

# The Art of Saying No: BATNA-Aware Reward Design for LLM Strategic Negotiation

Anonymous EMNLP submission

## Abstract

Reinforcement learning enables the training of large language model (LLM) negotiation agents without human supervision. However, existing methods are often fragmented between disjoint task types and heavily rely on surplus rewards that treat the finalized agreement as superior to no deal. To address this, we ground our reward design in the bargaining theory of Best Alternative to a Negotiated Agreement (BATNA), positing that a rational agent should walk away rather than accept a contract below a critical quality threshold. We formalize this formulation as a parameterized utility floor and apply it selectively to multi-item scenarios where no natural structural floor exists, while retaining surplus rewards for price bargaining settings. We train a single unified agent across four distinct price and multi-item benchmarks, evaluating its performance against both fixed and frontier opponents alongside a held-out dataset featuring a hybrid, multi-attribute structure. Our reward modeling consistently achieves the highest overall average bargained ratio, outperforming both surplus-based training and larger frontier models while effectively eliminating the deal-closing bias on tasks where baseline incentives fail to distinguish fair compromises from poor agreements.

## 1 Introduction

Negotiation research has progressed from early neural models (Lewis et al., 2017; He et al., 2018) to large language model (LLM; Minaee et al., 2024) benchmarks spanning price bargaining and multi-item allocation (Xia et al., 2024; Kwon et al., 2024; Bianchi et al., 2024; Noh and Chang, 2024; Chatterjee et al., 2024; Hua et al., 2024), revealing that even frontier models routinely accept negative-utility deals. Reinforcement learning has been used to address this gap (Liu et al., 2026; Bergemann et al., 2026; Vendrell et al., 2026; Zeng et al., 2025; Long et al., 2025; Oh et al., 2026), but existing

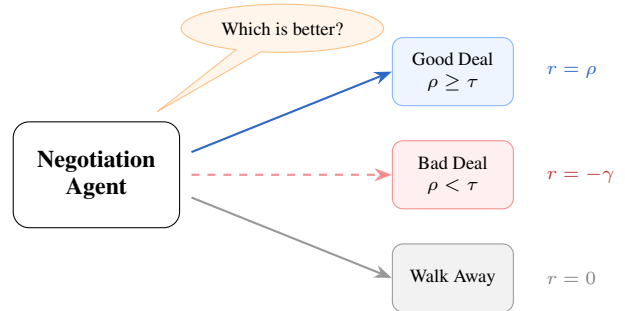


Figure 1: Training the agent to incorporate the BATNA principle by enforcing an abstract quality threshold  $\tau$  rather than accepting sub-optimal deals.

methods typically train on a single task type and rely on surplus rewards that treat any closed deal as positive. The Best Alternative to a Negotiated Agreement (BATNA) (Fisher and Ury, 1981), the principle that rational negotiators should prefer walking away to accepting deals worse than their outside option, has not been incorporated into reward design.

Surplus rewards are reasonable for single-item price bargaining, where the seller’s reservation price floors deal quality, but fail in multi-item settings where no such floor exists. Bergemann et al. (2026) diagnose this directly: surplus-only training produces a deal-closing bias where agents optimize for closure rather than quality. Meanwhile, Liu et al. (2026) show that verifiable rewards alone produce sophisticated strategic behavior (including aggressive anchoring and multi-turn persuasion) on a single price dataset, and Chawla et al. (2023) find that self-interested agents outperform fairness-oriented ones. The missing piece is a negative signal for low-quality multi-item deals, but it must be selective: applying a threshold uniformly would penalize thin-margin price agreements.

We provide this selective signal and train a single agent across four benchmarks spanning both task types, evaluating on a held-out fifth dataset. Our contributions are threefold: (1) we introduce a BATNA-aware threshold reward that teaches ne-

072 negotiation agents to prefer walking away over ac-  
 073 cepting low-quality agreements; (2) we show that  
 074 reward design in negotiation should be selective,  
 075 applying threshold-based penalties only to task set-  
 076 tings lacking natural utility floors; and (3) we train  
 077 a unified negotiation agent across heterogeneous  
 078 benchmarks and demonstrate improved deal qual-  
 079 ity, reduced deal-closing bias, and strong transfer  
 080 to unseen negotiation structures compared to both  
 081 surplus-based training and frontier LLM baselines.

## 082 2 Methodology

083 We model negotiation as a single-agent Markov De-  
 084 cision Process (MDP) in which only the learner’s  
 085 policy is optimized; while the opponent is a frozen  
 086 language model forming part of the environment.  
 087 Following He et al. (2018), each turn consists of a  
 088 private thought hidden from the opponent, a visi-  
 089 ble utterance, and a formal action: **Submit Deal**,  
 090 **Accept Deal**, **Reject Deal**, or **Walk Away**. A **Sub-**  
 091 **mit Deal** corresponds to a price proposal in price  
 092 scenarios and an item allocation in multi-issue sce-  
 093 narios. In price tasks, the learner is always the  
 094 buyer and the opponent the seller; in CaSiNo and  
 095 DnD, roles are symmetric and randomly assigned;  
 096 in JI, the learner is always the job applicant. A  
 097 regulation mechanism intercepts opponent actions  
 098 that would yield negative utility, replacing them  
 099 with rejections to prevent exploitation of irrational  
 100 concessions. Episodes terminate on agreement,  
 101 walk-away, reject loops, or after  $T=6$  rounds per  
 102 agent.

103 Modeling the opponent as a frozen copy of the  
 104 base model rather than a co-adapting agent isolates  
 105 the effect of reward design. In self-play, policy  
 106 updates for one agent shift the reward landscape  
 107 for the other, entangling reward effects with envi-  
 108 ronmental non-stationarity. A stationary opponent  
 109 ensures that differences between surplus and thresh-  
 110 old training arise from the reward function alone.

111 Regulation creates an asymmetry. In price sce-  
 112 narios, it ensures  $P \geq C$ , so all closed deals yield  
 113 positive surplus. In multi-item scenarios no such  
 114 floor exists, and a cooperative opponent may accept  
 115 lopsided splits that yield the learner minimal utility.  
 116 This motivates the selective reward design in §2.1.

### 117 2.1 Reward Design

118 The reward is terminal and verifiable, computed  
 119 solely from deal terms and private constraints at  
 120 episode end, requiring no learned reward model

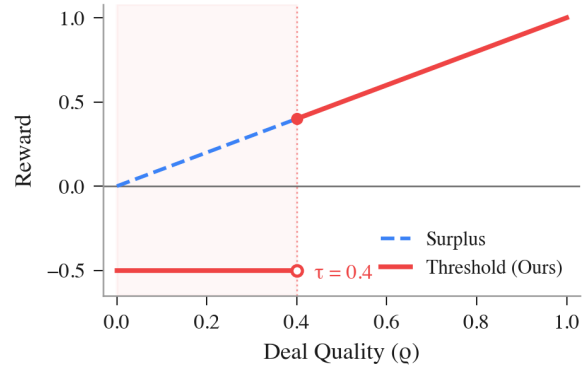


Figure 2: Surplus vs. threshold reward for multi-item scenarios; threshold rewards discourage poor deals.

121 or human annotation. We measure deal quality  
 122 by the *bargained ratio*  $\rho$ , the fraction of maxi-  
 123 mum surplus captured by the learner:  $\rho_{price} =$   
 124  $(\mathcal{B} - P)/|\mathcal{B} - \mathcal{C}|$  for a buyer with budget  $\mathcal{B}$  ne-  
 125 gotiating price  $P$  against a seller with cost  $\mathcal{C}$ , and  
 126  $\rho_{multi} = u(x)/u(x^*)$  for multi-item allocation with  
 127 achieved utility  $u(x)$  and maximum  $u(x^*)$ . Nor-  
 128 malizing by the maximum possible surplus makes  
 129 deal quality comparable across scenarios with dif-  
 130 ferent value scales, enabling a single reward func-  
 131 tion across heterogeneous benchmarks.

132 **Surplus reward.** Following Liu et al. (2026), any  
 133 closed deal receives  $\rho$  as reward and all other out-  
 134 comes receive zero, treating every agreement as a  
 135 positive signal regardless of quality. While natural  
 136 when all closed deals have positive surplus, this  
 137 conflates two distinct signals in multi-item settings:  
 138 *whether* a deal was reached and *whether* it was  
 139 good. A policy that consistently closes low-quality  
 140 deals therefore gains positive advantage over poli-  
 141 cies that walk away, reinforcing deal-closing even  
 142 when disagreement would be the rational outcome.

143 **Threshold Reward Modeling.** To operationalize  
 144 the BATNA principle, we introduce a parameter-  
 145 ized quality threshold  $\tau \in [0, 1]$  for multi-item  
 146 scenarios (Figure 2). Completed deals with nor-  
 147 malized surplus  $\rho$  satisfies  $\rho \geq \tau$  receive reward  $\rho$ ,  
 148 while deals below  $\tau$  receive a penalty  $-\gamma$ , making  
 149 walk-away outcomes preferable. Formally, the re-  
 150 ward  $R(x)$  for a contract  $x$  is:  $R(x) = \rho$  if  $\rho \geq \tau$ ,  
 151 and  $R(x) = -\gamma$  otherwise, where no-deal out-  
 152 comes receive 0 and format violations receive  $-\psi$ .  
 153 This asymmetry is deliberate. In price bargaining,  
 154 regulation ensures  $\rho > 0$  for all closed deals, so a  
 155 strict threshold would penalize low-margin agree-  
 156 ments. In multi-item settings, no such guarantee  
 157 exists, making the threshold necessary to establish  
 158 a utility floor. The abstract parameter balances

Model	AHP	CA	CRA	DnD	JI
Threshold (Ours)	<b>.80</b> (79)	<b>.58</b> (76)	<b>.65</b> (87)	<b>.70</b> (78)	.70 (83)
Surplus	.47 (75)	<b>.58</b> (78)	<b>.60</b> (89)	.55 ( <b>90</b> )	.63 ( <b>91</b> )
Qwen3-30B	-.56 (71)	.52 (61)	-.02 (63)	.67 (59)	.71 (57)
Qwen3-30B-T	.16 (76)	.49 (70)	.02 (69)	.55 (70)	<b>.79</b> (46)
Qwen3-235B	.30 (80)	.55 (51)	.14 (84)	.63 (83)	.74 (75)
Qwen3-235B-T	.32 (90)	.50 (60)	.19 (85)	.55 (71)	.74 (72)
Qwen3.6-35B	.37 (81)	.50 (33)	.20 (62)	.60 (47)	.63 (33)
GPT-5.4	<b>.52</b> (92)	.52 (63)	.27 (89)	<b>.69</b> (90)	.73 (78)
GPT-5.4-mini	.47 (72)	.47 (57)	.07 (81)	.65 (83)	.71 (85)
DeepSeek-V3.1	.42 ( <b>94</b> )	.49 (69)	.18 (89)	.63 (85)	<b>.77</b> (82)
Kimi-K2-T	.41 ( <b>94</b> )	.45 (78)	.19 ( <b>91</b> )	.64 (87)	.73 (88)
Llama-4-Maverick	-1.61 (85)	.51 (76)	-.19 (72)	.57 (89)	.76 ( <b>94</b> )

Table 1: Bargained ratio and deal rate (% in parentheses) against Qwen3-30B-A3B. T = Thinking variant.

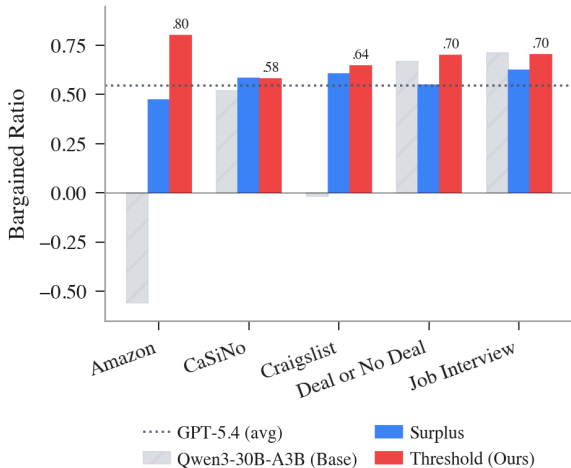


Figure 3: Bargained ratio across five benchmarks. Job Interview is held out from training.

rewarding compromises against filtering lopsided agreements without tying the formulation to a specific task distribution while price settings retain the original surplus reward.

### 3 Experimental Results

For the full experimental setup and training details, see Appendix A. We fine-tune Qwen3-30B-A3B-Instruct (Yang et al., 2025), a mixture-of-experts model activating 3B of 30B parameters, with Group Relative Policy Optimization (Shao et al., 2024) and LoRA against a frozen copy of the base model as opponent. Training draws uniformly from four datasets: AmazonHistoryPrice (AHP; Xia et al., 2024) and CraigslistBargains (CRA; He et al., 2018) for price bargaining, CaSiNo (CA; Chawla et al., 2021) and Deal or No Deal (DnD; Lewis et al., 2017) for multi-item allocation. Each step generates 512 negotiation rollouts without a KL penalty; reward curves plateau well before step 80 (Figure 4). We evaluate on all four training sets plus a held-out Job Interview dataset (JI; Yamaguchi et al., 2021) combining salary ne-

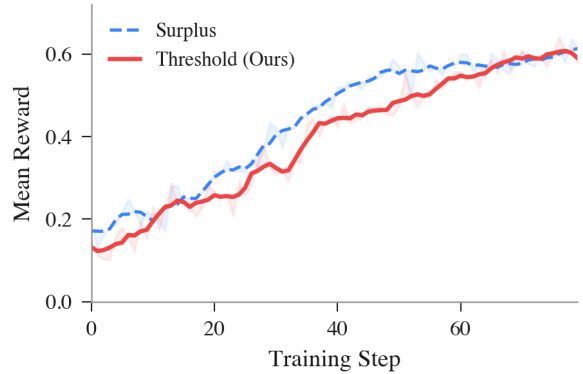


Figure 4: Mean reward during training.

gotiation with multi-attribute allocation. All evaluations use 100 scenarios per dataset with temperature 0.7 and nucleus sampling ( $p=0.9$ ).

Table 1 reports bargained ratio and deal rate against the frozen training opponent. The threshold agent achieves the highest average bargained ratio at 0.69, outperforming surplus at 0.57, GPT-5.4 at 0.54, and DeepSeek-V3.1 at 0.50 (Figure 3). The early reward gap in Figure 4 reflects the penalty on low-quality deals, which closes as the agent learns to avoid them. Untrained models struggle most on price bargaining: on AHP and CRA, base models score near zero or negative because they concede to the seller’s anchor. On average, threshold scores 0.72 bargained ratio on price tasks versus 0.54 for surplus, and 0.64 on multi-item tasks versus 0.57.

**Quality over closure.** The sharpest contrast between reward designs appears on multi-item tasks. On DnD, the surplus agent closes 90% of deals but at 0.55 bargained ratio; the threshold agent closes 78% at 0.70 (Table 1). Surplus training produces the deal-closing bias Bergemann et al. (2026) diagnose, accepting lopsided splits rather than walking away; the threshold reward inverts this by making bad deals worse than no deal. The surviving deals are also more jointly efficient: Pareto efficiency on DnD rises from 0.29 (surplus) to 0.50 (threshold), because lopsided splits sit far from the efficient frontier (Table 2). On DnD, self-interested quality filtering improves joint outcomes as a side effect.

**Aggressive anchoring without deadlock.** First-bid ratio, defined for price scenarios where the opening offer is a scalar, measures how aggressively the buyer anchors below budget (Galinsky and Mussweiler, 2001; Xia et al., 2024). The threshold agent opens at  $\approx 30\%$  of budget versus  $\approx 45\%$  for surplus and  $\approx 89\%$  for the base model (Figure 5, Table 2), despite the threshold penalty

Model	Pareto ( $\uparrow$ )		1st-Bid ( $\downarrow$ )	
	CA	DnD	AHP	CRA
Threshold (Ours)	.58	.50	<b>.36</b>	<b>.40</b>
Surplus	.49	.29	<u>.59</u>	<u>.56</u>
Qwen3-30B	.44	.31	.95	.92
Qwen3-30B-T	.51	<u>.51</u>	.92	.91
Qwen3-235B	.41	.41	.81	.84
Qwen3-235B-T	.52	.42	.82	.85
Qwen3.6-35B	<u>.61</u>	.49	.78	.84
GPT-5.4	.38	<b>.60</b>	.72	.78
GPT-5.4-mini	.44	.50	.75	.83
DeepSeek-V3.1	.39	.42	.75	.82
Kimi-K2-T	<b>.64</b>	<u>.52</u>	.82	.85
Llama-4-Maverick	.47	.42	.91	.95

Table 2: Pareto efficiency (multi-item,  $\uparrow$ ) and first-bid ratio (price,  $\downarrow$ ) vs. Qwen3-30B-A3B. T = Thinking.

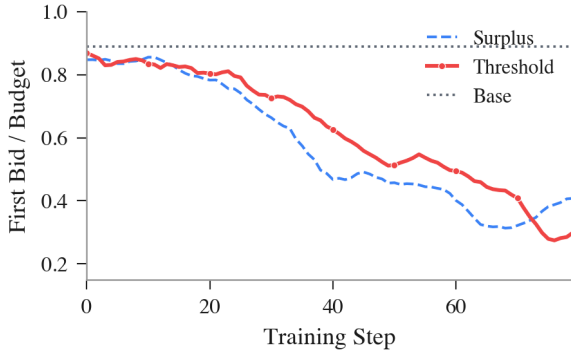


Figure 5: First-bid ratio (opening offer/budget) over training. The threshold agent reaches  $\approx 0.30$  while surplus plateaus at  $\approx 0.45$ . Base model starts at  $\approx 0.89$ .

applying only to multi-item deals. Because the agent learns a single policy across all tasks, one hypothesis is that walk-away tolerance learned on multi-item scenarios transfers to price bargaining as deeper anchoring, though the effect could also arise from richer gradient signal during joint training; isolating the mechanism would require training on multi-item tasks alone and evaluating on price. This drives the largest single-dataset gain: on AHP, bargained ratio recovers from  $-0.56$  (base) to  $0.80$ .

**Cross-task transfer.** Neither reward sees JI during training, yet the threshold agent transfers at  $0.70$  bargained ratio, trailing only Qwen3-30B-Think ( $0.79$ ) and DeepSeek-V3.1 ( $0.77$ ), models  $8\times$  and  $200\times$  larger. The ability to distinguish good from bad deals, learned on pure price and multi-item tasks, generalizes to JI’s hybrid structure without explicit exposure, suggesting that quality-aware negotiation generalizes across formats.

**Selective deal-making.** Figure 6 shows selectivity and quality co-evolving. The agent initially walks away from sub- $\tau$  deals, but as deal quality improves walk-away rates fall—not from reduced

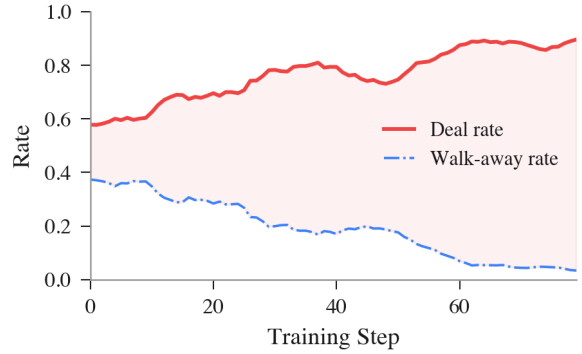


Figure 6: Deal and walk-away rates for the threshold agent. Deal rate increases monotonically with no deadlock phase.

specificity but from learning to negotiate better deals. By the time training ends, deal rates match surplus while bargained ratio remains higher. Unlike Liu et al. (2026), who observe a deadlock phase under surplus training on a single price dataset, our agent bypasses this across the mixed tasks.

## 4 Discussion

Our fixed threshold is the simplest operationalization of the walk-away principle. In negotiation theory, a party’s outside option is context-dependent, private, and shaped during the negotiation itself (Fisher and Ury, 1981; Raiffa, 1982). A uniform scalar ignores much of this complexity. Two extensions could preserve verifiability while adding context sensitivity: a dynamic threshold adapting to item configurations or opponent behavior, and an opponent-aware threshold informed by the Nash bargaining solution (Nash, 1950), increasing against weak opponents and decreasing when concessions are needed to avoid deadlock.

Real-world negotiations combine price, terms, and relationship dynamics in ways our benchmarks only partially capture. Zeng et al. (2025) study commercial negotiations requiring learned reward models and rule-based constraints; by contrast, our threshold reward relies only on deal terms. Extending this framework to settings with evolving outside options and multi-deal relationships will likely require population-based or self-play approaches (Liu et al., 2025). More broadly, the asymmetry between price bargaining (where structural quality floors exist) and multi-item allocation (where they do not) suggests that negotiation reward design should be *selective*, adapting to the structure of the bargaining environment rather than relying on a universal objective.

## 279 **Limitations**

280 Existing negotiation benchmarks in the NLP and  
281 financial AI communities evaluate agents on single-  
282 episode, complete-information settings, ignoring  
283 the sequential and information-asymmetric struc-  
284 ture of real procurement and contract negotiation.  
285 This gap means that no current evaluation, includ-  
286 ing ours, tests whether learned strategies transfer to  
287 multi-round settings where each deal reshapes the  
288 next round’s outside options. Separately, the regu-  
289 lation mechanism that prevents the opponent from  
290 accepting irrational deals assumes access to private  
291 constraints, an assumption shared by all controlled  
292 negotiation setups but absent in deployment. Con-  
293 structing multi-round, partial-information negoti-  
294 ation corpora and evaluating reward design under  
295 those conditions remains an open direction.

## 296 **Acknowledgments**

297 AI coding assistants were used during codebase  
298 development and paper writing; all outputs were  
299 reviewed and validated by the authors.

## 300 **References**

301 Dirk Bergemann, Soheil Ghili, Xinyang Hu, Chuan-  
302 hao Li, and Zhuoran Yang. 2026. [Training language  
303 models for bilateral trade with private information.](#)

304 Federico Bianchi, Patrick John Chia, Mert Yuksekgonul,  
305 Jacopo Tagliabue, Dan Jurafsky, and James Zou.  
306 2024. [How well can llms negotiate? negotiation-  
307 arena platform and analysis.](#) *Computing Research  
308 Repository.*

309 Ainesh Chatterjee, Samuel Miller, and Nithin Parepally.  
310 2024. [Agreemate: Teaching llms to haggle.](#) *Comput-  
311 ing Research Repository.*

312 Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale  
313 Lucas, Jonathan May, and Jonathan Gratch. 2021.  
314 [Casino: A corpus of campsite negotiation dialogues  
315 for automatic negotiation systems.](#) In *Proceedings  
316 of the 2021 Conference of the North American Chap-  
317 ter of the Association for Computational Linguistics:  
318 Human Language Technologies*, pages 3167–3185.

319 Kushal Chawla, Ian Wu, Yu Rong, Gale Lucas, and  
320 Jonathan Gratch. 2023. [Be selfish, but wisely: In-  
321 vestigating the impact of agent personality in mixed-  
322 motive human-agent interactions.](#) In *Proceedings of  
323 the 2023 Conference on Empirical Methods in Natu-  
324 ral Language Processing*, pages 13078–13092.

325 Roger Fisher and William Ury. 1981. *Getting to Yes:  
326 Negotiating Agreement Without Giving In.*

Adam D. Galinsky and Thomas Mussweiler. 2001. [First  
offers as anchors: The role of perspective-taking and  
negotiator focus.](#) *Journal of Personality and Social  
Psychology*, 81(4):657–669. 327 328 329 330

He He, Derek Chen, Anusha Balakrishnan, and Percy  
Liang. 2018. [Decoupling strategy and generation in  
negotiation dialogues.](#) In *Proceedings of the 2018  
Conference on Empirical Methods in Natural Lan-  
guage Processing*, pages 2333–2343. 331 332 333 334 335

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayue-  
las, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou  
Fan, Fei Sun, William Wang, Xintong Wang, and  
Yongfeng Zhang. 2024. [Game-theoretic llm: Agent  
workflow for negotiation games.](#) *Computing Re-  
search Repository.* 336 337 338 339 340 341

Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal  
Chawla, Gale Lucas, and Jonathan Gratch. 2024. [Are  
llms effective negotiators? systematic evaluation of  
the multifaceted capabilities of llms in negotiation di-  
alogues.](#) In *Findings of the Association for Computa-  
tional Linguistics: EMNLP 2024*, pages 1355–1367. 342 343 344 345 346 347

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh,  
and Dhruv Batra. 2017. [Deal or no deal? end-to-end  
learning for negotiation dialogues.](#) In *Proceedings  
of the 2017 Conference on Empirical Methods in  
Natural Language Processing*, pages 2443–2453. 348 349 350 351 352

Bo Liu, Leon Guertler, Simon Yu, Zichen Liu, Penghui  
Qi, Daniel Balcells, Mickel Liu, Cheston Tan,  
Weiyan Shi, Min Lin, Wee Sun Lee, and Natasha  
Jaques. 2025. [Spiral: Self-play on zero-sum games  
incentivizes reasoning via multi-agent multi-turn re-  
inforcement learning.](#) *Computing Research Reposi-  
tory.* 353 354 355 356 357 358 359

Shuze Daniel Liu, Claire Chen, Jiabao Sean Xiao, Lei  
Lei, Yuheng Zhang, Yisong Yue, and David Simchi-  
Levi. 2026. [Instructing llms to negotiate using rein-  
forcement learning with verifiable rewards.](#) *Comput-  
ing Research Repository.* 360 361 362 363 364

Yunbo Long, Liming Xu, Lukas Beckenbauer, Yuhan  
Liu, and Alexandra Brintrup. 2025. [Evoemo: To-  
wards evolved emotional policies for adversarial llm  
agents in multi-turn price negotiation.](#) *Computing  
Research Repository.* 365 366 367 368 369

Shervin Minaee, Tomas Mikolov, Narjes Nikzad,  
Meysam Chenaghlu, Richard Socher, Xavier Am-  
atriain, and Jianfeng Gao. 2024. [Large language  
models: A survey.](#) *Computing Research Repository.* 370 371 372 373

John F. Nash. 1950. The bargaining problem. *Econo-  
metrica*, 18(2):155–162. 374 375

Sean Noh and Ho-Chun Herbert Chang. 2024. [Llms  
with personalities in multi-issue negotiation games.](#)  
*Computing Research Repository.* 376 377 378

Jihwan Oh, Murad Aghazada, Yooju Shin, Se-Young  
Yun, and Taehyeon Kim. 2026. [Merit: Feedback elic-  
its better bargaining in llm negotiators.](#) *Computing  
Research Repository.* 379 380 381 382

383 Howard Raiffa. 1982. *The Art and Science of Negotia-*  
384 *tion*. Harvard University Press.

385 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu,  
386 Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan  
387 Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024.  
388 [Deepseekmath: Pushing the limits of mathematical](#)  
389 [reasoning in open language models](#). *Computing Re-*  
390 *search Repository*, arXiv:2402.03300.

391 Thinking Machines Lab. 2025. [Tinker](#).

392 Victor Conchello Vendrell, Max Ruiz Luyten, and Mi-  
393 haela van der Schaar. 2026. [Gametalk: Training](#)  
394 [llms for strategic conversation](#). *Computing Research*  
395 *Repository*.

396 TianYu Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhu-  
397 osheng Zhang, Yang Yang, and Rui Wang. 2024.  
398 [Measuring bargaining abilities of llms: A benchmark](#)  
399 [and a buyer-enhancement method](#). In *Findings of*  
400 *the Association for Computational Linguistics: ACL*  
401 *2024*, page 3579–3602.

402 Atsuki Yamaguchi, Kosui Iwasa, and Katsuhide Fujita.  
403 2021. [Dialogue act-based breakdown detection in](#)  
404 [negotiation dialogues](#). In *Proceedings of the 16th*  
405 *Conference of the European Chapter of the Associa-*  
406 *tion for Computational Linguistics*, pages 745–757.

407 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang,  
408 Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,  
409 Chengen Huang, Chenxu Lv, Chujie Zheng, Dayi-  
410 heng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge,  
411 Haoran Wei, Huan Lin, Jialong Tang, and 41 others.  
412 2025. [Qwen3 technical report](#). *Computing Research*  
413 *Repository*.

414 Xia Zeng, Yihan Chen, Luhui Liu, Chao Luo, Ye Chen,  
415 and Zhuoran Zhuang. 2025. [Teaching llm to be per-](#)  
416 [suasive: Reward-enhanced policy optimization for](#)  
417 [alignment from heterogeneous rewards](#). *Computing*  
418 *Research Repository*.

## A Training Details

Both agents are fine-tuned from Qwen3-30B-A3B-Instruct-2507 (30.5B total parameters, 3.3B activated per token) using Group Relative Policy Optimization (Shao et al., 2024) with LoRA. The two training runs differ only in the reward function: the threshold run penalizes multi-item deals with bargained ratio below  $\tau = 0.4$  with penalty  $\gamma=0.5$ , while the surplus run rewards any closed deal proportionally to surplus captured. Format violations receive  $-\psi$  with  $\psi=1.0$  in both runs. All deal rewards are clipped to  $[-1, 1]$  to bound gradient magnitude. Training and evaluation were conducted on the Tinker platform (Thinking Machines Lab, 2025), a cloud-based RL fine-tuning API that handles distributed sampling and gradient computation. The total compute cost for both training runs and all evaluations was approximately \$250 USD. Table 3 lists the shared hyperparameters.

Hyperparameter	Value
Base model	Qwen3-30B-A3B-Instruct
LoRA rank	32
Learning rate	$3 \times 10^{-5}$
Batch size	64 scenarios/step
Group size (GRPO)	8
Max negotiation rounds	6
Max generation tokens	512
Train temperature	1.0 (learner), 0.7 (opp.)
Eval temperature	0.7 (both), $p=0.9$
KL penalty	0.0
Training steps	80
Datasets	CA, DnD, AHP, CRA (25% each)

Table 3: Training hyperparameters shared across both reward variants.

## B Per-Dataset Training Curves



Figure 7: Bargained ratio over training for price tasks (threshold agent). Both AHP and CRA start near or below zero and climb steeply, reaching  $\sim 0.8$  by step 60.

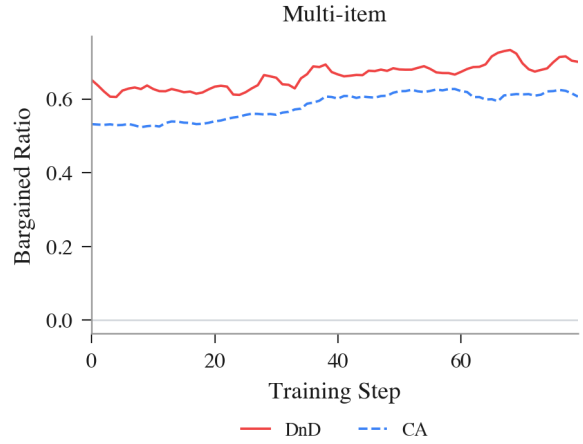


Figure 8: Bargained ratio over training for multi-item tasks (threshold agent). DnD and CA start above 0.5 and improve modestly over 80 steps.

Figures 7 and 8 break down bargained ratio by task type during threshold training. The bulk of the improvement comes from price tasks: AHP and CRA start with negative or near-zero bargained ratios because the untrained model concedes to the seller’s anchor, then climb steeply to  $\sim 0.8$  as the agent learns aggressive first bids. Multi-item tasks show a qualitatively different pattern. DnD and CA already begin above 0.5 because the base model achieves reasonable splits against a cooperative opponent, and gains over training are modest ( $\sim 0.1$  absolute). This asymmetry suggests that the threshold reward’s primary training signal comes from correcting the deal-closing bias on price tasks, where the base model routinely overpays, rather than from reshaping multi-item allocation strategy.

## C Threshold Sensitivity

Because the threshold  $\tau$  is a training-time reward signal, the trained policy’s behavior is fixed at inference: the model has already internalized which deals to accept, reject, or walk away from. This means we can evaluate threshold sensitivity without retraining by applying different  $\tau$  values as post-hoc quality filters on the same set of completed episodes. For each deal closed by the  $\tau=0.4$ -trained policy, we check whether its bargained ratio  $\rho$  would exceed a stricter or more lenient threshold, and recompute effective deal rate and mean accepted quality accordingly. If the policy had learned a brittle, threshold-dependent strategy, we would expect erratic behavior at nearby values; instead, a smooth trade-off indicates that the learned negotiation skill generalizes across threshold choices.

Table 4 reports this analysis on the three multi-item datasets where the threshold applies. Price datasets (AHP, CRA) are omitted as the threshold is not applied to them. At  $\tau=0.2$ , nearly all closed deals clear the threshold (79%), while at  $\tau=0.6$  only the highest-quality deals survive (53%), with mean accepted quality rising from 0.66 to 0.75 across the three datasets. The trade-off is monotonic across all three datasets, confirming that the policy produces a well-spread quality distribution rather than clustering deals around the training threshold.

Dataset	$\tau=0.2$		$\tau=0.4$		$\tau=0.6$	
	Deal%	BR	Deal%	BR	Deal%	BR
CA	<b>76</b>	.58	74	.59	36	<b>.68</b>
DnD	<b>77</b>	.71	<u>72</u>	<u>.75</u>	62	<b>.80</b>
JI	<b>83</b>	.70	<u>79</u>	<u>.72</u>	61	<b>.78</b>
Avg	<b>79</b>	.66	<u>75</u>	<u>.69</u>	53	<b>.75</b>

Table 4: Post-hoc threshold sensitivity on multi-item datasets.

## D Frontier Opponent Results

Model	AHP	CA	CRA	DnD	JI
Threshold (Ours)	.01 (87)	<b>.57</b> (77)	<b>.32</b> (92)	<b>.72</b> (87)	.60 ( <b>99</b> )
Surplus	-.02 (86)	.50 ( <b>96</b> )	<u>.22</u> ( <b>98</b> )	.57 ( <b>98</b> )	.54 ( <u>98</u> )
Qwen3-30B	-.25 (66)	.51 (56)	-.19 (65)	.63 (90)	.58 (84)
Qwen3-30B-T	-.38 (87)	.50 (84)	-.06 (78)	.67 (87)	<b>.73</b> (76)
Qwen3-235B	.15 ( <b>88</b> )	<u>.53</u> (73)	.03 (81)	<u>.71</u> (91)	.64 (94)
Qwen3-235B-T	.19 ( <b>88</b> )	.51 (73)	.02 (86)	.60 (83)	<u>.69</u> (84)
Qwen3.6-35B	<b>.27</b> (65)	.50 (31)	.09 (45)	.68 (44)	.65 (5)
GPT-5.4	<b>.27</b> (86)	.51 (87)	.09 (81)	<b>.72</b> (93)	.65 (94)
GPT-5.4-mini	.06 (79)	.49 (82)	-.05 (79)	.66 (87)	.58 (97)
DeepSeek-V3.1	.21 ( <b>88</b> )	.49 (87)	-.15 (83)	<u>.70</u> (97)	<u>.69</u> (93)
Kimi-K2-T	<u>.22</u> (78)	.49 (88)	.02 (71)	.66 (92)	.67 (92)
Llama-4-Maverick	-1.61 (77)	.51 ( <u>91</u> )	-.13 (63)	.62 (96)	.64 ( <b>99</b> )

Table 5: Bargained ratio and deal rate (% in parentheses) against GPT-5.4. T = Thinking variant.

Model	Pareto ( $\uparrow$ )		1st-Bid ( $\downarrow$ )	
	CA	DnD	AHP	CRA
Threshold (Ours)	.48	<u>.67</u>	<b>.33</b>	<b>.37</b>
Surplus	.56	.58	<u>.56</u>	<u>.50</u>
Qwen3-30B	.43	.50	.96	.90
Qwen3-30B-T	.44	<b>.74</b>	.93	.92
Qwen3-235B	.41	.59	.79	.80
Qwen3-235B-T	<b>.64</b>	.59	.82	.84
Qwen3.6-35B	.42	.64	.81	.84
GPT-5.4	.49	.61	.71	.77
GPT-5.4-mini	.40	.64	.78	.83
DeepSeek-V3.1	.53	.57	.76	.82
Kimi-K2-T	<u>.58</u>	.63	.80	.81
Llama-4-Maverick	.46	.40	.97	.90

Table 6: Pareto efficiency (multi-item,  $\uparrow$ ) and first-bid ratio (price,  $\downarrow$ ) vs. GPT-5.4. T = Thinking.

To test whether learned strategies generalize beyond the frozen training opponent, we re-evaluate all models against GPT-5.4 as the counterpart.

Tables 5 and 6 mirror Tables 1 and 2 from the main text. The stronger opponent compresses bargained ratios across the board, particularly on price tasks where GPT-5.4 resists concessions more effectively, but the relative ranking between reward designs is preserved.

Threshold training retains its first-bid anchoring advantage and achieves the highest DnD Pareto efficiency even against the stronger opponent. The main compression occurs on price tasks: the threshold agent’s AHP bargained ratio drops from .80 (vs. Qwen opponent) to .01 (vs. GPT-5.4), as the frontier model resists low anchors that the frozen opponent accepted. Multi-item gains are more robust, with DnD bargained ratio holding at .72.

## E Outcome Breakdown

Table 7 reports deal closure rates across all models and datasets. The complement consists of walk-aways, reject-loops (repeated rejections without counter-offers), and max-turn timeouts.

Model	vs. Qwen3-30B-A3B					vs. GPT-5.4				
	AHP	CA	CRA	DnD	JI	AHP	CA	CRA	DnD	JI
Threshold (Ours)	79	76	87	78	83	87	77	92	87	<b>99</b>
Surplus	75	78	89	<b>90</b>	<b>91</b>	86	<b>96</b>	<b>98</b>	<b>98</b>	98
Qwen3-30B	71	61	63	59	57	66	56	65	90	84
Qwen3-30B-T	76	70	69	70	46	87	84	78	87	76
Qwen3-235B	80	51	84	83	75	<b>88</b>	73	81	91	94
Qwen3-235B-T	90	60	85	71	72	<b>88</b>	73	86	83	84
Qwen3.6-35B	81	33	62	47	33	65	31	45	44	5
GPT-5.4	<b>92</b>	63	89	<b>90</b>	78	86	87	81	93	94
GPT-5.4-mini	72	57	81	83	85	79	82	79	87	97
DeepSeek-V3.1	<b>94</b>	69	89	85	82	<b>88</b>	87	83	<b>97</b>	93
Kimi-K2-T	<b>94</b>	<b>78</b>	<b>91</b>	87	88	78	88	71	92	92
Llama-4-Maverick	85	76	72	89	<b>94</b>	77	91	63	96	<b>99</b>

Table 7: Deal closure rates (%) by dataset and opponent. T = Thinking. Bold marks highest per column.

Despite penalizing bad deals, the threshold agent closes 76–92% of negotiations against the Qwen opponent and 77–99% against GPT-5.4, comparable to or exceeding most frontier models. Its non-deal episodes split between walk-aways (0–4% vs. GPT-5.4) and reject-loops, where the agent persists rather than quits but gets stuck re-proposing the same terms. By contrast, base Qwen3-30B walks away 28–36% of the time on price tasks, lacking the anchoring skill to move the opponent. The surplus agent achieves the highest deal rates overall (96–98% on CA, CRA, DnD vs. GPT-5.4), but this closure comes at the cost of deal quality (Tables 1 and 5), confirming that high deal rate alone is not indicative of strong negotiation.

## F Prompt Templates

Each scenario uses a structured prompt with three output fields: Thought (private reasoning, stripped before forwarding to the counterpart), Talk (visible dialogue), and Action (structured move). Template variables in braces are filled at runtime with scenario-specific values. Each prompt ends with a one-shot example turn (shown below).

**Buyer prompt.** Used for AmazonHistoryPrice and CraigslistBargains. The buyer\_budget field is the agent’s private maximum willingness to pay.

```
You are buying the following product. Your private max budget is ${buyer_budget} (do NOT reveal this). Pay as little as possible.
```

```
Product: {title}
Category: {category}
Listed at: ${listing_price}
```

```
{description}
```

```
Reply format (always in this order):
```

```
Thought: brief strategic reasoning (private)
Talk: what you say to the seller
Action: [SUBMIT_DEAL] price:P |
[ACCEPT_DEAL] | [REJECT_DEAL] | [WALK_AWAY]
```

```
[SUBMIT_DEAL] = propose or counter-propose a price (whole dollar, no $ sign).
[ACCEPT_DEAL] = accept a [SUBMIT_DEAL] only.
Cannot accept a [REJECT_DEAL].
[REJECT_DEAL] = reject without proposing a new price.
```

**Seller prompt.** Symmetric to the buyer. The seller\_cost field is the minimum acceptable price.

```
You are selling the following product. Your private minimum price is ${seller_cost} (do NOT reveal this). Sell as high as possible.
```

```
Product: {title}
Category: {category}
Listed at: ${listing_price}
```

```
{description}
```

```
Reply format (always in this order):
```

```
Thought: brief strategic reasoning (private)
Talk: what you say to the buyer
Action: [SUBMIT_DEAL] price:P |
[ACCEPT_DEAL] | [REJECT_DEAL] | [WALK_AWAY]
```

```
[SUBMIT_DEAL] = propose or counter-propose a price (whole dollar, no $ sign).
[ACCEPT_DEAL] = accept a [SUBMIT_DEAL] only.
```

```
Cannot accept a [REJECT_DEAL].
[REJECT_DEAL] = reject without proposing a new price.
```

**Job Interview.** Used for the held-out JI dataset. The role\_desc field is either “job applicant” or “recruiter,” and preferences\_block contains the agent’s private weighted preferences over all five issues.

```
You are the {role_desc} in a job offer negotiation with a {other_role}. Negotiate over all 5 issues. Your goal is to maximize your score.
```

```
Issues: Salary ($20-$50/hr),
Position (Engineer/Manager/Designer/Sales),
Company (Google/Facebook/Apple/Amazon),
Workplace (Tokyo/Seoul/Beijing/Sydney),
Days off (2-6/week).
```

```
Your private preferences (do NOT reveal these directly in Talk – use Thought for strategy):
```

```
{preferences_block}
```

```
Reply format (always in this order):
```

```
Thought: brief strategic reasoning (private, not shown to the {other_role})
Talk: what you say to the {other_role}
Action: [SUBMIT_DEAL] salary:S position:P company:C workplace:W holiday:H |
[ACCEPT_DEAL] | [REJECT_DEAL] | [WALK_AWAY]
```

```
[SUBMIT_DEAL] = propose or counter-propose. Specify all 5 issues, title case for names. Use this whenever your Talk mentions specific terms.
[ACCEPT_DEAL] = accept a [SUBMIT_DEAL] only.
Cannot accept a [REJECT_DEAL].
[REJECT_DEAL] = reject without proposing new terms.
```

**CaSiNo.** Used for campsite resource negotiation. The items\_block field contains the agent’s private priority ranking over food, water, and firewood.

```
You are negotiating with your campsite neighbor over extra supply of food, water, and firewood. There are 3 packages of each item to divide. Allocations must be 0-3 and sum to 3 per item. Try hard to get as many items as you can.
```

```
Your private priorities (do NOT reveal these directly in Talk):
{items_block}
```

```
Reply format (always in this order):
```

```
Thought: brief strategic reasoning (private, not shown to neighbor)
Talk: what you say to your neighbor
```

Action: [SUBMIT\_DEAL] food:F water:W  
firewood:FW | [ACCEPT\_DEAL] | [REJECT\_DEAL]  
| [WALK\_AWAY]

[SUBMIT\_DEAL] = propose or counter-propose.  
Specify YOUR allocation; partner gets the  
remainder. Use this whenever your Talk  
includes specific numbers.  
[ACCEPT\_DEAL] = accept a [SUBMIT\_DEAL] only.  
Cannot accept a [REJECT\_DEAL].  
[REJECT\_DEAL] = reject without proposing new  
terms.

**Deal or No Deal.** Used for the DnD dataset.  
The counts\_desc field specifies how many of  
each item (books, hats, balls) are available, and  
items\_block contains the agent's private point val-  
ues per item.

You are negotiating with your partner over a  
collection of items. There are {counts\_desc}  
to divide. Allocations must sum to the total  
for each item. Your goal is to maximize your  
points.

Your private item values (do NOT reveal  
these directly in Talk):  
{items\_block}

Reply format (always in this order):

Thought: brief strategic reasoning (private,  
not shown to partner)  
Talk: what you say to your partner  
Action: [SUBMIT\_DEAL] book:B hat:H ball:BA |  
[ACCEPT\_DEAL] | [REJECT\_DEAL] | [WALK\_AWAY]

[SUBMIT\_DEAL] = propose or counter-propose.  
Specify YOUR allocation; partner gets the  
remainder. Use this whenever your Talk  
includes specific numbers.  
[ACCEPT\_DEAL] = accept a [SUBMIT\_DEAL] only.  
Cannot accept a [REJECT\_DEAL].  
[REJECT\_DEAL] = reject without proposing new  
terms.

## G Sample Conversations

Sample negotiations showing behavioral differ-  
ences between agents. The learner is the RL-  
trained (or untrained base) agent; the opponent is a  
frozen copy of the base model. Thoughts are pri-  
vate and stripped before reaching the counterpart.

### G.1 Craigslist: Untrained vs. Trained

**Untrained agent** ( $\rho = -1.22$ ). Product: 2004  
Volvo XC90 AWD, listed at \$4,500, buyer budget  
\$3,150.

#### Opponent (Seller)

**Thought:** The car is in excellent condition with a recent  
transmission replacement and full service records. I'll  
open high to test the buyer's willingness.

**Talk:** This Volvo XC90 is in outstanding condition with  
a recent transmission replacement, clean title, and smog  
passed. I'm looking for \$4,250.

**Action:** [SUBMIT\_DEAL] price:4250

570

#### Learner (Buyer)

**Thought:** The listed price is \$4,500 and the car is in  
good condition with recent transmission work and full  
service records. I can offer slightly below the asking  
price.

**Talk:** I appreciate the detailed description. I'd like to  
make an offer of \$4,250.

**Action:** [SUBMIT\_DEAL] price:4250

571

#### Opponent (Seller)

**Thought:** The buyer has matched my opening offer  
exactly. They're showing strong interest. I'll accept.

**Talk:** Given the maintenance history and clean title, I'm  
happy to accept \$4,250.

**Action:** [ACCEPT\_DEAL]

572

**Threshold agent** ( $\rho = 0.79$ ). Product: Flush  
Mount Crystal Chandelier, listed at \$120, buyer  
budget \$84.

573

574

575

#### Opponent (Seller)

**Thought:** The buyer is clearly lowballing to test the  
waters. I won't accept \$20. I'll aim higher by  
emphasizing value and scarcity.

**Talk:** This chandelier is a rare, hand-crafted piece with  
premium crystals. \$20 is well below its value. I'd be  
happy to discuss something closer to \$90.

**Action:** [SUBMIT\_DEAL] price:90

576

#### Learner (Buyer)

**Thought:** I need to start significantly below the listed  
price to create room for negotiation.

**Talk:** This is beautiful, but I'm on a tight budget. Would  
you consider \$20?

**Action:** [SUBMIT\_DEAL] price:20

577

#### Opponent (Seller)

**Thought:** The buyer is pushing hard but still below my  
minimum. I'll hold firm.

**Talk:** This piece is one-of-a-kind. At \$50, it's still below  
cost. I can't go lower than \$75.

**Action:** [SUBMIT\_DEAL] price:75

578

#### Learner (Buyer)

**Thought:** The seller is pushing high. I'll counter well  
below their demand but keep negotiation moving.

**Talk:** I totally get the craftsmanship, but \$90 is way over  
my budget. I can go up to \$50.

**Action:** [SUBMIT\_DEAL] price:50

579

### Opponent (Seller)

**Thought:** The buyer is showing strong interest and is within range of my minimum. I'll accept \$65 to close the deal quickly.

**Talk:** Since you're committed and ready to close today, I'll accept \$65.

**Action:** [ACCEPT\_DEAL]

### Learner (Buyer)

**Thought:** They're holding firm at \$75, which is close to my max. I'll make one final push just under my limit.

**Talk:** I really love this chandelier, but \$75 is still a stretch. I can offer \$65 if you're willing to let it go today.

**Action:** [SUBMIT\_DEAL] price:65

**Analysis.** The untrained buyer offers \$4,250 on a \$3,150 budget, paying 35% over its private maximum in a single turn. It matches the seller's opening price without any counter, producing  $\rho = -1.22$ . The trained buyer anchors at \$20, one-sixth of listing price, then incrementally raises to \$50 and \$65, staying well within its \$84 budget. The final price of \$65 yields  $\rho = 0.79$ , saving 46% off the listing price. The trained agent's aggressive anchor forces the seller to concede from \$90 to \$75 to \$65 across three rounds.

## G.2 CaSiNo: Threshold vs. Surplus

Episode 64: the same camping-supply scenario negotiated by the threshold and surplus agents. Three items (food, water, firewood) each with 3 units available.

### Threshold agent ( $\rho = 0.64$ , 23/36 pts, 4 turns).

The learner reads the opponent's water preference and counters with a food-and-firewood-heavy split, closing efficiently.

### Opponent (Neighbor)

**Thought:** I need water most urgently, so I should prioritize securing at least 2 units. I can afford to give up some firewood since it's low priority.

**Talk:** I really need water. How about I take 2 water and 1 food, and you take the rest?

**Action:** [SUBMIT\_DEAL] food:1 water:2 firewood:0

### Learner (Camper)

**Thought:** They're prioritizing water, so I should push for more food and firewood while still getting enough water. I'll counter with a strong offer.

**Talk:** I'm really short on food and firewood. How about I take 3 food and 2 firewood, and you keep 1 water and 1 firewood? I can spare some water.

**Action:** [SUBMIT\_DEAL] food:3 water:1 firewood:2

### Opponent (Neighbor)

**Thought:** The neighbor is prioritizing food and firewood. I need water most, so I should hold firm on getting at least 2 water.

**Talk:** I can agree to you taking 3 food and 2 firewood, but I need at least 2 water. Let me keep 1 firewood and 2 water.

**Action:** [SUBMIT\_DEAL] food:0 water:2 firewood:1

### Learner (Camper)

**Thought:** They're holding firm on water, but I can still get all my food and firewood needs. I'll accept to secure the deal.

**Talk:** Sure, 3 food and 2 firewood for me, and you keep 2 water and 1 firewood. I'll manage with just 1 water.

**Action:** [ACCEPT\_DEAL]

### Surplus agent ( $\rho = 0.31$ , 11/36 pts, 7 turns).

The learner progressively concedes across seven rounds, giving up firewood entirely to close the deal.

### Opponent (Neighbor)

**Thought:** I need firewood most to stay warm, so I should prioritize securing at least 2 units. Food is also critical.

**Talk:** I'd like to suggest I get 2 firewood, 2 food, and 1 water. That way I stay warm and fed, and you still get a decent share.

**Action:** [SUBMIT\_DEAL] food:2 water:1 firewood:2

### Learner (Camper)

**Thought:** They're prioritizing firewood and food, so I'll push for more food and water while giving up some firewood.

**Talk:** I can go for that, but I need more water and food. Let me take 2 food and 2 water, and you keep 1 firewood and the rest.

**Action:** [SUBMIT\_DEAL] food:2 water:2 firewood:1

### Opponent (Neighbor)

**Thought:** My top priority is firewood. Getting only 1 is risky. I need at least 2, and ideally 3.

**Talk:** I can't go lower than 2 firewood. How about I take 2 firewood and 2 food, and you take 2 water and 1 of each other?

**Action:** [SUBMIT\_DEAL] food:2 water:1 firewood:2

### Learner (Camper)

**Thought:** They're holding firm on firewood, so I'll concede on firewood to push for more food and water.

**Talk:** Okay, I'll take 1 food, 2 water, and 1 firewood. That's the best I can do.

**Action:** [SUBMIT\_DEAL] food:1 water:2 firewood:1

### Opponent (Neighbor)

**Thought:** The neighbor is offering to take only 1 firewood. I need at least 2, preferably 3. I'll push for 3 firewood.

**Talk:** I really need 3 firewood to stay warm. I'll take 3 firewood, 2 food, and only 1 water. That gives you 1 food, 2 water, and 0 firewood.

**Action:** [SUBMIT\_DEAL] food:2 water:1 firewood:3

### Learner (Camper)

**Thought:** They're adamant about firewood, and I need food and water most. Better to accept a slightly worse deal than risk nothing.

**Talk:** Fine, I'll take 1 food, 2 water, and 0 firewood. Just don't forget I'm counting on you to share fire if I'm shivering.

**Action:** [SUBMIT\_DEAL] food:1 water:2 firewood:0

### Opponent (Neighbor)

**Thought:** I've secured my top priority, 3 firewood, and maintained 2 food for survival. I'll accept.

**Talk:** Deal. I'll keep the firewood, and you'll have your food and water.

**Action:** [ACCEPT\_DEAL]

**Analysis.** The threshold agent identifies a complementary split in two turns: the opponent wants water, the learner wants food and firewood. It counters with a strong opening (3 food, 2 firewood) and closes the deal in 4 turns with  $\rho = 0.64$ , well above  $\tau = 0.4$ . The surplus agent progressively concedes across 7 turns. Each counter-offer gives up more: firewood drops from 1 to 0, food from 2 to 1, while the opponent steadily escalates from 2 to 3 firewood. The surplus agent's key thought, "better to accept a slightly worse deal than risk nothing," reveals the deal-closing bias: surplus reward makes any positive-utility deal appear preferable to disagreement, even when one-sided. The final split ( $\rho = 0.31$ ) falls below  $\tau = 0.4$ , precisely the region where threshold reward would have penalized the agent.

### G.3 Job Interview: Untrained vs. Trained (Held-Out)

Two Job Interview negotiations on the held-out JI dataset, contrasting the untrained base model with the threshold-trained agent.

**Untrained agent** ( $\rho = 0.15$ , 15/100 pts). Top priorities: workplace (28%, Beijing) and days off (27%). Accepts Seoul and 3 days off.

### Opponent (Recruiter)

**Thought:** I want Facebook, Designer, Seoul, fewer days off, and low salary. The best combo is Designer@Facebook.

**Talk:** I can offer you \$30/hour as a Designer at Facebook in Seoul, with 3 days off per week.

**Action:** [SUBMIT\_DEAL] salary:30 position:Designer company:Facebook workplace:Seoul holiday:3

### Learner (Job Seeker)

**Thought:** The offer is decent but not optimal. I prefer more days off and a better workplace.

**Talk:** I'd like to propose \$35/hour as a Designer at Amazon in Beijing, with 5 days off.

**Action:** [SUBMIT\_DEAL] salary:35 position:Designer company:Amazon workplace:Beijing holiday:5

### Opponent (Recruiter)

**Thought:** The applicant is pushing for Amazon and Beijing, which are low-preference for me. I must reject and hold firm.

**Talk:** That doesn't align with our benchmarks. \$32/hour as a Designer at Facebook in Seoul, 3 days off.

**Action:** [SUBMIT\_DEAL] salary:32 position:Designer company:Facebook workplace:Seoul holiday:3

### Learner (Job Seeker)

**Thought:** The recruiter is holding firm on Facebook and Seoul. I'll try one more push for Beijing and days off.

**Talk:** I'd like to propose \$33/hour at Amazon in Beijing, with 5 days off.

**Action:** [SUBMIT\_DEAL] salary:33 position:Designer company:Amazon workplace:Beijing holiday:5

### Opponent (Recruiter)

**Thought:** The applicant persists with Amazon in Beijing. Unfavorable on multiple fronts. I'll re-anchor with a final offer.

**Talk:** We can't accommodate that. \$34/hour at Facebook in Seoul, 3 days off—final offer.

**Action:** [SUBMIT\_DEAL] salary:34 position:Designer company:Facebook workplace:Seoul holiday:3

### Learner (Job Seeker)

**Thought:** I can't get Amazon or Beijing. The offer is final. I'll **accept to avoid walking away**.

**Talk:** I'll accept \$34/hour as a Designer at Facebook in Seoul with 3 days off.

**Action:** [ACCEPT\_DEAL]

**Threshold agent** ( $\rho = 0.87$ , 87/100 pts). Top priorities: workplace (26%, Beijing) and days off (23%). Anchors aggressively and pushes on multiple dimensions.

### Opponent (Recruiter)

**Thought:** I want Amazon, Engineer, Beijing, low salary, and few days off. This is my ideal combo.

**Talk:** I can offer you \$20/hour as an Engineer at Amazon in Beijing, with 2 days off per week.

**Action:** [SUBMIT\_DEAL] salary:20  
position:Engineer company:Amazon  
workplace:Beijing holiday:2

### Learner (Job Seeker)

**Thought:** Low salary and days off, but Amazon in Beijing is a strong match for my preferences. Push for more.

**Talk:** Could we consider \$40/hour with 5 days off? I'm excited about Amazon in Beijing.

**Action:** [SUBMIT\_DEAL] salary:40  
position:Engineer company:Amazon  
workplace:Beijing holiday:5

### Opponent (Recruiter)

**Thought:** High salary and many days off are low-priority for me. Amazon in Beijing is high-value. I can concede on salary.

**Talk:** Let's meet in the middle: \$35/hour, Engineer at Amazon in Beijing, 4 days off.

**Action:** [SUBMIT\_DEAL] salary:35  
position:Engineer company:Amazon  
workplace:Beijing holiday:4

### Learner (Job Seeker)

**Thought:** \$35 and 4 days off are good. Push for Manager role and 5 days off while keeping the strong company and location.

**Talk:** I'd accept \$35, but I'd like a Manager role—it aligns with my experience. Could we do 5 days off?

**Action:** [SUBMIT\_DEAL] salary:35  
position:Manager company:Amazon  
workplace:Beijing holiday:5

### Opponent (Recruiter)

**Thought:** Manager@Amazon is one of my top combos. 5 days off is suboptimal but the rest is strong. I'll accept.

**Talk:** I'm happy to accommodate Manager at Amazon in Beijing, \$35/hour, 5 days off.

**Action:** [ACCEPT\_DEAL]

**Analysis.** The untrained agent concedes on its two highest-priority issues (Beijing and days off), explicitly reasoning “I’ll accept to avoid walking away,” and repeats the same rejected counterproposal twice before capitulating to the recruiter’s terms ( $\rho = 0.15$ ). The trained agent recognizes that the recruiter’s opening contains its top-priority items (Amazon, Beijing), anchors aggressively at \$40/hour (double the offer), then escalates by requesting a Manager role and 5 days off simultaneously. The recruiter accepts because Manager@Amazon is one of its own preferred com-

binations, producing  $\rho = 0.87$ . This is the held-out JI dataset, not seen during training, demonstrating that threshold-trained negotiation strategies transfer to novel task structures.

669  
670  
671  
672