Deal With It! Diagnosing Post-Training in Bargaining, from Fluency to Surplus Extraction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Remember how GPT-3 was fluent in English but notoriously lacked symbolic reasoning ability? Whilst reaching expert level coding and theorem proving abilities, in multi-turn bargaining, modern LLMs disappoint to extract economic value. Yet, although the domain's game theoretic basis is ideal for theoretical reference points, most evaluations only compare LLMs with each other, resembling leaderboards. In this paper we introduce theoretical upper and lower bounds, playing against scripted opponents for sets of simulated episodes. Upper: The optimal reward achievable with full knowledge of the opponent's behavior and reservation value. Lower: The highest reward attainable for a policy that acts unconditional to the opponent's past actions (*open-loop* analogue). A policy undershooting the lower bound, fails to convert cues from past opponent offers into economic surplus. In our prepared zero-sum multi-turn bargaining environment, a learner's achievable maximum lies in $[\bar{R}_{\text{open-loop}}, \bar{R}^{*}_{\text{full information}}] = [0.350, 0.543]$. The subsequent evaluation of a frontier model underperformed the lower bound while producing fluent counteroffers, corroborating that fluency does not need to imply surplus capitalization. Finally, attempting to leverage the benchmarked environment for post-training, a Qwen3.5-4B policy failed to exploit rollouts efficiently on Vanilla GRPO. Despite seemingly stable training ($\bar{R}\colon -0.43 \rightarrow 0.35$), by step 50, over 60% of sampled rollouts suffered advantage collapse, in line with recent adjacent RLVR research.

1 Introduction

GPT-3 was fluent in grammar, yet its arithmetic abilities collapsed as operands grew [Brown et al., 2020]. Post-training has since turned such objectively verifiable domains into model strengths [DeepSeek-AI, 2025]. Yet, in other domains suited to verifiable rewards, like multi-turn bargaining, contemporary models still exhibit GPT-3 era behavior. As adjacent work across different settings in the last months has shown, models negotiate fluently but capture far less of the economic value than is available [Cosentino et al., 2026, Zhang et al., 2026, Lei, 2026].

Part of the reason the underperformance persists is model evaluation. Recent evaluations compare models against each other or against prompted baselines, which produces a ranking rather than a scale [Liu et al., 2026, Zhu et al., 2026]. In recent bargaining benchmarks, the comparison set is mostly other LLMs [Bianchi et al., 2024, Davidson et al., 2024, Zhu et al., 2026]. And normalizing outcomes by available surplus [Liu et al., 2026] is objective but still not a reference, as it measures what a model *captured* rather than what was *capturable* against an opponent. References for optimal play are rare because optimal play is not computable in rich negotiation settings. We deliberately created a small environment. A single price bargain against deterministic scripted opponents, which makes bounding reference values computable exactly per episode.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

We contribute a two part post-training diagnosis using the same environment. First, we bracket the learner's achievable optimum between an exact full information best response (upper bound) and the best of 2,616 grid searched open-loop policies, whose offers are independent of the opponent's past actions (lower bound). Under our reward function (normalized to $[-1, 1]$) the bounds are 0.543 and 0.350. Both references are computed per episode and replayable from seeds. On identical episodes, a frontier model underperforms the lower bound on both buyer and seller roles. Second, we run GRPO post-training on the same environment. It lifts mean reward from -0.43 to 0.35, near the open-loop bound, without evidence that it capitalizes on the opponent's past offer behavior. Training appears stable while, at step 50, more than 60% of the batch's samples collapse to zero advantage, matching reports from adjacent RLVR work [Yu et al., 2025, Li et al., 2025, Le et al., 2025].

2 Related Work

Research on RLVR environments [DeepSeek-AI, 2025] in domains like bargaining is on the rise in 2026. Similar to us, TERMS-Bench [Zhang et al., 2026] reports scores relative to a dynamic programming oracle and to bots that concede a fixed fraction per offer. We are aiming for improved interpretability and added rigor. Deterministic opponents make our backward induction exact rather than an expectation over the opponent's hidden type, and we introduce a directly interpretable lower bound as the optimum over 2,616 parameterizations. Moreover, GameTalk [Vendrell et al., 2026] compares its trained 3B model's deals to the solution maximizing both players' gains (Nash bargaining solution). In positive-sum games a gain in joint welfare need not mean the agent took surplus from the opponent [Cosentino et al., 2026]. Thus, in our zero-sum games rewards can solely be associated with surplus extraction, being well-defined (minimax) for efficient RL training. We did not find a published post-training environment in bargaining that comes with computed references (full information best response and searched open-loop bound).

3 Environment and Reference Bounds

The game is a single price bilateral bargain. A seller with private cost c and a buyer with private value v (their respective reservation values) alternate offers for up to T rounds; each turn a player may counteroffer a price, accept the standing offer, or walk. The surplus is $v - c$ and prices inside $[c, v]$ (called Zone Of Possible Agreement / ZOPA), profit both roles. It is the standard alternating offers setup [Rubinstein, 1982] under two-sided private information [Myerson and Satterthwaite, 1983]. Episodes draw from $c \sim U[40, 60]$, $v \sim U[90, 110]$, $T \sim U[2, 6]$, guaranteeing a ZOPA width of at least 30 and walking never optimizing any player's surplus.

The reward is defined as the normalized surplus split centered at the ZOPA midpoint:

$$r_{\text{seller}} = \text{clip}\left(\frac{2p - (c + v)}{v - c},\ -1,\ 1\right), \qquad r_{\text{buyer}} = -r_{\text{seller}} \tag{1}$$

No deal (walk or expiry) returns 0 to both and a format violation (after one retry) forfeits at -1. Capturing the whole surplus yields $+1$, conceding it -1, and a deal at the midpoint 0.

A player acts given its own reservation value, the public prior ranges, the normalized amount of rounds passed, the dialogue history, and the current standing offer (Appendix B). As an RLVR domain it differs from math or code, since the objective is adversarial and information asymmetric. As the relevant unknown is the opponent's reservation value, the environment rewards Bayesian bargaining; that is, updating a probabilistic estimate of the opponent's reservation value from the offers made and acting accordingly [Fudenberg and Tirole, 1983].

The opponents are three deterministic scripted bot families from automated negotiation literature [Faratin et al., 1998, Baarslag et al., 2013]. *Boulware* (stays close to opening price, concedes late) and *Conceder* (concedes early) are the two shapes of the standard time dependent concession (power function $t^{1/e}$ with randomized hidden parameters) [Faratin et al., 1998]. Lastly, the "take it or leave it" bot follows a constant price [Güth et al., 1982]. Each accepts any standing offer at least as good as what it was about to demand (as in `AC_next` [Baarslag et al., 2014]). Consequently, each demands a different best response. Episodes draw the family uniformly, and Appendix D reports both bounds per family for reweighting.

Upper bound. Supplied with the opponent's hidden parameters, a dynamic programming solver works backward from the final round and returns the exact full information best response for that episode. Averaging over 16,000 simulated episodes (3 opponent families $\times$ 2 roles $\times \approx$ 2,667 draws) yields **0.543**. The result is far from 1.0 since no amount of foresight makes a Boulware bot concede past its reservation. The learner has no access to the opponent's parameters, so 0.543 is not a training target.

Lower bound. At the other end of information disclosure sit policies that act independent of the negotiation dialogue (open loop). Such a policy has a predetermined offer strategy depending on its own value, the public priors and the horizon. Each turn it compares its predetermined concession value with the current standing offer to accept or walk. We grid searched 2,616 parameterizations and reran the top 20 on 8,000 fresh episodes to reduce selection bias (Appendix C). The winner averages **0.350** pooled.

Bound interpretation. A learner sees strictly more than the open-loop family and strictly less than the solver, so its achievable optimum lies in $[0.350, 0.543]$. Underperforming the lower bound on identical episodes is highly informative, since such a policy extracts less surplus than (the best) searched scripted bot that acts totally independent of opponents' past actions.

4 Frontier Model Evaluation

We evaluated gpt-5.6-terra (medium reasoning) zero-shot on 128 episodes. Because draw luck dominates outcomes at this sample size, both references were recomputed on the exact same 64 episodes per role.

The model underperforms the open-loop winner on both roles (Figure 1; Table 4, Appendix F). The midpoint reward design is not driving this result, since under the simplistic realized surplus share the differences per episode are also negative (incl. 95% CIs).

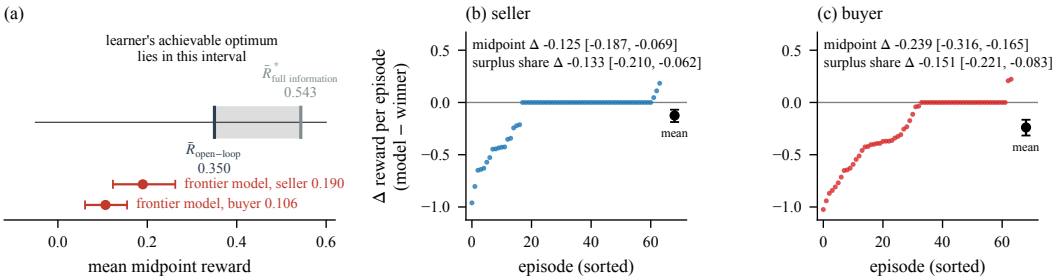

Figure 1: (a) Both bounds (8,000 episode set) and the model's means per role. (b, c) Differences per episode vs. the open-loop winner, both reward conventions, 95% CIs.

Despite profit guaranteeing ZOPA, it walked in 42% of its seller and 27% of its buyer episodes. Nevertheless, the model produced well formulated counteroffers with zero rule violations, exhibiting fluency without domain competence, akin to GPT-3.

5 GRPO Post-Training

With the reference bounds as the diagnostic, we turned the environment to its intended use of post-training. The policy is Qwen3.5-4B (thinking disabled, 768 token cap), trained with GRPO (group size 16, batch 512) [Shao et al., 2024] via LoRA, against the same scripted opponent mix as before. Two sanity check runs (few dollars of compute) preceded the two main runs we report below.

Run 1. On draws $c \sim U[40, 80]$, $v \sim U[50, 100]$, $T \sim U[2, 6]$, RLVR solved the format issues faced with the base 4B model, reducing both forfeit rate (from 30%) and truncation rate (from 76%) to 0%. However, the draw distribution allowed episodes with negative surplus ($v \leq c$), where walking is the only rational move, and the policy generalized walking behavior. Our training termination script triggered at step 25, when (a) the share of walked episodes where deals were rational and (b)

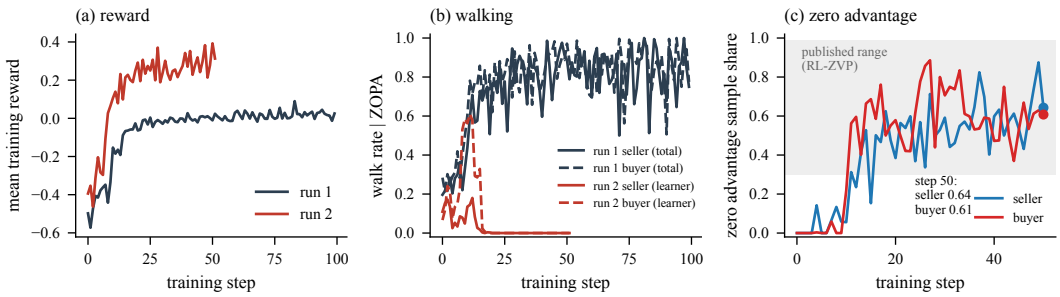

Figure 2: (a) Mean training reward, both runs. (b) Walk rate on episodes with a ZOPA, per role. (c) Zero advantage share, run 2; RL-ZVP's published range shaded.

the share of zero advantage groups both exceeded 50%. Run 2 therefore sampled only from the ZOPA guaranteeing distribution of Section 3.

Run 2. As hoped, by step 20, the training resulted in a walk rate of 0%. However, the policy converged to fixed openings that ignore its own private draw, both as seller and buyer, with counteroffers confined to three and four prices, respectively (Appendix E). At step 50, 62% of the rollouts concluded in zero advantage (Figure 2). Yet, archived mean rewards rose ($\bar{R}_{\text{seller}}$: $-0.41 \rightarrow 0.30$, $\bar{R}_{\text{buyer}}$: $-0.45 \rightarrow 0.41$; step 1 to 50, $n \in [29, 35]$ per cell) and forfeit and walk rate remained at 0%.

Similar advantage collapse incidents have been measured in adjacent domains [Yu et al., 2025], and RL-ZVP reports zero variance prompts at 30-99% of each batch [Le et al., 2025]. No published measurement was found for continuous price multi-turn bargaining. The likely cause is zero variance groups. Whether a different advantage estimator than GRPO avoids them has not been tested yet, but adjacent evidence [Wang et al., 2026, Jung, 2026, Liu et al., 2025, Yuan et al., 2025] is detailed in Appendix A.

6 Limitations, Selection History, and Outlook

The limitations follow from design choices. Our upper bound is privileged by choice, and the reference a learner should ideally be measured against, a Bayes-optimal policy [Ghavamzadeh et al., 2015] with the learner's own observation space, is unbuilt. Our environment is based on a single price and zero sum game against three scripted opponent families, far away from a realistic negotiation scenario. Within it, our empirical evidence is limited: one 4B policy trained against bots rather than in self-play, and one frontier model evaluated at $n = 128$, exploratory rather than preregistered (Appendix F). Whether any of this transfers to richer settings is untested, but several adjacent results are concordant with the frontier finding [Cosentino et al., 2026, Lei, 2026].

Planned extensions for external validity are stochastic opponents, multi-issue bargaining, multi model evaluations, and self-play.

7 Conclusion

Whether post-training delivers target behavior is difficult to answer solely from reward curves. Training a 4B policy on GRPO fixed forfeit and format issues, yet under rising reward ($-0.43 \rightarrow 0.35$), at step 50, 60% of rollouts had collapsed to zero advantage. The computed reference points reveal that it was still comparable to an open-loop bot rather than a model conditioned on its environment draw (corroborated by the policy's invariance in opening offers). Computed bounds in (solvable) RLVR environments make that distinction quantitatively interpretable, so we propose them as tools for diagnosing post-training.

References

Tim Baarslag, Katsuhide Fujita, Enrico H. Gerding, Koen Hindriks, Takayuki Ito, Nicholas R. Jennings, Catholijn Jonker, Sarit Kraus, Raz Lin, Valentin Robu, and Colin R. Williams. Evaluating practical negotiating agents: Results and analysis of the 2011 international competition. *Artificial Intelligence*, 198:73–103, 2013.

Tim Baarslag, Koen Hindriks, and Catholijn Jonker. Effective acceptance conditions in real-time automated negotiation. *Decision Support Systems*, 60:68–77, 2014.

Tim Baarslag, Reyhan Aydoğan, Koen V. Hindriks, Katsuhide Fujita, Takayuki Ito, and Catholijn M. Jonker. The automated negotiating agents competition, 2010–2015. *AI Magazine*, 36(4):115–118, 2015.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 235*, pages 3935–3951, 2024. arXiv:2402.05863.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. arXiv:2005.14165.

Romain Cosentino, Sarath Shekkizhar, Adam Earle, and Silvio Savarese. Counterparty modeling is not strategy: The limits of llm negotiators. *arXiv preprint arXiv:2605.16575*, 2026.

Tim R. Davidson, Veniamin Veselovsky, Martin Josifoski, Maxime Peyrard, Antoine Bosselut, Michal Kosinski, and Robert West. Evaluating language model agency through negotiations. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2401.04536; codebase name LAMEN.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Drew Fudenberg and Jean Tirole. Sequential bargaining with incomplete information. *The Review of Economic Studies*, 50(2):221–247, 1983.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 8(5–6):359–483, 2015.

Werner Güth, Rolf Schmittberger, and Bernd Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3(4):367–388, 1982.

Wooil Jung. Dropout-grpo: Variational stochasticity for continuous latent reasoning. *arXiv preprint arXiv:2606.10184*, 2026.

Thanh-Long V. Le et al. No prompt left behind: Exploiting zero-variance prompts in llm reinforcement learning via entropy-guided advantage shaping. *arXiv preprint arXiv:2509.21880*, 2025.

Yingjie Lei. Prefbench: Evaluating zero-shot llm agents in hidden-preference personalized pricing negotiations. *arXiv preprint arXiv:2605.22855*, 2026.

Chen Li et al. Adaptive group policy optimization: Towards stable training and token-efficient reasoning. *arXiv preprint arXiv:2503.15952*, 2025.

Bo Liu et al. Spiral: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. *arXiv preprint arXiv:2506.24119*, 2025.

Shuze Daniel Liu et al. Instructing llms to negotiate using reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2604.09855*, 2026.

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

Wenhua Nie et al. Equilibrium residuals expose three regimes of matrix-game strategic reasoning in language models. *arXiv preprint arXiv:2605.10410*, 2026.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Victor Conchello Vendrell, Max Ruiz Luyten, and Mihaela van der Schaar. Gametalk: Training llms for strategic conversation. *arXiv preprint arXiv:2601.16276*, 2026.

Minzheng Wang, Run Luo, Yanbo Wang, Zichen Liu, Yuqiao Tan, Tao Tan, Xu Nan, Yinhe Zheng, and Wenji Mao. Breaking the impasse: Dual-scale evolutionary policy training for social language agents. *arXiv preprint arXiv:2605.08721*, 2026.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring bargaining abilities of llms: A benchmark and a buyer-enhancement method. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. arXiv:2402.15813.

Qiying Yu et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Huining Yuan et al. Marshal: Incentivizing multi-agent reasoning via self-play with strategic llms. *arXiv preprint arXiv:2510.15414*, 2025.

Erica Zhang et al. Terms-bench: Diagnosing llm negotiation agents beyond deal rate. *arXiv preprint arXiv:2605.13909*, 2026.

Chris Zhu et al. Piearena: Ranking and profiling language agents in realistic negotiation scenarios. *arXiv preprint arXiv:2602.05302*, 2026.

A Extended Related Work

Other reference bounds in negotiation training. Xia et al. [2024] measures bilateral LLM bargaining by deal rate and by profit normalized by the ZOPA, and improves the buyer through prompting rather than RL training; neither metric is a computed optimum. NegotiationArena [Bianchi et al., 2024] and Davidson et al. [2024] fall short of objective bounds, as they rank models amongst each other. Moreover, computable negotiation bots are no recent invention. Teams in the automated negotiation competitions have tuned scripted strategies against benchmark scenarios since 2010 [Baarslag et al., 2013, 2015]. What we add is computed references inside an RLVR post-training environment.

Potential inefficiencies with GRPO as estimator. Whether a different advantage estimator could have avoided the advantage collapse of Section 5 is not clear from previous research (Table 1).

Table 1: Published results regarding the collapse of Section 5, ordered by distance to our environment.

Paper	Environment	Comparison	Result
DEPT [Wang et al., 2026]	3 multi-turn LLM games, 2 negotiations	per role dual EMA vs. GRPO, whole systems	collapse avoided (EMA plus reshaping); no estimator isolating ablation
SPIRAL [Liu et al., 2025]	3 zero-sum LLM games, incl. a simple negotiation	per (game, role) EMA; group mean unmentioned	no group mean comparison available
MARSHAL [Yuan et al., 2025]	6 bilateral games, none bargaining	per role group mean vs. SPIRAL's EMA	group mean wins on 5 of 6 games
Dropout-GRPO [Jung, 2026]	single-turn mathematical reasoning	group mean, 32 rollouts vs. running EMA, 1 rollout	group mean 29.0 vs. 25.1; plain GRPO stalls at the SFT baseline
Nie et al. [2026]	zero-sum matrix games, theory only	both roles in one group vs. one role per group	role merged group mean is exactly zero, by cancellation
This paper	LLM bargaining, multi-turn	vanilla GRPO, one role per prompt	61–64% zero advantage at step 50, from groups with identical rewards, not from cancellation

The only result in our own domain is not an isolating ablation, and in adjacent domains evidence points both ways. A related degeneracy is proven for role merged grouping, where both roles' rollouts share one normalization group and their opposite rewards cancel each other out [Nie et al., 2026, App. B, Thm. 2]. As each prompt contains one role, a group can never contain both and the cancellation cannot arise. Further isolating ablation comparisons are needed for negotiation post-training.

B Environment and Reward Specification

Each player receives one system prompt stating role, private value, public priors, horizon, rules, reward, and action format. Subsequent turns append "Round k. Opponent: [message]". Buyer and seller prompts only differ in ROLE and the direction of the midpoint in SCORING. Illustrative seller prompt for the draw $c = 50$, $v = 100$, $T = 6$ with priors $c \sim U[40, 60]$, $v \sim U[90, 110]$; v and the opponent's family and parameters are drawn but not revealed. Verbatim:

```
ROLE: SELLER. Your cost c = 50 (private).
PRIORS (public): c ~ U[40,60], v ~ U[90,110], rho = 0. P(no-ZOPA) = 0.00.
HORIZON: revealed, T = 6 rounds. Current round in each message.
RULES: alternate turns, integer prices 0-150. ACCEPT binds at opponent's last
  offer. WALK ends at (0,0). Invalid output: 1 retry, then forfeit.
SCORING (public): a deal at price p scores (2p - v - c)/(v - c), clamped to
  [-1, 1], for the seller and the negative for the buyer. Your score is
  positive only if p is above the midpoint of c and v. If v <= c (no ZOPA), a
  deal scores +1/-1 by the same midpoint rule (0 exactly at it). WALK,
  deadline, and breakdown score 0 for both. Forfeiting by invalid output
  scores you -1, the opponent +1. A deal can score worse than walking away.
Your reply MUST end with one line that is exactly "#### OFFER(N)" (N replaced
by your integer price), "#### ACCEPT", or "#### WALK". Nothing may follow that
line. Replies without such a final line waste your retry, then forfeit.
```

In the priors line, `rho` is the correlation between the c and v draws and `P(no-ZOPA)` the probability that $v \leq c$, computed from the draw bounds.

A reply is valid iff, after removing think tags, exactly one action line remains: `#### OFFER(n)` with $n \in \{0, \ldots, 150\}$, `#### ACCEPT` against a standing offer, or `#### WALK`. Two consecutive invalid replies on the same turn forfeit the episode (-1 to the forfeiter and $+1$ to the opponent), and a rollout truncated by the token budget ends with reward 0. Figure 3 shows the reward structure.

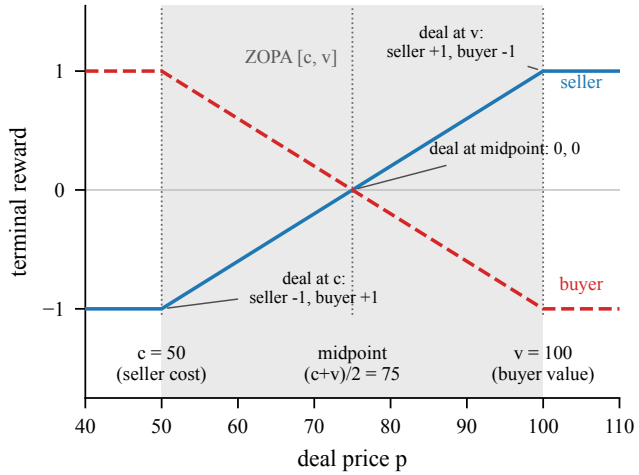

Figure 3: Mechanism and reward schematic.

c, v, and T are integer uniform draws and offers are integer prices. The learner opens every episode, therefore the scripted opponent replies between learner turns.

Opponent parameters are randomized per episode. Boulware draws its concession exponent uniformly from $[0.1, 0.5]$ and Conceder from $[2.0, 4.0]$. Opening prices are drawn from 10 below to 30 above the opposing player's priors: integer offset u drawn uniformly from $\{-10, \ldots, 30\}$ (seller at $v_{hi} + u$, buyer at $c_{lo} - u$), with the draw range truncated to the price range. The take it or leave it bot draws its constant price uniformly between its own reservation and the opposing player's prior bound (v_{hi} for the seller, c_{lo} for the buyer). Figure 4 illustrates the three families on one example episode.

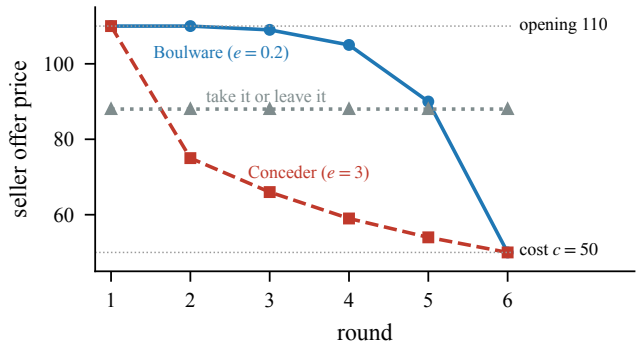

Figure 4: The three opponent families' offers per round, illustrated for the seller role ($c = 50$, opening price 110), where $T = 6$. The take it or leave it bot posts one constant price (here 88).

C Open-Loop Reference Computation

Each candidate policy is a time dependent concession bot with an acceptance rule depending on the standing offer. Either

1. accepting when the standing offer is at least as favorable as the next planned price (offset by a margin m), or
2. at a revealed deadline accept an offer at least as favorable as the final planned price and otherwise walk, or
3. in every other turn offer the planned price.

The concession family follows $\text{frac} = t^{1/e}$ over normalized time t, with four parameters: exponent $e \in \{0.02, 0.05, 0.1, 0.2, 0.4, 0.7, 1.0, 2.0, 4.0\}$; opening price offset from the opponent's priors, `anchor_off` $\in \{-40, -30, -20, -10, 0, 10, 20, 30, 50\}$; acceptance `margin` $\in \{-5, 0, 5, 10\}$; and `floor_frac` $\in \{0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 2.0\}$ which interpolates the final concession value between the own reservation (0) and the expected midpoint (1), values above 1 extrapolating past the midpoint. The four grids give $9 \times 9 \times 4 \times 8 = 2{,}592$ concession candidates. Additionally, two camper families are added: 15 fixed price campers on $\{40, 45, \ldots, 110\}$ (every reservation value possible), and 9 campers that hold out for a fixed profit $k \in \{0, 5, \ldots, 40\}$ over their own reservation. Total 2,616.

Selection of the best candidate policy runs in two stages with identical episodes across candidates (same draws, opponents and seeds). All 2,616 are screened on 400 episodes per role, the top 20 are then rerun on 8,000 fresh episodes per role with disjoint parameter seeds (because selecting a maximum over 2,616 candidates on the screening set is upwardly biased). The winner is a bot with Boulware type concession behavior ($e = 0.02$, `anchor_off` $= -20$, `margin` $= -5$, `floor_frac` $= 1.25$) reaching a pooled mean of 0.3502 (seller 0.3502, buyer 0.3501); Figure 5 shows its offers per round. This mean is still a maximum over 20 finalists, so a small residual selection bias remains. Extending the grid further does not improve the open-loop bound. With $T \leq 6$, every $e \leq 0.02$ yields the identical integer outputs, and among the refined finalists performance above `floor_frac` $= 1.25$ declines.

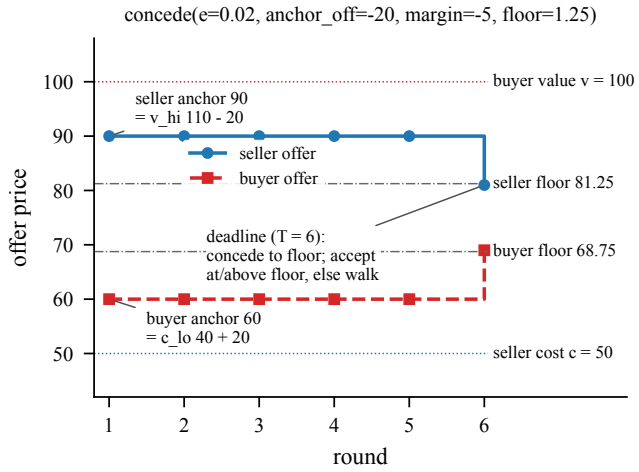

Figure 5: The searched winner's predetermined prices per round, illustrated for buyer ($v = 100$, opponent priors $[c_{lo}, c_{hi}] = [40, 60]$) and seller ($c = 50$, opponent priors $[v_{lo}, v_{hi}] = [90, 110]$) role, where $T = 6$.

D Results per Opponent Family

Table 2 reports both reference bounds separately per opponent family (mean reward $\pm$ standard error). Both are computed on the same 8,000 episodes per role of Appendix C, the family is drawn uniformly (such that the cells on average represent 2,667 episodes). Both ends are Monte Carlo means (per family standard errors are ≤ 0.010).

The last row is the mean over all episodes with counts realized from uniform draw. To evaluate a different opponent mix, reweight the three family rows (both bounds move together since they share the episodes).

Table 2: Both bounds per opponent family. Mean reward $\pm$ standard error.

	Upper bound		Open-loop winner	
Opponent family	Seller	Buyer	Seller	Buyer
Boulware	0.607 ± 0.010	0.590 ± 0.010	0.372 ± 0.007	0.366 ± 0.007
Conceder	0.823 ± 0.005	0.828 ± 0.005	0.545 ± 0.004	0.548 ± 0.004
Take it or leave it	0.202 ± 0.006	0.211 ± 0.006	0.131 ± 0.005	0.139 ± 0.005
All families	0.545	0.541	0.350	0.350

The mean reward differences between the families show how the maximum reward achievable depends on the opponent (not only playing the best strategy yourself). Regardless of optimal strategy, it is nearly impossible to extract surplus against a take it or leave it bot, but comparably easy against a Conceder bot.

The comparison of Section 4 uses neither mean and instead replays the full-information best response and the open-loop winner on the model's own episodes, reporting differences per episode.

E Training Configuration

Both training runs share the same configuration: Qwen3.5-4B; hosted RFT (prime-rl trainer); batch 512 as 32 prompts $\times$ group size 16; role partitioned groups (separate seller and buyer prompts); cap of 768 tokens with thinking disabled; temperature 1.0; LoRA at platform defaults (alpha 16). No KL term and no advantage normalization beyond the group mean. The platform drops zero variance groups from the gradient update and logs their share at every step. Run 1: 100 steps, roughly \$17. Run 2: 52 steps, roughly \$7 (draw distributions in Section 5).

Both runs had preregistered stopping and success criteria. In run 1 the walk collapse criterion triggered at step 25 (the buyer walked in 82 percent of episodes with a ZOPA, with zero advantage share at 0.60). The run continued to 100 steps, but all run 1 numbers are reported from step 25. During run 2, no stopping criterion triggered, but the success criterion, a zero advantage share ≤ 0.5, failed at step 50 (seller 0.64, buyer 0.61).

Table 3: Zero advantage sample share at the archived steps of run 2.

Step	Seller entry	Buyer entry
1	0.00	0.00
5	0.00	0.00
25	0.57	0.78
50	0.64	0.61

Table 3 reports the zero advantage sample share at the archived steps of run 2. The share initially turns nonzero at step 4 and stays above 0.5 from roughly step 14 onward.

Figure 6 plots opening price against the player's own private draw, showing direct evidence that converged openings ignore the draw.

At step 50 the trained policy's mean reward sits on either side of the open-loop bound (seller 0.30, buyer 0.41 vs. 0.350; different episode sets). Exceeding the bound says nothing about opponent inference, since the open-loop winner reaches it independent of the opponent's past actions. The openings ignoring the policy's own draw (Figure 6) point the same way.

F Frontier Evaluation

The frontier evaluation ran through an API script with: temperature 1.0, $\leq 4{,}096$ completion tokens, reasoning effort at the provider default (medium), model identifier `gpt-5.6-terra`, accessed 2026-08. Per role, episodes are the first 64 rows of a frozen 512-episode set drawn from the Section 3 distribution, with the same games in both roles. The policy played all 128 episodes with zero forfeits,

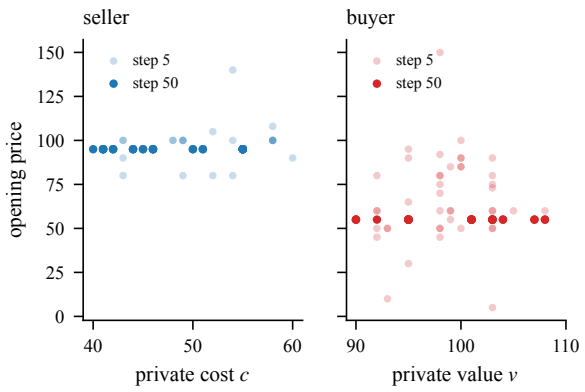

Figure 6: Opening price vs. own private draw, final archived episodes.

and at least one counteroffer in 44 of 64 seller and 37 of 64 buyer episodes, walks in 42 and 27 percent.

The comparison scripts replay the open-loop winner bot and compute the full-information best response on identical episodes; differences are model minus reference, with percentile bootstrap confidence intervals ($B = 10{,}000$). Table 4 reports all four cells; W/L/T counts wins, losses, ties against the open-loop winner. Against the full-information best response the model never wins (0/29/35 seller, 0/38/26 buyer; ties are episodes where both end without a deal above the midpoint). Surplus share is the acting role's realized share of the surplus on deals, $(p - c)/(v - c)$ for the seller and $(v - p)/(v - c)$ for the buyer, and 0 otherwise.

Table 4: Frontier model vs. searched open-loop winner and full-information best response on identical episodes ($n = 64$ per role). Bootstrap 95% CIs, $B = 10{,}000$.

Cell	Model	Open-loop	Full-info	Δ open-loop [95% CI]	Δ full-info [95% CI]	W/L/T
Seller, midpoint reward	0.190	0.315	0.481	-0.125 $[-0.187, -0.069]$	-0.291 $[-0.387, -0.199]$	3/17/44
Buyer, midpoint reward	0.106	0.346	0.528	-0.239 $[-0.316, -0.165]$	-0.422 $[-0.528, -0.315]$	2/33/29
Seller, surplus share	0.290	0.423	0.530	-0.133 $[-0.210, -0.062]$	-0.239 $[-0.329, -0.156]$	3/17/44
Buyer, surplus share	0.327	0.478	0.584	-0.151 $[-0.221, -0.083]$	-0.258 $[-0.337, -0.185]$	2/33/29

Figure 7 shows the episode outcome composition per role on the same episodes.

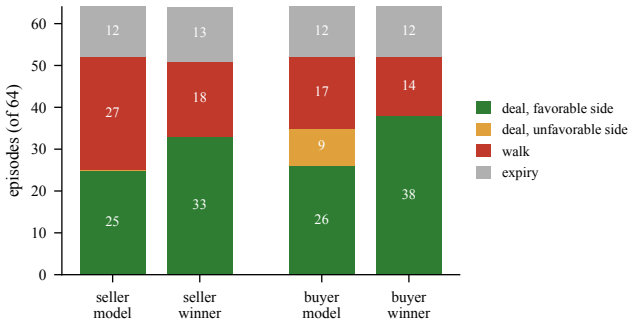

Figure 7: Episode outcome composition per role (deal above/below midpoint, walk, expiry), model vs. open-loop winner on the same episodes.

Caveats beyond Section 6: The bounds differ in Table 4 (full-information mean 0.481 seller, 0.528 buyer) as they are recomputed on exactly the same 128 episode seeds as the frontier model during benchmarking, where 64 draws per role are few. Finally, comparisons between the bounds and the trained 4B policy's mean rewards are limited, as it did not run on the same frozen episode seed during training.