly rather than optimizing the two objectives in tension.

7 Conclusion

7.1 Summary

This paper presents a structured bilateral bargaining environment that serves two purposes: a three-dimensional evaluation framework for frontier LLMs and a training environment for fine-tuning open-weight models. The evaluation framework measures individual rationality, strategic effectiveness, and allocative efficiency through automated metrics computed from machine-parseable negotiation traces. A round-robin benchmark of five frontier models reveals that aggressive anchoring with calibrated concession dominates accommodating strategies, and that strategic competence manifests as consistency across price tiers. A two-stage training pipeline (SFT followed by GRPO) identifies a tension between the selectivity instilled by behavioral cloning and the deal-closing bias introduced by online RL—a diagnosis that points to specific reward design improvements.

7.2 Limitations

Benchmark scope. The benchmark covers a single negotiation domain (consumer goods), one protocol (alternating offers with structured tool calls), and five frontier models evaluated at a single point in time. Results may not transfer to multi-issue bargaining, multi-party settings, or future model generations. The round-robin design captures pairwise dynamics but does not test against human counterparts, whose strategic and communicative behavior may differ qualitatively from LLM-vs-LLM play.

Training scale. The training experiments are a pilot study: two model sizes (8B, 14B), one training opponent (GPT-4.1), and one RL algorithm (GRPO). The SFT–RL tension we identify may be addressable with better reward design, but the current study diagnoses the problem without resolving it. The evaluation against only two frontier counterparts (GPT-4.1, o3-mini) limits the generalizability of transfer conclusions.

Statistical precision. Some results rest on limited sample sizes: 80 scenarios per price-tier bracket, 200 NGFT scenarios per pairing in the frontier benchmark. Bracket-level violation rates (e.g., 3.75% = 3 negotiations out of 80) should be interpreted as directional rather than precise estimates.

7.3 Future Directions

Reward design for RL. The central training finding—that RL erodes SFT's selectivity—traces to a reward structure that assigns $R_{\text{utility}} = 0$ to both rational walk-aways and irrational acceptances (Section 5.4.4). An explicit penalty for deals below the reservation price would break this symmetry. More broadly, a staged

reward curriculum that first establishes IR compliance and then optimizes surplus may prevent RL from trading rationality for deal completion.

Diverse training opponents. Training against a single fixed opponent risks learning opponent-specific strategies. Training against a population of models—or against earlier checkpoints of the trainee itself—would test whether the learned improvements reflect general strategic competence rather than exploitation of a particular counterpart's behavioral profile.

Multi-issue and multi-party extensions. The current framework negotiates a single price for a single item. Extending to multiple negotiable attributes (delivery terms, warranties, bundled items) would increase the strategic complexity and create opportunities for integrative bargaining, where agents create value by trading across issues rather than dividing a fixed surplus. Multi-party extensions (e.g., procurement auctions with multiple suppliers) would connect to a broader set of operations research applications.

Human–AI negotiation. The behavioral patterns documented here—anchoring effects, concession dynamics, first-mover advantage—are well-established in human negotiation research. Deploying the agent system with human counterparts would test whether these patterns transfer from LLM-vs-LLM play, and whether humans respond to LLM strategies in the same way that other LLMs do.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599, 2024.

Zishan Ahmad, Suman Saurabh, Vaishakh Menon, Asif Ekbal, Roshni Ramnani, and Anutosh Maitra. Ina: An integrative approach for enhancing negotiation strategies with reward-based dialogue agent. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2536–2549, 2023.

Tim Baarslag, Koen Hindriks, Catholijn Jonker, Sarit Kraus, and Raz Lin. The first automated negotiating agents competition (anac 2010). In *New Trends in agent-based complex automated negotiations*, pp. 113–135. Springer, 2012.

Tim Baarslag, Mark JC Hendrikx, Koen V Hindriks, and Catholijn M Jonker. Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*, 30(5):849–898, 2016.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Max H Bazerman and Margaret Ann Neale. *Negotiating Rationally*. Free Press, 1992.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis. *arXiv preprint arXiv:2402.05863*, 2024.

Ho-Chun Herbert Chang. Multi-issue bargaining with deep reinforcement learning. *arXiv preprint arXiv:2002.07788*, 2020.

Ainesh Chatterjee, Samuel Miller, and Nithin Parepally. Agreemate: Teaching llms to haggle. *arXiv preprint arXiv:2412.18690*, 2024.

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

Tim R Davidson, Veniamin Veselovsky, Martin Josifoski, Maxime Peyrard, Antoine Bosselut, Michal Kosinski, and Robert West. Evaluating language model agency through negotiations. *arXiv preprint arXiv:2401.04536*, 2024.

DeepSeek-AI. DeepSeek-V3 technical report. Technical report, DeepSeek, 2025. Available at: `https://arxiv.org/abs/2412.19437`.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*, 2023.

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

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Etan A Green and E Barry Plunkett. The science of the deal: Optimal bargaining on ebay using deep reinforcement learning. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 1–27, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. Technical report, DeepSeek, 2025. arXiv preprint arXiv:2501.12948.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of EMNLP*, pp. 2333–2343, 2018a. doi: 10.18653/v1/D18-1256.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. *arXiv preprint arXiv:1808.09637*, 2018b.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2021. arXiv:2106.09685.

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

Simon Keizer, Markus Guhe, Heriberto Cuayáhuitl, Ioannis Efstathiou, Klaus-Peter Engelbrecht, Mihai Dobre, Alexandra Lascarides, and Oliver Lemon. Evaluating persuasion strategies and deep reinforcement learning methods for negotiation dialogue agents. In *The 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 480–484. Association for Computational Linguistics, 2017.

Dexin Kong, Xu Yan, Ming Chen, Shuguang Han, Jufeng Chen, and Fei Huang. Fishbargain: An llm-empowered bargaining agent for online fleamarket platform sellers. *arXiv preprint arXiv:2502.10406*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention, 2023. arXiv:2309.06180.

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*, 2017.

Kenneth Li, Yiming Wang, Fernanda Viégas, and Martin Wattenberg. Dialogue action tokens: Steering language models in goal-directed dialogue with a multi-turn planner. *arXiv preprint arXiv:2406.11978*, 2024.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 2024.

Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981. doi: 10.1287/moor.6.1.58.

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

John F. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950. doi: 10.2307/1907266.

Jihwan Oh, Murad Aghazada, Se-Young Yun, and Taehyeon Kim. Llm agents for bargaining with utility-based feedback. *arXiv preprint arXiv:2505.22998*, 2025.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. o3 and o4-mini system card. Technical report, OpenAI, 2025. Available at: https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Qwen Team. Qwen3 technical report. Technical report, Alibaba Cloud, 2025. Available at: https://arxiv.org/abs/2505.09388.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models, 2020. arXiv:1910.02054.

John Riley and Richard Zeckhauser. Optimal selling strategies: When to haggle, when to hold firm. *The Quarterly Journal of Economics*, 98(2):267–289, 1983.

Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109, 1982. doi: 10.2307/1912531.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. Introduces GRPO.

Joel Sobel and Ichiro Takahashi. A multistage model of bargaining. *The Review of Economic Studies*, 50(3): 411–426, 1983.

Leigh Thompson and Dennis Hrebec. Lose–lose agreements in interdependent decision making. *Psychological Bulletin*, 120(3):396–409, 1996.

Siddharth Verma, Justin Fu, Mengjiao Yang, and Sergey Levine. Chai: A chatbot ai for task-oriented dialogue with offline reinforcement learning. *arXiv preprint arXiv:2204.08426*, 2022.

Volcano Engine. VERL: Volcano engine reinforcement learning for LLM, 2024. https://github.com/volcengine/verl.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring bargaining abilities of llms: A benchmark and a buyer-enhancement method. *arXiv preprint arXiv:2402.15813*, 2024.

Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. Improving dialog systems for negotiation with personality modeling. *arXiv preprint arXiv:2010.09954*, 2020.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

A Game Environment Details

This appendix provides implementation details for the bilateral bargaining environment described in Section 3.

A.1 Example Product Listing

Title: Dyson Supersonic Hair Dryer Nickel/Copper

Category: Other
Description:

- The Dyson Supersonic hair dryer is engineered to protect hair from extreme heat damage for fast drying and precision styling.
- Fast drying: The small, powerful Dyson digital motor V9 combined with Air Multiplier technology produces a high-pressure, high-velocity jet of controlled air, for fast drying and precision styling.
- Intelligent heat control: Unlike some others, the Dyson Supersonic hair dryer measures air temperature over 40 times a second, and regulates the heat. This prevents extreme heat damage, to help protect your hair's shine.
- Engineered for different hair types: After rigorously testing different hair types in our laboratories, Dyson has engineered a range of attachments designed to style different types of hair.
- Five attachments to style different hair types: Flyaway attachment hides flyaways under longer hair for a smooth, shiny finish. Styling concentrator For precision styling. Diffuser Helps reduce frizz and create defined curls and waves. Gentle air attachment Fast yet gentle styling for fine hair and sensitive scalps. Wide-tooth comb Helps shape and lengthen curly and textured hair as it dries.

Price History:

- Highest: $429.99 (Sep 07, 2023)
- Lowest: $365.49 (Oct 12, 2023)

Figure 20: Example listing used in the negotiation simulation environment.

A.2 Tool Specification

Table 12 lists the complete tool specification available to each agent. All tools take an `agent_name` parameter ("buyer" or "seller") identifying the caller; this parameter is omitted from the main text for brevity. The main text (Table 1) lists the six primary tools; the full specification below includes an additional tool (`wait_for_time_period`) used to model tactical delays.

Tool	Parameters	Effect
`make_offer`	`price`, `side_offer` (opt.)	Proposes a formal price; becomes pending offer that the counterpart must respond to. Price must be positive.
`respond_to_offer`	`response` (Boolean)	Accepts (`True`) or rejects (`False`) the pending offer. Acceptance concludes the negotiation at the offered price.
`send_message`	`content` (string)	Sends a free-form natural-language message to the counterpart. No direct effect on negotiation state.
`search_price`	(none)	Returns historical high and low prices for the item. Information only.
`quit_negotiation`	(none)	Exercises the outside option (BATNA). Negotiation terminates immediately with no deal; both utilities are zero.
`wait_for_response`	(none)	Yields the turn to the counterpart. Must be the last tool called in a turn.
`wait_for_time_period`	`duration` (seconds)	Advances the simulation clock by the specified duration. Models tactical delays in real-world negotiation.

Table 12: Complete tool specification for the bilateral bargaining environment. Each turn permits at most 3 tool calls. The main text (Table 1) lists the six primary tools; `wait_for_time_period` is an additional tool used to model tactical delays.

A.3 Event Protocol

The simulation engine uses an event-driven architecture managed by an `EventManager` class. The core components are:

- *Event queue:* a min-heap priority queue sorted by timestamp.
- *Simulation clock:* tracks current simulation time.
- *Event batching:* groups simultaneous events from the same agent into a single observation for the counterpart.

Each tool call creates an event with four fields: timestamp, type (matching the tool name), acting agent, and tool-specific data. The event processing loop proceeds as follows:

1. Pop the next event batch from the queue (events sharing the same timestamp and agent).
2. Advance the simulation clock to the event timestamp.
3. Convert events to a natural-language observation (e.g., "Seller proposed $1,500"; "Buyer rejected your offer and proposed $1,200").
4. Route the observation to the counterpart agent.
5. The counterpart generates a `Thought:` block (internal reasoning, hidden from the opponent) followed by a `Code:` block (tool calls).
6. The simulator parses and executes the tool calls, creating new events.
7. Check termination conditions; if not met, return to step 1.

A.4 Example Negotiation Transcripts

GFT example. Item: "Used Laptop" (historical high $1,500, historical low $800). Buyer reservation price $b = \$1{,}200$; seller reservation price $s = \$900$. ZOPA $= [900, 1200]$; total surplus $= \$300$.

Round	Agent	Action
1	Seller	`make_offer(1400)`
2	Buyer	`send_message("Your price is too high for a used laptop")`; `make_offer(950)`
3	Seller	`send_message("I can't go that low, but I can offer a discount")`; `make_offer(1150)`
4	Buyer	`make_offer(1050)`
5	Seller	`respond_to_offer(True)`

Table 13: GFT example: deal at $p^* = \$1{,}050$. Buyer utility $U_B = 150$, seller utility $U_S = 150$. Surplus shares: 50/50.

NGFT example. Same item. Buyer reservation price $b = \$850$; seller reservation price $s = \$1{,}100$. No ZOPA exists (NGFT scenario).

Round	Agent	Action
1	Seller	`make_offer(1300)`
2	Buyer	`make_offer(800)`
3	Seller	`send_message("I can't go below $1,100")`
4	Buyer	`send_message("I can't go above $850")`
5	Buyer	`quit_negotiation()`

Table 14: NGFT example: no deal (rational outcome). Both utilities are zero. The buyer correctly identifies the infeasible negotiation and exercises the outside option.

B Training Implementation Details

This appendix provides implementation details for the training pipeline described in Section 5.

B.1 SFT Data Pipeline and Hyperparameters

Data pipeline. The SFT data pipeline has four stages:

1. *Demonstration generation.* DeepSeek-R1 self-play negotiations are generated within the simulator using the same system prompt template, tool definitions, and item catalog used in the benchmark (Section 4.1).
2. *Reasoning-trace cleaning.* `<think>...</think>` reasoning blocks are removed from input prompts (but retained in target completions), as described in Section 5.2.
3. *Turn-level decomposition.* Each multi-turn trajectory is decomposed into autoregressive training samples of increasing context length (Section 5.2).
4. *Format conversion.* Samples are converted to input–output pairs, where the input concatenates the system prompt and conversation history and the output is the target agent response. Loss is computed only on output tokens.

SFT hyperparameters. Table 15 lists the SFT training hyperparameters for both model sizes. Training uses the TRL SFTTrainer with DeepSpeed ZeRO Stage 3 (Rajbhandari et al., 2020) on 2 GPUs.

Parameter	8B Model	14B Model
Optimization		
Learning rate	2×10^{-5}	1×10^{-5}
LR schedule	Cosine	Cosine
Warmup ratio	0.1	0.1
Number of epochs	3	3
Weight decay	0.01	0.01
Max gradient norm	0.1	0.1
Per-device batch size	2	1
Gradient accumulation steps	8	8
Precision	bfloat16	bfloat16
Sequence		
Max sequence length	8,192	8,192
Memory efficiency		
Fine-tuning mode	Full	Full
Gradient checkpointing	✓	✓
DeepSpeed ZeRO stage	3	3
Number of GPUs	2	2

Table 15: SFT training hyperparameters for 8B and 14B models. Both models are trained with full parameter updates (no LoRA). The effective batch size is per-device batch × accumulation steps × GPUs: 32 for 8B, 16 for 14B.

B.2 GRPO Architecture and Key Parameters

Hardware and GPU allocation. All training runs use a single node with 8×H100 80 GB GPUs. We separate inference and training onto dedicated GPU sets so that rollout generation and gradient computation overlap:

- *Inference cluster (GPUs 0–5):* runs vLLM (Kwon et al., 2023) with tensor parallelism TP = 2 and data parallelism DP = 3, processing multiple negotiations in parallel with up to 16K-token contexts. Prefix caching is enabled to reuse KV states across rollouts sharing the same system prompt.
- *Training cluster (GPUs 6–7):* runs DeepSpeed ZeRO Stage 3 (Rajbhandari et al., 2020) for memory-efficient gradient computation. Optimizer states and model parameters are sharded across both GPUs.

Key hyperparameters. Table 16 lists the GRPO training hyperparameters for both model sizes.

Parameter	8B Model	14B Model
Optimization		
Learning rate	3×10^{-6}	2×10^{-6}
Warmup steps	20	20
Max training steps	200	200
Per-device batch size	1	1
Gradient accumulation steps	16	16
Effective batch size	128	128
Precision	bfloat16	bfloat16
GRPO		
Rollouts per prompt (G)	8	8
Clipping range (ϵ)	0.2	0.2
KL penalty (β)	0.05	0.005
Stability constant (ϵ_{stable})	10^{-4}	10^{-4}
Memory efficiency		
LoRA rank	— (full)	64
Gradient checkpointing	✓	✓
DeepSpeed ZeRO stage	3	3
Generation		
Max sequence length	16,384	16,384
Temperature	1.0	1.0
vLLM tensor parallelism	2	2
Weight sync frequency	Every 8 grad. steps	Every 8 grad. steps

Table 16: GRPO training hyperparameters for 8B and 14B models. The 8B model is trained with full parameters; the 14B model uses LoRA (rank 64, <1% trainable parameters) due to memory constraints.

B.3 Context Length Growth

Context length grows over the course of a negotiation. At round t, the total context length is

$$L_t = L_{\text{init}} + \sum_{i=1}^{t} \big(L^i_{\text{prompt}} + L^i_{\text{response}} + L^i_{\text{tool}} + L^i_{\text{result}} \big), \tag{12}$$

where the per-component token counts are approximately:

- $L_{\text{init}} \approx 2{,}500$ tokens (role assignment, constraints, instructions).
- $L^i_{\text{prompt}} \approx 500$ tokens (per-turn context update).
- $L^i_{\text{response}} \approx 300$ tokens (agent-generated code).
- $L^i_{\text{tool}} \approx 100$ tokens (tool execution results).
- $L^i_{\text{result}} \approx 200$ tokens (environment feedback).

This totals ∼1,100 tokens per round, exhausting a 16K context window in roughly 12–13 rounds. The maximum context length for training is set to 13,312 tokens.

B.4 Gradient-Safe Overflow Handling

When context overflow occurs (i.e., $L_t > 13{,}312$), GRPO still requires valid logprobs for the episode to form a well-shaped training batch. The overflow handling proceeds as follows:

1. Generate a synthetic `quit_negotiation` response (with `Thought`/`Code` blocks matching the expected format).

2. Attach synthetic logprobs using the PAD token (token ID 151643, `<|endoftext|>` in Qwen models) with a nominal logprob of -0.1.
3. Execute the synthetic response through the environment, terminating the negotiation with no deal.
4. Assign reward $R = 0$.

The choice of the PAD token is critical. An early implementation used token ID 1 (the double-quote character `"` in Qwen's vocabulary), which created real gradient updates: with a negative advantage from the failed negotiation, GRPO would systematically decrease the probability of generating double quotes after long contexts. The PAD token avoids this because $\pi_\theta(\text{PAD} \mid s) \approx 0$ under normal generation—the model never produces this special token during inference, so gradient updates to its probability are harmless noise. Formally, the gradient contribution from an overflowed episode is

$$\nabla_\theta \mathcal{L} = -\hat{A}^g \cdot \nabla_\theta \log \pi_\theta(\text{PAD} \mid s), \tag{13}$$

where $\hat{A}^g$ is near zero (the reward of zero is close to the group mean) and $\nabla_\theta \log \pi_\theta(\text{PAD} \mid s)$ points in a direction irrelevant to normal text generation. Both factors ensure negligible parameter updates.

The context overflow rate is approximately 3.2% at the start of training and decreases as the model learns to conclude negotiations within fewer rounds.

B.5 Weight Synchronization Protocol

Synchronizing the vLLM inference server with the training process after every gradient step creates prohibitive overhead. We update weights every `gradient_accumulation_steps` $\times$ `num_iterations` $= 8$ gradient steps (Table 16), reducing synchronization overhead while maintaining acceptable policy lag. The protocol proceeds as follows:

1. The training process computes gradients over accumulated batches.
2. After 8 gradient steps, it initiates weight synchronization.
3. The vLLM server atomically swaps model weights.
4. Subsequent rollouts use the updated policy immediately.

B.6 Training Dynamics

The training process naturally progresses through three phases:

1. **Tool mastery (steps 1–20).** The agent learns to generate well-formed tool calls and execute valid code. Rewards are dominated by R_{parsing} and $R_{\text{execution}}$ (Equation (8) and Table 8), as successful negotiations are rare.
2. **Constraint awareness (steps 20–40).** The agent begins respecting IR constraints, learning to reject offers that yield negative utility. IR violation rates drop from >30% to <5%.
3. **Strategic optimization (steps 40–58).** With basic competence established, the agent develops negotiation strategies: anchoring with aggressive opening offers, progressive concessions, and strategic outside-option exercise.

C Frontier Benchmark: Supplementary Results

This appendix provides supplementary figures and tables for the frontier model benchmark (Section 4).

C.1 Aggregate Tables

The tables below provide the full numerical values underlying the figures in Section 4.2.

C.1.1 IR Compliance

		GFT Violation (%)		NGFT Violation (%)	
Model	**Role**	Self	Opp.	Self	Opp.
DeepSeek-V3	Buyer	0.45	0.15	0.10	0.40
	Seller	0.70	0.45	0.60	0.00
Gem.-2.5-Flash	Buyer	0.20	0.25	0.40	0.00
	Seller	0.10	0.10	**1.90**	0.00
Gem.-2.5-Pro	Buyer	0.10	0.20	**0.00**	**0.00**
	Seller	**0.00**	**0.00**	0.70	0.00
GPT-4.1	Buyer	0.25	0.15	0.20	0.10
	Seller	0.10	0.20	**2.30**	0.20
o3	Buyer	**0.05**	0.20	0.10	**5.00**
	Seller	0.05	**0.30**	**0.00**	0.60

Table 17: IR compliance for frontier models. Self: the model's own violation rate (deals yielding negative utility for itself). Opp.: the opponent's violation rate when facing this model. All values averaged over all five opponents. Bold highlights best and worst values. In NGFT, any deal is a violation for at least one party; o3 as buyer induces 5.00% opponent violations despite only 0.10% self-violations.

o3 as Buyer vs.	**o3 Self-Viol. (%)**	**Seller Opp.-Viol. (%)**
GPT-4.1	0.5	**11.5**
Gem.-2.5-Flash	0.0	**9.5**
Gem.-2.5-Pro	0.0	3.5
DeepSeek-V3	0.0	0.5
o3	0.0	0.0

Table 18: NGFT pairwise breakdown for o3 as buyer. Self-violation = o3's own irrational deals; opponent-violation = seller's irrational deals when facing o3. The irrationality comes almost entirely from opponents: GPT-4.1 and Gemini-2.5-Flash violate at 11.5% and 9.5%, while o3 itself violates at most 0.5%.

C.1.2 Strategic Effectiveness and Behavioral Drivers

	Surplus (%)		Deal Rate (%)		Init. Aggr.		Conc. Rate		Patience	
Model	B	S	B	S	B	S	B	S	B	S
DeepSeek-V3	35.1	49.6	94.4	92.1	0.34	1.97	0.52	0.50	6.65	5.98
Gem.-2.5-Flash	24.3	57.6	86.4	94.9	0.37	2.14	0.67	0.54	5.72	6.02
Gem.-2.5-Pro	44.9	72.6	96.7	91.7	0.40	2.06	0.51	0.43	7.07	7.06
GPT-4.1	34.1	57.7	94.0	95.2	0.36	2.07	0.53	0.52	6.07	6.08
o3	47.2	77.0	98.0	95.6	0.39	3.04	0.47	0.55	6.50	6.86

Table 19: Outcome metrics and behavioral drivers for frontier models (GFT scenarios), averaged over all five opponents. B = buyer, S = seller. Surplus: fraction of ZOPA captured conditional on agreement. Deal rate: fraction of negotiations reaching agreement. Init. Aggr. (buyer): fraction of the seller's asking price cut by the buyer's first counter-offer (higher = more aggressive). Init. Aggr. (seller): ratio of first offer to reservation price (higher = more aggressive). Conc. Rate: fraction of remaining surplus conceded per round. Patience: mean number of rounds to termination (max 10).

C.1.3 Radar-Plot Profile Tables

Model	IR Compl. (%)	Surplus (%)	Deal Rate (%)	Anchor.	Consistency (%)	Pressure Res. (%)
o3	99.95	47.2	98.0	0.39	93.9	99.90
Gem.-2.5-Pro	99.90	44.9	96.7	0.40	94.0	100.00
GPT-4.1	99.75	34.1	94.0	0.36	87.5	99.80
DeepSeek-V3	99.55	35.1	94.4	0.34	93.6	99.90
Gem.-2.5-Flash	99.80	24.3	86.4	0.37	87.2	99.60

Table 20: Buyer radar-plot profile for frontier models. IR Compl. = 100 − GFT self-violation rate. Surplus: buyer surplus share (GFT). Deal Rate: buyer GFT deal rate. Anchor.: fraction of the seller's asking price cut by the buyer's first counter-offer. Consistency = 100 − quintile surplus spread (pp). Pressure Res. = 100 − NGFT self-violation rate. All axes: higher is better.

Model	IR Compl. (%)	Surplus (%)	Deal Rate (%)	Anchor.	Consistency (%)	Pressure Res. (%)
o3	99.95	77.0	95.6	3.04	96.0	100.00
Gem.-2.5-Pro	100.00	72.6	91.7	2.06	94.0	99.30
GPT-4.1	99.90	57.7	95.2	2.07	96.2	97.70
DeepSeek-V3	99.30	49.6	92.1	1.97	90.0	99.40
Gem.-2.5-Flash	99.90	57.6	94.9	2.14	93.4	98.10

Table 21: Seller radar-plot profile for frontier models. IR Compl. = 100 − GFT self-violation rate. Surplus: seller surplus share (GFT). Deal Rate: seller GFT deal rate. Anchor.: ratio of first offer to reservation price. Consistency = 100 − quintile surplus spread (pp). Pressure Res. = 100 − NGFT self-violation rate. All axes: higher is better.

C.2 Pairwise Heatmaps

Each heatmap shows the metric value for all 25 buyer–seller pairings (rows = buyer models, columns = seller models) in GFT scenarios unless otherwise noted.

C.2.1 Violation Rate

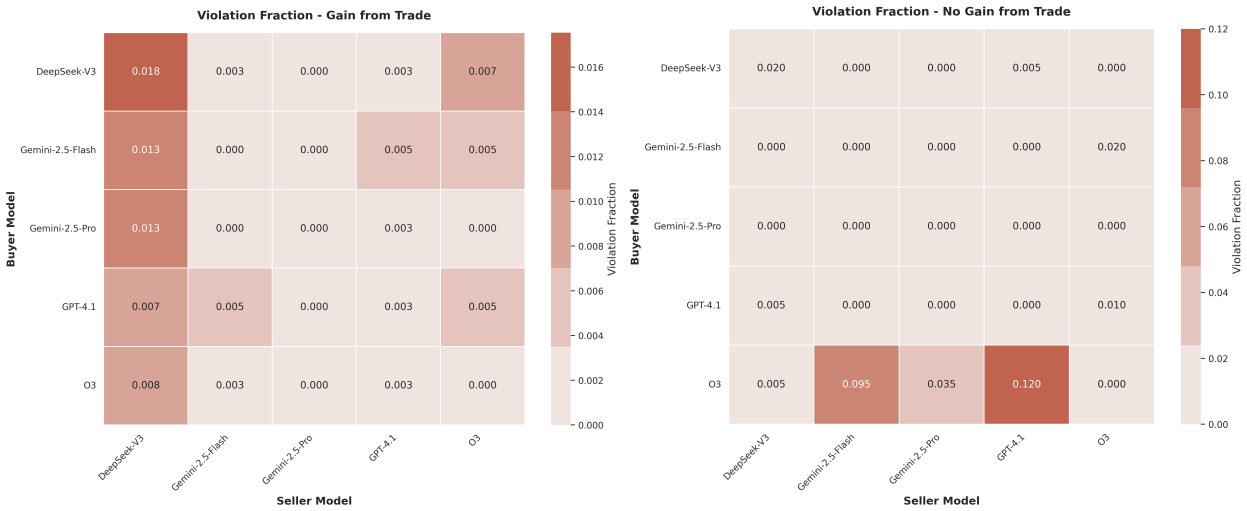

Figure 21: Violation rate heatmaps for all 25 buyer–seller pairings. Left: GFT scenarios (violations arise from deals struck outside the ZOPA). Right: NGFT scenarios (every deal is a violation for at least one party, so violation rate = deal rate).

C.2.2 Average Utility

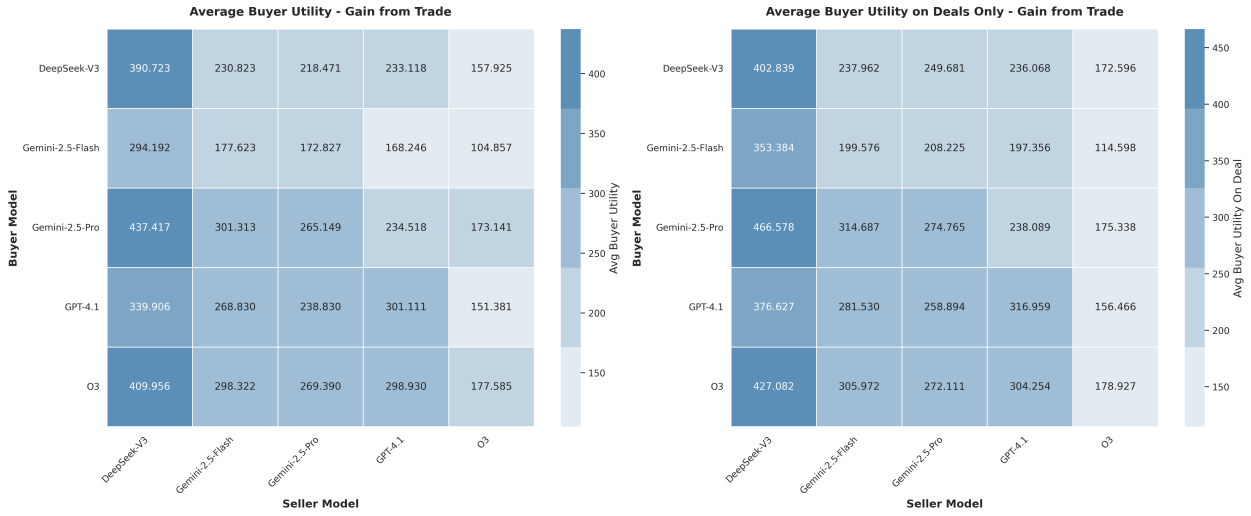

Figure 22: Average buyer utility in GFT scenarios. Left: averaged over all negotiations (including no-deal outcomes at utility = 0). Right: conditional on agreement.

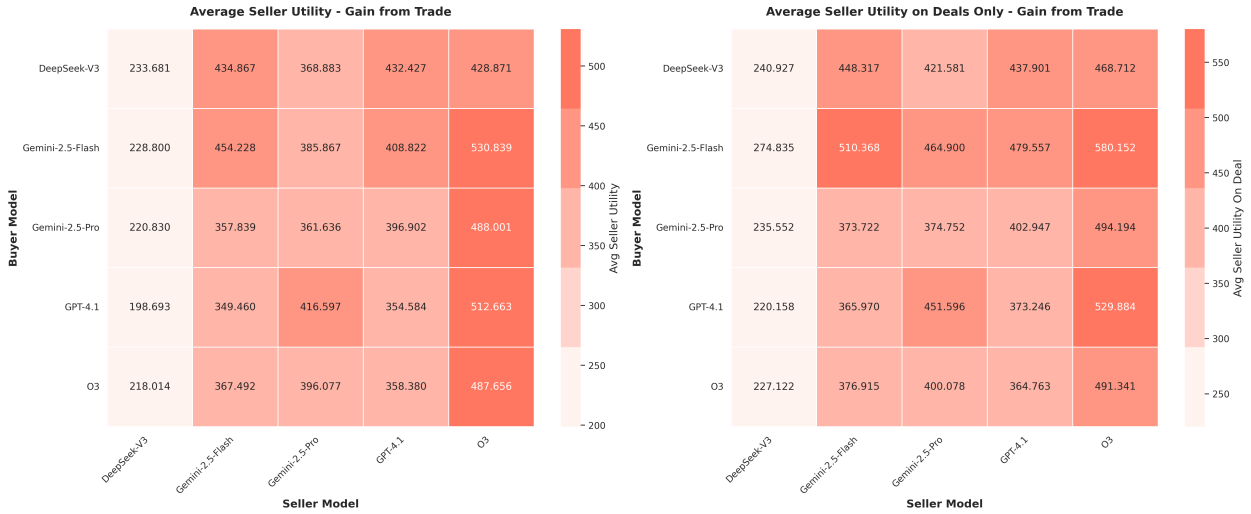

Figure 23: Average seller utility in GFT scenarios. Left: averaged over all negotiations. Right: conditional on agreement.

C.2.3 Surplus Share

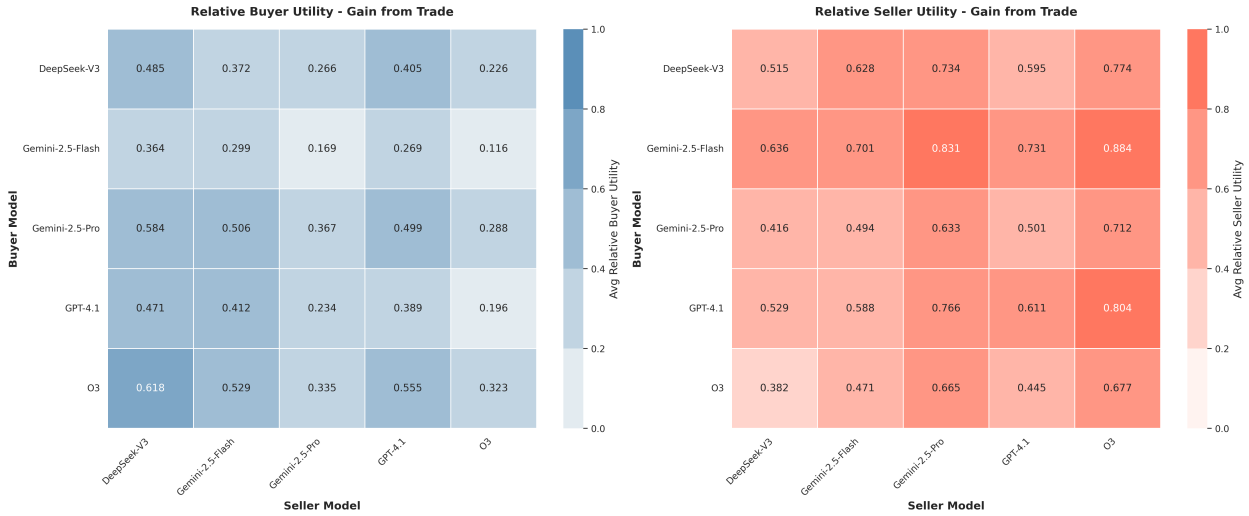

Figure 24: Surplus share heatmaps for all 25 buyer–seller pairings (GFT scenarios). Left: buyer surplus share. Right: seller surplus share. The two shares sum to one for each cell.

C.2.4 Deal Rate

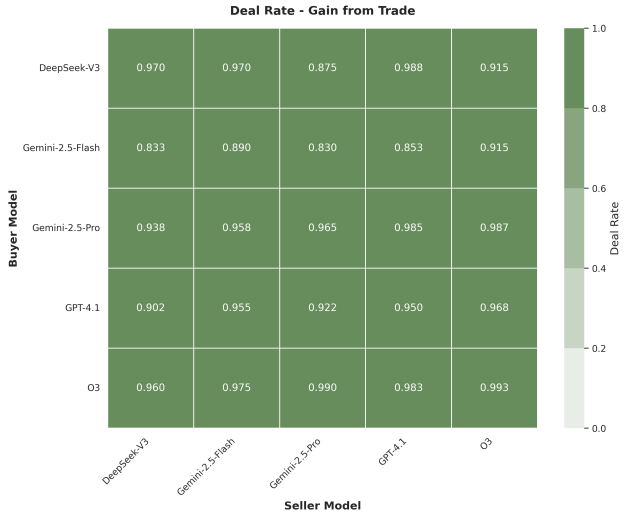

Figure 25: Deal rate for all 25 buyer–seller pairings (GFT scenarios). NGFT deal rates equal violation rates (Figure 21) and are omitted.

C.2.5 Initial Aggressiveness

Figure 26: Seller initial aggressiveness (first offer / reservation price). Higher values indicate more aggressive opening offers.

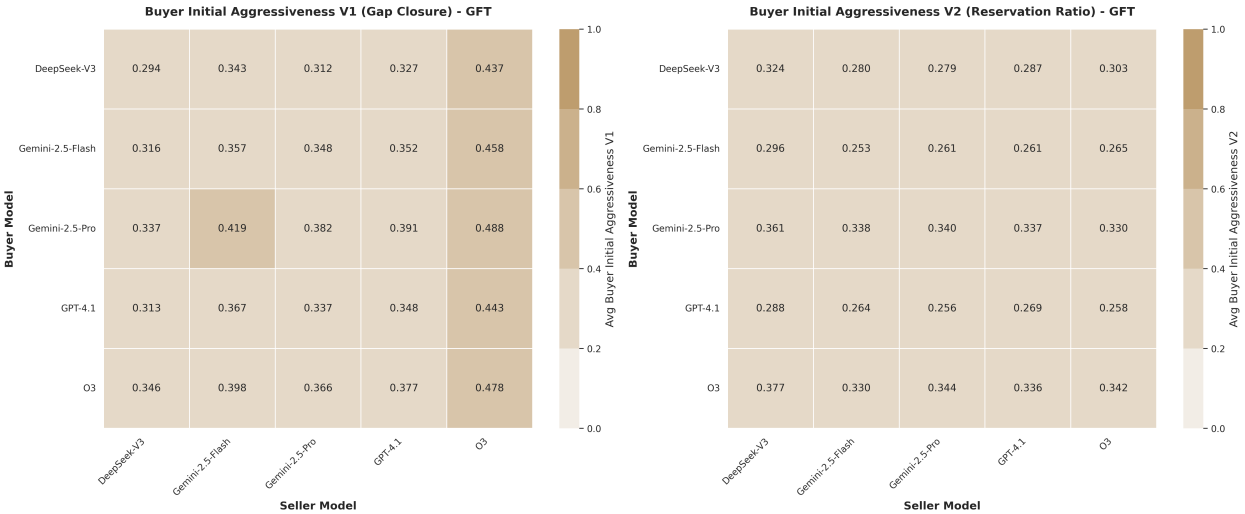

Figure 27: Buyer initial aggressiveness. Left: V1 (gap closure). Right: V2 (reservation ratio). Lower values indicate more aggressive opening offers.

C.2.6 Concession Rate

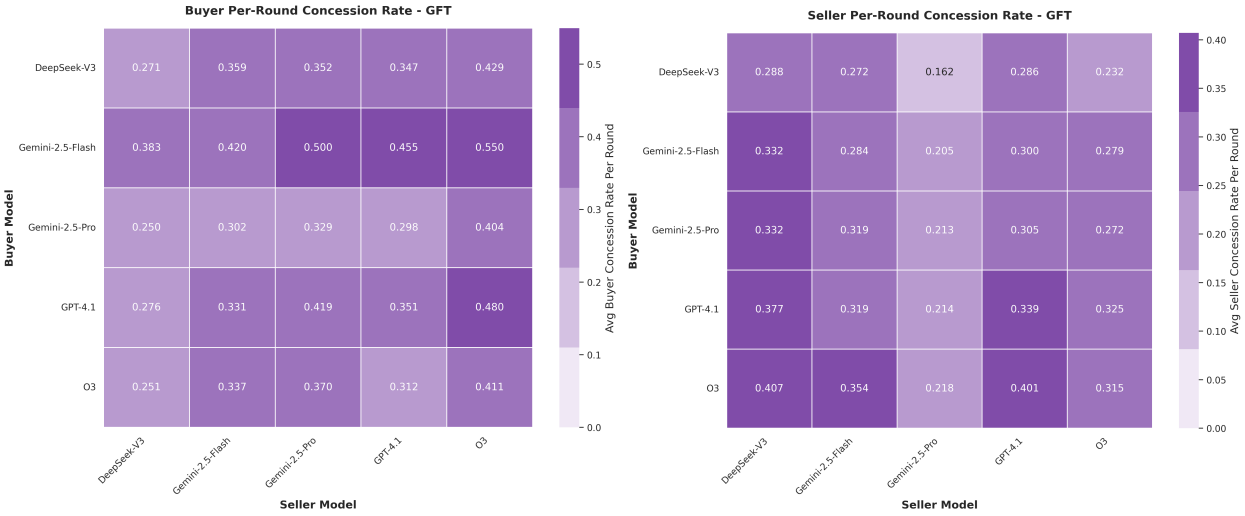

Figure 28: Per-round concession rate. Left: buyer concession rate. Right: seller concession rate. Higher values indicate faster concessions.

C.2.7 Temporal Patience

Figure 29: Average negotiation rounds for all 25 buyer–seller pairings (GFT scenarios). The protocol allows a maximum of 10 rounds.

C.3 Price-Tier Decomposition

This subsection provides price-tier decomposition results: aggregate tables (Tables 22 to 24) underlying Section 4.2.2, followed by per-opponent breakdowns. All tables report GFT scenarios with reservation price quintiles (Q1 = lowest, Q5 = highest).

C.3.1 Aggregate Bracket Tables

	Buyer Surplus by Buyer Value (%)						Seller Surplus by Seller Cost (%)					
Model	Q1	Q2	Q3	Q4	Q5	Spread	Q1	Q2	Q3	Q4	Q5	Spread
o3	44.5	49.9	49.6	48.2	43.8	6.1	79.3	76.4	75.5	75.3	78.7	4.0
Gem.-2.5-Pro	42.8	45.0	47.6	47.1	41.6	6.0	75.0	69.6	73.4	69.6	75.6	6.0
GPT-4.1	26.1	34.1	34.5	38.6	36.9	12.5	59.7	55.9	56.3	58.2	58.5	3.8
DeepSeek-V3	31.6	35.9	38.0	33.6	36.3	6.4	54.9	50.8	51.0	46.3	44.9	10.0
Gem.-2.5-Flash	16.7	23.0	25.6	26.8	29.5	**12.8**	59.6	56.5	54.8	55.9	61.4	6.6

Table 22: Surplus share by reservation price quintile, averaged across all 5 opponents (GFT). Left panel: buyer surplus by buyer-value bracket. Right panel: seller surplus by seller-cost bracket. Spread = max − min across quintiles. Strong models (o3, Gem.-Pro) show 4–6 pp spreads; weak buyers (Gem.-Flash) show 13 pp.

	Buyer Deal Rate by Buyer Value (%)						Seller Deal Rate by Seller Cost (%)					
Model	Q1	Q2	Q3	Q4	Q5	Spread	Q1	Q2	Q3	Q4	Q5	Spread
o3	97.3	98.8	97.2	98.3	98.5	1.6	96.0	94.5	95.7	96.5	95.0	2.0
Gem.-2.5-Pro	97.0	96.0	96.3	96.8	97.3	1.3	89.5	89.5	93.8	93.5	92.0	4.3
GPT-4.1	94.0	91.3	94.2	94.5	95.8	4.5	92.3	92.5	97.0	97.0	97.0	4.7
DeepSeek-V3	92.3	95.0	95.7	95.8	93.0	3.5	90.8	90.5	94.2	92.0	92.8	3.7
Gem.-2.5-Flash	87.5	79.8	87.8	87.5	89.5	**9.7**	92.0	94.3	95.0	96.8	96.8	4.8

Table 23: Deal rates by reservation price quintile, averaged across all 5 opponents (GFT). Left panel: buyer deal rate by buyer-value bracket. Right panel: seller deal rate by seller-cost bracket. Flash as buyer shows the largest spread (9.7 pp); all other models stay within 1–5 pp.

	As Buyer (%)						As Seller (%)					
Model	Q1	Q2	Q3	Q4	Q5	Spread	Q1	Q2	Q3	Q4	Q5	Spread
o3	0.50	0.25	0.25	0.25	0.00	0.50	0.25	0.75	0.00	0.50	0.25	0.75
Gem.-2.5-Pro	0.50	0.00	0.25	0.25	0.50	0.50	0.00	0.00	0.00	0.00	0.00	**0.00**
GPT-4.1	0.50	0.00	0.50	0.75	0.25	0.75	0.50	0.25	0.25	0.25	0.25	0.25
DeepSeek-V3	1.00	0.50	0.50	0.75	0.25	0.75	1.25	0.25	1.51	1.75	1.00	1.50
Gem.-2.5-Flash	1.00	0.50	0.00	0.75	0.00	1.00	0.25	0.50	0.00	0.25	0.00	0.50

Table 24: Combined violation rates (either party) by reservation price quintile, averaged across all 5 opponents (GFT). Left panel: by buyer-value bracket (model as buyer). Right panel: by seller-cost bracket (model as seller). All rates are below 2% with no systematic bracket trend. Gemini-2.5-Pro as seller: 0% in every quintile.

C.3.2 Per-Opponent Buyer Surplus Share

vs. Seller	Q1	Q2	Q3	Q4	Q5	Avg
o3	26.8	37.1	35.0	34.5	28.0	32.3
GPT-4.1	54.6	60.6	55.7	54.4	52.2	55.5
Gem.-Pro	27.7	37.6	37.4	36.2	28.3	33.4
Gem.-Flash	59.7	52.0	55.3	52.5	45.4	53.0
DeepSeek	53.7	62.2	64.5	63.6	65.2	61.8

Table 25: o3 as buyer: surplus share (%) by value quintile and opponent.

vs. Seller	Q1	Q2	Q3	Q4	Q5	Avg
o3	28.0	28.3	29.9	33.7	24.1	28.8
GPT-4.1	53.6	52.3	49.2	49.6	44.8	49.9
Gem.-Pro	30.4	38.1	42.0	40.8	31.6	36.6
Gem.-Flash	50.6	50.9	56.8	51.7	43.0	50.6
DeepSeek	51.6	55.6	60.3	59.5	64.7	58.3

Table 26: Gemini-2.5-Pro as buyer: surplus share (%) by value quintile and opponent.

vs. Seller	Q1	Q2	Q3	Q4	Q5	Avg
o3	12.9	18.4	17.8	25.6	23.3	19.6
GPT-4.1	32.1	38.3	42.2	39.8	42.2	38.9
Gem.-Pro	12.8	23.7	25.3	28.8	26.3	23.4
Gem.-Flash	35.7	42.7	40.6	48.7	38.2	41.2
DeepSeek	36.9	47.3	46.5	50.3	54.7	47.1

Table 27: GPT-4.1 as buyer: surplus share (%) by value quintile and opponent.

vs. Seller	Q1	Q2	Q3	Q4	Q5	Avg
o3	19.6	20.8	26.1	23.5	22.7	22.5
GPT-4.1	42.3	43.2	43.5	30.6	42.6	40.4
Gem.-Pro	18.5	26.8	30.6	30.1	26.5	26.5
Gem.-Flash	33.8	41.0	40.6	36.1	34.9	37.3
DeepSeek	43.6	47.5	49.1	47.5	54.9	48.5

Table 28: DeepSeek-V3 as buyer: surplus share (%) by value quintile and opponent.

vs. Seller	Q1	Q2	Q3	Q4	Q5	Avg
o3	6.0	9.9	10.7	14.0	17.5	11.6
GPT-4.1	20.9	29.9	25.4	28.6	29.2	26.8
Gem.-Pro	9.4	15.7	17.4	20.2	21.5	16.8
Gem.-Flash	28.1	26.3	35.8	28.4	30.3	29.8
DeepSeek	19.0	33.3	38.9	42.7	48.8	36.5

Table 29: Gemini-2.5-Flash as buyer: surplus share (%) by value quintile and opponent. Flash shows the strongest Q1→Q5 increase, especially vs. DeepSeek (+29.8 pp).

C.3.3 Per-Opponent Buyer Deal Rate

vs. Seller	Q1	Q2	Q3	Q4	Q5	Avg
o3	100.0	100.0	97.5	100.0	98.8	99.3
GPT-4.1	98.8	98.8	97.5	98.8	97.5	98.3
Gem.-Pro	96.3	100.0	100.0	98.8	100.0	99.0
Gem.-Flash	93.8	97.5	98.8	97.5	100.0	97.5
DeepSeek	97.5	97.5	92.4	96.3	96.3	96.0

Table 30: o3 as buyer: deal rate (%) by value quintile and opponent.

vs. Seller	Q1	Q2	Q3	Q4	Q5	Avg
o3	100.0	98.8	97.5	98.8	98.8	98.8
GPT-4.1	98.8	97.5	98.8	100.0	97.5	98.5
Gem.-Pro	95.0	98.8	98.8	96.3	93.8	96.5
Gem.-Flash	97.5	91.3	96.3	96.3	97.5	95.8
DeepSeek	93.8	93.8	90.0	92.5	98.8	93.8

Table 31: Gemini-2.5-Pro as buyer: deal rate (%) by value quintile and opponent.

vs. Seller	Q1	Q2	Q3	Q4	Q5	Avg
o3	98.8	91.3	96.3	97.5	100.0	96.8
GPT-4.1	93.8	92.5	95.0	96.3	97.5	95.0
Gem.-Pro	92.5	88.8	91.3	90.0	98.8	92.3
Gem.-Flash	96.3	96.3	94.9	96.3	93.8	95.5
DeepSeek	88.8	87.5	93.8	92.5	88.8	90.3

Table 32: GPT-4.1 as buyer: deal rate (%) by value quintile and opponent.

vs. Seller	Q1	Q2	Q3	Q4	Q5	Avg
o3	90.0	91.3	95.0	95.0	86.3	91.5
GPT-4.1	97.5	98.8	98.8	98.8	100.0	98.8
Gem.-Pro	83.8	88.8	91.3	90.0	83.8	87.5
Gem.-Flash	93.8	97.5	95.0	98.8	100.0	97.0
DeepSeek	96.3	98.8	98.7	96.3	95.0	97.0

Table 33: DeepSeek-V3 as buyer: deal rate (%) by value quintile and opponent.

vs. Seller	Q1	Q2	Q3	Q4	Q5	Avg
o3	97.5	87.5	91.3	91.3	90.0	91.5
GPT-4.1	78.8	78.8	87.5	87.5	93.8	85.3
Gem.-Pro	88.8	70.0	82.5	82.5	91.3	83.0
Gem.-Flash	87.5	81.3	92.5	90.0	93.8	89.0
DeepSeek	85.0	81.3	85.0	86.3	78.8	83.3

Table 34: Gemini-2.5-Flash as buyer: deal rate (%) by value quintile and opponent. Flash shows the strongest bracket sensitivity, with Q1→Q5 gains of up to 15 pp (vs. GPT-4.1).

C.3.4 Combined View vs. GPT-4.1

	Deal Rate (%)				**Buyer Surplus Share (%)**			
Quintile	o3	Pro	DeepSeek	Flash	o3	Pro	DeepSeek	Flash
Q1 (lowest)	98.8	98.8	97.5	78.8	54.6	53.6	42.3	20.9
Q2	98.8	97.5	98.8	78.8	60.6	52.3	43.2	29.9
Q3	97.5	98.8	98.8	87.5	55.7	49.2	43.5	25.4
Q4	98.8	100.0	98.8	87.5	54.4	49.6	30.6	28.6
Q5 (highest)	97.5	97.5	100.0	93.8	52.2	44.8	42.6	29.2

Table 35: Per-model price-tier decomposition: buyer deal rate and buyer surplus share vs. GPT-4.1 (GFT scenarios), partitioned by buyer value quintile. Strong negotiators (o3, Pro) show flat performance across brackets; weaker negotiators (Flash) show bracket-dependent deal rates.

D Training Results: Supplementary Tables

This appendix provides the full numerical tables underlying the figures in Section 5.4.

D.1 Violation Rates

Model	Stage	Role	vs GPT-4.1 (%)		vs o3-mini (%)		Average (%)	
			Self	Opp.	Self	Opp.	Self	Opp.
Qwen3-8B	Base	Buyer	0.50	0.00	1.00	0.00	0.75	0.00
		Seller	**3.75**	0.25	1.00	0.00	**2.38**	0.13
	SFT	Buyer	0.00	0.50	1.25	0.75	0.63	0.63
		Seller	**0.25**	0.75	0.50	0.25	**0.38**	0.50
	RL	Buyer	0.25	0.00	0.75	0.25	0.50	0.13
		Seller	**3.25**	0.50	**2.50**	0.00	**2.88**	0.25
Qwen3-14B	Base	Buyer	0.25	0.00	0.25	0.00	0.25	0.00
		Seller	1.00	0.50	0.25	0.25	0.63	0.38
	SFT	Buyer	0.00	0.00	1.00	0.50	0.50	0.25
		Seller	0.00	0.50	0.50	0.00	0.25	0.25
	RL	Buyer	0.25	0.00	0.25	0.00	0.25	0.00
		Seller	1.25	0.50	0.25	0.75	0.75	0.63

Table 36: GFT violation rates across training stages. Self: the Qwen model accepts a deal yielding negative utility. Opp.: the frontier counterpart accepts such a deal when facing Qwen. GPT-4.1 is the RL training opponent (in-distribution); o3-mini is unseen (out-of-distribution). Bold highlights key self-violation values discussed in text.

Model	Stage	Role	vs GPT-4.1 (%)		vs o3-mini (%)		Average (%)	
			Self	Opp.	Self	Opp.	Self	Opp.
Qwen3-8B	Base	Buyer	0.00	0.00	0.25	0.00	0.13	0.00
		Seller	**4.00**	0.00	**4.00**	0.00	**4.00**	0.00
	SFT	Buyer	0.25	0.50	0.25	0.00	0.25	0.25
		Seller	**0.50**	0.00	**0.25**	0.00	**0.38**	0.00
	RL	Buyer	0.00	0.00	0.00	0.00	0.00	0.00
		Seller	**4.00**	0.25	**6.50**	0.00	**5.25**	0.13
Qwen3-14B	Base	Buyer	—	—	—	—	—	—
		Seller	0.00	0.25	0.25	0.00	0.13	0.13
	SFT	Buyer	0.00	0.00	0.00	0.00	0.00	0.00
		Seller	0.25	0.00	0.25	0.25	0.25	0.13
	RL	Buyer	0.50	0.00	0.00	0.00	0.25	0.00
		Seller	0.25	0.00	0.00	0.00	0.13	0.00

Table 37: NGFT violation rates across training stages. Self: the Qwen model strikes a deal in a no-gains-from-trade scenario (irrational for at least one party). Opp.: the frontier counterpart does so when facing Qwen. Bold highlights key self-violation values discussed in text. Dash indicates missing data.

D.2 Strategic Effectiveness and Allocative Efficiency

			Deal Rate (%)		Surplus Share (%)	
Model	**Stage**	**Role**	vs 4.1	vs o3	vs 4.1	vs o3
Qwen3-8B	Base	Buyer	70.00	66.75	22.41	11.16
		Seller	80.00	70.25	32.53	24.80
	SFT	Buyer	86.25	79.25	40.77	36.12
		Seller	70.25	56.50	60.56	64.79
	RL	Buyer	69.75	69.50	20.39	11.60
		Seller	78.25	68.50	31.69	26.06
Qwen3-14B	Base	Buyer	97.25	95.25	23.31	12.96
		Seller	97.25	94.50	39.41	33.22
	SFT	Buyer	71.75	64.25	44.77	36.83
		Seller	49.25	41.00	58.04	60.99
	RL	Buyer	95.50	93.50	21.97	13.01
		Seller	99.00	95.50	40.18	37.77

Table 38: Strategic effectiveness and allocative efficiency across training stages (GFT scenarios). Deal rate: fraction of negotiations reaching agreement. Surplus share: fraction of ZOPA captured conditional on agreement. GPT-4.1 is the RL training opponent (in-distribution); o3-mini is unseen (out-of-distribution).

D.3 Price-Tier Decomposition Tables

		Buyer Violation by Buyer Value (%)						Seller Violation by Seller Cost (%)					
Model–Stage	**Opp.**	Q1	Q2	Q3	Q4	Q5	Spread	Q1	Q2	Q3	Q4	Q5	Spread
14B-Base	4.1	0.00	0.00	1.25	0.00	0.00	1.25	2.50	0.00	2.50	1.25	1.25	2.50
	o3	0.00	0.00	1.25	0.00	0.00	1.25	1.25	0.00	1.25	0.00	0.00	1.25
14B-SFT	4.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.25	0.00	0.00	1.25	1.25
	o3	1.25	1.25	3.75	0.00	1.25	3.75	1.25	1.25	0.00	0.00	0.00	1.25
14B-RL	4.1	0.00	0.00	0.00	1.25	0.00	1.25	2.50	1.25	1.25	3.75	0.00	3.75
	o3	1.25	0.00	0.00	0.00	0.00	1.25	0.00	0.00	2.50	2.50	0.00	2.50
8B-Base	4.1	1.25	0.00	1.25	0.00	0.00	1.25	2.50	2.50	**7.50**	**7.50**	0.00	**7.50**
	o3	1.25	1.25	1.25	1.25	0.00	1.25	0.00	1.25	1.25	0.00	2.50	2.50
8B-SFT	4.1	0.00	0.00	0.00	0.00	2.50	2.50	1.25	1.25	0.00	1.25	1.25	1.25
	o3	2.50	0.00	1.25	5.00	1.25	5.00	0.00	1.25	0.00	3.75	2.50	3.75
8B-RL	4.1	1.25	0.00	0.00	0.00	0.00	1.25	2.50	1.25	**5.00**	**6.25**	3.75	**5.00**
	o3	2.50	0.00	1.25	0.00	1.25	2.50	0.00	3.75	2.50	3.75	2.50	3.75

Table 39: Combined violation rate by reservation-price quintile and training stage, split by opponent. Bold highlights 8B seller violations exceeding 5%.

		Buyer Surplus by Buyer Value (%)						Seller Surplus by Seller Cost (%)					
Model–Stage	**Opp.**	Q1	Q2	Q3	Q4	Q5	Spread	Q1	Q2	Q3	Q4	Q5	Spread
14B-Base	4.1	18.9	24.3	18.7	20.0	34.1	15.4	42.8	38.4	39.0	37.0	40.0	5.8
	o3	4.7	7.9	12.4	14.0	26.5	21.8	35.0	32.1	31.3	37.5	30.2	7.3
14B-SFT	4.1	**49.3**	**48.9**	**40.7**	**43.3**	**42.5**	**8.6**	**64.2**	**51.4**	**58.1**	**63.1**	**56.0**	12.8
	o3	**30.7**	**37.7**	**38.7**	**37.7**	**39.3**	**8.6**	**60.1**	**66.4**	**59.4**	**56.8**	**61.6**	9.6
14B-RL	4.1	14.0	20.3	18.1	22.9	34.7	20.7	42.2	37.4	44.0	33.8	43.4	10.2
	o3	2.9	9.4	11.1	15.0	25.3	22.4	36.5	42.3	35.8	36.8	37.5	6.5
8B-Base	4.1	12.0	22.7	18.7	21.6	36.6	24.6	32.9	31.7	32.1	35.2	31.1	4.1
	o3	1.8	4.3	9.0	11.7	26.3	24.5	22.2	30.3	19.8	26.2	25.0	10.5
8B-SFT	4.1	**37.9**	**44.5**	**43.5**	**36.9**	**41.0**	**7.6**	**64.5**	**56.2**	**58.7**	**57.4**	**66.1**	9.9
	o3	**31.5**	**34.2**	**34.7**	**39.4**	**40.3**	**8.8**	**65.4**	**63.2**	**66.1**	**68.7**	**61.3**	7.4
8B-RL	4.1	11.8	22.7	16.7	20.9	28.9	17.1	32.2	28.8	35.3	31.1	31.5	6.5
	o3	6.9	5.7	6.5	10.8	28.3	22.6	23.9	29.4	20.0	32.8	25.4	12.8

Table 40: Surplus share by reservation-price quintile and training stage, split by opponent. SFT compresses buyer-side spreads to 7–9pp; Base and RL show 15–25pp Q1→Q5 gradients.

		Buyer Deal Rate by Buyer Value (%)						Seller Deal Rate by Seller Cost (%)					
Model–Stage	**Opp.**	Q1	Q2	Q3	Q4	Q5	Spread	Q1	Q2	Q3	Q4	Q5	Spread
14B-Base	4.1	97.5	97.5	96.3	95.0	100.0	5.0	96.3	97.5	97.5	98.8	96.3	2.5
	o3	96.3	97.5	96.3	95.0	91.3	6.2	93.8	93.8	95.0	93.8	96.3	2.5
14B-SFT	4.1	73.8	62.5	77.5	71.3	73.8	15.0	48.8	56.3	45.0	35.0	61.3	26.3
	o3	62.5	72.5	55.0	62.5	68.8	17.5	36.3	42.5	45.0	36.3	45.0	8.7
14B-RL	4.1	96.3	96.3	96.3	93.8	95.0	2.5	100.0	98.8	100.0	97.5	98.8	2.5
	o3	87.5	92.5	96.3	95.0	96.3	8.8	96.3	96.3	93.8	96.3	95.0	2.5
8B-Base	4.1	68.8	71.3	73.8	67.5	68.8	6.3	77.5	82.5	77.5	77.5	85.0	7.5
	o3	66.3	55.0	67.5	72.5	72.5	17.5	68.8	75.0	68.8	66.3	72.5	8.7
8B-SFT	4.1	83.8	88.8	88.8	90.0	80.0	10.0	70.0	75.0	67.5	67.5	71.3	7.5
	o3	73.8	75.0	82.5	82.5	82.5	8.7	57.5	58.8	60.0	46.3	60.0	13.7
8B-RL	4.1	68.8	67.5	72.5	63.8	76.3	12.5	76.3	78.8	73.8	80.0	82.5	8.7
	o3	66.3	68.8	71.3	73.8	67.5	7.5	73.8	70.0	68.8	60.0	70.0	13.8

Table 41: Deal rate by reservation-price quintile and training stage, split by opponent. 14B-SFT as seller shows the largest bracket spread (26.3pp vs. GPT-4.1).

E Welfare Efficiency Computation

This appendix describes the computation of the theoretical welfare benchmarks used in Section 6: the first-best expected gains from trade (GFT^{FB}), the Myerson-Satterthwaite second-best benchmark (GFT^{SB}), and the theoretical efficiency ratio η^{theory}.

E.1 Normalization

Items span heterogeneous dollar scales; a used laptop and a designer handbag are not directly comparable in nominal terms. We normalize both reservation prices by each item's historical highest price H_i:

$$c_i = \frac{S_i}{H_i}, \qquad v_i = \frac{B_i}{H_i}. \tag{14}$$

We normalize by H_i rather than the midpoint $M_i = (H_i + L_i)/2$ for the following reason. The reservation price sampling procedure (Section 3.4) draws seller costs proportionally to L_i and buyer values proportionally to H_i. The ratio $r_i = H_i/L_i$ varies from approximately 1 to 7.5 across items. Using M_i creates a spurious dependence between normalized seller types and r_i: items with the same underlying seller type S_i/L_i map to different S_i/M_i values depending on r_i. Among single-scalar normalizers, H_i yields the tightest combined support for both sides and makes the buyer distribution nearly r-independent (correlation between v_i and r_i: -0.06).

The observed normalized supports after pooling GFT and NGFT scenarios are $c_i \in [0.095,\ 0.996]$ and $v_i \in [0.191,\ 1.220]$.

E.2 Empirical Type Distributions

We use sampled reservation prices as private-type proxies: $S_i \approx C_i$ and $B_i \approx V_i$. The empirical CDFs are

$$\widehat{F}_C(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{c_i \le x\}, \qquad \widehat{F}_V(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{v_i \le x\}. \tag{15}$$

Both GFT and NGFT scenarios are pooled before fitting. NGFT scenarios ($v_i < c_i$) are valid observations of the joint type distribution and are retained.

For visualization, we also fit a K-component lognormal mixture to each side, selecting $K \in \{2, 3, 4\}$ by BIC:

$$X \sim \sum_{j=1}^{K} \pi_j \,\text{LogNormal}(\mu_j, \sigma_j^2), \quad \sum_j \pi_j = 1. \tag{16}$$

BIC selects $K = 4$ for both sides on the Qwen item pool. Table 42 reports the fitted parameters. The $K = 4$ mixture substantially reduces KS statistics relative to a single lognormal (seller: 0.022 vs. 0.097; buyer:

0.032 vs. 0.109). This continuous approximation is used for smooth CDF/PDF plots only; all theoretical benchmarks use the nonparametric distributions described below.

Side	j	π_j	μ_j	σ_j	KS
Seller C	1	0.123	−1.152	0.424	0.022
	2	0.390	−0.910	0.091	
	3	0.422	−0.691	0.081	
	4	0.065	−0.360	0.160	
Buyer V	1	0.419	−0.903	0.120	0.032
	2	0.328	−0.404	0.243	
	3	0.057	−1.293	0.389	
	4	0.195	0.022	0.108	

Table 42: Fitted $K = 4$ lognormal mixture parameters for seller cost C and buyer value V (Qwen item pool, H_i normalization). KS: Kolmogorov-Smirnov statistic against empirical CDF. Used for visualization only; theoretical benchmarks use nonparametric KDE.

E.3 First-Best Benchmark

The first-best allocation trades whenever $v \geq c$:

$$\mathrm{GFT}^{FB} = \mathbb{E}[(V - C)^+] = \iint (v - c)^+ \, d\widehat{F}_V(v) \, d\widehat{F}_C(c). \tag{17}$$

We evaluate this integral via Monte Carlo with 200,000 draws from the product of empirical marginals $\widehat{F}_V \otimes \widehat{F}_C$, using boundary-corrected kernel density estimation to smooth the discrete empirical distributions before sampling.

E.4 Myerson-Satterthwaite Second-Best Benchmark

The second-best benchmark is the expected surplus under the optimal incentive-compatible, individually rational, and budget-balanced direct mechanism (Myerson & Satterthwaite, 1983). The pointwise objective uses virtual valuations with Lagrange multiplier α on the budget-balance constraint:

$$\Psi(v, c;\, \alpha) = (v - c) - \alpha \left[\frac{1 - F_V(v)}{f_V(v)} + \frac{G_C(c)}{g_C(c)} \right]. \tag{18}$$

The second-best allocation trades when $\Psi(v, c;\, \alpha^\star) \geq 0$, and

$$\mathrm{GFT}^{SB} = \mathbb{E}[(V - C)\, \mathbf{1}\{\Psi(V, C;\, \alpha^\star) \geq 0\}]. \tag{19}$$

We solve for $\alpha^\star$ numerically via the following procedure.

1. *Grid construction.* Truncate to near the 0.01% and 99.9% quantiles of the KDE-smoothed distributions. Build dense 1D grids for v and c.
2. *Hazard evaluation.* Compute $(1 - F_V(v))/f_V(v)$ and $G_C(c)/g_C(c)$ on the grids using boundary-corrected KDE densities.
3. *Ironing.* For each candidate α, compute raw virtual terms $\Phi_{\mathrm{raw}}(v; \alpha)$ and $\Gamma_{\mathrm{raw}}(c; \alpha)$ and apply weighted isotonic regression (PAVA) to obtain monotone $\bar{\Phi}$ and $\bar{\Gamma}$. Ironing is required because lognormal mixture densities need not be log-concave.
4. *Bisection.* Find $\alpha^\star$ such that the revenue residual $R(\alpha) = \iint \big[(v - c) - \alpha(\text{hazard sum})\big] \cdot \mathbf{1}\{\bar{\Phi} \geq \bar{\Gamma}\} \cdot f_V(v)\, g_C(c)\, dv\, dc = 0$.
5. *Welfare evaluation.* Compute GFT^{SB} by integrating $(v - c)$ over the second-best trade region implied by $\alpha^\star$.

E.5 Caveats

Three factors limit strict comparability between LLM negotiation outcomes and the theoretical benchmarks.

No distributional knowledge. The MS bound requires agents to know the counterparty's type distribution F_V or G_C. LLM agents know only their own reservation price. Without this knowledge they cannot compute virtual valuations and therefore cannot implement the MS trade rule, which withholds some GFT trades (where $v \geq c$ but virtual surplus is negative) to satisfy budget balance.

Cooperative rather than strategic behavior. A purely strategic agent may refuse to close a GFT deal to extract better terms, forgoing some efficient trades to improve its average surplus. This is the source of the first-best–second-best wedge. Top frontier models, particularly o3, do not behave this way: o3 achieves a 98.0% GFT deal rate as buyer and 95.6% as seller, closing nearly all GFT negotiations rather than holding out. This cooperativeness pushes realized surplus toward GFT^{FB}.

Independence assumption. The benchmarks are computed on the product measure $\widehat{F}_V \otimes \widehat{F}_C$ of the pooled marginals, imposing independence between buyer values and seller costs. In the data, v_i and c_i are observed outcomes of a joint marketplace pricing process; positive dependence between item value and cost would cause the product measure to misrepresent the true joint distribution and misstate GFT^{SB}.

Finite-sample noise. Benchmarks are estimated from $N = 600$–800 scenarios. The KDE-smoothed GFT^{SB} inherits estimation noise of approximately ± 1–2%, so small empirical exceedances of η^{theory} are within this margin.

E.6 Full Pairwise Welfare Efficiency Results

Tables 43 and 44 report all individual buyer–seller pairings underlying the role-averaged tables in Sections 6.2 and 6.3.

Buyer	**Seller**	$\widehat{\text{GFT}}^{\text{emp}}$	η^{emp}
o3	o3	0.2406	0.940
o3	Gem.-2.5-Pro	0.2394	0.935
Gem.-2.5-Pro	o3	0.2390	0.933
DeepSeek-V3	GPT-4.1	0.2389	0.933
o3	Gem.-2.5-Flash	0.2384	0.931
DeepSeek-V3	Gem.-2.5-Flash	0.2382	0.930
Gem.-2.5-Pro	GPT-4.1	0.2376	0.928
o3	GPT-4.1	0.2367	0.924
o3	DeepSeek-V3	0.2365	0.924
Gem.-2.5-Pro	Gem.-2.5-Pro	0.2359	0.921
DeepSeek-V3	DeepSeek-V3	0.2358	0.921
Gem.-2.5-Pro	Gem.-2.5-Flash	0.2333	0.911
GPT-4.1	o3	0.2302	0.899
GPT-4.1	Gem.-2.5-Flash	0.2296	0.897
DeepSeek-V3	Gem.-2.5-Pro	0.2293	0.896
Gem.-2.5-Pro	DeepSeek-V3	0.2280	0.891
DeepSeek-V3	o3	0.2240	0.875
GPT-4.1	GPT-4.1	0.2239	0.875
GPT-4.1	Gem.-2.5-Pro	0.2160	0.844
Gem.-2.5-Flash	o3	0.2068	0.808
Gem.-2.5-Flash	Gem.-2.5-Flash	0.2040	0.797
GPT-4.1	DeepSeek-V3	0.2016	0.787
Gem.-2.5-Flash	DeepSeek-V3	0.1901	0.743
Gem.-2.5-Flash	GPT-4.1	0.1871	0.731
Gem.-2.5-Flash	Gem.-2.5-Pro	0.1796	0.701

Table 43: All 25 frontier buyer–seller pairs ranked by empirical welfare efficiency $\eta^{\text{emp}} = \widehat{\text{GFT}}^{\text{emp}}/\text{GFT}^{FB}$. $\text{GFT}^{FB} = 0.2560$. Theoretical ceiling: $\eta^{\text{theory}} = 0.929$.

Buyer	**Seller**	$\widehat{\text{GFT}}^{\text{emp}}$	η^{emp}
GPT-4.1	Qwen3-14B-RL	0.1792	0.891
Qwen3-14B	GPT-4.1	0.1779	0.885
o3-mini	Qwen3-14B-RL	0.1777	0.884
Qwen3-14B-RL	GPT-4.1	0.1776	0.883
GPT-4.1	Qwen3-14B	0.1774	0.882
Qwen3-14B	o3-mini	0.1743	0.867
o3-mini	Qwen3-14B	0.1725	0.858
Qwen3-14B-RL	o3-mini	0.1698	0.845
Qwen3-8B-SFT	GPT-4.1	0.1591	0.791
GPT-4.1	Qwen3-8B	0.1484	0.738
Qwen3-8B-SFT	o3-mini	0.1483	0.738
GPT-4.1	Qwen3-8B-RL	0.1476	0.734
Qwen3-8B	GPT-4.1	0.1466	0.729
o3-mini	Qwen3-8B	0.1368	0.681
Qwen3-8B-RL	GPT-4.1	0.1346	0.670
Qwen3-14B-SFT	GPT-4.1	0.1339	0.666
Qwen3-8B-RL	o3-mini	0.1330	0.662
o3-mini	Qwen3-8B-RL	0.1319	0.656
GPT-4.1	Qwen3-8B-SFT	0.1299	0.646
Qwen3-8B	o3-mini	0.1266	0.630
Qwen3-14B-SFT	o3-mini	0.1224	0.609
o3-mini	Qwen3-8B-SFT	0.1124	0.559
GPT-4.1	Qwen3-14B-SFT	0.0868	0.432
o3-mini	Qwen3-14B-SFT	0.0777	0.387

Table 44: All 24 Qwen-variant buyer–seller pairs ranked by empirical welfare efficiency $\eta^{\text{emp}} = \widehat{\text{GFT}}^{\text{emp}}/\text{GFT}^{FB}$. $\text{GFT}^{FB} = 0.2010$. Theoretical ceiling: $\eta^{\text{theory}} = 0.924$.