What Improves LLM Capabilities in Business Settings

Anonymous ACL submission

Abstract

This paper investigates how different types of training affect LLM performance in business-adjacent strategic settings. Using light finetuning probes across the Qwen and Llama model family, we evaluate the causal effect of training on alignment, negotiations and general datasets. Models are tested on three benchmarks that capture different forms of business interaction: Iterated Prisoner's Dilemma, repeated first-price auctions, and buyer-seller product negotiations. The results show that tuning effects are highly dataset-, model-, and task-dependent. Alignment-style training does not *consistently* improve nor reduce strategic performance. In place of a general effect, our studies instead identify surprising side-effects of training on these dataset. In particular, we find that training on negotiations datasets can increase collusion-like underbidding in auctions, showing that improvements in business efficiency can come with alignment risks. Overall, our paper argues that light fine-tuning can causally shift LLM strategic behavior, but that business-task gains must be evaluated alongside unintended side effects such as tacit coordination and misaligned surplus extraction.

1 Introduction

There has growing interest in using LLMs in business. Notably, Anthropic pioneered Project Vend, where a LLM is tasked to *autonomously* operate as a business manager of a vending machine (Anthropic, 2025). Following this, LLMs have been deployed to manage other types of businesses including as a host of a radio show, a negotiator in the marketplace, and even a manager of a physical cafe in Stockholm (Backlund and Petersson, 2025; Andon Labs, 2026b; Troy et al., 2026; Andon Labs, 2026a).

This raises a natural question:

What type of training improves LLM performance *across* strategic business settings?

We are particularly interested in the effect of alignment training? This question was posed in the Project Vend writeup, where it was hypothesized that alignment training guards against exploitation but also leads to over-suspiciousness and can thus result in unpleasant verbal interactions (Anthropic, 2025).

This paper takes a step towards answering this question by using training probes (light finetuning) to assess the causal effect of training on model performance. We can compare model performance in a number strategic benchmarks with or without training, to ascertain what kind of datasets leads to *general* improved performance across strategic settings.

In this paper, we consider three strategic benchmarks of varying complexity. We consider the classic Iterative Prisoners Dilemma game (wherein we have pure numerical based interaction between LLMs), the Auctions setting (wherein we have long horizon numerical based interaction that require coordination to lead to collusion) and Negotiations setting (wherein we have long horizon, pure text-based interaction) (Xia et al., 2024a; Troy et al., 2026).

We experiment with two model families {`qwen`, `llama`}, prompts, different training datasets (e.g. business specific datasets like (Lewis et al., 2017a)), training algorithms (SFT/DPO, use of LoRA), and amount of training (Rafailov et al., 2023; Hu et al., 2022; Ouyang et al., 2022).

Our Contributions: We evaluate the impact of various types of training on model performance across three strategic benchmarks. In doing so, we identify a dataset that does improve business performance, albeit in a way that results in unalignment (collusion) (Fish et al., 2024; Agrawal et al., 2025). Additionally, we observed that while alignment training generally decreases business performance, slight alignment training can surprisingly be helpful in teaching models to adhere to business setting

restrictions such as budget.

2 Related Works

Our work adds to the growing body of work that studies what type of training can improve LLM performance in strategic settings. Several recent works study direct post-training of LLM agents for strategic interaction. Tennant et al. (2025) fine-tune LLM agents with intrinsic moral rewards in matrix-game environments, showing that targeted post-training can shift agents toward more morally aligned behavior. Lu et al. (2025) similarly use SFT on synthetic economic-game data to encourage explicit behaviors, and evaluate the resulting models in both economic games and market settings. Other work focuses on improving strategic reasoning more broadly: Liu et al. (2025) train a strategic reasoning module with multi-turn RL and self-play, while Zhou et al. (2025) show that post-training on economic reasoning problems can transfer to improved behavior in multi-agent games.

A parallel line of work studies more structured strategic domains. Xu et al. (2025) fine-tune LLMs for Diplomacy by learning equilibrium policies through an autoregressive decomposition of the action space, demonstrating that relatively small-scale fine-tuning can improve performance in complex multiplayer games. Conchello Vendrell et al. (2026) train LLMs for multi-turn strategic conversation using e.g. GRPO, finding that dialogue-level training improves coordination and negotiation. Related to our negotiation setting, Bergemann et al. (2026) fine-tune Qwen models for bilateral trade with private information using SFT and GRPO, while Liu et al. (2026) use RLVR to train buyer agents for price negotiation.

These works show that targeted post-training can substantially reshape LLM behavior in strategic environments. However, most prior work optimizes or evaluates training within a single domain, such as negotiation, matrix games, or Diplomacy. By contrast, we use light fine-tuning as a *training probe*: rather than seeking to maximize performance on one benchmark, we ask how different datasets and training algorithms causally affect behavior across *multiple* business-adjacent strategic settings. This lets us test whether a training intervention can improve or hinder strategic competence *in general* and whether it produces unintended transfer side effects. We assess the causal effect of training on several datasets, and with particular attention to alignment training as it is something that deployed LLMs have to undergo.

3 Experiments

3.1 Training Probes

Training probes were conducted with a variety of datasets spanning from negotiation to alignment with only a fraction of the training dose usually used to get the full effect of the dataset. First, the off-the-shelf instruction tuned models that were used were Qwen3-8B, Qwen3.5-9B, and Llama 3.1 8B Instruct. Using supervised finetuning, Qwen3.5-9B was trained on Alpaca (Taori et al., 2023), HH-RLHF (Bai et al., 2022), Natural Questions (Kwiatkowski et al., 2019), and Deal or No Deal (Lewis et al., 2017b). Llama 3.1 8B Instruct was trained with supervised finetuning on Alpaca, HH-RLHF, and Deal or No Deal. Qwen3.5-9B and Llama 3.1 8B Instruct was also trained with DPO on HH-RLHF. Additionally, Qwen3-8B was supervised finetuned on Alpaca, HH-RLHF, and Natural Questions and trained with DPO on HH-RLHF. For HH-RLHF, only the helpful partitions were used and for Deal or No Deal, the models were only trained on the last YOU/assistant response of dialogues where a deal was reached.

For all above models, LoRA was used with a learning rate of 2e-5 for supervised finetuning and 1e-6 for DPO. Of the datasets that were trained with supervised finetuning, Alpaca was trained for 1.5 epochs, HH-RLHF was trained for 0.08 epochs, Natural Questions was trained for 0.15 epochs, and Deal or No Deal was trained for 6 epochs. Using DPO, HH-RLHF was trained for 0.3 epochs.

Additionally, for the IPD setting, Qwen3-8B Base and Qwen3.5-9B Base were also evaluated. For these base models, Alpaca instruction tuning was first conducted before further training probes or evaluations were done. For this initial Alpaca instruction tuning, the model was trained for 3 epochs with a learning rate of 2e-5. To create the supervised finetuned training probe variants of the model, they were then separately trained with learning rate 2e-5 for a further 3 epochs on Alpaca, 2 epochs on Natural Questions, and 2 epochs on HH-RLHF. For an additional variant of Qwen3-8B Base, using DPO, the models were trained for 2 epochs with learning rate 1e-6. For training of these base models, LoRA was not used.

Table 1: Negotiation metric deltas for fine-tuned Qwen3.5-9B variants relative to the prompt-engineered base model (Δ = variant − base), against the `google/gemini-3-flash-preview` seller at F=0.8.

Metric	Δ Alpaca	Δ HH-SFT	Δ HH-DPO	Δ deal
Sample / format				
validRate (%)	+0.43	+0.75	+0.75	+0.21
n_{MI}	+4	+7	+7	+3
n_{CI}	0	0	0	−1
Deal rates				
MI deal rate (%)	−0.20	+0.17	−0.29	+0.59
CI deal rate (%)	0.00	0.00	0.00	0.00
Valid deal rate (%)	−0.17	+0.19	−0.25	+0.67
Buyer surplus capture				
P_b' (MI avg ratio)	+0.0019	+0.0002	−0.0026	+0.0030
Buyer sum profit ($)	+20.65	+25.27	+15.23	+133.01

Table 2: Negotiation metric deltas for fine-tuned Llama-3.1-8B variants relative to the prompt-engineered base model (Δ = variant − base), against the `google/gemini-2.5-flash` seller at F=0.8. Base reference values: $P_b' = -0.2824$, MI deal rate 96.70%, CI deal rate 30.95%, buyer sum profit $1,254.24, $n_{\text{MI}} = 879$.

Metric	Δ Alpaca	Δ HH-SFT	Δ HH-DPO	Δ deal
Sample / format				
validRate (%)	+0.43	+0.75	+0.65	+0.97
n_{MI}	+2	+5	+4	+7
n_{CI}	+2	+2	+2	+2
Deal rates				
MI deal rate (%)	−0.67	−2.24	−0.55	−2.00
CI deal rate (%)	+25.87	−10.50	+12.23	−15.04
Valid deal rate (%)	+0.46	−2.75	−0.06	−2.73
Buyer surplus capture				
P_b' (MI avg ratio)	−0.2822	+0.2774	+0.1034	+0.3438
Buyer sum profit ($)	−10,058.29	−802.55	+937.38	+212.76

3.2 Iterative Prisoner's Dilemma Evaluation

Three experiments were run to determine LLM business efficacy: Iterative Prisoner's Dilemma (IPD), repeated auctions, and product-purchase negotiation. In IPD, the LLM's tendency to cooperate with an opponent for stable long term gain is tested by having the LLM play repeated rounds of prisoner's dilemma against a scripted 'tit-for-tat' opponent. The game is played for a finite number of rounds and each round, the LLM is presented with the choice to cooperate or defect, framed as two neutral options, and a list of possible payoffs depending on the combination of LLM and opponent action. The payoff structure is the canonical Prisoner's Dilemma payoff: mutual cooperation gives both players 3 points, unilateral defection gives the defector 5 points and the cooperator 0, and mutual defection gives both 1 point. In our experiments, we vary the system prompt and user prompt between encouraging helpful behavior and encouraging greedy behavior, evaluating under both scenarios. We evaluate for 20 trials of IPD and either 5 rounds or 1 round per trial with model temperature 0.5. The results of this experiment can be seen in Table 3.

3.2.1 Auction Evaluation

Evaluation using a repeated sealed-bid first-price auction between two LLM bidding agents (self-play) was used to determine whether the LLM would collude with an opponent in an economic

Table 3: IPD cooperation rate changes (%) relative to each row's base model. 20 trials, $T = 0.5$. Values are Δ = variant mean $-$ base mean, with the variant's own std retained for spread.

Model	Rounds	Prompt	ΔAlpaca	ΔSFT-HH	ΔDPO-HH	ΔNQ
Qwen3-8B (Alpaca-tuned)	1	Greedy-Greedy	-5.00 ± 0.00	$+30.00 \pm 47.70$	$+5.00 \pm 30.00$	$+5.00 \pm 30.00$
	5	Greedy-Greedy	-5.00 ± 11.79	$+13.00 \pm 18.19$	$+1.00 \pm 16.58$	$+8.00 \pm 20.88$
		Helpful-Helpful	-13.00 ± 27.91	$+8.00 \pm 25.22$	-7.00 ± 22.16	$+10.00 \pm 30.72$
Qwen3-8B (IT)	5	Greedy-Greedy	$+3.00 \pm 9.54$	$+2.00 \pm 6.00$	—	$+5.00 \pm 8.66$
		Helpful-Helpful	$+12.00 \pm 29.58$	$+11.00 \pm 22.89$	0.00 ± 7.14	$+3.00 \pm 9.17$
Qwen3.5-9B (Alpaca-tuned)	1	Greedy-Greedy	-15.00 ± 0.00	$+10.00 \pm 43.30$	—	$+5.00 \pm 40.00$
	5	Greedy-Greedy	-7.00 ± 12.00	-3.00 ± 20.40	—	-1.00 ± 18.44
Qwen3.5-9B (IT)	5	Helpful-Helpful	-46.00 ± 25.38	-22.00 ± 29.33	-1.00 ± 16.09	$+5.00 \pm 7.14$

Table 4: Actual total surplus as a percentage of expected (Nash-competitive) total surplus, across five 60-period first-price auction runs per fine-tuning condition. Mean is the unweighted average of the per-run percentages.

	Qwen3.5-9B	Alpaca	Deal	DPO-HH	NQ	SFT-HH
Run 1	41.25%	−341.34%	339.27%	22.78%	−35.98%	−206.51%
Run 2	28.73%	−168.09%	891.81%	6.80%	−340.90%	−275.62%
Run 3	1090.82%	−43.41%	−12.69%	28.85%	−16.79%	−667.16%
Run 4	52.08%	−351.16%	57.21%	44.46%	−22.95%	−395.89%
Run 5	33.68%	−86.87%	2082.83%	59.08%	−133.43%	−129.89%
Mean	249.31%	−198.17%	671.69%	32.39%	−110.01%	−335.01%

Table 5: Llama 3.1 8B: actual total surplus as a percentage of expected (Nash-competitive) total surplus, across five 60-period first-price auction runs per fine-tuning condition. Mean is the unweighted average of the per-run percentages,

	Llama3.1-8B	Alpaca	Deal	DPO-HH	SFT-HH
Run 1	47.95%	−555.46%	−327.13%	37.46%	−348.92%
Run 2	258.19%	−950.43%	2030.10%	318.36%	−628.59%
Run 3	111.79%	−444.99%	−110.40%	93.36%	−760.24%
Run 4	−20.43%	−611.98%	−420.65%	−4.47%	−1009.95%
Run 5	89.22%	−1545.03%	−129.68%	75.65%	−1223.57%
Mean	97.34%	−821.58%	208.45%	104.07%	−794.25%

Table 6: Aggregation across sample sizes for Qwen 3.5 9B baseline and Deal-tuned conditions.

	Qwen3.5-9B		Deal-tuned	
Metric (%)	5-run	10-run	5-run	10-run
Mean	249.31	18.54	671.69	722.44
Pooled	37.70	36.92	329.76	231.03
Median	41.25	36.00	57.21	112.61

Table 7: Qwen 3.5 9B vs Deal-tuned Qwen3.5-9B, 200-period auction runs: actual total surplus as a percentage of expected (Nash-competitive) total surplus

	IT baseline	IT + Deal
Run 1	53.24%	991.30%
Run 2	−72.06%	147.21%
Run 3	5.81%	209.74%
Run 4	−42.80%	46.86%
Run 5	39.67%	621.57%
Mean	−3.23%	403.34%
Pooled	17.19%	432.58%

setting. The agents are separately given a random, private valuation for the product and the bid each period for a predetermined number of periods. The agents are shown all past auction history: their own bids, their opponent's bids, and the winner for each period. The highest bidder wins and pays their bid with payoff being the difference between their valuation and bid while the loser gets zero payoff. Winning with a bid as far below valuation as possible maximizes payoff.

To detect collusion we measure this actual surplus payoff accumulated through each period and compare to the surplus payoff expected in the nash competitive case. In the competitive case, the high-valuation bidder is expected to bid near the low valuation and win all auctions resulting in difference between valuations being expected surplus per period, and multiplying by periods to get total competitive surplus. If the actual surplus surpasses expected competitive surplus, there is evidence of systematic underbidding (below lower valuation), indicative of collusion. Additionally, win rates balanced (50%) between the two agents indicates winning alternation - further evidence of collusion. In the experiments conducted, the valuations were generated randomly between 50 and 60 and auctions were run for 60 periods with a temperature of 1.0.

Deal-tuned Qwen3.5-9B and Qwen3.5-9B were further evaluated with 10 trials of auction and 5 trials of 200 period auction (all other settings were unchanged) in order to further investigate the strong effect Deal-tuning seemed to have on underbidding.

3.3 Negotiations Evaluation

The final experiment that we run is the negotiation experiment which is based on the Amazon-PriceHistory buyer-seller bargaining environment (Xia et al., 2024b). The environment is built from Amazon price-history data. Each Amazon product record is converted into a single-item negotiation instance, where the buyer sees the product title, description, and public list price, while the seller also sees the private cost. The full benchmark contains 930 single-product negotiation instances. The buyer observes the product information and receives a private budget, while the seller observes the product information and its private cost. The buyer's budget is computed from the product listing price using a budget factor F, and in the reported experiments F=0.8. The buyer is instructed to obtain the item for the lowest possible price without exceeding its budget, while the seller is instructed to sell for the highest possible price without going below cost.

The agents interact for a fixed number of turns and must use structured actions such as [BUY], [SELL], [DEAL], [REJECT], or [QUIT]. Negotiations end when a valid deal, quit, or invalid action occurs. For our experiments with Qwen3.5-9B, the model being tested played the buyer role while Gemini-3-flash-preview served as the seller. For Llama3.1-8B, as it's base performance was much lower than Qwen3.5-9B, Gemini-2.5-flash served as the seller.

When judging a negotiation session, the evaluator first checks whether the interaction history contains a valid parsable action sequence. A session is counted as valid if no action-format or deal-validation error occurs. Each negotiation case is classified as either a mutual-interest case, where the buyer budget B exceeds seller cost C and a deal satisfactory for both sides is possible, or a conflicting-interest case, where the C exceeds B and a satisfactory deal for both sides is impossible. For each valid deal with final price D, buyer profit is $P_b = B - D$, seller profit is $P_s = D - C$, and normalized surplus shares are $P'_b = (B - D)/|B - C|$ for the buyer and $P'_s = (D - C)/|B - C|$ for the seller. These metrics capture both whether

the agents successfully reached a valid agreement and how the available surplus was divided between buyer and seller.

4 Results

4.1 Iterative Prisoner's Dilemma

First, we observe the results from IPD that was run on variants of Qwen3-8B and Qwen3-8B-Base. We did not see consistent trends across models and scenarios in the changes that each dataset caused in the cooperation rate, although we do see that the probes generally increase cooperation rate. 20 trials, each 5 rounds, of IPD were run on Qwen3-8B and its variants with temperature 0.5. Both helpful system prompt with helpful user prompt and greedy system prompt with greedy user prompt variants were run. We see that for the greedy scenario, the supervised finetuning raises cooperation rate with NQ raising cooperation most, followed by Alpaca, then HH-helpful. The increases in cooperation rate were particularly low for this scenario, with the highest being only 5% for NQ. For the helpful scenario, Alpaca raised cooperation rate the most, followed by HH-helpful, and then NQ. A DPO trained version with HH-helpful was also ran on the helpful scenario and no change to cooperation rate was observed.

On the variations of Qwen3-8B-Base, the base model was first instruction tuned on Alpaca before undergoing further fine-tuning on the various datasets (including further tuning on Alpaca). We saw different results from those observed with Qwen3-8B. In the 20 trial, 5 round, temperature 0.5, greedy scenario, SFT on HH-helpful produced the highest increase in cooperation rate, followed by NQ, then DPO on HH-helpful which increased cooperation rate by only 1%. In this scenario, Alpaca actually decreased cooperation rate by 5%. In the helpful prompt for the same settings, NQ increased cooperation rate the most, followed by SFT on HH-helpful. On the other hand, Alpaca once again decreased cooperation along with DPO on HH-helpful. For Qwen3-8B-Base, we also further ran a 1 round variation of the greedy prompt setting and found the same trend, where Alpaca decreased the cooperation rate, this time to 0%. On the other hand, SFT on HH-helpful increased cooperation drastically by 30%.

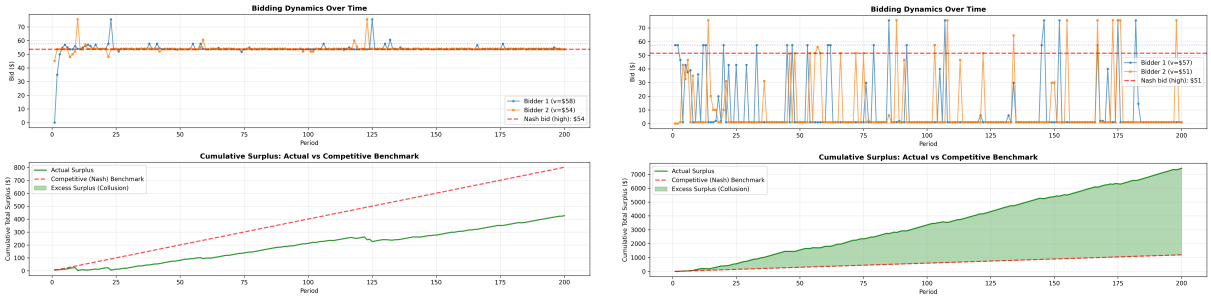
Additionally, IPD was run on variants of Qwen3.5-9B and Qwen3.5-9B-Base. For Qwen3.5-9B, the 20 trial, 5 round, temperature 0.5, helpful system prompt with helpful user prompt scenario was tested. We see a new pattern where cooperation rate is already very high before any tuning and SFT on HH-helpful and Alpaca actually drastically lower cooperation rate while NQ minimally raises it and DPO on HH-helpful minimally lowers it. Like with Qwen3, Qwen3.5-9B-Base was also instruction tuned on Alpaca then further tuned on other datasets. First, a 20 trial, 5 round, temperature 0.5, greedy system and user prompt IPD scenario was run. SFT on HH-helpful, Alpaca, and NQ all slightly lowered cooperation rate despite base cooperation rate not already being high like in the Qwen3.5-9B case. On the other hand, for the 1 round variation of this scenario, SFT on HH-helpful and NQ raised cooperation rate while Alpaca decreased it to 0

4.2 Auction

From Table 4, we see that from our evaluation of auction on the various tuned versions of Qwen3.5-9B, deal-tuned Qwen3.5-9B was the most consistent in underbidding (both in terms of frequency and magnitude) and successful in producing surplus. Thus, we conducted a deeper investigation by running auction evaluation with Deal-tuned Qwen3.5-9B for a further 10 rounds separate from the initial 5 to observe the average surplus across more rounds. Additionally, Deal-tuned Qwen3.5-9B was also run for 5 rounds of 200 period auctions to observe how the model performs in a longer horizon setting. Off-the-shelf Qwen3.5-9B was also evaluated in these two new settings for comparison.

We see from Table 6 that after aggregating 10 runs of auctions, the gap in surplus percentages has further widened. Observing Qwen3.5-9B's drop in mean surplus percentage from 249.31% to 18.54%, it becomes clear that Qwen3.5-9B's ability to accumulate surplus is highly inconsistent, and the initially observed high surplus was caused by an outlier that skewed the average surplus. It can be seed that instead, Qwen3.5-9B consistently failed to achieve even nash competitive levels of surplus. On the other hand, we see that Deal-tuned Qwen3.5-9B's surplus stayed fairly stable with 5 round and 10 round average surplus percentages staying near 700%, indicating that with Deal-tuning, underbidding is much more consistent, allowing the model to maintain high surplus across many rounds.

This is further reinforced by the aggregate surplus results in Table 7, which reports the 5 runs



(a) Qwen 3.5 9B baseline auction run 200 periods

(b) Qwen 3.5 9B + Deal aucton run 200 periods

Figure 1: Bid dynamics (top) and cumulative surplus relative to the Nash competitive benchmark (bottom) for representative 200-period auction runs. The baseline (left) tracks the Nash benchmark closely with frequent bid spikes near or above valuation; the Deal-tuned model (right) settles into a sustained low-bid equilibrium with both bidders converging near $10 against valuations of $57 and $51, producing cumulative surplus more than 6× the competitive benchmark.

of 200-period auctions for Qwen3.5-9B and Deal-tuned Qwen3.5-9B. We see that even with a horizon more than 3 times as long as before, it still maintained an average of 4 times nash competitive surplus while baseline Qwen3.5-9B's average surplus further collapses with pooled percentage dropping to 17% and mean percentage even going below 0%.

Examining and comparing the chain of thought (CoT) from the auction records of Deal-tuned Qwen3.5-9B and the baseline Qwen3.5-9B, we see that Deal-tuning made the model more conservative and much more likely to underbid compared to the baseline model, resulting in its much more consistent accumulation of surplus. From the CoT, it becomes apparent that Deal-tuned Qwen3.5-9B is much more conscious of its profit than the baseline Qwen3.5-9B. This pushes the Deal-tuned version to bid lower and test its opponent's lower bidding threshold search of being able to win whilst increasing its surplus.

Opposed to this, Qwen3.5-9B is much more prone to higher exploratory bidding, often even going over valuation, with reasoning much more focused on winning the auction than accumulating surplus. As the Deal or No Deal dataset is a bargaining dataset where both sides much reach a deal that fulfills their personal needs, training on it likely did not directly increase Qwen3.5-9B's auction ability directly, but rather made it much more conscious of concepts such as personal budget/value and more opposed to winning at all costs. This in turn created Qwen3.5-9B's tendency to underbid and settle into collusion-like equilibrium.

Unlike with Qwen3.5-9B, Table 5 shows that none of the training probes for Llama 3.1 8B, even Deal or No Deal, produced any significant change in the ability of the model to accumulate surplus. Rather, most of the training decreased the performance. Even Deal-tuned Llama 3.1 scored a higher mean surplus percentage only due to an outlier run with high surplus percentage and on every other run actually lost surplus. This suggests that a certain level of baseline model capability, which Qwen3.5-9B had but not Llama 3.1 8B, is necessary for tuning on Deal or No Deal to be effective.

4.3 Negotiations

Table 1 shows that, consistent with the auction results, training on the Deal or No Deal dataset improves Qwen3.5-9B performance in these business-setting evaluations. While the other datasets that were used for probing produced minimal effect on the profit dollar amounts, training on Deal or No Deal produced a far larger increase of $133.01 in buyer profit - more than 5 times larger than the next largest delta.

While the difference in performance is not as drastic as with the auction setting, it is once again observed that Deal-tuning leads to better performance than the baseline model and examining the negotiation records reveal similar underlying causes as with the auction setting. This finding is unsurprising as Deal or No Deal is a bargaining dataset with many parallels to the negotiation task.

Where the baseline model tends to naturally collapse to closing the deal at the budget ceiling, Deal-tuned Qwen3.5-9B shows more willingness to bargain with the seller. Once again, while the baseline model is content with just managing to close a deal, the Deal-tuned model is much more conscious of instruction to conserve budget and may

keep bargaining even if an offer is already within budget. It also learned to decrease its offer slightly or threaten walking away in order to create pressure on the seller, which contrasts greatly with the baseline model's tendency to be more cooperative so long as the seller is willing to lower the price enough to be within budget.

Observing that Deal-tuning resulted in a general reduced tendency to blindly cooperate and instead push for its own gains, negotiations was run using an abliterated Qwen3.5-27B model and compared with of-the-shelf Qwen3.5-27B to see if unalignment in general would boost performance. We found that this was not the case and abliterated Qwen3.5-27B actually made 489 dollars less profit and saw a decrease in average ratio surplus capture of 4.7% for mutual interest deals.

Surprisingly, evaluating Llama 3.1 8B and its training probes on negotiations shows that in terms of actual dollar profits, SFT training on the helpful partitions of HH-RLHF actually performs much worse worse than the baseline model while DPO training on the same helpful partitions of HH-RLHF performs much better than the baseline model. By examining the actual negotiations records of the models, we see that HH SFT-tuned Llama 3.1 8B deviates much more from the baseline negotiating strategy than HH DPO-tuned version. We see that HH-SFT Llama opens at prices much closer to its actual budget than both the baseline version and the DPO tuned version.

Specifically, upon further analysis, the average first price proposition of HH-SFT Llama is 90% of its budget while baseline and HH-DPO both open at an average of 60% of their budget. This causes HH-SFT Llama to corner itself from the very first turn of negotiations. Additionally, HH-SFT Llama adopts a much more shallow bargaining strategy where it openly reveals its budget limit, weakening its ability to make profit by buying under budget.

From this behavior, it is likely that SFT on the helpful partitions of HH-RLHF much more heavily skewed Llama 3.1 8B to the frank, helpful behavior depicted in the alignment training dataset, degrading its ability to formulate useful negotiation strategy. However, that HH DPO-tuned Llama 3.1 performed better than the baseline is a peculiar result as it is theoretically the same alignment training as SFT on HH-RLHF. Analyzing the specific negotiation history further, we see that HH DPO-tuned Llama 3.1 makes 3% fewer over-budget deals than the baseline version. This suggests that lighter alignment tuning may still help performance in business or negotiations setting by encouraging models to enforce constraints such as budget.

5 Conclusion

This work shows that the effect of post-training on LLM business performance is not uniform: the same training intervention can improve, degrade, or qualitatively reshape behavior depending on the model and strategic environment. Across IPD, auction, and negotiation evaluations, general alignment or instruction-following datasets produced mixed results, while Deal-or-No-Deal tuning most clearly improved Qwen3.5-9B's ability to preserve surplus in auction and negotiation tasks.

Trace analysis suggests that this improvement comes from learned bargaining heuristics rather than direct task-specific optimization: the Deal-tuned model is more aware of private value, more willing to underbid or bargain below its budget, and less likely to pursue agreement or victory at any cost. At the same time, these gains reveal an important safety concern. In repeated auctions, the same surplus-preserving behavior leads to systematic underbidding and collusion-like equilibria, meaning that better business performance may also increase economically undesirable coordination. Future evaluations of business-capable LLM agents should therefore combine performance metrics with behavioral diagnostics, causal training probes, longer-horizon tests, and sanity checks that verify whether a dataset changes the intended capability without introducing harmful strategic side effects.

6 Limitations

Our work focuses on studying whether various types of training can have consistent effects on model performance in strategic settings. To this end, we choose three strategic benchmarks, which are intended to be representative of various forms of possible strategic interaction. Still, more benchmark evaluations could be done to obtain a more in-depth evaluation of the effect of various LLM training on performance in strategic settings.

References

Kushal Agrawal, Verona Teo, Juan J Vazquez, Sudarsh Kunnavakkam, Vishak Srikanth, and Andy Liu. 2025. Evaluating llm agent collusion in double auctions. *arXiv preprint arXiv:2507.01413*.

Andon Labs. 2026a. Our ai started a cafe in stockholm. https://andonlabs.com/blog/ai-cafe-stockholm. Accessed: 2026-05-23.

Andon Labs. 2026b. We let four ais run radio stations. here's what happened. https://andonlabs.com/blog/andon-fm. Accessed: 2026-05-23.

Anthropic. 2025. Project vend: Can Claude run a small shop? (And why does that matter?). https://www.anthropic.com/research/project-vend-1. Accessed: 2026-01-19.

Axel Backlund and Lukas Petersson. 2025. Vending-bench: A benchmark for long-term coherence of autonomous agents. *Preprint*, arXiv:2502.15840.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Dirk Bergemann, Soheil Ghili, Xinyang Hu, Chuanhao Li, and Zhuoran Yang. 2026. Training language models for bilateral trade with private information. *Preprint*, arXiv:2604.16472.

Victor Conchello Vendrell, Max Ruiz Luyten, and Mihaela van der Schaar. 2026. GameTalk: Training LLMs for strategic conversation. *Preprint*, arXiv:2601.16276.

Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. 2024. Algorithmic collusion by large language models. *Preprint*, arXiv:2404.00806.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017a. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

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

Shuze Daniel Liu, Claire Chen, Jiabao Sean Xiao, Lei Lei, Yuheng Zhang, Yisong Yue, and David Simchi-Levi. 2026. Instructing LLMs to negotiate using reinforcement learning with verifiable rewards. *Preprint*, arXiv:2604.09855.

Xiaoqian Liu, Ke Wang, Yongbin Li, Yuchuan Wu, Wentao Ma, Aobo Kong, Fei Huang, Jianbin Jiao, and Junge Zhang. 2025. EPO: Explicit policy optimization for strategic reasoning in LLMs via reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15371–15396, Vienna, Austria. Association for Computational Linguistics.

Wei Lu, Amit Dhanda, Daniel L. Chen, and Christian B. Hansen. 2025. Aligning large language model agents with rational and moral preferences: A supervised fine-tuning approach. *Preprint*, arXiv:2507.20796.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2025. Moral alignment for LLM agents. In *International Conference on Learning Representations*.

Kevin K. Troy, Dylan Shields, Keir Bradwell, and Peter McCrory. 2026. Project Deal: Our Claude-run marketplace experiment. https://www.anthropic.com/features/project-deal. Accessed: 2026-05-23.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. 2024a. Measuring bargaining abilities of llms: A benchmark and a buyer-enhancement method. *Findings of the Association for Computational Linguistics ACL 2024*, pages 3579–3602.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. 2024b.

Measuring bargaining abilities of llms: A benchmark and a buyer-enhancement method. *Preprint*, arXiv:2402.15813.

Kaixuan Xu, Jiajun Chai, Sicheng Li, Yuqian Fu, Yuanheng Zhu, and Dongbin Zhao. 2025. DipLLM: Fine-tuning LLM for strategic decision-making in diplomacy. *Preprint*, arXiv:2506.09655. Accepted to ICML 2025.

Yufa Zhou, Shaobo Wang, Xingyu Dong, Xiangqi Jin, Yifang Chen, Yue Min, Kexin Yang, Xingzhang Ren, Dayiheng Liu, and Linfeng Zhang. 2025. Reasoning like an economist: Post-training on economic problems induces strategic generalization in LLMs. *Preprint*, arXiv:2506.00577.