Natural Language Processing and Game Theory Converge: An Investigation of Rational Bidding in Second-Price Auctions

Roy Ludan (ID. 032736233), Ofri Hefetz (ID. 209028067)

Submitted as final project report for the NLP course, IDC, 2024

1 Introduction

Auctions represent a fascinating domain where economic theory and strategic decision-making converge. The second-price auction holds a unique position among various auction types due to its distinctive rules and implications for bidder behavior. In a second-price auction, each bidder submits a bid without knowing the others' bids; the highest bidder wins, but crucially, they pay the amount of the second-highest bid. This mechanism theoretically simplifies the bidding strategy: Participants should bid their true valuation of the item, as overbidding does not increase the chance of winning but can result in an overpayment, and underbidding risks losing the auction despite a willingness to pay more than the winning price. The intersection of game theory and NLP allows us to explore whether Large Language Models (LLMs) can simulate rational agents in such settings. Specifically, it aims to investigate whether these models can exhibit rational bidding behavior consistent with economic theory and understand the decision-making process through Chain-of-Thought (CoT) prompting.

2 Literature Review

1. In the paper *"Put Your Money Where Your Mouth Is: Evaluating Strategic Planning and Execution of LLM Agents in an Auction Arena"* by Chen et al., the authors constructed a novel simulation environment for evaluating LLMs within auctions, a setting chosen for its unpredictability and the requirement of skills related to resource and risk management. They conducted several controlled simulations using state-of-the-art LLMs as bidding agents. Their conclusions regarding LLM agency are largely tied to the type of LLM, with some models showing greater planning and adaptability traits, while GPT-4 is notably identified as the superior model.
2. In *"STEER: Assessing the Economic Rationality of Large Language Models,"* Raman et al. introduce the 'STEER' benchmark to quantitatively assess LLMs' performance as economic decision-making agents, evaluating traits like prompting, introspection, and chain-of-thought reasoning. They found that the GPT-4 (Turbo) model was the most accurate overall. The paper also discusses specific evaluation metrics, including a second-price bid element, though its overall importance in the assessment is not clearly defined.

3 Methodology

1. Our experimentation framework included 5 different LLMs:
   (a) Mistral (Mistral-7B-v0.1)
   (b) LLAMA (Meta-Llama-3.1-8B)
   (c) T5 (flan-t5-large)
   (d) Gemini
   (e) GPT-4o
2. All LLMs have been given a prompt with either a zero-shot bid request or a one-shot bid request, which means that each prompt contains an example of a possible response of what a typical bid should look like.
3. All LLMs have been prompted with 4 types of items depending on their price: small, medium, large, and extra-large.
4. To handle model hallucinations, we ran each experiment as many times as needed until we secured at least 30 consistent responses.
5. All experiments were conducted in sets of 10 consecutive prompts.
6. After each prompt execution the experiment was restarted, thus there was no context window effect between the various prompts/experiments.
7. We used the following constant parameters for each model:
   (a) max_new_tokens = 150
   (b) do_samples = True
   (c) temperature = 0.7
   (d) top_k=50
   (e) top_p=0.9
8. Link to the code used to conduct the experiments can be found in the references section of this paper.
9. Responses have either:
   (a) Qualified and then classified as either:
      i. Dominant strategic response: If the responded bid is equal to the item's value given in the prompt AND the CoT reasoning for the bid included a rational reason for the value given.
      ii. Non-dominant strategic response: If the responded bid does NOT equal to the item's value in the prompt OR if the CoT reasoning for the bid does not include a rational reason for the given bid.
   (b) Disqualified due to a hallucination/incorrect response, such as (but not limited to):
      i. Details fabrication.
      ii. Irrational use of formulas.
      iii. Repeated response with no CoT reasoning
      iv. Answering a question

4 Results

4.1 Non-descriptive prompts

4.1.1 Zero-shot prompts

1. Most models have performed well for the task of bidding for various items using no example in the prompt (e.g. "*Your value for the item is 1 USD. This is all the information you have (1) How much will you bid? Just return your bid and no other info. (2) Briefly explain your reasoning.*"), although we can also see in figure 1 that some models have been more successful and consistent than others throughout the experiments with model GPT-4o being the notable leader on both response results and CoT reasoning.

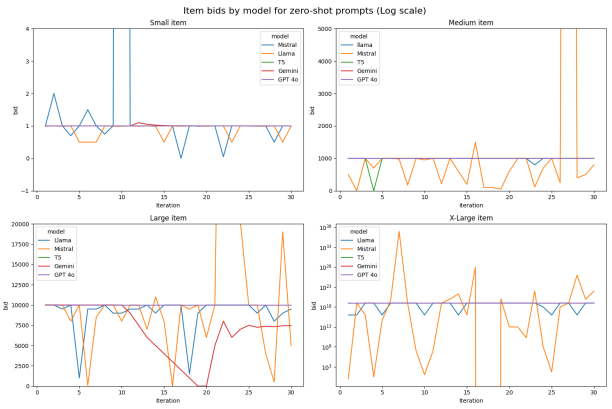

Figure 1: Item bids by model for zero-shot prompts (Log scale)

2. We found that some variability exists between model responses as we increase the price of the item, and we assume that this is probably due to an insufficient amount of training data that contains items in the 100 Billion USD range.

3. When we look at the number of responses categorized as dominant strategic responses with rational bidding behavior, the image becomes even clearer, with once again the GPT-4o model being the notable winner and the Gemini model with some good bid and reasoning for some categories.

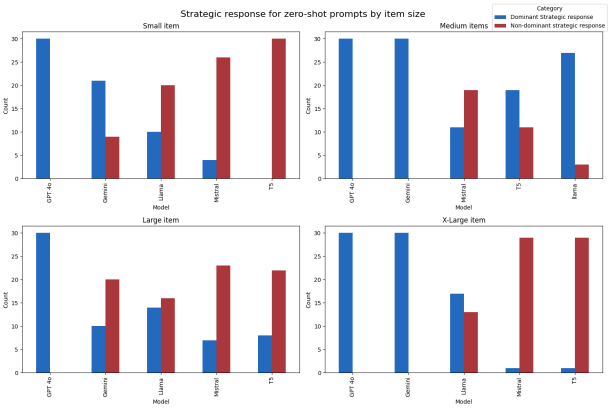

Figure 2: Strategic response for zero-shot prompts by item size

4.1.2 One-shot prompts

1. One-shot prompts improve model responses drastically, as we can see almost no variability between models except for the X-Large item section.

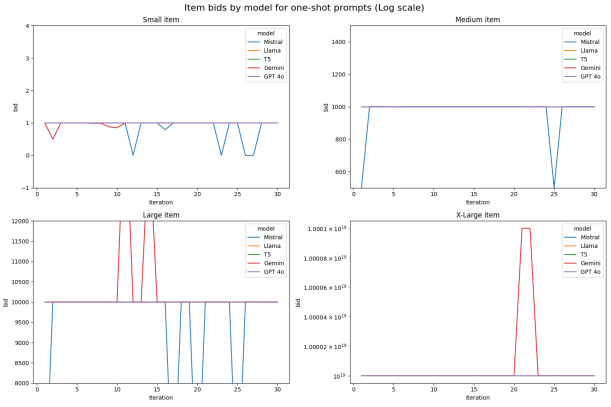

Figure 3: Item bids by model for one-shot prompts (Log scale)

2. Here, we can also see that the GPT-4o model is notably the best model as it returns a set of consistent results with a rational bidding behavior in its CoT reasoning.

4.1.3 Comparing zero-shot and one-shot results

1. In order to find whether a true improvement has been made between the zero-shot and the one-shot experiments, we used a T-test statistic that will allow us to quantitatively measure whether there's a significant statistical difference between the two samples.

item size	t-statistic	p-value
Small	1.00529	0.315575
Medium	0.815589	0.415387
Large	-0.665192	0.506442
X-Large	1	0.318122

Table 1: T-test results between the zero-shot and one-shot experiments by item size

Table 1 shows us that in no category the p-value is significantly smaller than 0.05, and thus, we cannot reject the null hypothesis of no significant difference between the two samples.

This finding can also be supported when we inspect the difference in STD of the categories between the two samples in Table-2.

Category	GPT 4o	Gemini	Llama	Mistral	T5
Small Zero-shot	0	0.0203463	0.201774	18.0876	0
Medium Zero-shot	0	0	10836.4	182.574	36.5031
Large Zero-shot	0	3162.59	2178.83	14113.5	0
X-Large Zero-shot	0	0	4.44094e+18	1.82574e+36	0
Small One-shot	0	0.0958261	0	0.345079	0
Medium One-shot	0	0	0	126.836	0
Large One-shot	0	1149.21	0	1728.7	0
X-Large One-shot	0	2.53699e+14	0	0	0

Table 2: Standard deviation of bids by model

4.2 Descriptive prompts

By using descriptive prompts (e.g. "*The item: Plackers Micro Mint Dental Floss Picks with Travel Case, 12 Count. Your value for the item is 1 USD. This is all the information you have: (1) How much will you bid? Just return your bid and no other info. (2) Briefly explain your reasoning.*") We wanted to test whether the addition of more information to the auctioned item would cause the models to respond with more dominant strategic responses. This experiment was also conducted with zero-shot and one-shot prompts. We selected 4 items that will serve as the appropriate auctioned items and added a description to each prompt.

4.2.1 Zero-shot prompts

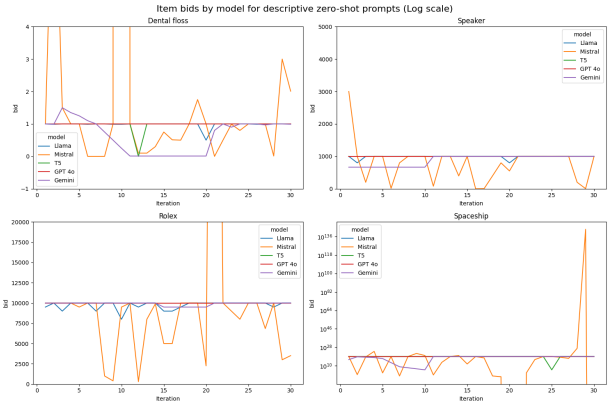

Figure 5: Item bids by model for descriptive zero-shot prompts (Log scale)

1. Figure 5 shows an interesting insight where most models generally deviate from the item's true value downwards, meaning most of them

4.2.3 Comparing descriptive zero-shot and descriptive one-shot results

1. Table 3 only shows a meaningful statistical difference in the 'Large' category between the two prompt types.

item size	t-statistic	p-value
Small (Dental floss)	0.998085	0.319048
Medium (Speaker)	-2.18894	0.029377
Large (Rolex)	0.266684	0.789897
X-Large (Spaceship)	1	0.318122

Table 3: T-test results between the descriptive zero-shot and descriptive one-shot results

Category	GPT 4o	Gemini	Llama	Mistral	T5
Dental floss Zero-shot	0	0.507258	0.0910507	368.287	0.182574
Speaker Zero-shot	0	159.82	50.732	576.896	0
Rolex Zero-shot	0	203.419	486.394	20787.3	0
Spaceship Zero-shot	0	4.49655e+18	0	3.65148e+142	1.82574e+18
Dental floss One-shot	0	0	0	0.351107	0
Speaker One-shot	0	113.427	0	126.854	0
Rolex One-shot	0	0	0.479463	1525.63	1525.64
Spaceship One-shot	0	0	0	1.82574e+30	0

Table 4: Standard deviation of bids by model

2. Table 4 also shows an improvement in overall STD as we can see a decrease in overall STD values across the different categories.

---

5 Responses analysis

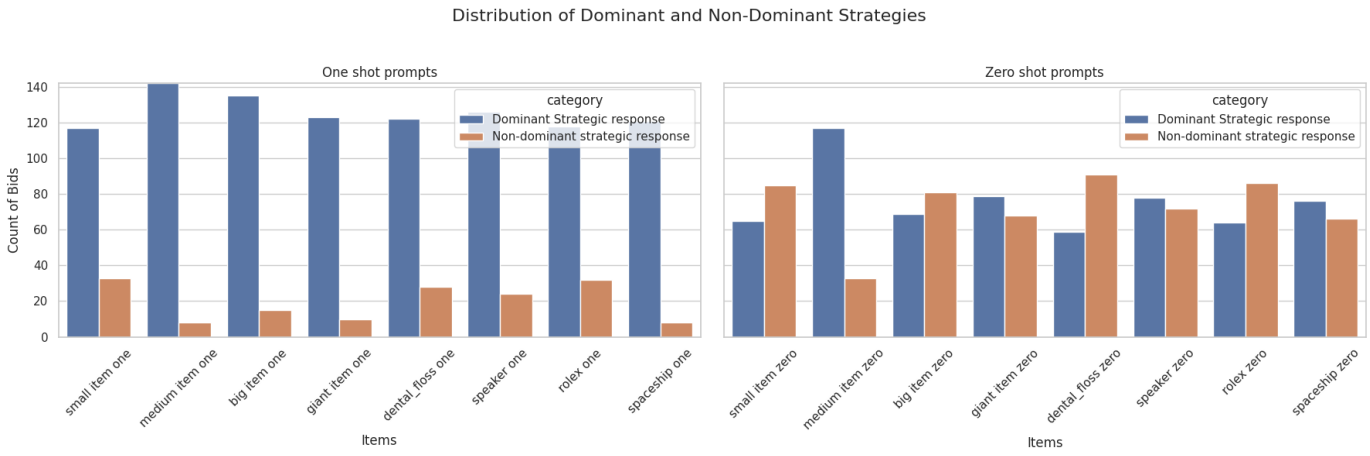

Figure 9: Distribution of dominant and non-dominant strategies among all the items.

1. From Figure 9, we can conclude some key points:

(a) Effectiveness of One-Shot Prompts: The graph on the left, representing one-shot prompts, shows a strong preference for dominant strategic responses across all items. This suggests that when the LLMs are given an example, they are more likely to follow the dominant strategy: bid their true valuation in a second-price auction. This indicates that the models better understand the auction dynamics when provided with an example.

(b) In contrast, the graph on the right, representing zero-shot prompts, shows a more balanced distribution between dominant and non-dominant strategies. This indicates that LLMs need an example to follow the dominant strategy consistently. The variability in responses suggests that LLMs may struggle to apply the dominant strategy consistently when they must infer the appropriate behavior without explicit guidance.

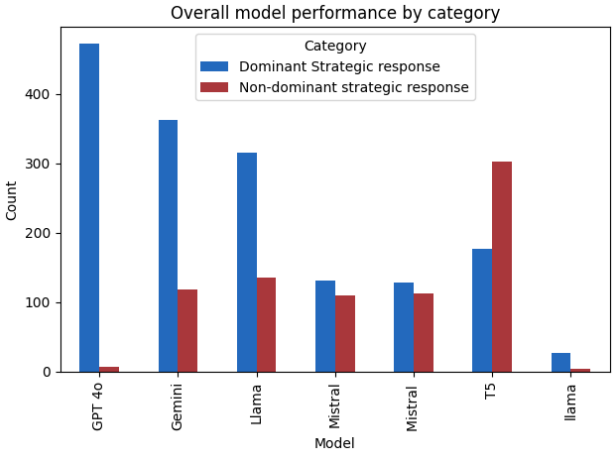

Figure 10: Overall model performance by category

2. From figure 10 we can conclude:

(a) Dominance of Strategic Bidding: The GPT 4o model demonstrates a clear tendency towards strategic bidding across all categories. This suggests that GPT 4o is better at understanding the underlying game theory principles and adapting its bidding strategy to maximize its potential gains.

(b) Llama's Conservatism: The Llama model shows a notable inclination towards underbidding. This conservative approach might stem from a risk-averse strategy or potentially a less sophisticated understanding of the auction dynamics.

(c) Model-Specific Biases: The observed biases in bidding behavior suggest that different language models possess distinct risk profiles and decision-making processes.

Top five	1	2	3	4	5
small item one	bid	usd	bidding	price	second
medium item one	bid	second	price	valuation	bidding
big item one	bidding	bid	second	valuation	price
giant item one	bid	price	bidding	auction	win
small item zero	bid	usd	second	bidding	value
medium item zero	bid	valuation	second	bidding	auction
giant item zero	bid	valuation	second	true	auction
dental_floss zero	bid	item	usd	value	auction
spaceship zero	bid	valuation	second	auction	bidding
dental_floss one	second	bid	usd	bidding	true
speaker one	bid	second	bidding	valuation	true
rolex one	bid	second	bidding	true	highest
spaceship one	bid	second	true	bidding	auction

Table 5: Top Five Most Frequent Bidding Terms by Category

3. Table 5 provides valuable insights that enhance our understanding of the response analysis (the full results can be found in the project git). An interesting observation is the absence of specific item names like "Rolex," "Dental Floss," "Spaceship," and "Speaker" from the top 5 most frequent words (and even in the top 10) for these respective items. This absence suggests that the LLMs may prioritize reasoning based on general auction principles rather than focusing on the unique attributes of each item. This could be because the models are trained on a wide variety of text data and may, therefore, generalize across different contexts, emphasizing universal auction-related terms such as "valuation," "bid," and "price" instead of specific item names.

6 Conclusion

In this paper, we have investigated the potential of Large Language Models (LLMs) to simulate rational bidding behavior in second-price auctions. Through a series of experiments, we have explored the bidding strategies of five different LLMs, varying in size and complexity, and evaluated their ability to exhibit strategic behavior consistent with economic theory.

Our results indicate that, overall, the GPT-4o model (Figure 10) showed the most consistent and rational bidding behavior across all types of prompts, followed by the Gemini model. We also found that providing contextual information about the item improved all models' rational bidding behavior, suggesting that richer contextual information helps LLMs to better assess the item's value and formulate more informed bidding strategies.

We also noted some model-specific biases and inclinations towards underbidding or risk-averse behavior, highlighting the need for carefully assessing model performance in different domains and contexts.

Overall, our study suggests that the intersection of game theory and NLP holds significant potential for understanding and improving strategic decision-making in various domains, with auctions being just one example.

7 Future directions

1. Token Importance and Bid Correlation: Investigate the relationship between the importance or attention weights assigned to different tokens in the prompt and the resulting bid. We can better understand the internal decision-making process by analyzing how the model's attention distribution affects its bidding decisions. This analysis could also guide the optimization of prompt structures to produce more accurate and strategically sound bids.

2. Fine-grained Analysis: A more detailed analysis could explore the correlation between model size, training data, and bidding behavior.

3. Real-World Applications: Exploring the potential of integrating these models into real-world auction platforms could enhance bidding efficiency and fairness. By automating decision-making, reducing biases, and personalizing strategies, these models can make auctions more accessible and equitable. However, care-

ful attention to transparency, regulatory compliance, and technical implementation is essential for successful integration.

4. Experiment with different types of prompts:**(e.g., few-shot or more complex scenarios) to see how they influence the LLMs' strategic choices. Additionally, analyzing why certain items in the zero-shot scenario still resulted in dominant strategies could provide insights into the underlying factors influencing LLMs' decision-making.
5. Pronoun Analysis: Examines the use of pronouns in the model's reasoning to determine whether the model refers to itself or another entity when making a bid. Specifically, we can analyze (1) First-Person Pronouns (e.g., "I," "my," "we") and (2) Third-Person References (e.g., "the bidder," "they," "he/she"). By comparing bidding strategies in these two contexts, we can identify whether the model behaves differently when it is the bidder versus when advising another bidder. This could reveal underlying biases or variations in the model's decision-making process based on its perceived role.
6. Ensemble Methods: 1) Majority voting: Aggregate the bids or strategies proposed by different models and choose the most common or average bid. This method can help smooth out inconsistencies and capitalize on the collective wisdom of multiple models. 2) Weighted averaging: Assign weights to each model based on their past performance or confidence levels in certain types of auctions. Bids from higher-performing models would have more influence on the final decision.

8 References

1. Put Your Money Where Your Mouth Is: Evaluating Strategic Planning and Execution of LLM Agents in an Auction Arena by Chen et al. https://openreview.net/pdf?id=crMMk4I8Wy
2. STEER: Assessing the Economic Rationality of Large Language Models by Raman et al. https://arxiv.org/abs/2402.09552
3. NLP Final project experiments notebook.
4. Code used to run the experiments in this paper