

Experimental Settings and Additional Simulation Results

EC.1. Experimental Settings

This section provides the experimental settings for the nine experiments in Table 1, including game introduction, hypotheses and simulated scales of experiments.

EC.1.1. Chen et al. (2013)

EC.1.1.1. Core Hypothesis Chen et al. (2013) investigate the role of payment scheme in the newsvendor setting. They define two types of payment schemes (Table EC.1). Under the O payment scheme, participants pay for the items at the point they are ordered and then get a per-unit payment if they are sold. Under the C payment scheme, participants get the margin at the point items are ordered and then are refunded the per unit price for items unsold. In summary, the two payment schemes are: O-payment scheme ($-c$ per unit ordered, $+p$ per unit sold) and C payment scheme ($+(p - c)$ per unit ordered, $-p$ per unit leftover). The two are mathematically equivalent, but differ in the timing of the payments.

Table EC.1 Payment schemes in Chen et al. (2013)

Payment scheme	Order-time payments	Demand-time payments	
	Per unit ordered	Per unit sold	Per unit leftover
O	$-c$	$+p$	0
C	$+(p - c)$	0	$-p$

In this study, we replicate the O and C treatments. The hypothesis is whether subjects order higher quantities under O payment scheme compared to the C payment scheme.

EC.1.1.2. Task Structure Consistent with the replication study by Davis et al. (2023), we set $p = \$2$, $c = \$1$ and use an LLM to simulate 50 subjects (25 per condition). Each simulated subject acts as a business manager making order decisions across 25 rounds.

The experiment is framed as a multi-round Newsvendor problem repeated for $T = 25$ periods. In each period t , the LLM agent received prompt includes the current balance B_t (with initial endowment $B_1 = 100$) and historical outcomes (decision and balance in the past rounds), and output the order quantity q_t . The demand D_t is drawn from a discrete distribution generated by the sum of three independent six-sided dice (i.e., $D_t \in \{3, 4, \dots, 18\}$ with an expected value of $\mu = 10.5$).

Mathematically, given the order quantity q_t and the realized demand D_t , the single-period profit is formulated as $\pi_t = 2 \min(q_t, D_t) - q_t$. Unsold inventories are discarded at the end of each period, and the agent's objective is to maximize the final wealth B_{26} .

EC.1.1.3. Data Analysis We conducted Mann-Whitney test to validate whether data under two conditions show significant difference. p -value from the data generated by all the LLMs, except for GPT-4o mini, is much lower than 0.01, whereas GPT-4o mini has a p -value of 0.806.

In addition, we analyze the average order quantity under aforementioned two conditions. Aside from the hypothesis, Chen et al. (2013) also demonstrated that order quantity under scheme O is higher than optimal value derived by expected profit maximization model (10.5), while order quantity under scheme C is lower than 10.5. Our generated data showed that average order quantity under scheme O by different LLMs are LLaMA-70B (9.95), Qwen-72B (9.98), Qwen-32B (9.97), GPT-3.5 (10.78), GPT-4 (11.448). Average order quantity under scheme C are LLaMA (9.50), Qwen-72B (9.06), Qwen-32B (7.49), GPT 3.5 (10.25), GPT-4 (9.87). Thus, GPT-4 showcases more similar average order quantity compared with original data.

EC.1.2. Croson and Donohue (2006)

EC.1.2.1. Core Hypothesis Croson and Donohue (2006) study the bullwhip effect in a multi-tier supply chain consisting of a factory, distributor, wholesaler and retailer. In two studies, they vary whether inventory information is or is not shared and show that inventory information sharing helps to alleviate (but not eliminate) the bullwhip effect.

Our core hypothesis is sharing dynamic inventory information across the supply chain will decrease the level of order oscillation. Specifically, we will conduct the same two treatments in their original study: Beer Game with and without information sharing.

EC.1.2.2. Task Structure We simulate LLMs as 11 groups of managers to conduct the beer distribution game for each treatment. Each group consists of 4 roles, Retailer, Wholesaler, Distributor, and Factory, operating sequentially from downstream to upstream.

The supply chain operates over $T = 48$ weeks. Each member of the chain is responsible for maintaining beer inventory and fulfilling orders from their immediate downstream partner. Crucially, the system features significant delays: a 2-week order transmission delay and a 2-week shipment delay (with the factory subject to a 3-week overall production lead time). We initialize the system in a steady state, where each role starts with 12 kegs of beer in on-hand inventory, alongside orders and shipments of 4 kegs in each pipeline position.

In each period t , the LLM agent managing echelon i observes the newly arrived shipment $S_{i,t}^{in}$ and the incoming order $D_{i,t}$ from its downstream partner. The agent fulfills the order from its available inventory. Unfulfilled orders are not lost but kept as backlogs. The agent then makes a single decision: the order quantity $O_{i,t}$ to be sent to its upstream supplier (or the production quantity, if factory). The ending inventory (or backlog) state updates recursively as $I_{i,t} = I_{i,t-1} + S_{i,t}^{in} - S_{i,t}^{out}$.

Under “Base” treatment, agents operate under decentralized information; they can access their own local history (past orders, shipments, and inventory levels) but cannot observe the states of other echelons or the true consumer demand (except the Retailer); Under “Share” treatment, except for their own information, each

agent can observe the *inventory position* of all team members. The inventory position is defined as the sum of the on-hand inventory and the pipeline inventory (i.e., beer that has been ordered but not yet received). Consumer demand is simulated by the computer and drawn uniformly from an integer set $\{0, 1, \dots, 8\}$.

The objective of each agent is to minimize the cumulative costs incurred by the team over the 48 weeks. The single-period cost $C_{i,t}$ consists of a \$0.50 per-unit holding cost for positive inventory ($I_{i,t} > 0$) and a \$1.00 per-unit backlog penalty for negative inventory ($I_{i,t} < 0$).

EC.1.2.3. Data Analysis Focusing on the hypothesis first, We conduct one-sided Mann-Whitney test towards variance of order quantity between base condition (benchmark) and share condition (inventory information is revealed to four different roles every time they make decisions). Amongst all the LLMs, LLaMA-70B ($U = 1208.5$, $p\text{-value} = 0.0042$), GPT-3.5 ($U=1372$, $p\text{-value} = 6.97\text{e-}05$), GPT-4o mini ($U=1689.5$, $p\text{-value} = 9.24\text{e-}11$) showed that variance under base condition are significantly greater than variance under share condition, while GPT-4 ($U = 7.48$, $p\text{-value} = 0.96$), Qwen-32B ($U = 479$, $p\text{-value} = 0.99$) and Qwen-72B ($U=1099$, $p\text{-value} = 0.13$) could not conclude that variances of two conditions are different.

EC.1.3. Davis et al. (2011)

EC.1.3.1. Core Hypothesis Davis et al. (2011) investigate how auctioneers set reserve prices in auctions. Subjects play the role of a seller in a second-price auction with n potential buyers. The authors vary the number of bidders as well as the distribution from which the buyers' private values are drawn.

The core hypothesis is that the seller's chosen reserve price is increasing in the number of bidders, specifically when the number of bidders is 1, 4, 7 and 10. The Drop-Out price for bidders are sampled from cumulative distribution given by $F(v) = (v/100)^{1/3}$, which we refer to as the cube-root distribution.

EC.1.3.2. Task Structure We simulate an auction environment using LLMs to represent 40 independent subjects acting as sellers. Each simulated seller participates in a sequence of 60 independent auctions. In each round, the number of prospective buyers is randomly and uniformly drawn from the set $\{1, 4, 7, 10\}$.

The auction mechanism is a second-price sealed-bid auction with a seller-defined reserve price. The buyer with the highest drop-out price (i.e., valuation) wins the auction, provided their valuation meets or exceeds the seller's reserve price. The seller's profit in a given round is determined as follows:

- **Profit = Second-Highest Drop-Out Price**, if the second-highest drop-out price is greater than or equal to the reserve price.
- **Profit = Reserve Price**, if the highest drop-out price exceeds the reserve price, but the second-highest drop-out price falls below it.
- **Profit = 0**, if all drop-out prices are strictly below the reserve price (i.e., the item remains unsold).

At the end of each period, the LLM agent receives historical feedback, including their chosen reserve price, the realized number of buyers, the winning bid (if applicable), and the profit earned in past rounds.

Although determining the optimal reserve price requires complex calculations involving order statistics and probability distributions, we do not provide the LLM agents with any auxiliary statistical information. Specifically, agents are not given the explicit probabilities of the item remaining unsold, selling exactly at the reserve price, or selling above the reserve price.

EC.1.3.3. Data Analysis We conduct regression analysis and pairwise t-test to verify hypothesis that the seller's chosen reservation price is increasing in the number of bidders. The dependent variable is the observed reserve price in a period, and the independent variables are (a) the number of bidders minus the average number of bidders and (b) the number of decision period minus the average number of periods, where we could interpret constant as average reservation price and can be compared to risk-neutral theoretical optimal reserve levels (42 for Cube-root). Among all the LLMs, coefficients on $n - \bar{n}$ conducted by LLaMA-70B, Qwen-72B, Qwen-32B, GPT-3.5, GPT-4 and GPT-4o mini are 1.888, 0.935, 1.026, 0.57, 0.599, 0.479, indicating the amount of price LLM simulated sellers increased for each additional bidder. Constants are respectively 19.049, 24.253, 17.736, 27.379, 24.725, 26.695, which are all significantly lower than optimal level: 42. However, by conducting t-test, GPT 3.5 shows no significant different between reservation price under $n = 7$ (28.85) and $n = 10$ (29.11); GPT-4o mini shows no significant different among $n = 4$ (27.33), $n = 7$ (27.49) and $n = 10$ (27.71).

EC.1.4. Engelbrecht-Wiggans and Katok (2008)

EC.1.4.1. Core Hypothesis Engelbrecht-Wiggans and Katok (2008) investigate the role of regret and feedback in bidding behavior in first-price sealed-bid auctions. They define two types of feedback. Under “Loser’s Regret” subjects receive feedback on the winning price and how large their missed opportunity to win was (their resale value – winning bid, or zero if they won). Under “Winner’s regret” subjects receive feedback on the second highest bid and how much money was left on the table (their bid – second highest bid, or zero if they did not win).

We replicate 2 treatments: “Loser’s Regret” vs. “Both” (Both loser’s and winner’s regret feedback). The core hypothesis is that “Both” will lead to lower average bids than “Loser’s Regret.”

EC.1.4.2. Task Structure We simulate an auction environment where LLMs act as 20 independent buyers (bidders) competing against two computerized opponents to purchase a fictitious asset. The auction mechanism is a first-price sealed-bid auction: the bidder who submits the highest bid wins the asset and pays exactly their submitted bid amount.

The experiment consists of a total of 100 bidding decisions. For each decision, the agent’s bid is applied to 10 consecutive, independent auctions. While the agent’s bid and resale values remain constant across these 10 auctions, the computerized competitors draw new resale values and generate new bids for each individual auction.

The agent's resale value for the asset is assigned from the set $\{50, 60, 70, 80, 90\}$. The agent holds a constant valuation for 20 consecutive bidding decisions (i.e., 200 individual auctions) before it changes to a new value from the set, presented in a randomized order. The two computerized competitors' valuations are drawn independently from a discrete uniform distribution of integers between 1 and 100. The competitors are pre-programmed to bid in a manner that maximizes their expected earnings against similarly programmed opponents. (2/3 of their resale values, but the LLM agent would not know.)

The agent's profit in each individual auction is determined strictly by their bid (b) and their assigned resale value (v):

- **Profit** = $v - b$, if the agent's bid is the highest among all three bidders (i.e., the agent wins the auction).
- **Profit** = 0, if the agent does not place the highest bid.

Bidding above one's resale value ($b > v$) results in a negative profit (loss) if the auction is won.

At the end of each block of 10 auctions, the agent receives detailed historical feedback for that block. This includes their own bid, the winning price and second-highest bid for each of the 10 auctions, and their realized profit. Additionally, under the "Both regret" treatment, the agent receives two calculated metrics to evaluate their bidding efficiency:

1. *Money left on the table*: Calculated as the agent's bid minus the second-highest bid (if the agent won), representing how much less they could have bid while still winning. It is 0 if the agent lost.
2. *Missed opportunity to win*: Calculated as the agent's resale value minus the winning bid (if the agent lost, but the winning bid was lower than the agent's resale value). It is 0 if the agent won or if the winning bid exceeded their resale value.

while under the "Loser's regret", the LLM agent could only get the information of *Missed opportunity to win*.

EC.1.4.3. Data Analysis We compute bid/value for each subject and conduct one-sided t-test. LLaMA-70B ($t = 8.07$, $p\text{-value} = 4.45\text{e-}10$), Qwen-72B ($t = 30.86$, $p\text{-value} = 8.15\text{e-}29$), Qwen-32B ($t = 4$, $p\text{-value} = 0.00013$) are consistent with the conclusion; while GPT-3.5 ($p\text{-value} = 0.21$), GPT-4o mini ($p\text{-value} = 0.92$) are not consistent with the original conclusion.

EC.1.5. Ho and Zhang (2008)

EC.1.5.1. Core Hypothesis Ho and Zhang (2008) investigate a supply chain contracting problem in which they test two-part tariffs as a way to increase supply chain efficiency. They vary the framing of the two-part tariff as either a fixed fee (TPT) or a quantity discount (QD). They find that supply chain efficiency is higher under the quantity discount framing. The core hypothesis we replicate is that supply chain efficiency is higher when a two-part tariff is framed as a quantity discount as opposed to a fixed fee.

EC.1.5.2. Task Structure We use an LLM to simulate 250 subjects per condition, each participating in 11 consecutive rounds. Half of the simulated subjects are assigned the role of Player A (the manufacturer), while the other half act as Player B (the retailer).

- **In the TPT condition**, Player A determines both a per-unit wholesale price (PRICE A) and a lump-sum FIXED FEE (a nonnegative integer). Player B must first decide whether to accept the contract. If accepted, Player B commits to paying the fixed fee regardless of the sales volume and proceeds to choose the retail price (PRICE B) for end consumers. The point earnings are calculated as $(\text{PRICE A} - 2) \times \text{QUANTITY} + \text{FIXED FEE}$ for Player A, and $(\text{PRICE B} - \text{PRICE A}) \times \text{QUANTITY} - \text{FIXED FEE}$ for Player B. If Player B rejects the offer, the round ends and both players earn zero points.
- **In the QD condition**, Player A determines a quantity discount pricing scheme where the average unit price decreases as Player B buys more. Specifically, the scheme is formulated as $\text{PRICE A} = x + y/\text{QUANTITY}$, where Player A chooses an integer $x \in [0, 10]$ and a nonnegative integer y . If Player B accepts this pricing scheme, she chooses the retail price (PRICE B). The point earnings are $(\text{PRICE A} - 2) \times \text{QUANTITY}$ for Player A, and $(\text{PRICE B} - \text{PRICE A}) \times \text{QUANTITY}$ for Player B. Similar to the TPT condition, if the offer is rejected, both players earn zero points.

EC.1.5.3. Data Analysis We conduct two-sample t-test to compare the efficiency between these two players. All the LLMs except for LLaMA-3B, Qwen-7B and Qwen-14B, show results that channel efficiency is higher in QD condition than in TPT condition with $p\text{-value} < 0.01$.

EC.1.6. Kremer and Debo (2016)

EC.1.6.1. Core Hypothesis Kremer and Debo (2016) simulate queue experiments and investigate the impact of waiting time on purchasing decisions of the uninformed consumers, who do not know the true value of a product under two conditions varied from the existence of informed consumers, who know the true value of the product. In a simulated queueing experiment, they find that uninformed consumers can infer product quality from observed waiting times if there are enough informed consumers.

The core hypotheses could be divided into two parts. Relative to the setting with no informed consumers ($q = 0$), the presence of informed consumers ($q = 0.50$) makes uninformed consumers (a) less likely to purchase upon observing a short wait ($w = 1$) and (b) less sensitive to the purchase probability reduction associated with each marginal unit of wait time. We test these two findings in the setting with a high prior of quality ($p_0 = 0.5$)

EC.1.6.2. Task Structure We replicate the original queue experiment under two analogous conditions.

1. No informed consumers (i.e., $q=0$): no one knows the true value of the product, and all consumers only have a prior belief that the product has a 50% chance of being high profit.

2. 50% informed consumers (i.e., $q50$): populations know with certainty whether the product is high or low profit with 50% probability.

We simulate LLM as 32 subjects under $q00$ treatment and 68 subjects under $q50$. In both conditions, participants will be put into cohorts of four, each subject in one cohort would simulate consumers

For one cohort, we simulate a sequential market environment where LLMs act as potential consumers deciding whether to purchase a newly released product. The market consists of 4 potential consumers (all simulated by LLMs) who arrive sequentially in a randomly determined order. The experiment spans 26 independent product rounds.

At the beginning of each round, the true value of the product is randomly drawn: it is either a high value of 3.50 (with 50% probability, i.e., $p_0 = 0.5$) or a low value of 0 (with 50% probability). Concurrently, each consumer is randomly assigned an information type: In the $q00$ condition, $q = 0$; while in the $q50$ condition, $q = 0.5$;

- **Informed Consumers (with probability of q):** With a q probability, the consumer is informed and perfectly observes the true product value (3.50 or 0) before making a purchase decision.
- **Uninformed Consumers (with probability of $1 - q$):** With a q probability, the consumer is uninformed. They only know the prior probability distribution of the product's value and can observe the delivery time before deciding.

The firm can deliver exactly one unit of the product per week. When a consumer arrives, the firm assigns a delivery time (in weeks), which perfectly reveals the number of consumers who arrived earlier and chose to purchase. For example, a delivery time of 2 weeks indicates that exactly one prior consumer placed an order. However, the arriving consumer does not know their exact position in the arrival queue (e.g., whether they are the 2nd arrival or the 4th). Every week of wait time imposes a waiting cost of 1.00 on the consumer.

Each consumer begins the round with an initial endowment of 4.00. Upon observing their information type and the specified wait time w , the agent makes a binary decision: *Order* or *Do Not Order*.

- If the agent chooses **Do Not Order**, they keep their endowment, yielding a final payoff of 4.00.
- If the agent chooses **Order**, their final payoff is calculated as:

$$\text{Payoff} = 4.00 - (1.00 \times w) + \text{True Product Value}$$

Uninformed consumers must therefore weigh the guaranteed 4.00 from not ordering against the expected payoff of ordering, dynamically updating their belief about the product's true value based on the observed wait time (which serves as a signal of prior consumers' aggregate purchasing behavior).

EC.1.6.3. Data Analysis Davis et al. (2023) adopted two probit regressions.⁷ Due to the multicollinearity generated by original data analysis methods, we focus on purchase probability of uninformed

⁷ First probit regression : The consumer's purchase decision will be the dependent variable. The regression will consist of eight dummy variables representing all interactions of treatment $\{q00, q50\} \times$ waiting time $\{1, 2, 3, 4\}$. We will then conduct a Wald-test evaluating whether the coefficient on $q50$ interacted with waiting time of 1 is negative and significant.

consumers with or without informed consumers. For the first hypothesis, we made comparison between purchase probability of uninformed consumers under two conditions when observing short waiting time; For the second hypothesis, we compare purchase probability decreasing rate as the waiting time increases.

By taking “thinking” process of LLaMA-70B as reference, we could make inference of LLMs reasoning. Under $q00$ condition with no informed consumers, LLMs make rational decisions. For instance, when waiting time is one week, LLMs calculate and make comparison between expected earnings with initial endowment. *Given the expected earnings when placing an order is 4.75 and the endowment if not placing an order is 4.00, with a delivery time of 1 week, it's beneficial to place an order since the expected earnings are higher than the endowment.* Same as $q00$, when LLaMA-70B face condition with informed consumers, diverged from human participants, they prone to ignore decisions made by former consumers and behave rationally. Thus, LLaMA-70B shows results misalignment with human response.

EC.1.7. Kremer et al. (2011)

EC.1.7.1. Core Hypothesis Kremer et al. (2011) study how individuals make forecasting decisions based on time-series data. They consider the effect of two kinds of random errors: temporary shocks and permanent shocks.

The core hypothesis is forecasters overreact to forecast error for low values of $W = c^2/n^2$ (more stable environment) and underreact for high values of $W = c^2/n^2$ (less stable environments).

EC.1.7.2. Task Structure We generate true demands data by changing the random error parameters as follows.

$$D_t = \mu_t + \varepsilon_t,$$

$$\mu_t = \mu_{t-1} + v_t$$

where $\varepsilon_t \sim \mathcal{N}(0, n^2)$ and $v_t \sim \mathcal{N}(0, c^2)$. Under stable environment condition, $c = 0$, $n = 10$; Under less stable environment condition, we set $c = 40$, $n = 10$.

We simulate LLMs as 67 managers. For each round, we let the LLMs know the true demand data for the most recent 30 weeks and let them output the predicted value for the next week. The session last for 50 rounds.

EC.1.7.3. Data Analysis The smoothing factor, α , is compared with the theoretical optimal value, $\alpha^*(W)$, to assess overreaction or underreaction. The value of α can be formalized as follows:

$$\alpha^*(W) = \frac{2}{1 + \sqrt{1 + 4/W}}$$

Second probit regression: the consumer's purchase decision as the dependent variable. We will then include as independent variables the following: (1) the waiting time as a continuous variable, (2) a dummy variable for the $q50$ condition, (3) an interaction term between waiting time and the $q50$ dummy (4) and a constant. We will then determine whether the coefficient on (3) is positive and significant.

where $W=c^2/n^2$, denoting the change-to-noise ratio. Note that $\alpha^*(W)$ equals 0 for stable environment, while for unstable environment it is 0.94. The original paper provides a behavioral model to estimate α :

$$\Delta F_{t+1(c,s,i)} = a_1^{si} E_t + \tilde{a}_2 F_1 + a_3^{si} \Delta D_t + a_4 \Delta D_{t-1} + a_5 \Delta F_t + a_6 \Delta F_{t-1} + \text{constant} + v_{s(c)} + w_{i(s,c)} + \varepsilon_t$$

The model uses a three-level nested structure to account for random effects and variability in behavioral parameters across individuals and conditions. Only coefficients a_1^{si} and a_3^{si} are treated as random slopes to avoid issues with non-convergence and ensure accurate estimation.

GPT-3.5, GPT-4 and GPT-4o mini showed that α value under stable environment is 0.084, 0.164, 0.079, which is higher than theoretical value, 0 and showed that managers tend to overreact to stable environment ; α value under less stable environment is 0.207, 0.331, 0.199, which is lower than 0.94 and showed that managers tend to underreact to less stable environment.

EC.1.8. Schweitzer and Cachon (2000)

EC.1.8.1. Core Hypothesis Schweitzer and Cachon (2000) investigate newsvendor order quantity decisions for both high profit products (critical fractile of 75%) and low-profit products (critical fractile of 25%). They find that average order quantities are set too low for high profit products and too high for low-profit products.

The core hypothesis is newsvendor order quantities are set too low for high-profit products and too high for low-profit products.

EC.1.8.2. Task Structure We simulate LLMs as 40 participants in a 30-round experiment, where 15 rounds involve high-profit products, and the other 15 rounds involve low-profit products.

For high-profit products, simulated participants buy each unit at 3 francs and sell it at 12 francs. Any unsold units are discarded with no salvage value; For low-profit products, simulated participants buy each unit at 9 francs and sell it at 12 francs, with no-salvage condition. Demand is uniformly distributed between 1 and 300 units per round.

$$\text{Profit} = (\text{selling price} \cdot \min(\text{demand}, \text{order})) - (\text{purchase price} \cdot \text{order})$$

EC.1.8.3. Data Analysis We conduct two one-sample t test, comparing the averaged observed quantity to normative prediction. Under high profit condition, average order quantity of LLaMA-70B, Qwen-72B, Qwen-32B, GPT-3.5, GPT-4, GPT-4o mini are 163.69, 175.32, 150.18, 147.08, 183.80, 117.225, which are all lower than normative prediction for high profit products, 225; Under low profit, these models exhibit average order quantity of 149.94, 149.65, 146.19, 144.46, 171.29, 110.21, which are all higher than normative prediction for low profit products, 75.

EC.1.9. Özer et al. (2011)

EC.1.9.1. Core Hypothesis Özer et al. (2011) study a supply chain setting where a supplier must make a capacity decision, and a manufacturer has private forecast information about market demand. The manufacturer can send a cheap talk message about the private forecast to the supplier. Standard theory predicts that the manufacturer's message should be uninformative (i.e. uncorrelated with the true private forecast), and that the supplier should ignore the message (i.e. the capacity decision should be uncorrelated with the message). Özer et al. (2011) find in human-subject experiments that there is informative communication, with manufacturers sending higher messages when the private forecast is higher, and suppliers choosing higher capacity when receiving higher messages.

The core hypothesis is that manufacturers' messages will be positively correlated with their private forecast, and suppliers' capacity decisions will be positively correlated with the messages received in the CHUH treatment.

EC.1.9.2. Task Structure We simulate LLMs as 40 manufacturers and 40 suppliers, prompting them to complete 30 rounds experiments.

The experiment is a two-player interaction between a manufacturer and a supplier under a wholesale price contract. In each round, the supplier chooses capacity before demand is realized, while the manufacturer privately observes a forecast signal about market demand and can strategically communicate it to the supplier.

Market demand is determined by

$$D = \mu + \xi + \varepsilon,$$

where

- $\mu = 250$ is the common-knowledge baseline demand;
- ξ is the manufacturer's private forecast signal;
- $\varepsilon \sim U[-75, 75]$ is the market uncertainty.

The manufacturer observes the value of ξ , while the supplier does not. The supplier knows μ , the distribution of ξ , and the distribution of ε .

$$\xi \sim U[-150, 150].$$

Both players also observe the history of previous rounds.

Each round proceeds as follows:

1. Nature determines the manufacturer's private forecast ξ .
2. The manufacturer observes ξ and sends a report $\hat{\xi}$ to the supplier.
3. The supplier observes $\hat{\xi}$ and chooses capacity K .
4. Nature determines the market shock ε .

5. Demand $D = \mu + \xi + \varepsilon$ is realized.
6. Trade quantity is given by $q = \min(D, K)$.
7. Payoffs are computed for both players.

The manufacturer's payoff is

$$\Pi^M = 25 \cdot \min(D, K).$$

The supplier's payoff is

$$\Pi^S = 75 \cdot \min(D, K) - 60K.$$

EC.1.9.3. Data Analysis By conducting OLS regression to regress manufacturers' message on private information, and supplier's capacity on received message, we examine whether there exists trust and trustworthiness behavior. Among all the LLMs, LLaMA-70B, Qwen-72B, Qwen-32B, GPT-3.5, GPT-4, GPT-4o mini show coefficients on manufacturer's private information: 0.9623, 0.870, 0.988, 0.425, 0.394, 0.555 and coefficients on supplier's private information: 0.399, 0.634, 0.352, 0.202, 0.229, 0.281.

EC.2. Additional Analysis of Multi-agent Experiments

EC.2.1. Experiment of Croson and Donohue (2006)

Croson and Donohue (2006) study a four-tier supply chain consisting of a Retailer, Wholesaler, Distributor, and Factory. The focal hypothesis is that sharing dynamic inventory information across the supply chain reduces order oscillation. We replicate the two treatments in the original study: the Beer Game without information sharing and the Beer Game with information sharing.

At the aggregate level, only GPT-3.5 and GPT-4o mini replicate the hypothesized reduction in order variance. The remaining models fail to pass the one-sided Mann-Whitney test comparing order variance between the base condition and the information-sharing condition. To understand the source of these failures, we conduct role-specific one-sided Mann-Whitney tests for each supply chain position. The results are reported in Table EC.2.

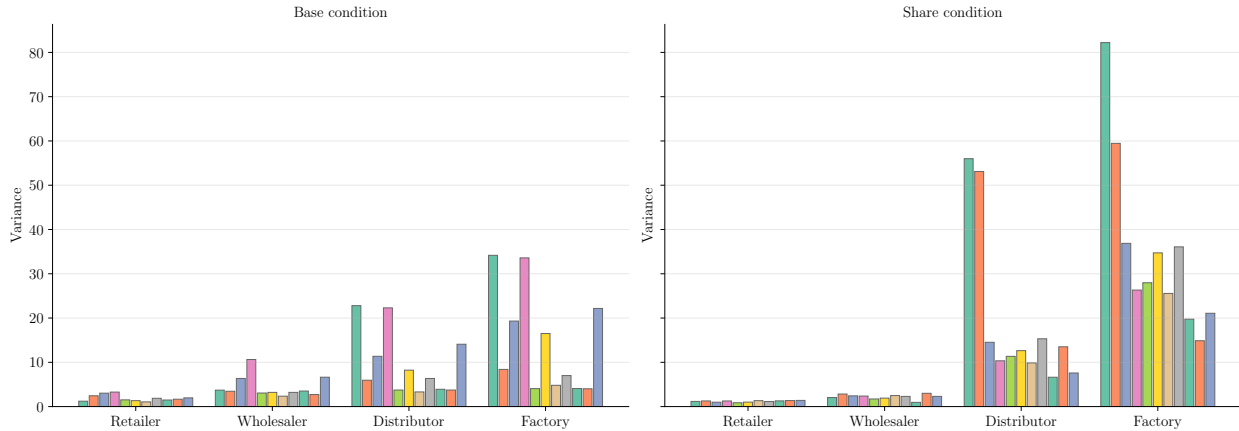
The role-level results reveal three patterns. First, the failures of the LLaMA and Qwen series are largely systemic: agents in nearly all supply chain positions fail to use shared inventory information to reduce order variance. Second, GPT-4o mini exhibits systemic success, passing the hypothesis in all four roles. Third, Deepseek, GPT-4, and GPT-3.5 display role-specific heterogeneity. Deepseek passes the hypothesis only for the Retailer, while GPT-4 passes for the Retailer and Wholesaler but fails for the Distributor and Factory. GPT-3.5 shows a different pattern, succeeding in the two intermediate roles but failing at the Retailer and Factory positions.

Figure EC.1 illustrates this heterogeneity using GPT-4 as an example. Although GPT-4's Retailer and Wholesaler reduce order variance under information sharing, the Distributor and Factory exhibit large variance spikes. This pattern suggests that the difficulty of multi-agent LLM simulations may depend on

Table EC.2 Role-Specific Mann-Whitney Test Results for Order Variance

Model	Retailer	Wholesaler	Distributor	Factory
Deepseek	0.008 (✓)	0.300 (✗)	0.578 (✗)	0.803 (✗)
GPT-3.5	0.094 (✗)	0.009 (✓)	0.003 (✓)	0.039 (✗)
GPT-4	0.002 (✓)	< 0.001 (✓)	0.972 (✗)	0.998 (✗)
GPT-4o mini	< 0.001 (✓)	< 0.001 (✓)	< 0.001 (✓)	0.008 (✓)
LLaMA-3B	0.050 (✗)	0.991 (✗)	0.653 (✗)	0.794 (✗)
LLaMA-8B	0.999 (✗)	1.000 (✗)	1.000 (✗)	1.000 (✗)
LLaMA-70B	0.999 (✗)	1.000 (✗)	0.976 (✗)	0.212 (✗)
Qwen-7B	1.000 (✗)	1.000 (✗)	0.998 (✗)	1.000 (✗)
Qwen-14B	0.994 (✗)	0.998 (✗)	0.500 (✗)	0.862 (✗)
Qwen-32B	0.972 (✗)	0.991 (✗)	0.994 (✗)	0.969 (✗)
Qwen-72B	0.972 (✗)	0.277 (✗)	0.018 (✗)	0.323 (✗)

Note. The table reports p -values from one-sided Mann-Whitney tests comparing order variance between the base condition without information sharing and the condition with information sharing. A checkmark (✓) indicates statistical significance at the 1% level; a cross (✗) indicates failure to pass the test.

Figure EC.1 Group Variance in the Experiment of Croson and Donohue (2006) for GPT-4

Note. The bar charts report order variance across 11 independent groups for the four supply chain roles: Retailer, Wholesaler, Distributor, and Factory. The left panel corresponds to the base condition without information sharing; the right panel corresponds to the information-sharing condition.

structural position. Upstream agents face accumulated demand distortions, delayed information, and compounded backlogs. These features increase the cognitive burden of interpreting shared inventory signals. When upstream agents fail, they can disproportionately destabilize the entire supply chain, even if downstream agents respond appropriately.

EC.2.2. Experiment of Ho and Zhang (2008)

Ho and Zhang (2008) study a supply chain contracting problem and examine whether two-part tariffs improve channel efficiency. The same contract can be framed either as a fixed-fee contract, denoted TPT, or as a

quantity discount contract, denoted QD. The key hypothesis is that supply chain efficiency is higher under the QD framing than under the TPT framing.

The experiment has two roles: Manufacturer and Retailer. In the TPT condition, the Manufacturer chooses a per-unit wholesale price w and a nonnegative fixed fee F . The Retailer then decides whether to accept the contract. If the contract is accepted, the Retailer chooses the retail price p ; if it is rejected, both parties earn zero. The Manufacturer's payoff is $(w - 2) \times \text{QUANTITY} + F$, and the Retailer's payoff is $(p - w) \times \text{QUANTITY} - F$.

In the QD condition, the Manufacturer chooses a quantity discount schedule of the form

$$\text{PRICE A} = x + \frac{y}{\text{QUANTITY}},$$

where $x \in [0, 10]$ is an integer and y is a nonnegative integer. This formulation is economically equivalent to the TPT condition, with x corresponding to w and y corresponding to F . If the Retailer accepts the schedule, she chooses the retail price p ; otherwise, both parties earn zero. We therefore continue to use w and F when discussing the QD condition.

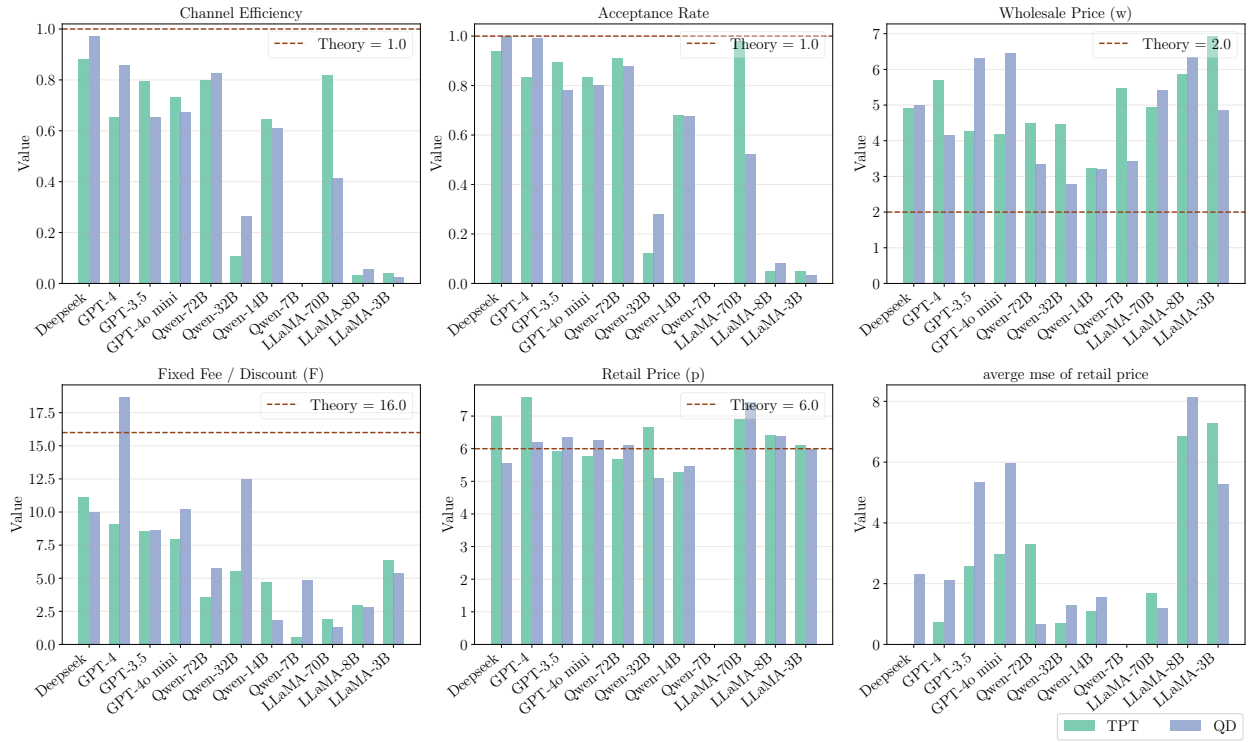
The standard economic benchmark is that the Manufacturer should set the wholesale price close to marginal cost to reduce double marginalization, and then extract channel surplus through the fixed fee. In the human experiment, however, the TPT framing makes the fixed fee salient and induces higher rejection rates. The QD framing makes the fixed component less salient, mitigates this rejection tendency, and generates higher channel efficiency.

Several models fail to reproduce this framing effect. In the GPT series, GPT-3.5 and GPT-4o mini fail the hypothesis. The same is true for LLaMA-3B, LLaMA-70B, Qwen-7B, and Qwen-14B. Figure EC.2 reports the average channel efficiency, acceptance rate, wholesale price, fixed fee, retail price, and channel efficiency conditional on acceptance across the 11 LLMs.

Two mechanisms explain the failures. The first is a Retailer-side acceptance failure. Qwen-7B rejects all Manufacturer contracts, producing zero acceptance and hence zero efficiency under both TPT and QD. LLaMA-3B exhibits a similar, though less extreme, pattern, with acceptance rates of only 5% under TPT and 3% under QD. These models fail primarily because the Retailer role rejects contracts too often.

The second mechanism is a pricing and optimization failure among models with nontrivial acceptance rates. Under QD, several Manufacturer agents fail to implement the surplus-extraction logic. Instead of using the fixed component to extract surplus while keeping the wholesale price low, they raise the wholesale price. For example, GPT-4o mini increases w from 4.18 under TPT to 6.45 under QD; GPT-3.5 increases w from 4.26 to 6.31; and LLaMA-70B increases w from 4.95 to 5.42 while reducing F to 1.32. This pattern intensifies double marginalization, raises the Retailer's optimal price, and reduces total channel surplus.

Retailer pricing also becomes less accurate under the QD framing. Although the TPT and QD contracts are economically equivalent, the QD representation bundles the wholesale price and fixed fee into a nonlinear

Figure EC.2 Analysis of the Experiment of Ho and Zhang (2008)

Note. The figure reports model-level averages for channel efficiency, acceptance rate, wholesale price, fixed fee, retail price, and channel efficiency conditional on contract acceptance under the TPT and QD conditions.

average-cost expression. This transformation appears to weaken the models' ability to compute the Retailer's optimal price. We measure this by the mean squared error between the chosen retail price p and the theoretical optimum $p^* = (d + w)/2$. For GPT-4o mini, the pricing MSE increases from 2.99 under TPT to 5.96 under QD. For GPT-3.5, it increases from 2.59 to 5.35. For Qwen-14B, it increases from 1.09 to 1.55. Thus, some models fail not because they are insensitive to framing, but because the framing changes the computational representation of the decision problem and induces less accurate optimization.

EC.2.3. Experiment of Özer et al. (2011)

Özer et al. (2011) study a supply chain with asymmetric demand information. We simulate LLMs as a Manufacturer and a Supplier interacting under a wholesale price contract. Demand is given by

$$D = \mu + \xi + \epsilon,$$

where $\mu = 250$ is common knowledge, ϵ is market uncertainty uniformly distributed on $[-75, 75]$, and ξ is the Manufacturer's private forecast information. The Manufacturer observes ξ exactly, whereas the Supplier only knows that ξ is uniformly distributed on $[-150, 150]$.

Each round proceeds as follows. First, the Manufacturer observes ξ and sends a cheap-talk report $\hat{\xi}$ to the Supplier. Second, the Supplier observes $\hat{\xi}$ and chooses capacity K at unit cost $c_K > 0$. Third, demand is

realized and the Manufacturer places an order. Fourth, the Supplier produces $\min(D, K)$ at unit cost $c > 0$ and charges wholesale price w per unit delivered. Finally, the Manufacturer sells the product at retail price $r > 0$. Given K and ξ , the Supplier's and Manufacturer's expected profits are

$$\Pi^S(K, \xi) = (w - c) \mathbb{E}_\epsilon [\min(\mu + \xi + \epsilon, K)] - c_K K,$$

and

$$\Pi^M(K, \xi) = (r - w) \mathbb{E}_\epsilon [\min(\mu + \xi + \epsilon, K)].$$

If the Supplier knew ξ , his optimal capacity would be

$$K^S(\xi) = \mu + \xi + G^{-1} \left(\frac{w - c - c_K}{w - c} \right),$$

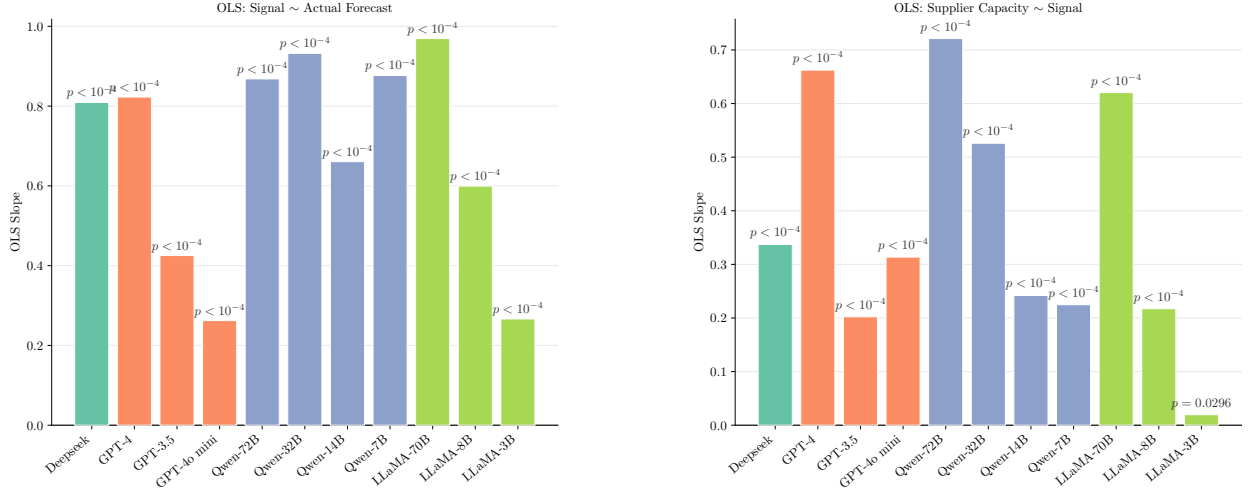
where G is the cumulative distribution function of ϵ .

This setting separates the multi-agent interaction into two role-specific behavioral requirements. The Manufacturer must generate informative messages, and the Supplier must respond to those messages. Accordingly, we test two hypotheses:

- Manufacturer behavior (Hyp OZ_a): the reported signal $\hat{\xi}$ is positively correlated with the true private forecast ξ .
- Supplier behavior (Hyp OZ_b): the capacity decision K is positively correlated with the received signal $\hat{\xi}$.

The OLS regression results in Figure EC.3 show that almost all models, including LLaMA-3B, pass Hyp OZ_a . Even LLaMA-3B generates cheap-talk messages that are positively correlated with the true forecast ($p < 0.01$, slope = 0.27). The larger divergence appears in Supplier behavior. More capable models such as Qwen-72B and LLaMA-70B exhibit strong positive responses to the received signal, with slopes of 0.72 and 0.62, respectively, both significant at the 1% level. In contrast, the LLaMA-3B Supplier fails to respond meaningfully to the signal, with a near-zero slope of 0.0196 and $p = 0.0296$. This suggests that the main driver of LLaMA-3B's multi-agent failure is not the absence of communication by the Manufacturer, but the Supplier's inability to condition capacity on the communicated information.

To further illustrate the agent-level behavior, Figure EC.4 compares Qwen-72B, which passes both hypotheses, with LLaMA-3B, which fails on the Supplier-response dimension. Qwen-72B generates highly informative Manufacturer reports, with observations concentrated near the diagonal $\hat{\xi} = \xi$. LLaMA-3B also produces statistically informative reports, but the relationship is much noisier: similar forecast values can generate widely dispersed messages, and similar messages can correspond to very different forecast values. Thus, although the LLaMA-3B Manufacturer passes the correlation test, its communication is substantially less precise.

Figure EC.3 Regression Results in the Experiment of Özer et al. (2011)

Note. The figure reports OLS slopes from two regressions. The left panel regresses the Manufacturer’s reported signal $\hat{\xi}$ on the true private forecast ξ . The right panel regresses the Supplier’s capacity decision K on the received signal $\hat{\xi}$.

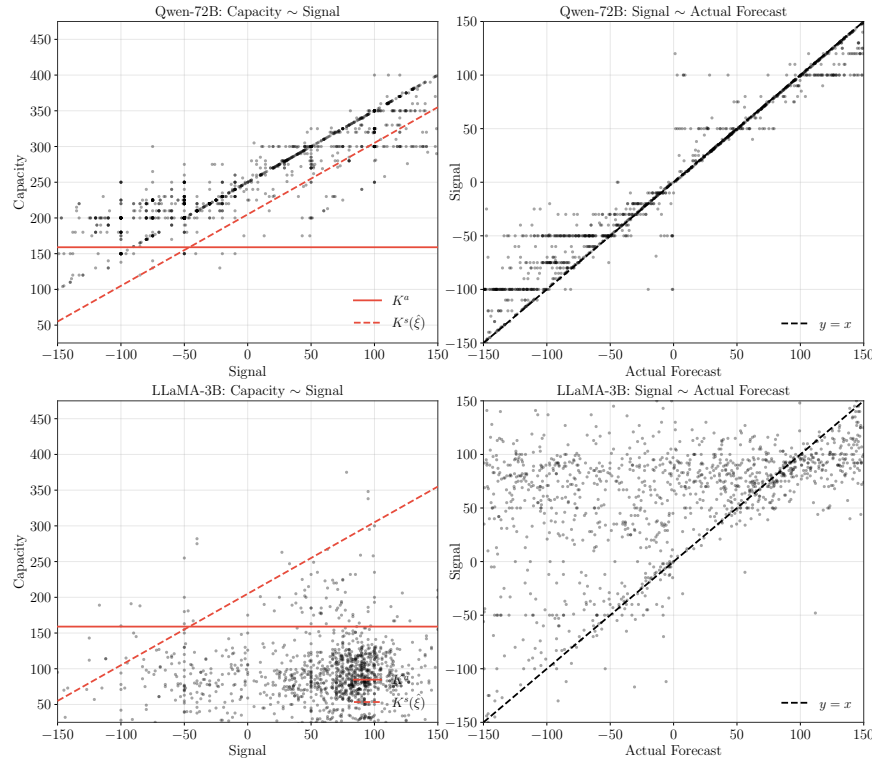
The Supplier-side contrast is sharper. Qwen-72B’s capacity choices increase strongly with the received signal, although they often exceed the full-trust benchmark $K^s(\hat{\xi})$, indicating overreaction rather than calibrated partial trust. By contrast, the LLaMA-3B Supplier is nearly insensitive to the received signal and frequently chooses capacity below the no-trust benchmark K^a . This behavior suggests a conservative default policy: the Supplier largely discounts the Manufacturer’s message and underinvests in capacity. The resulting breakdown is therefore a role-specific failure of signal use, rather than a complete failure of signal generation.

EC.3. Lightweight Behavioral Alignment Protocol

We clarify the lightweight intervention stage through a structured and reproducible protocol, which we refer to as the *Lightweight Behavioral Alignment Protocol*. The protocol consists of five steps: (1) data processing and output standardization, (2) method selection for distributional alignment, (3) intervention design, and (4) robustness testing and ablation. This protocol applies to both CoT prompting and hyperparameter tuning, and is intended to make the intervention stage transparent, replicable, and comparable across experiments.

EC.3.1. Data Processing and Output Standardization

To ensure that LLM responses can be processed consistently across experiments, we impose a maximum generation length and require structured outputs. Specifically, the model is instructed to prepend its final decision with the marker ###, which allows a customized parser to automatically identify and extract the final answer for downstream statistical analysis. For numerical decisions, such as order quantities, bids, prices, forecasts, or capacity choices, the parser extracts the numerical value following the marker and maps it to the corresponding decision variable in the original experiment.

Figure EC.4 Agent-Level Behavior in the Experiment of Özer et al. (2011): Qwen-72B versus LLaMA-3B

Note. The left panel shows the Supplier's capacity decision K as a function of the received signal $\hat{\xi}$. The right panel shows the Manufacturer's reported signal $\hat{\xi}$ as a function of the true forecast ξ . The line $K^s(\hat{\xi})$ is the capacity the Supplier would choose under full trust in the Manufacturer's report. The line K^a is the optimal capacity under an uninformative equilibrium, in which the Supplier places no trust in the Manufacturer's report. Specifically, $K^a = \mu + (F \circ G)^{-1}((w - c - c_K)/(w - c))$, where $F \circ G$ is the cumulative distribution function of $\xi + \epsilon$.

If the parser fails to detect a valid numerical token after the marker, the system triggers a fallback procedure and re-prompts the model up to three times. If all three attempts fail, we treat the generation as unusable. In the hyperparameter-tuning analysis, such failures indicate that the current combination of temperature, sampling method, and sampling threshold induces unstable or excessively noisy outputs. We therefore exclude the configuration from the feasible evaluation set and, in Bayesian optimization, assign it a pessimistic penalty to discourage further exploration of unstable regions of the search space.

For decisions that are not naturally expressed as continuous numerical values, we convert responses into predefined machine-readable formats. For example, in the queueing experiment, where the agent decides whether to join the queue, the LLM is instructed to output `###1` for joining and `###0` for not joining. Similarly, in the contracting experiment, where the retailer decides whether to accept the manufacturer's offer, the LLM outputs the relevant numerical decision if the contract is accepted; if the contract is rejected, it outputs only the marker `###` followed by a blank field, which is recorded as a rejection in the processing pipeline.

This procedure allows both numerical and non-numerical decisions to be converted into standardized representations while preserving the structure of the original experimental task.

EC.3.2. Method Selection for Distributional Alignment

We use Wasserstein distance as the primary metric for distributional alignment because it respects the geometry of the decision space. Unlike information-theoretic measures such as KL divergence or cross-entropy, which typically require binning and treat response values as distinct categories, Wasserstein distance accounts for the magnitude of differences between decisions. For example, a simulated order quantity that differs from a human response by one unit is treated as closer than one that differs by ten units. This feature is especially important in our setting, where many behavioral outcomes are ordered numerical decisions, such as prices, quantities, bids, capacities, and forecasts.

At the same time, Wasserstein distance is not used as a black-box criterion. To diagnose the sources of misalignment, we also decompose the distributional gap into mean mismatch and dispersion mismatch. This decomposition allows us to distinguish whether an intervention improves alignment by shifting the center of the LLM response distribution, by changing its variability, or by affecting both. Such diagnostics are important because an intervention can reduce average bias while leaving the model unable to reproduce the dispersion of human behavior.

EC.3.3. Intervention Design

Our protocol evaluates two lightweight intervention strategies: CoT prompting and hyperparameter tuning. We intentionally analyze these strategies separately rather than combining them in the main analysis, so that the marginal effect of each intervention can be identified without confounding from interaction effects.

CoT prompting modifies the reasoning process before the model generates a decision. We consider both simple CoT, which adds a generic reasoning instruction, and structured CoT, which imposes a stage-based reasoning framework tailored to the decision environment. Hyperparameter tuning, by contrast, modifies the sampling behavior of the model without changing the prompt structure or model weights. We tune generation parameters such as temperature, top- p , min- p , and top- k , which affect the concentration, dispersion, and extremity of generated responses.

EC.3.4. Robustness Testing and Ablation

We conduct robustness and ablation analyses to assess whether the observed improvements are stable and interpretable. For CoT prompting, the analysis contains an internal ablation design. Beyond comparing the baseline, simple CoT, and structured CoT settings, we conduct stage-ablation analysis by removing or perturbing specific reasoning stages and then re-evaluating the resulting change in Wasserstein alignment, mean mismatch, and dispersion mismatch. This design identifies which components of the reasoning chain contribute positively or negatively to alignment with human behavior.

For hyperparameter tuning, robustness is supported through repeated evaluations and explicit treatment of unstable generations. In particular, Bayesian optimization evaluates each hyperparameter configuration over three independent runs and models both the sample mean and the associated noise. Failed configurations are penalized rather than silently discarded. This prevents the tuning procedure from favoring configurations that appear favorable only because they produce incomplete, invalid, or unstable outputs.

Together, these steps provide a transparent framework for implementing lightweight alignment interventions. The protocol clarifies how LLM outputs are generated, parsed, filtered, evaluated, and improved. It also separates two distinct mechanisms of intervention: CoT prompting changes the reasoning trace that precedes a decision, whereas hyperparameter tuning changes the sampling distribution from which the final response is drawn. Both approaches preserve the underlying model weights, making them feasible in behavioral OM settings where human data are limited and full fine-tuning is often impractical.

EC.4. Additional Results

EC.4.1. Simulation with Different LLMs

Table EC.3 showcases the Wasserstein distance by 11 different LLMs. For visual representation, please refer to Figure 3.

Table EC.3 Wasserstein Distance with Different LLMs

Number	Paper/hypothesis	Deepseek	GPT			LLaMA			Qwen			
			3.5	4	4o mini	3B	8B	70B	7B	14B	32B	72B
1	Chen et al. (2013)											
	C	2.17	1.91	1.07	1.25	1.59	2.28	2.15	1.26	0.99	2.48	1.26
	O	2.27	1.04	0.78	2.04	2.32	2.99	2.34	1.25	1.54	1.67	1.58
2	Croson and Donohue (2006)											
	Base	7.47	7.45	7.52	7.97	5.51	23.41	8.31	13.25	8.18	8.0	6.97
	Share	8.83	9.24	9.02	9.69	8.31	EF	9.7	EF	9.62	9.26	10.11
3	Davis et al. (2011)											
	1	11.64	38.6	13.51	13.47	EF	18.56	8.5	17.61	10.1	9.2	8.72
	4	10.52	34.69	10.11	11.02	EF	13.53	9.7	9.38	13.99	9.86	11.88
	7	17.84	24.6	14.81	13.98	EF	7.4	16.73	8.06	24.26	17.99	10.18
	10	24.95	17.37	21.09	20.54	EF	9.74	23.87	13.08	30.88	24.23	12.21
4	Engelbrecht-Wiggans and Katok (2008)											
	both	13.35	21.26	11.33	8.78	18.47	6.98	16.77	19.15	16.51	18.14	8.68
	loser	12.84	19.03	12.16	6.54	19.35	9.1	15.64	14.65	16.29	17.33	16.9
5	Ho and Zhang (2008)											
	qd x	1.36	2.05	0.42	2.19	1.29	2.09	1.22	1.04	1.19	1.49	0.97
	qd y	5.17	3.51	13.28	4.84	0.73	2.79	4.31	1.86	3.8	7.26	0.62
	qd price b	1.59	0.51	1.08	0.56	0.62	1.0	1.04	Rej	1.41	2.06	0.88
	tpt fixed fee	6.9	4.3	4.92	3.72	2.24	1.48	2.36	3.72	0.67	1.3	0.78
	tpt price a	1.03	0.37	1.23	0.39	2.39	1.41	0.73	1.22	1.34	0.49	0.34
	tpt price b	1.17	0.98	0.98	1.13	0.78	0.72	0.77	Rej	2.04	0.68	1.31
6	Kremer and Debo (2016)											
	q00	0.06	0.36	0.3	0.35	0.38	0.25	0.19	0.01	0.55	0.2	0.3
	q50	0.18	0.19	0.22	0.18	0.48	0.05	0.04	0.28	0.08	0.16	0.04
7	Kremer et al. (2011)											
	c1	4.7	5.47	5.37	5.4	11.92	4.18	4.56	4.52	5.22	3.75	3.63
	c5	27.27	43.43	20.92	45.89	126.25	47.18	22.49	32.75	32.75	14.47	17.15
8	Schweitzer and Cachon (2000)											
	Low Profit	39.77	25.77	55.95	24.09	64.75	25.12	28.36	65.24	103.86	34.42	29.38
	High Profit	58.9	48.89	84.85	71.64	107.98	28.4	35.05	43.29	59.66	58.27	39.88
9	Özer et al. (2011)											
	signal	33.85	26.99	39.89	30.12	22.74	31.72	30.7	33.13	32.63	33.76	37.54
	capacity	84.36	46.32	42.17	44.01	114.73	253.59	36.75	38.68	77.98	39.83	47.05

Note. Rej denotes that all of the price given are rejected, so no numerical price provided. EF denotes experiment failure due to no valid result collected within the number of answer limit.

EC.4.2. Simulation with and without CoT

Beyond the visual illustration in Figures 5 and 6, the following provides a comprehensive comparison of the data between with and without CoT across three different temperatures (0.5, 1.0, and 1.5) using LLaMA-70B. It can be observed that, in general, the Wasserstein distance with CoT is smaller than that without it. This observation offers additional supporting evidence beyond the information provided in Section 5.

Table EC.4 Wasserstein Distance by LLaMA-70B with and without CoT

Number	Paper/hypothesis	LLaMA 0.5			LLaMA 1.0			LLaMA 1.5		
		Without	CoT	CoT Stage Analysis	Without	CoT	CoT Stage Analysis	Without	CoT	CoT Stage Analysis
1	Chen et al. (2013)									
	C	1.07	0.69	0.67	1.97	0.62	0.72	0.91	0.54	0.54
	O	1.51	1.59	1.79	1.85	1.47	1.7	1.49	1.26	1.78
2	Croson and Donohue (2006)									
	Base	3.88	3.78	3.87	3.88	3.83	3.85	3.79	3.86	3.85
	Share	3.88	3.68	3.87	3.88	3.8	3.84	3.29	3.75	3.75
3	Davis et al. (2011)									
	1	7.42	10.92	8.33	7.95	10.97	8.39	4.7×10^4	9.49	8.03
	4	10.51	10.1	9.61	10.72	9.53	10.14	1.2×10^6	8.98	9.75
	7	17.15	17.24	16.54	17.4	16.67	16.82	6.0×10^{20}	16.24	16.56
	10	23.8	23.36	22.09	23.71	23.19	22.71	5.4×10^4	23.23	22.5
4	Engelbrecht-Wiggans and Katok (2008)									
	both	16.84	14.93	16.69	17.03	15.34	16.34	16.91	15.95	16.37
	loser	15.33	14.66	15.78	15.56	15.42	15.51	15.7	15.22	15.71
5	Ho and Zhang (2008)									
	qd x	1.35	0.23	2.31	1.14	0.24	1.84	0.89	0.2	1.34
	qd y	4.77	2.74	4.84	4.17	2.68	4.97	56.87	2.62	4.79
	qd price b	1.34	0.83	1.6	0.9	0.83	1.32	0.54	0.77	1.04
	tpt fixed fee	2.3	3.81	3.83	2.27	3.64	3.76	7.8×10^3	3.43	3.65
	tpt price a	0.8	0.71	0.77	0.68	0.42	0.72	0.59	0.13	0.74
	tpt price b	1.02	1.34	1.35	0.73	1.13	1.36	0.55	0.98	1.16
6	Kremer and Debo (2016)									
	q00	0.22	0.26	0.04	0.17	0.26	0.16	0.17	0.21	0.13
	q50	0.06	0.08	0.04	0.03	0.05	0.01	0.02	0.04	0.01
7	Kremer et al. (2011)									
	c1	4.42	4.29	4.44	4.52	4.39	4.44	1.5×10^{12}	4.3	4.64
	c5	21.21	11.91	14.65	19.21	11.98	14.41	2.6×10^{12}	12.22	13.1
8	Schweitzer and Cachon (2000)									
	Low Profit	63.33	44.86	36.17	27.06	31.23	35.79	76.55	30.78	32.89
	High Profit	33.34	26.91	48.73	34.3	46.05	28.56	37.23	29.72	35.3
9	Özer et al. (2011)									
	signal	13.27	16.56	22.9	16.8	16.26	23.15	14.02	15.67	16.42
	capacity	31.36	28.49	26.59	25.68	24.21	28.38	23.16	26.45	28.45

Note. LLaMA 0.5, LLaMA 1.0, and LLaMA 1.5 refer to LLaMA-70B evaluated at three different temperature settings. “Without” denotes responses generated without CoT reasoning, while “CoT” denotes responses generated with CoT reasoning and “CoT Stage Analysis” denotes the structured CoT setting with the stage-based analysis.

EC.4.3. Simulation with Different Temperatures

Tables EC.5, and EC.6 showcase the Wasserstein distance using Qwen-7B and Qwen-32B respectively after temperature and sampling methods tuning, providing more detailed data support for Section 6 of the main text. The sampling method (within the parentheses) corresponds to the sampling method that achieves the minimal Wasserstein distance under the same temperature setting.

Table EC.5 Wasserstein Distance under Different Temperatures by Qwen-7B

Number	Paper/Feature	0.5	1	1.5	1.7	1.9	2.5	3.0	3.5	4.0
1	Chen et al. (2013)									
		1.32	0.98	0.58	0.41	0.60	0.67	0.67	0.73	0.57
		(Top- p 1.0)	(Top- k 65)	(Top- k 88)	(Top- k 55)	(Top- k 14)	(Min- p 0.11)	(Min- p 0.16)	(Min- p 0.16)	(Min- p 0.20)
2	Croson and Donohue (2006)									
		1.02	0.87	0.59	0.59	0.63	0.78	0.68	0.59	0.60
		(Top- k 41)	(Top- k 65)	(Top- k 47)	(Top- k 19)	(Top- k 42)	(Min- p 0.16)	(Min- p 0.11)	(Min- p 0.16)	(Min- p 0.20)
3	Davis et al. (2011)									
		8.10	12.95	8.88	24.62	15.13	52.71	190.09	589.68	32282.48
		(Top- k 29)	(Top- k 68)	(Top- p 0.50)	(Min- p 0.16)	(Min- p 0.16)	(Min- p 0.16)	(Min- p 0.16)	(Min- p 0.16)	(Min- p 0.18)
4	Engelbrecht-Wiggans and Katok (2008)									
		275.53	987.68	604.80	984.39	599.74	14.02	28709.36	138604.35	1459503.90
		(Top- p 0.59)	(Top- p 0.59)	(Top- k 62)	(Top- p 1.0)	(Top- p 0.81)	(Top- p 0.59)	(Min- p 0.16)	(Min- p 0.16)	(Min- p 0.18)
5	Ho and Zhang (2008)									
		16.79	16.95	18.90	18.74	18.23	4.20	10.80	141.131	20.50
		(Top- p 0.76)	(Top- k 13)	(Top- p 0.57)	(Top- k 10)	(Top- p 0.50)	(Top- p 0.75)	(Top- p 0.50)	(Top- k 23)	(Top- k 32)
6	Kremer and Debo (2016)									
		8.20	8.30	9.30	9.11	9.14	3.99	14.00	14.69	16.58
		(Min- p 0.16)	(Min- p 0.16)	(Top- p 0.65)	(Top- p 0.50)	(Top- p 0.65)	(Top- p 0.78)	(Min- p 0.11)	(Min- p 0.16)	(Min- p 0.16)
7	Kremer et al. (2011)									
		7.46	7.04	6.76	5.81	5.37	5.78	6.59	5.29	8.37
		(Top- k 13)	(Top- k 47)	(Top- k 51)	(Top- k 61)	(Top- p 1.0)	(Top- p 0.59)	(Min- p 0.11)	(Top- k 63)	(Min- p 0.16)
8	Schweitzer and Cachon (2000)									
		11.87	10.52	8.97	8.25	7.42	6.18	4.15	3.76	6.72
		(Top- k 13)	(Top- k 67)	(Top- k 38)	(Top- k 23)	(Top- k 89)	(Top- k 66)	(Top- k 66)	(Top- k 51)	(Top- k 13)
9	Özer et al. (2011)									
		17.77	18.00	17.94	16.72	16.93	14.33	17.97	12.78	13.78
		(Top- p 0.77)	(Top- p 0.65)	(Top- p 1.0)	(Top- k 55)	(Top- k 14)	(Top- k 26)	(Min- p 0.11)	(Min- p 0.05)	(Min- p 0.11)
10	Schweitzer and Cachon (2000)									
		15.21	16.00	15.80	15.92	15.18	14.23	16.93	15.08	14.33
		(Min- p 0.16)	(Top- p 0.81)	(Min- p 0.16)	(Top- p 0.59)	(Top- k 13)	(Top- k 13)	(Min- p 0.16)	(Min- p 0.09)	(Min- p 0.11)
11	tpt fixed fee									
		1.43	1.12	0.69	0.58	0.53	0.58	0.63	0.67	0.69
		(Top- k 40)	(Top- k 66)	(Top- k 88)	(Top- k 19)	(Top- p 0.94)	(Min- p 0.11)	(Min- p 0.16)	(Min- p 0.20)	(Min- p 0.20)
12	tpt price a									
		2.22	1.60	1.01	0.84	0.70	0.68	0.71	0.70	0.70
		(Top- k 40)	(Top- k 32)	(Top- k 43)	(Top- k 25)	(Top- k 43)	(Min- p 0.05)	(Min- p 0.05)	(Min- p 0.16)	(Min- p 0.20)
13	q00									
		3.66	3.64	3.34	3.24	3.06	2.92	3.14	3.09	3.53
		(Top- p 0.93)	(Top- k 11)	(Top- k 36)	(Top- p 1.0)	(Top- p 1.0)	(Top- p 0.65)	(Min- p 0.11)	(Min- p 0.16)	(Min- p 0.16)
14	q50									
		1.64	1.21	0.99	1.00	0.93	0.99	0.99	0.95	1.04
		(Top- k 37)	(Top- k 36)	(Top- k 45)	(Top- k 65)	(Min- p 0.16)	(Top- k 27)	(Top- k 47)	(Min- p 0.16)	(Min- p 0.11)
15	c1									
		0.14	0.02	0.02	0.05	0.11	-	-	-	-
		(Top- p 1.0)	(Top- k 27)	(Top- k 16)	(Top- p 1.0)	(Top- p 0.65)	-	-	-	-
16	c5									
		0.21	0.22	0.22	0.22	0.22	-	-	-	-
		(Min- p 0.16)	(Top- p 0.50)	(Top- p 0.59)	(Top- p 0.50)	(Top- p 0.51)	-	-	-	-
17	Low Profit									
		4.77	4.53	4.28	3.98	3.93	4.23	3.53	2.72	2.61
		(Top- k 42)	(Top- k 78)	(Top- k 71)	(Top- p 1.0)	(Top- k 16)	(Top- p 0.56)	(Min- p 0.05)	(Min- p 0.05)	(Min- p 0.07)
18	High Profit									
		40.90	40.10	41.50	42.21	40.09	44.04	45.77	49.93	52.57
		(Top- k 11)	(Min- p 0.16)	(Top- p 0.81)	(Min- p 0.16)	(Top- p 1.0)	(Top- k 76)	(Min- p 0.20)	(Min- p 0.11)	(Min- p 0.16)
19	signal									
		58.58	56.60	56.49	51.78	52.66	51.80	50.71	54.08	53.11
		(Top- k 40)	(Min- p 0.11)	(Top- k 25)	(Top- k 54)	(Top- k 20)	(Min- p 0.11)	(Top- k 63)	(Min- p 0.11)	(Min- p 0.11)
20	capacity									
		35.45	36.75	32.53	32.29	22.76	24.08	17.34	21.27	17.41
		(Top- k 40)	(Min- p 0.11)	(Top- p 0.81)	(Top- p 1.0)	(Top- p 1.0)	(Top- k 27)	(Top- k 73)	(Min- p 0.11)	(Min- p 0.11)
21	capacity									
		28.19	27.03	28.21	30.84	30.82	-	48.95	58.30	-
		(Top- k 39)	(Top- k 74)	(Top- p 0.65)	(Top- p 0.65)	(Top- p 0.65)	-	(Min- p 0.16)	(Min- p 0.16)	-
22	capacity									
		44.32	35.88	28.46	23.03	20.72	-	23.45	13.20	-
		(Top- k 67)	(Top- k 19)	(Top- k 10)	(Top- k 33)	(Top- k 32)	-	(Min- p 0.16)	(Min- p 0.16)	-

Table EC.6 Wasserstein Distance under Different Temperatures by Qwen-32B

Number	Paper/Feature	0.5	1	1.5	1.7	1.9	2.5	3.0	3.5	4.0
1	Chen et al. (2013)									
	C	3.06 (Top-k 32)	2.30 (Top-k 62)	2.27 (Top-k 80)	2.15 (Top-k 69)	2.07 (Top-k 70)	2.22 (Min-p 0.16)	2.09 (Min-p 0.05)	1.75 (Min-p 0.05)	1.83 (Min-p 0.16)
	O	1.88 (Top-p 0.59)	1.89 (Top-p 0.65)	1.94 (Top-k 74)	1.95 (Top-p 0.50)	1.89 (Top-p 0.59)	2.12 (Top-k 12)	2.12 (Min-p 0.16)	-	-
2	Croson and Donohue (2006)									
	Base	8.04 (Top-k 32)	7.95 (Top-k 64)	7.77 (Top-k 85)	7.61 (Top-k 18)	7.47 (Top-k 66)	6.64 (Top-k 12)	6.98 (Min-p 0.07)	5.93 (Min-p 0.07)	5.28 (Min-p 0.11)
	Share	9.24 (Top-k 32)	8.87 (Top-k 66)	7.26 (Top-k 90)	6.69 (Top-k 20)	6.10 (Top-k 67)	6.97 (Min-p 0.16)	6.93 (Min-p 0.16)	27.69 (Min-p 0.16)	134.67 (Min-p 0.20)
3	Davis et al. (2011)									
	1	7.68 (Min-p 0.11)	7.99 (Min-p 0.05)	7.33 (Top-p 1.0)	7.22 (Top-k 55)	6.88 (Top-k 17)	5.64 (Top-k 39)	4.84 (Top-k 28)	3.74 (Top-k 32)	3.16 (Top-k 16)
	4	10.04 (Top-k 40)	9.63 (Top-p 0.81)	9.15 (Top-p 1.0)	9.14 (Top-k 66)	8.83 (Top-k 16)	8.24 (Top-k 49)	7.24 (Top-k 41)	6.48 (Top-k 11)	5.27 (Top-k 66)
	7	17.69 (Top-p 0.81)	17.22 (Top-k 16)	16.49 (Top-k 40)	16.69 (Top-k 90)	16.33 (Top-k 32)	15.50 (Top-k 65)	13.78 (Top-k 19)	12.69 (Top-k 23)	11.34 (Top-k 21)
	10	23.56 (Top-p 0.81)	23.60 (Min-p 0.11)	22.53 (Top-k 60)	22.80 (Top-k 46)	22.32 (Top-k 18)	21.09 (Top-k 73)	19.23 (Top-k 26)	17.74 (Top-k 23)	16.01 (Top-k 12)
4	Engelbrecht-Wiggans and Katok (2008)									
	both	17.65 (Min-p 0.16)	17.65 (Top-p 0.50)	17.72 (Top-p 1.0)	17.36 (Top-p 1.0)	17.08 (Top-p 1.0)	16.64 (Top-k 33)	16.37 (Top-k 65)	16.21 (Top-k 25)	16.15 (Min-p 0.05)
	loser	17.08 (Top-p 0.59)	17.03 (Top-p 0.59)	16.99 (Top-k 85)	16.81 (Top-k 19)	16.76 (Top-k 66)	16.57 (Top-k 40)	14.82 (Top-p 0.84)	15.67 (Top-k 88)	15.34 (Top-k 38)
5	Ho and Zhang (2008)									
	qd x	1.53 (Top-k 10)	1.43 (Top-p 1.0)	1.33 (Top-p 1.0)	1.28 (Top-k 11)	1.23 (Top-k 66)	1.08 (Top-k 42)	0.87 (Top-k 65)	0.76 (Top-k 67)	0.65 (Top-k 26)
	qd y	6.79 (Top-p 0.59)	7.91 (Top-k 26)	7.42 (Top-k 89)	7.13 (Top-k 56)	7.07 (Top-k 67)	4.98 (Top-p 0.97)	1.50 (Top-p 1.0)	1.41 (Top-p 0.81)	7.89 (Min-p 0.16)
	tpt fixed fee	0.81 (Top-p 0.67)	0.83 (Top-p 0.59)	0.81 (Top-p 0.60)	0.88 (Top-p 0.56)	0.93 (Top-p 0.56)	0.70 (Top-p 1.0)	0.45 (Top-p 1.0)	1.80 (Top-k 46)	1.58 (Top-k 87)
	tpt price a	0.61 (Top-k 40)	0.46 (Top-k 20)	0.39 (Top-p 0.75)	0.38 (Top-k 55)	0.37 (Top-k 40)	0.35 (Top-k 11)	0.33 (Top-k 14)	0.37 (Top-k 17)	0.37 (Min-p 0.11)
6	Kremer and Debo (2016)									
	q00	0.22 (Top-k 10)	0.19 (Top-k 29)	0.16 (Top-k 44)	0.15 (Top-k 67)	0.14 (Top-k 10)	0.12 (Top-k 43)	0.12 (Min-p 0.05)	0.12 (Min-p 0.11)	0.10 (Min-p 0.05)
	q50	0.14 (Top-p 0.59)	0.15 (Top-k 27)	0.14 (Min-p 0.11)	0.15 (Top-k 65)	0.14 (Top-k 65)	0.15 (Top-p 1.0)	0.14 (Top-p 0.50)	0.14 (Min-p 0.16)	0.14 (Min-p 0.05)
7	Kremer et al. (2011)									
	c1	3.86 (Top-k 40)	3.78 (Top-k 20)	3.63 (Top-p 1.0)	3.55 (Top-k 90)	3.51 (Top-p 1.0)	3.23 (Top-k 21)	2.89 (Top-k 31)	3.21 (Min-p 0.05)	3.22 (Min-p 0.16)
	c5	13.64 (Top-k 40)	13.55 (Top-k 44)	13.47 (Min-p 0.16)	13.23 (Top-k 51)	13.10 (Top-k 56)	12.69 (Top-k 15)	11.81 (Top-k 36)	10.62 (Top-k 74)	12.76 (Min-p 0.16)
8	Schweitzer and Cachon (2000)									
	Low Profit	33.93 (Top-k 40)	29.68 (Top-p 0.52)	22.75 (Top-k 45)	28.80 (Top-k 19)	23.23 (Top-k 16)	22.71 (Top-k 10)	22.06 (Top-k 67)	19.31 (Top-k 16)	20.83 (Top-k 10)
	High Profit	44.90 (Top-k 32)	45.33 (Top-k 64)	45.51 (Top-k 14)	44.71 (Top-k 13)	43.49 (Top-k 67)	43.37 (Top-k 14)	31.97 (Top-k 26)	39.11 (Top-k 45)	37.85 (Min-p 0.16)
9	Özer et al. (2011)									
	signal	32.32 (Top-k 32)	31.03 (Top-k 84)	28.58 (Top-k 32)	29.04 (Top-k 55)	26.63 (Top-k 32)	24.56 (Top-k 33)	24.34 (Top-k 40)	23.22 (Min-p 0.16)	20.87 (Min-p 0.11)
	capacity	31.86 (Top-k 77)	35.57 (Top-p 0.65)	36.36 (Min-p 0.05)	36.78 (Top-p 0.65)	35.19 (Top-p 0.72)	40.86 (Top-p 0.50)	49.68 (Min-p 0.16)	49.82 (Min-p 0.16)	55.58 (Min-p 0.16)

EC.4.4. LLaMA-70B Word Output Distribution

We employ the CoT prompting of LLaMA-70B; accordingly, Figures EC.5 and EC.6 depict the distributions of human and LLM decision frequencies across six temperature settings in nine BOM experiments. These results primarily highlight the distributional differences between human- and LLM-generated data. While the outputs of LLMs tend to concentrate around a limited set of values, human responses exhibit a more dispersed pattern. Furthermore, we observe that the output distribution varies with temperature, which motivates our quantitative analysis using the Wasserstein distance and subsequent calibration through temperature tuning. For some experiments (e.g., Kremer and Debo (2016) in EC.5), we only report distributions under three temperature settings (0.5, 1.0, and 1.5) due to inappropriate outputs in the remaining settings within the three available attempts.

Figure EC.5 LLaMA-70B Word Probability Distribution: Replication Results of Conditions from Chen et al. (2013), Croson and Donohue (2006), Davis et al. (2011) and Engelbrecht-Wiggans and Katok (2008)

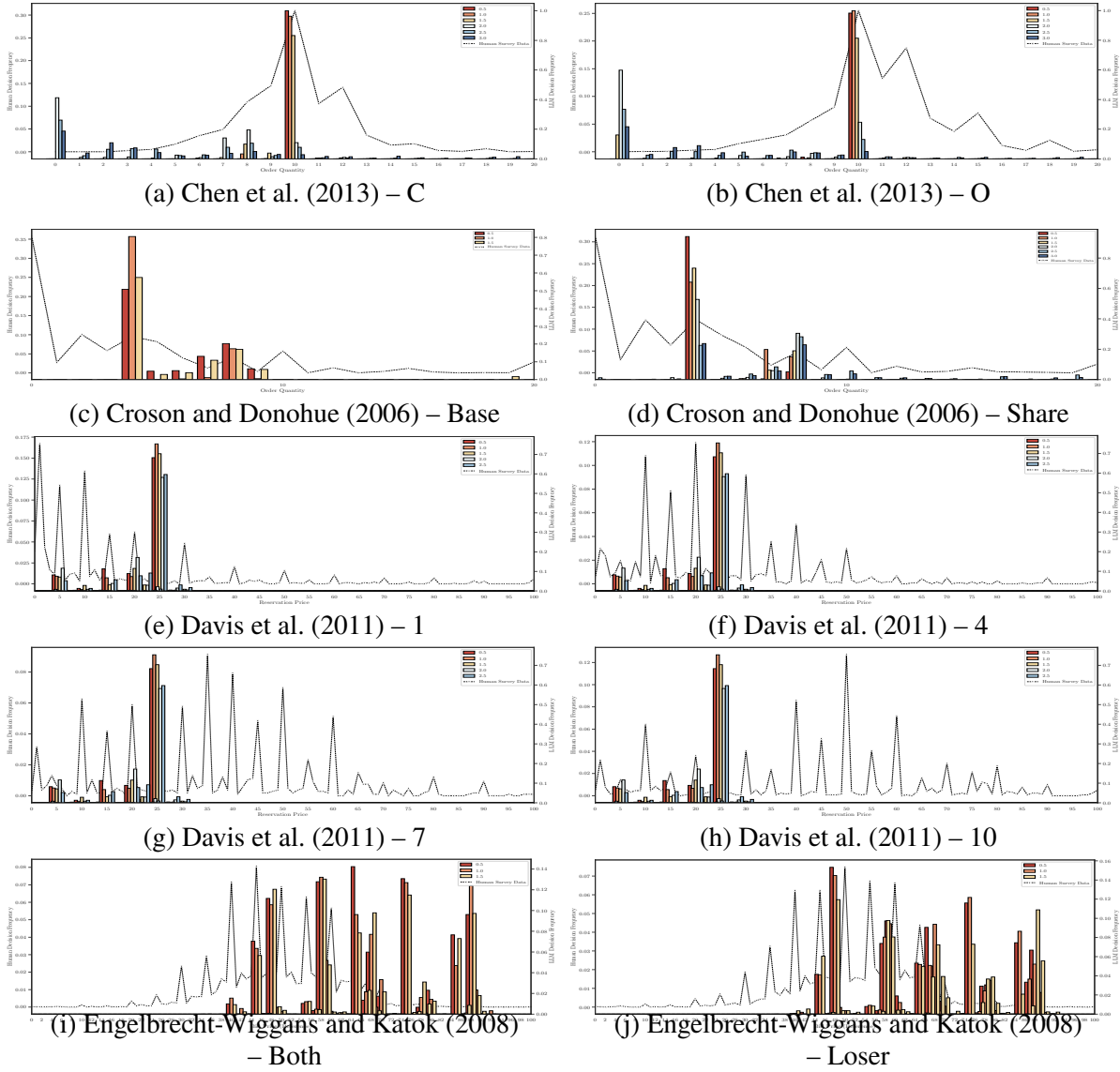


Figure EC.6 LLaMA-70B Word Probability Distribution: Replication Results of Conditions from Ho and Zhang (2008), Kremer and Debo (2016), Kremer et al. (2011), Schweitzer and Cachon (2000) and Özer et al. (2011)

