

LaMP-Val: Large Language Models Empower Personalized Valuation in Auction

Jie Sun^{1,2}, Tianyu Zhang¹, Houcheng Jiang¹, Kexin Huang¹,
Xiang Shu³, Zhibo Zhu⁴, Lintao Ma⁵, Xingyu Lu⁶,
Jun Zhou⁷, Junkang Wu¹, Chi Luo¹, An Zhang¹,
Jiancan Wu^{1*}, Xiang Wang^{1*}

¹ University of Science and Technology of China ² Shanghai Innovation Institute
³ East China Normal University ⁴ Xi'an Jiaotong University
⁵ Ocean University of China ⁶ Nanjing University ⁷ Zhejiang University
{sunjie2019, tianyuz1, janghc, huangkx}@mail.ustc.edu.cn

Abstract

Auctions are a vital economic mechanism used to determine the market value of goods or services through competitive bidding within a specific framework. However, much of the current research primarily focuses on the bidding algorithms used within auction mechanisms. This often neglects the potential benefits of incorporating individual users' unique preferences into the valuation process. Our theoretical and empirical analysis demonstrates that valuation errors can significantly impact the overall utility. To bridge this gap, we propose a personalized valuation framework, namely Large Language Models-powered Personalized Valuation (LaMP-Val), which integrates Large Language Models to incorporate personalized semantic preference into users valuation process. LaMP-Val integrating three components: data, learning, and evaluation. The data component tackles the challenge of building a novel dataset specifically for LLMs fine-tuning in personalized valuation modeling. The learning component introduces a diversity template to enhance LLMs' capacity for modeling fine-grained personal valuation patterns. The evaluation component establishes a closed-loop system where LLM-generated valuations interact with bidding strategies and auction. It proposes two novel metrics to quantify valuation precision and bidding intention accuracy in personalized scenarios. Extensive experiments show that LaMP-Val more accurately captures personalized values and achieves greater profits than baseline approaches.

1 Introduction

An auction, covering valuation and bidding, is a crucial economic mechanism that helps determine the market value of commodities or services

*Corresponding authors.

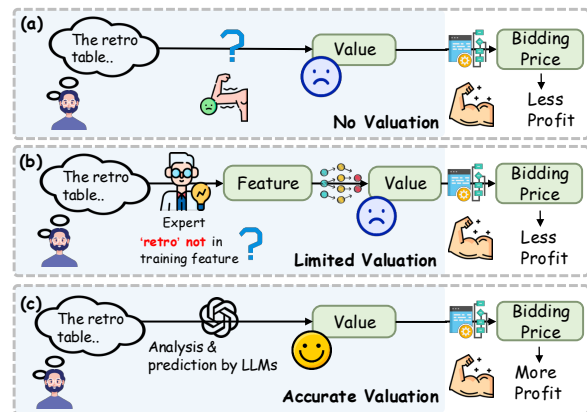


Figure 1: (a) Existing works mainly focus on bidding algorithms (from value to bidding price) but neglect the valuation process (from user needs to determine value). (b) Existing works use experts to generate features for predicting values, but are limited to fixed features (e.g., “retro” not in training feature). (c) Using LLMs to analyze semantic information to predict value, accurately capturing user preferences.

through competitive bidding (Weber, 2003; Aggarwal and Badanidiyuru, 2024; Chen and Nabi, 2023). Valuation is the process by which bidders assess an item’s worth of an item based on their individual needs and the product description (Zhang and Niu, 2023). Bidding, on the other hand, refers to a competitive process in which participants try to win ownership of the items and maximize their profits by developing strategic algorithms that comply with specific auction rules (Klemperer, 1999). The rise of online exchange platforms has broadened the use of auctions beyond traditional advertising. Nowadays, auctions are commonly utilized for selling data (Janin and Qin, 2020; Travizano and Sarraute, 2018) and second-hand goods (Han and Yin, 2020, 2019; Yamaura and Kanemaki, 2019). This rapid expansion underscores the need for improved methods that help users achieve fair value based on

their individual needs.

As in Figure 1(a), most auction studies (Schlosser and Boissier, 2018; Balseiro and Deng, 2021; Bachrach and Talgam-Cohen, 2022; Golrezaei and Sahoo, 2024; Hajiaghayi and Lahaie, 2024b) focus on theoretical bidding strategies and mechanisms but overlook individual valuation, despite their theoretical robustness. However, our preliminary experiments (Sec 4.1) indicates that 1% valuation errors result in approximately 10% utility losses, demonstrating that valuation errors significantly affect final utility.

While existing studies (Zhang and Niu, 2023; Leme and Pál, 2016; Peri and Curry, 2021) have investigated feature-based personalized valuation, these efforts rely on expert-engineered features for training valuation models. Such approaches face challenges in scenarios requiring fine-grained user preference modeling, particularly when processing unstructured user-generated content. These synthetic, feature-based methods inherently struggle with representation inadequacy when managing user-provided textual descriptions that go beyond predefined feature boundaries as shown in Figure 1(b). This limitation leads to valuation errors, which ultimately reduce the final profit. Moreover, traditional advertising auction metrics, such as utility and value (Lv and Zhang, 2022), tend to focus solely on economic gains. They often overlook the emotional factors that influence user decisions, like collectible value and sentimental attachment. To our knowledge, there is currently no systematic text-based method that addresses individual preferences for item valuation, nor are there established evaluation metrics that correspond to these preferences.

The advanced capabilities of Large Language Models (LLMs) in semantic comprehension and following instructions (Brown and Mann, 2020; OpenAI, 2023) make them promising for capturing personalized preferences, such as interpreting ambiguous descriptions (e.g., “retro style”). State-of-the-art closed-source models, such as Gemini (Team and Georgiev, 2024) and o3 (Pfister and Jud, 2025), demonstrate strong reasoning abilities. However, their dependency on API-based access necessitates data transmission to cloud servers, which poses significant risks to privacy. In personal auction scenarios, the potential misuse of sensitive user information, such as transaction histories, severely limits the use of closed-source models in contexts where privacy is a concern. Therefore,

local deployment is essential to meet the requirements for privacy preservation. While open-source models like DeepSeek R1 (DeepSeek-AI, 2025) show similar reasoning performance, their substantial computational resource requirements lead us to seek more lightweight and efficient models. Central to this investigation is the crucial question:

How can we use LLMs to model personalized preferences and achieve accurate product valuation?

In this paper, we introduce a framework called Large Language Models-powered Personalized Valuation (LaMP-Val) to address this problem. LaMP-Val consists of three main components: data, learning, and evaluation. The data module addresses the critical challenge of constructing a novel dataset for LLM fine-tuning in personalized valuation modeling, a domain previously hindered by three fundamental limitations: the value-price paradox (Kehoe, 1989), preference distribution skewness (De Langhe and Fernbach, 2016), and rationale absence in economic decision traces (Wei and Wang, 2022). The learning component develops a diversity template to fine-tune LLMs that enables LLMs to model nuanced personal valuation patterns, overcoming traditional approaches’ inability to model item valuation influenced by personalized preference. The evaluation module pioneers a closed-loop system where LLM-generated personalized valuations dynamically interact with bidding strategies and market environments. To address the inadequacy of conventional metrics in personalized scenarios, we propose Personalized Utility (PU) and Personalized Value (PV) These metrics are designed to assess both the precision of users’ personalized valuations and the accuracy of their decision-making regarding bidding intentions. Extensive experiments demonstrate that our methods can achieve significant profits compared to baseline approaches. Our codes are available at <https://github.com/sunjie279/LaMP-Val>.

The following sections introduce the preliminary material, our proposed framework, present and analyze the empirical results, related works, and finally conclude the paper.

2 Preliminary

The Vickrey auction, a distinctive variant of the sealed-bid auction (Liu and Wu, 2021), is widely used due to its incentive-compatible design. This auction model has garnered significant interest from both academia (M and Paul, 2006; Zhang

and Zhang, 2015; Huang et al., 2025) and industry (Sureka and Wurman, 2002; Bikhchandani and De Vries, 2011), as it encourages bidders to place honest bids based on their item’s valuation, as the winning bidder pays the second-highest bid, effectively mitigating the “winner’s curse” of potentially overpaying (Karl, 2016).

We will now define the key notations related to personalized valuation. Consider a user participating in an auction system with M items and a total budget constraint of B . Let B_m represent the remaining budget when valuing the m -th item. Each item m is linked to a personalized preference signal s_m , which captures relevant information and reviews about the item.

The valuation model \mathcal{V} processes these preference signals to jointly generate dual outputs:

$$\hat{f}_m, \hat{v}_m = \mathcal{V}(s_m), \quad m \in \{1, \dots, M\}, \quad (1)$$

where $\hat{f}_m \in \{0, 1\}$ represents the predicted preference of user and \hat{v}_m denotes the estimated valuation. These outputs drive two subsequent processes: the bidding algorithm \mathcal{A} computes the bid price $b_m = \mathcal{A}(\hat{f}_m, \hat{v}_m, B_m)$, followed by the auction mechanism \mathcal{E} that determines the allocation outcome:

$$z_m, p_m = \mathcal{E}(b_m, b_m^o), \quad (2)$$

where $z_m \in \{0, 1\}$ indicates the winning status, p_m is the actual payment, and b_m^o denotes the highest competing bid for the m -th item.

The total profits of the users, specifically their utility and value (Liu and Wu, 2021), are given by the following equations:

$$u = \sum_{m=1}^M z_m \cdot (v_m - p_m), \quad v = \sum_{m=1}^M z_m \cdot v_m. \quad (3)$$

The optimization objective integrates utility with preference modeling accuracy:

$$\begin{aligned} \max_{\mathcal{V}} \quad & u + \lambda_1 \cdot v - \lambda_2 \cdot D(f, \hat{f}), \\ \text{s.t.} \quad & \sum_{m=1}^M z_m \cdot p_m \leq B, \end{aligned} \quad (4)$$

where $D(f, \hat{f})$ measures the discrepancy between true preferences f and predicted scores \hat{f} using weighted F1 score (Tao and Yi, 2013). The parameters λ_1 and λ_2 balance economic gain against users’ preferences.

3 Method

In this section, we present the LaMP-Val framework, which integrates data, learning, and evaluation for personalized valuation. Section 3.1 introduces our innovative dataset through LLM-driven data augmentation. Building upon this foundation, Section 3.2 details the learning paradigm that enables LLMs to capture personalized valuation and preference from users’ needs. Finally, Section 3.3 presents the evaluation environment and the proposed personalized metrics specifically designed for this scenario.

3.1 Data: Desensitive, Reasonable, Consistent Data Augmentation

To address the issue of personalized valuation, we need data that includes preferences for item valuation. However, there is a scarcity of datasets that encompass semantic descriptions, user preferences, and item valuations in this field. To overcome this challenge, we implemented an LLM augmentation method to create a fine-tuning dataset. The Epinions* (Zhao and McAuley, 2014) dataset is a classic shopping dataset containing 508k product rating information. This rich semantic diversity supports comprehensive preference and valuation modeling while maintaining privacy compliance through de-identification. We utilized Epinions, as our primary data source, construct the desensitized, reasonable, consistent valuation dataset.

Desensitized. Each data sample includes item name, item prices, item reviews, and consumer ratings, as well as sensitive information like user_id. As illustrated in Figure 2(a), we first drop out the privacy parts, leaving only item name, transaction price, item reviews, and consumer ratings.

Reasonable. To create a suitable valuation dataset, we tackle the significant challenge of developing a new dataset for fine-tuning LLMs in personalized valuation modeling, which has been previously impeded by three key limitations. First, in auction scenarios, the final selling price often diverges from the bidder’s valuation (*a.k.a.*, perceived value) (Kehoe, 1989). Existing works (Satterthwaite and Shneyerov, 2008; Easley and Kleinberg, 2010; Virág, 2010) show that in large-scale markets, the price mechanism can achieve optimal allocation of resources, and prices tend to converge to buyers’ accurate valuations. Thus, we filter the

*<https://www.shopping.com/>, ©2019-2025 eBay Inc. All Rights Reserved

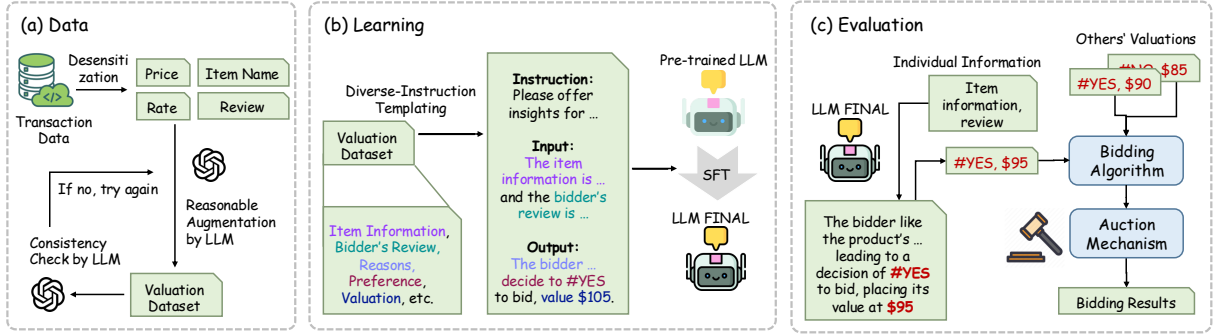


Figure 2: Overview of our method: (1) **Data**: Transaction data undergoes desensitization processing, extracting information such as price, item name, rating, and review. LLMs are employed to analyze individual preferences, complete product descriptions, user preferences, valuations, and their justifications. Then, check the consistency and rationality of the generated data. If they are not met, retry. (b) **Learning**: Utilize diverse instruction templates to template item information and user reviews in the valuation dataset into a fine-tuning dataset. Train the pre-trained LLMs via SFT to form the final model (LLM FINAL). (c) **Evaluation**: LLM FINAL generates bidding decisions (e.g., preference #YES, valuation of \$95) based on item information and reviews. Combined with other users’ information, the final bidding result is determined through the bidding algorithm and auction mechanism.

items with fewer than 15 user purchase records to ensure sufficient market participation. Second, we address the prevalent rating bias (De Langhe and Fernbach, 2016) where bidders’ average ratings systematically diverge from objective quality metrics. Building on recent work (Guo and Kong, 2025a) demonstrating the effectiveness of extreme rating conversion, we transform the highest and lowest scores into binary preference labels (“liked” vs. “disliked”). This approach reduces rating bias from inconsistent rating scales while preserving essential signals for modeling subjective valuations in auction environments. Lastly, drawing from existing literature (Wei and Wang, 2022), it has been illustrated that the organization of thoughts can significantly augment the capability of LLMs to tackle complex reasoning challenges (Liyanae and Gokani, 2024; Fang and Lee, 2023; Peng and Li, 2023). Thus, we use a strong reasoning model to clarify the derived preferences and valuations of personalized preference and item information.

Consistent. We conduct a validation process using LLM-based verification to ensure that the generated rationales accurately reflect the alignment between user preferences and item value. If an instance fails this verification, we initiate an iterative refinement process to regenerate the reasoning until we achieve consistent and satisfactory results.

Notably, our approach differs fundamentally from knowledge distillation. GPT-4 is used exclusively for rationale generation to enhance training data quality, not for distilling valuation labels. This process synthesizes explanatory rationales from

item descriptions and user needs, following established augmentation strategies for prediction tasks. We have created a refined dataset that includes 923 unique item types and contains a total of 23,065 individual instances. These instances have been randomly divided into training, validation, and testing sets in a ratio of 6:1:3. An example of the Valuation Dataset can be found in Appendix C.

3.2 Supervised Fine-Tuning with Diversity-Instruction Generation

As depicted in Figure 2(b), our methodology constructs the SFT dataset through structured instruction templates that encapsulate the valuation information. Each training sample consists of three components: “Instruction” specifying the task objective, “Input” containing item attributes (item information, reviews), and “Output” presenting augmented reasons, user preference, and valuation.

The instruction design philosophy is based on insights from (Guo and Kong, 2025b), which demonstrate that diversity in instruction significantly enhances model robustness by improving worst-case performance. We use LLM to generate 30 distinct instruction templates that convey the same semantic content but use different phrasing to implement this. Human annotators then validate these templates to ensure their accuracy. Subsequently, these instructions were randomly assigned to each instance. Moreover, to consistently guide LLMs outputting valid preferences and valuations, we precede these preferences and values with specific markers: guide signs “#” for preferences and “\$”

for valuations (Dinh and Zeng, 2022).

Through this approach, we create a diverse-instruction training that encourages the model to develop a generalized understanding of underlying task requirements rather than overfitting to specific phrasings. With the constructed SFT dataset, we perform supervised fine-tuning on the pre-trained LLM, ultimately deriving the optimized *LLM FINAL* through this diverse-instruction learning paradigm.

3.3 Evaluation: Personalized Preference Evaluation on Real Data

This section presents the evaluation pipeline architecture, then discusses conventional metrics’ limitations (e.g., utility/value) and introduces a personalized methodology addressing these issues.

Pipeline. As illustrated in Figure 2(c), LaMP-Val’s evaluation framework employs the supervised fine-tuned model *LLM FINAL* to generate personalized valuations \hat{v}_m and preference indicators \hat{f}_m . This output subsequently drives a two-stage evaluation process combining strategic bidding and auction mechanism simulation on real data from Epinions.

As illustrate in Algorithm 1, the Individual Pacing (IP) algorithm (Balseiro and Gur, 2019) optimizes bidding under budget constraints by targeting a spending rate $\rho = B/M$. With maximum valuation \bar{v} , the multiplier upper bound is set as $\bar{\lambda} \geq \bar{v}/\rho$. Initializing $\lambda_1 \in [0, \bar{\lambda}]$ and residual budget $B_1 = B$, the algorithm computes the bid $b_m = \min(\frac{v_m}{1+\lambda_m}, B_m)$ for each item m . The Vickrey mechanism (Liu and Wu, 2021) resolves competition via sealed bids: allocation $z_m = \mathbf{1}_{\{b_m > b_m^o\}}$ and payment $p_m = z_m \cdot b_m^o$. The multiplier updates via clipped stochastic gradient descent:

$$\lambda_{m+1} = \text{clip}_{[0, \bar{\lambda}]}(\lambda_m - \epsilon(\rho - p_m)), \epsilon = 1/\sqrt{M}, \quad (5)$$

while the budget decrements as $B_{m+1} = B_m - p_m$.

Personalized Evaluation Metrics. Auction performance analysis traditionally employs *utility* and *value* metrics (Lv and Zhang, 2022), defined in Equation 3. These conventional measures effectively quantify corporate profits in auction scenarios by focusing on monetary transactions. However, the assumption of uniform item desirability across all bidders limits their applicability to individual users, who inherently possess heterogeneous preference structures and subjective valuations.

To address this limitation, we develop Personalized Utility (PU) and Personalized Value

Algorithm 1 Individual Pacing Algorithm in Vickrey mechanism

- 1: **Input:** Number of items M , budget B , maximum possible valuation \bar{v} , target spending rate $\rho = B/M$, upper bound of the multiplier $\bar{\lambda} \geq \bar{v}/\rho$, item value $\{v_m\}$, highest bidding of other bidders $\{b_m^o\}$, $\epsilon = 1/\sqrt{M}$. $\triangleright \bar{\lambda} = \bar{v}/\rho$
 - 2: Initialize $\lambda_1 \leftarrow 0, B_1 \leftarrow B$
 - 3: **for** $m = 1$ to M **do**
 - 4: Bid $b_m \leftarrow \min(\frac{v_m}{1+\lambda_m}, B_m)$
 - 5: Observe $(z_m, p_m) \leftarrow (\mathbf{1}_{\{b_m > b_m^o\}}, z_m b_m^o)$
 - 6: Update $\lambda_{m+1} \leftarrow \text{clip}_{[0, \bar{\lambda}]}(\lambda_m - \epsilon(\rho - p_m))$
 - 7: Deduct $B_{m+1} \leftarrow B_m - p_m$
 - 8: **end for**
-

(PV) constructs that integrate preference-awareness into auction evaluation. The personalized utility metric specifically quantifies the net economic gain for preferred items, expressed through the double-filtered summation:

$$\text{PU} = \sum_{m=1}^M \mathbf{1}_{\{f_m=1 \cap z_m=1\}} \cdot (v_m - p_m), \quad (6)$$

where M represents the total item count, $\mathbf{1}_{\{f_m=1 \cap z_m=1\}}$ serves as a joint indicator requiring both preference declaration ($f_m = 1$) and successful acquisition ($z_m = 1$), v_m indicates personal valuation, and p_m denotes transaction price. This two-stage filtration ensures that the PU exclusively considers desired and obtained items.

Conversely, the personalized value metric captures the maximum potential satisfaction derived from preferred acquisitions, independent of payment considerations:

$$\text{PV} = \sum_{m=1}^M \mathbf{1}_{\{f_m=1 \cap z_m=1\}} \cdot v_m. \quad (7)$$

The PV formulation maintains the valuation perspective, allowing for a focus on pure benefit perception. These two metrics work together to characterize personal preferences: PU reflects tangible economic outcomes, while PV represents ideal fulfillment scenarios. This establishes a comprehensive framework for the individualized assessment of the auction.

4 Experiments

This section begins with the preliminary experiment. Then we present an overview of the experi-

Model	Personalized Profit		Traditional Metrics		
	PU \uparrow	PV \uparrow	weighted F1 \uparrow	MAE \downarrow	RMSLE \downarrow
LLaMA	-1072	92787	0.6493	2251	2.6781
Mistral	1199	84231	0.6692	2463	2.5653
GPT-3.5	2231	100680	0.8652	2431	2.1146
GPT-4	896	79488	0.8784	2203	1.7756
LaMP-Val	5872	102004	0.9084	536	0.4818

Table 1: Diverse evaluations are conducted on our datasets: Personalized Utility (PU) and Personalized Value (PV) for personalized profits, weighted F1 for preference, Mean Absolute Error (MAE) and Root Mean Squared Logarithmic Error (RMSLE) for value. These metrics are applied to a test set comprising 7,515 samples, with the auction conducted on 900 items and a budget of 100,000. Arrows indicate the desired direction for each metric: \uparrow signifies that higher values are better, while \downarrow indicates that lower values are preferable.

mental setup, including baseline methods, evaluation metrics, and implementation details. In addition, we demonstrate the experimental results and conduct an ablation study. The results of these experiments aim to address the following questions:

(Q1) Does LaMP-Val capture personalized preferences and valuations more accurately?

(Q2) Does LaMP-Val yield greater profit for the user in real auction scenarios?

(Q3) How do the base model and instruction template affect the performance of LaMP-Val?

This section addresses the above three questions in order.

4.1 Preliminary Experiment

We simulate an auction system involving 20 bidders ($N = 20$) competing for 500 items ($M = 500$), with each bidder operating under a budget constraint of 50 ($B = 50$). The true values (v) that bidders assign to the items are uniformly distributed between 0 and 1 and are assumed to be independent. We incorporate three different levels of Gaussian noise (ε) with standard deviations of $\sigma = 0, 0.01$, and 0.1 into the true valuations, resulting in noisy values (\tilde{v}). To calculate the utility (u for $\sigma = 0$ and \tilde{u} for $\sigma \neq 0$) for each condition, we implement an individual pacing algorithm in the Vickrey mechanism. Utility is defined as the total value gained from the difference between the true item values and the prices paid for winning bids. The results presented in Table 2 indicate that as the standard deviation of the noise increases, the utility decreases. Particularly, at the highest noise level ($\sigma = 0.1$), the utility becomes negative, suggesting that the individual pacing strategy is less effective under conditions of high noise.

Noise std (σ)	Utility	Utility Decrease (%) \downarrow
0.00	0.4385	0.0%
0.01	0.3964	9.6%
0.10	-0.6659	251.9%

Table 2: Utility results for different standard deviations of noise. The utility decrease is relative to the base case.

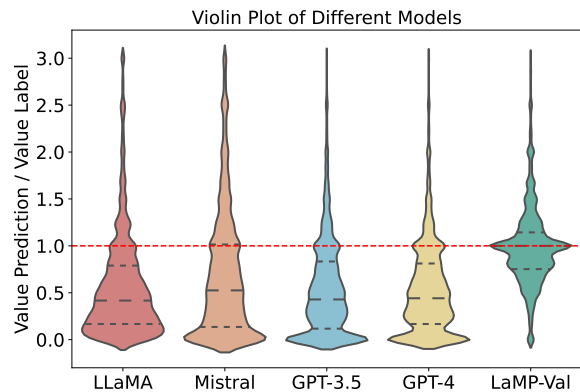


Figure 3: A violin plot showing the ratio $v_{\text{pred}}/v_{\text{label}}$ from various models. The density of the violin, primarily around the red line, indicates a better valuation.

We present a theorem asserting that the long-run average utility in the absence of noise strictly exceeds that in the presence of noise. A detailed proof and further discussion of this theorem are provided in Appendix A.1.

4.2 Experimental Setup

Baselines. By using prompt-based methods, we can effectively achieve personalized preferences and valuations based on user text descriptions. This experiment will employ these methods as baselines, focusing on the LLaMA-3-8B-Instruct (Grattafiori and Dubey, 2024), Mistral-7B-Instruct-v0.2 (Jiang and Sablayrolles, 2023), GPT-3.5 (gpt-3.5-turbo-0613) (Koon and Cicecki, 2023), and GPT-

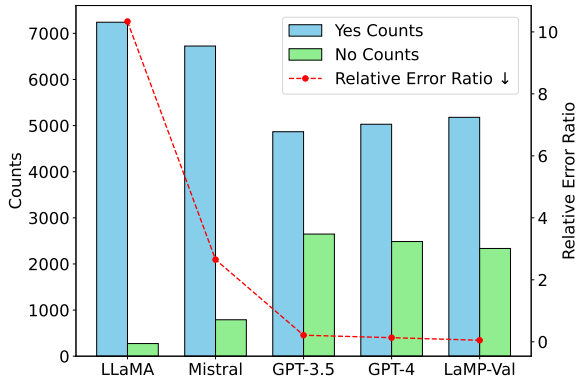


Figure 4: The number of “YES” and “NO” predicted by each model. The lower Relative Error Ratio indicates stronger label ratio alignment and improved preference distribution accuracy.

4 (OpenAI, 2023), with prompt sample in Appendix C. After gathering the preferences and valuations, we will implement a bidding algorithm and conduct an auction, as illustrated in Figure 2(c).

Evaluation Metric. Our evaluation metrics fall into two categories. The first category, personalized profit, focuses on assessing the model’s ability to generate profits during the auction bidding phase. This is represented by metrics such as PU and PV, with higher values indicating better performance. The second category, traditional metrics, aims to evaluate the model’s accuracy in capturing bidder preferences and item values during the instruction tuning phase. This includes metrics such as the weighted F1 score (Tao and Yi, 2013) of preference, Mean Absolute Error (MAE) (Willmott and Matsuura, 2005), and Root Mean Squared Logarithmic Error (RMSLE) (Jadon et al., 2022) for valuation. A higher weighted F1 score is better in this category, while smaller values for MAE and RMSLE are preferred.

Implementation Details. In the data phase, GPT-4 generates preference and valuation reasons and does consistency checks. In the learning phase, we employ the one-shot in-context learning (Dong and Li, 2024) approach for the prompt-based baselines. This method allows LLMs to assess the adaptability and predictive capacity of the models when confronted with new, unseen data. Specifically, we randomly select one sample from the validation set and use it to construct prompts. For our proposed LaMP-Val methods, we utilize Mistral-7B-Instruct as base model, zero-shot prompting (Li, 2023) to predict valuations and preferences. More details can be found in Appendix F. In the evalua-

Budget	Model	#item=600		#item=900	
		PU \uparrow	PV \uparrow	PU \uparrow	PV \uparrow
100 K	LLaMA	-1046	89694	-1072	92787
	Mistral	1204	83765	1199	84231
	GPT-3.5	<u>2233</u>	100820	<u>2231</u>	<u>100680</u>
	GPT-4	1040	72660	896	79488
	LaMP-Val	5870	<u>99161</u>	5872	102004
1 M	LLaMA	-1046	121694	-1066	132772
	Mistral	<u>1668</u>	<u>362709</u>	<u>1548</u>	<u>379556</u>
	GPT-3.5	1258	125845	1149	134481
	GPT-4	1015	73084	871	80512
	LaMP-Val	15223	489633	15115	501537

Table 3: Comparison of model performance on PU and PV metrics across different budgets and item quantities, bolding the best and underlining the second-best.

tion phase, we incorporate budget constraints (Balseiro and Kroer, 2023; Chen and Wang, 2023) to mirror realistic market scenarios, with the budget set at $B = 100,000$.

4.3 Main Results

Key Metric Improvements. Table 1 presents the key results of various methods. The weighted F1 score reflects the accuracy of personalized preference, where LaMP-Val achieves a score of 0.9084, surpassing all other methods. In addition, MAE indicates valuation accuracy; here, LaMP-Val also outperforms the competition with a score of 536. Moreover, RMSLE demonstrates benefits in handling prediction tasks across a broad range of values, and our method yields a significantly lower result. These results suggesting that LaMP-Val more accurately captures user preferences and item valuations. To further illustrate that LaMP-Val generates greater personalized profit, we assess the PU and PV of the various methods. The results indicate that LaMP-Val’s PU and PV are higher than those of the other methods, showcasing its superior capability in generating personalized profit.

Valuation Distribution. The comparative evaluation of predicted values across multiple models is illustrated in Figure 3, which only includes values less than 3 for better display. The violin plots indicate that models such as LLaMA-3-8B-Instruct, Mistral-7B-Instruct, GPT-3.5, and GPT-4 tend to underestimate actual item values. This suggests a systemic bias that could lead to missed opportunities for bidders. In contrast, the LaMP-Val model shows a more balanced distribution of predictions, hovering around the true values without consistently overestimating or underestimating. Addi-

Base Model	Method	Personalized Profit				Traditional Metrics		
		PU ₆₀₀ ↑	PV ₆₀₀ ↑	PU ₉₀₀ ↑	PV ₉₀₀ ↑	weighted F1 ↑	MAE ↓	RMSLE ↓
LLaMA	LaMP-Val (w/o G)	5233	105233	5233	105233	0.7965	1198	0.6326
	LaMP-Val (w/o R)	1686	99864	1686	99864	0.8984	1260	0.5799
	LaMP-Val	5233	105233	5233	105233	0.8985	635	0.5045
Mistral	LaMP-Val (w/o G)	5367	97669	5368	100181	0.8810	555	0.5002
	LaMP-Val (w/o R)	5577	97962	5678	99535	0.8843	551	0.5054
	LaMP-Val	5870	99161	5872	102004	0.9084	536	0.4818

Table 4: Ablation study with varying templates and base models. The subscripts of the PU and PV denote the number of items.

tionally, LLaMA, Mistral, GPT-3.5, and GPT-4 often assign a significant number of items a zero valuation, indicating a bias towards assigning zero value when these models perceive a lack of interest in specific items. However, the LaMP-Val model does not display this bias, highlighting its superior reliability in value estimation.

Preference Ratio. The counts of “YES” and “NO” preferences are illustrated in Figure 4. Additionally, we calculate the Relative Error Ratio (RER) to show how closely the “YES” to “NO” ratio aligns with the label. The RER is defined as $RER = (|\text{ratio}_{\text{pred}} - \text{ratio}_{\text{label}}|) / \text{ratio}_{\text{label}}$, where the ratio is calculated as $\#YES / \#NO$. A lower RER indicates a better alignment of the preference ratio. From Figure 4, we can see that LaMP-Val has the smallest RER, demonstrating its superior ability to capture personalized preferences. In contrast, the RER is higher for the LLaMA and Mistral models, indicating that these pre-trained models are less capable of predicting personalized preferences.

4.4 Stability Analysis of Profit

To investigate the stability of the profit conferred by our learned valuations across varying scenarios, we conduct experiments under different budget constraints (100K and 1M) to examine the changes in both PU and PV. The results presented in Table 3 reveal that our proposed method, LaMP-Val, consistently approaches optimal performance levels under differing budget limitations, with the sole exception being when the budget is set at 100K, and the number of items (#items) is 600, where it marginally trails behind GPT-3.5. Intriguingly, we observe that GPT-4’s performance does not surpass that of GPT-3.5, a discrepancy potentially attributable to GPT-3.5 having encountered similar training data during its development. Furthermore, it is noteworthy that under a more generous budget of 1M, Mistral exhibits commendable performance.

This phenomenon may be elucidated by the fact that, as depicted in Figure 3, Mistral’s percentile lines align more closely with the line of 1, indicative of its valuation capability in this context.

4.5 Ablation Study

We investigate the effects of various modifications and different base models on the performance of our method. Initially, we augment the dataset with a reasoning component and use guide signs to direct LLMs in producing the specific information we need. To assess the effectiveness of these two modifications, we create two distinct datasets: $D_{w/o R}$, which lacks the reasoning component, and $D_{w/o G}$, which does not utilize guide signs. Samples from $D_{w/o R}$ and $D_{w/o G}$ can be found in Appendix B. Additionally, we adapt our model to a different foundational architecture, LLaMA3-8B-Instruct, to evaluate whether base model changes significantly influence our approach’s performance. This investigation helps us understand how varying the underlying model architecture can affect overall performance metrics.

Based on the results presented in Table 4, we can draw the following observations: 1) The LaMP-Val (Mistral) model demonstrates superior performance across most metrics, particularly in the weighted F1 score (0.9084) and PU scores of 5870 and 5872 for PU₆₀₀ and PU₉₀₀, respectively. 2) Including reasoning and guide signs in the LaMP-Val model (using either the LLaMA or Mistral as the base model) generally improves metrics. However, these enhancements are not consistently observed across all metrics. 3) The impact of different components varies depending on the base model used; for the LLaMA base model, adding these components significantly reduces the MAE, while the impact is less pronounced with the Mistral-based LaMP-Val model.

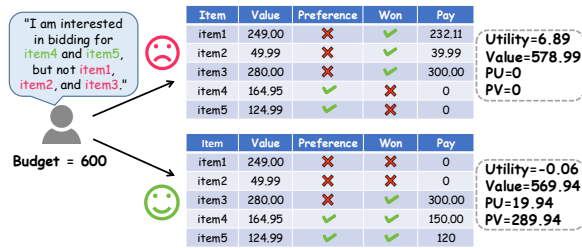


Figure 5: The discrepancy between a bidder’s preferences and the auction outcomes, with the budget constraints, utility, value, PU, and PV.

5 Case Study

To further validate the efficacy of our proposed evaluation metrics, PU and PV, in comparison to the conventional utility metric, we conducted a case study using five randomly selected samples from our experimental data, as illustrated in Figure 5. This study examines a specific bidder’s preferences for auction items under a fixed budget constraint of \$600, necessitating prioritization due to the infeasibility of bidding on all items of interest.

In the first scenario, despite achieving higher overall utility and value, the bidder fails to acquire the desired items, resulting in zero PU and PV scores. In contrast, the second scenario demonstrates that even with lower utility and total value, the bidder successfully secures items of interest, leading to higher PU and PV scores. These results suggest that in real-world auction settings, where bidders prioritize obtaining specific items, PU and PV metrics more accurately indicate success than traditional measures of utility and value.

This case study highlights the relevance of PU and PV as more effective metrics for evaluating auction outcomes, particularly in scenarios where bidders’ preferences and budget constraints critically shape their bidding strategies. By focusing on the acquisition of desired items, these metrics align more closely with bidder satisfaction, reflecting the true nature of auction where personal preferences significantly influence bidding decisions.

6 Related work

LLMs meet Auction. While recent works on integrating LLMs with auction have flourished, they mainly focus on general bidding strategies rather than personalized valuation modeling for individual participants. [Chen and Yuan \(2023\)](#) model strategic interactions via LLMs in budget-constrained bidding scenarios. [Dütting and Mir-](#)

[rokni \(2024\)](#) jointly generate text through word-by-word bidding with multi-agent LLMs. [Zhu and Horton \(2024\)](#) simulate auctions via LLM bidding agents, revealing human-aligned behavioral patterns. [Dubey and Feng \(2024\)](#) analyze ad integration in LLM summaries via layered auction-module architecture. [Hajiaghayi and Lahaie \(2024a\)](#) embed ads in LLM outputs via retrieval-augmented auctions for efficiency-fairness balance. [Yin \(2025\)](#) study signaling effects on strategic behavior using LLM-simulated disclosure in multi-agent auctions. [Shah et al. \(2025\)](#) examines LLMs’ behavior in simulated auctions, demonstrating their potential as cost-effective proxies for human participants in experimental economics research. [Huang et al. \(2024\)](#) proposes a transformer-based method called Auctionformer to efficiently solve the equilibrium of various auction games in a unified framework.

Bidding Algorithm. Most auction research ([Schlosser and Boissier, 2018](#); [Balseiro and Deng, 2021](#); [Bachrach and Talgam-Cohen, 2022](#); [Golrezaei and Sahoo, 2024](#); [Hajiaghayi and Lahaie, 2024b](#)) advance theoretical auction mechanisms spanning dynamic pricing, reserve price optimization, and worst-case robust designs, yet omit personalized valuation considerations in their strategic frameworks. Other existing studies ([Zhang and Niu, 2023](#); [Leme and Pál, 2016](#); [Peri and Curry, 2021](#)) leverage expert-designed features (dynamic campaign profiles, reserve price strategies, and human-guided allocation patterns) to learn personalized valuations, enhancing fairness, welfare, and performance in automated bidding systems.

7 Conclusion

This study focuses on the valuation challenge considering user semantic preference, including accurately capturing user preference and valuation, and gaining more profit. We propose the LaMP-Val, a learning-based framework containing three essential parts: data, learning, and evaluation. LaMP-Val builds a valuation dataset for preference learning, fine-tuning LLMs, and proposes personalized evaluation metrics. Through comprehensive evaluations across diverse auction scenarios, our methodology has proven effective in delivering accurate valuations and reflecting user preferences, achieving significant profit gains compared to baseline models in real-world auctions.

Limitations

A worthwhile direction for future work is to extend the LaMP-Val framework to integrate broader open-sourced auction mechanisms and bidding algorithms, alongside developing more qualified semantic-rich datasets to evaluate personalized valuation methods. We envision that the LaMP-Val framework signifies a pioneering step towards integrating semantic analysis into bid valuation processes.

Acknowledgement

This research is supported by the National Natural Science Foundation of China (92270114, 62302321).

References

- Gagan Aggarwal and Ashwinkumar Badanidiyuru. 2024. Auto-bidding and auctions in online advertising: A survey. *arXiv preprint*, arXiv:2408.07685.
- Nir Bachrach and Inbal Talgam-Cohen. 2022. Distributional robustness: From pricing to auctions. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, page 150, Boulder, CO.
- Santiago R. Balseiro and Yuan Deng. 2021. Robust auction design in the auto-bidding world. In *Advances in Neural Information Processing Systems 34*, pages 17777–17788, Virtual Event.
- Santiago R. Balseiro and Yonatan Gur. 2019. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science*, 65(9):3952–3968.
- Santiago R. Balseiro and Christian Kroer. 2023. Contextual standard auctions with budgets: Revenue equivalence and efficiency guarantees. *Management Science*, 69(11):6837–6854.
- Sushil Bikhchandani and Sven De Vries. 2011. An ascending vickrey auction for selling bases of a matroid. *Operations research*, 59(2):400–413.
- Tom B. Brown and Benjamin Mann. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, Virtual Event.
- Jiangjie Chen and Siyu Yuan. 2023. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *arXiv preprint*, arXiv:2310.05746.
- Ming Chen and Sareh Nabi. 2023. Advancing ad auction realism: Practical insights & modeling implications. *arXiv preprint*, arXiv:2307.11732.
- Yurong Chen and Qian Wang. 2023. Coordinated dynamic bidding in repeated second-price auctions with budgets. In *Proceedings the 40th International Conference on Machine Learning*, pages 5052–5086, Honolulu, HI.
- Bart De Langhe and Philip M Fernbach. 2016. Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research*, 42(6):817–833.
- DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*, abs/2501.12948.
- Tuan Dinh and Yuchen Zeng. 2022. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *arXiv preprint*, arXiv:2206.06565.
- Qingxiu Dong and Lei Li. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, FL.
- Avinava Dubey and Zhe Feng. 2024. Auctions with LLM summaries. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 713–722, Barcelona, Spain.
- Paul Dütting and Vahab Mirrokni. 2024. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, pages 144–155, Singapore, Singapore.
- David Easley and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.
- Luyang Fang and Gyeong-Geon Lee. 2023. Using gpt-4 to augment unbalanced data for automatic scoring. *arXiv preprint*, arXiv:2310.18365.
- Negin Golrezaei and Sourav Sahoo. 2024. Bidding in uniform price auctions for value maximizing buyers. *arXiv preprint*, arXiv:2406.03674.
- Aaron Grattafiori and Abhimanyu Dubey. 2024. The llama 3 herd of models. *arXiv preprint*, arXiv:2407.21783.
- Yongkang Guo and Yuqing Kong. 2025a. Mitigating the participation bias by balancing extreme ratings. In *Proceedings of the ACM on Web Conference 2025*, pages 1441–1455, Sydney, Australia.
- Yongkang Guo and Yuqing Kong. 2025b. Mitigating the participation bias by balancing extreme ratings. In *Proceedings of the ACM on Web Conference 2025*, pages 1441–1455, Sydney, Australia.
- MohammadTaghi Hajiaghayi and Sébastien Lahaie. 2024a. Ad auctions for llms via retrieval augmented generation. In *Advances in Neural Information Processing Systems 38*, Vancouver, Canada.

- MohammadTaghi Hajiaghayi and Sébastien Lahaie. 2024b. Ad auctions for llms via retrieval augmented generation. *arXiv preprint*, arXiv:2406.09459.
- Liang Han and Zhaozheng Yin. 2019. Vision-based price suggestion for online second-hand items. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1988–1996, Nice, France.
- Liang Han and Zhaozheng Yin. 2020. Price suggestion for online second-hand items with texts and images. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2784–2792, Seattle, WA.
- Kexin Huang, Ziqian Chen, Xue Wang, Chongming Gao, Jinyang Gao, Bolin Ding, and Xiang Wang. 2024. Auctionformer: A unified deep learning algorithm for solving equilibrium strategies in auction games. In *Proceedings the 41st International Conference on Machine Learning*, Vienna, Austria.
- Kexin Huang, Ziqian Chen, Xue Wang, Chongming Gao, Jinyang Gao, Bolin Ding, and Xiang Wang. 2025. Learning bayesian nash equilibrium in auction games via approximate best response. In *Proceedings the 42nd International Conference on Machine Learning*, Vancouver, Canada.
- Kuo-Han Hung and Ching-Yun Ko. 2025. Attention tracker: Detecting prompt injection attacks in llms. In *Findings of the Association for Computational Linguistics*, pages 2309–2322, Albuquerque, NM.
- Aryan Jadon, Avinash Patil, and Shruti Jadon. 2022. A comprehensive survey of regression based loss functions for time series forecasting. *arXiv preprint*, abs/2211.02989.
- Simon Janin and Kaihua Qin. 2020. Filebounty: Fair data exchange. In *IEEE European Symposium on Security and Privacy Workshops, EuroS&P Workshops 2020*, pages 357–366, Genoa, Italy.
- Albert Q. Jiang and Alexandre Sablayrolles. 2023. Mistral 7b. *arXiv preprint*, arXiv:2310.06825.
- Christian K Karl. 2016. Investigating the winner’s curse based on decision making in an auction environment. *Simulation & Gaming*, 47(3):324–345.
- Timothy J. Kehoe. 1989. *Microeconomic Theory (Yoshihiko Otani and Mohamed El-Hodiri)*, volume 31.
- Paul Klemperer. 1999. Auction theory: A guide to the literature. *Journal of Economic Surveys*, 13(3):227–286.
- Jan Kocon and Igor Cichecki. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.
- Renato Paes Leme and Martin Pál. 2016. A field guide to personalized reserve prices. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1093–1102, Montreal, Canada.
- Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, page 641–647, Varna, Bulgaria.
- Weiwei Liu and Huaming Wu. 2021. Aucswap: A vickrey auction modeled decentralized cross-blockchain asset transfer protocol. *Journal Of Systems Architecture*, 117:102102.
- Chandreen R Liyanage and Ravi Gokani. 2024. Gpt-4 as an x data annotator: Unraveling its performance on a stance classification task. *Public Library of Science*, 19(8):e0307741.
- Hongtao Lv and Zhilin Zhang. 2022. Utility maximizer or value maximizer: Mechanism design for mixed bidders in online advertising. *arXiv preprint*, abs/2211.16251.
- Ausubel Lawrence M and Milgrom Paul. 2006. The lovely but lonely vickrey auction. *Combinatorial auctions*, 17:22–26.
- OpenAI. 2023. GPT-4 technical report. *arXiv preprint*, arXiv:2303.08774.
- Baolin Peng and Chunyuan Li. 2023. Instruction tuning with gpt-4. *arXiv preprint*, arXiv:2304.03277.
- Neehar Peri and Michael J. Curry. 2021. Preferencenet: Encoding human preferences in auction design with deep learning. In *Advances in Neural Information Processing Systems 34*, pages 17532–17542, Virtual Event.
- Rolf Pfister and Hansueli Jud. 2025. Understanding and benchmarking artificial intelligence: Openai’s o3 is not agi. *arXiv preprint*, abs/2501.07458.
- Mark Satterthwaite and Artyom Shneyerov. 2008. Convergence to perfect competition of a dynamic matching and bargaining market with two-sided incomplete information and exogenous exit rate. *Games and Economic Behavior*, 63(2):435–467.
- Rainer Schlosser and Martin Boissier. 2018. Dynamic pricing under competition on online marketplaces: A data-driven approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 705–714, London, UK.
- Anand Shah, Kehang Zhu, Yanchen Jiang, Jeffrey G. Wang, Arif K. Dayi, John J. Horton, and David C. Parkes. 2025. Learning from synthetic labs: Language models as auction participants. *arXiv preprint*, abs/2507.09083.
- Ashish Sureka and Peter R Wurman. 2002. Applying the generalized vickrey auction to pricing reliable multicasts. In *International Workshop on Quality of Future Internet Services*, pages 283–292, Zurich, Switzerland.

- Peng Tao and Huang Yi. 2013. A method based on weighted f-score and svm for feature selection. In *Proceedings of the 25th Chinese control and decision conference*, pages 4287–4290.
- Gemini Team and Petko Georgiev. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*, abs/2403.05530.
- Matias Trivizano and Carlos Sarraute. 2018. Wibson: A decentralized data marketplace. *arXiv preprint*, abs/1812.09966.
- Gábor Virág. 2010. Competing auctions: Finite markets and convergence. *Theoretical Economics*, 5:241–274.
- Larry Wasserman. 2004. *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media.
- Robert J. Weber. 2003. Auction theory: By vijay krishna. academic press, 2002. *Games and Economic Behavior*, 45(2):488–497.
- Jason Wei and Xuezhi Wang. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*, pages 24824–24837, New Orleans, LA.
- Cort J. Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30(1):79–82.
- Yusuke Yamaura and Nobuya Kanemaki. 2019. The resale price prediction of secondhand jewelry items using a multi-modal deep model with iterative co-attention. *arXiv preprint*, abs/1907.00661.
- Yue Yin. 2025. Infobid: A simulation framework for studying information disclosure in auctions with large language model-based agents. *arXiv preprint*, abs/2503.22726.
- Haoqi Zhang and Lvyin Niu. 2023. A personalized automated bidding framework for fairness-aware on-line advertising. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5544–5553, Long Beach, CA.
- XP Zhang and XN Zhang. 2015. Research on railway project bidding based on vickrey theory. *WIT Transactions on the Built Environment*, (157).
- Tong Zhao and Julian J. McAuley. 2014. Leveraging social connections to improve personalized ranking for collaborative filtering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 261–270, Shanghai, China.
- Kehang Zhu and John Joseph Horton. 2024. Evidence from the synthetic laboratory: Language models as auction participants. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

A Theorem

In this part, we present a theorem asserting that the long-run average utility in the absence of noise strictly exceeds that in the presence of noise, followed by its proof.

A.1 Valuation Error Theorem

Now we denote the following variables. In an auction involving M items, a bidder’s precise valuation for the m -th item is denoted as v_m , and their estimated value is $\tilde{v}_m = v_m + \varepsilon_m$, where ε represents an independent zero-mean random noise. Then the algorithm integrated into the auction environment computes the biddings b_m and \hat{b}_m based on the valuations and remaining budget, with the latter considering the noisy scenario. The auction environment then determines if the bidder wins the m -th item through the binary variable z_m (\tilde{z}_m), based on the bidder’s bid b_m (\hat{b}_m) and the highest bid from other bidders b_m^o , with p_m (\tilde{p}_m) being the price to be paid. The utility obtained by the bidder is denoted by u_m (\tilde{u}_m). The tilde denotes the scenario with noise.

Mathematically, consider the auction of m items as a measurable space Ω , with $v, \varepsilon, b, z, u, p$ being random variables defined on Ω . The σ -algebra of noise $\sigma(\varepsilon)$ is independent of $\sigma(v, b, z, u, p)$, and $\mathbb{E}[\varepsilon] = 0$. Thus, $v_m, \varepsilon_m, b_m, z_m, u_m, p_m$ are observations of $v, \varepsilon, b, z, u, p$ during the m -th random trial. Each trial is independent and identically distributed (i.i.d.) by definition.

Before proving the main theorem, we will first establish a lemma that provides a fundamental property of the product of two sequences.

Lemma 1. For two real sequences $\{\alpha_n\}_{n=1}^{\infty}, \{\beta_n\}_{n=1}^{\infty}$,

$$\text{if } \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \alpha_n > 0$$

$$\text{and } \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \beta_n > 0, \text{ then}$$

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \alpha_n \beta_n > 0.$$

Proof. We prove this lemma under more general conditions that $\{\alpha_n\}_{n=1}^{\infty}, \{\beta_n\}_{n=1}^{\infty}$ are two i.i.d random sequences and are independent with each other. Moreover, we assume $E[\alpha_1^2] < \infty, E[\beta_1^2] < \infty$. It’s obvious that real numbers satisfy these assumptions.

Thus by Strong Law of Large Numbers ([Wasser-](#)

man, 2004), almost surely we have

$$\mathbb{E}[\alpha_1] = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \alpha_n > 0, \quad (8)$$

$$\mathbb{E}[\beta_1] = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \beta_n > 0. \quad (9)$$

According to assumptions above, $\{\alpha_n \beta_n\}_{n=1}^{\infty}$ is a i.i.d. random sequence and $E[|\alpha_1 \beta_1|] \leq \sqrt{E[\alpha_1^2]E[\beta_1^2]} < \infty$.

Thus by Strong Law of Large Numbers (Wasserman, 2004), almost surely we have

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \alpha_n \beta_n = \mathbb{E}[\alpha_1 \beta_1] = \mathbb{E}[\alpha_1] \mathbb{E}[\beta_1] > 0,$$

where the last equality holds because of independence between α_1 and β_1 .

For the special case where $\{\alpha_n\}_{n=1}^{\infty}, \{\beta_n\}_{n=1}^{\infty}$ are two real sequences without randomness, the proposition to be proved holds. \square

We have now proved the lemma 1. This lemma will be the key to proving the main theorem A.1, because it provides a conclusion about the positivity of the average of the product of two series, which will help us analyze the difference between the long-term average utilities. The statement of the main theorem is as follows.

Theorem A.1. *With notations above, denote $P_1 = \mathbb{P}\left[\frac{v}{1+\lambda} \leq B \leq \frac{v+\varepsilon}{1+\lambda}\right]$ and $P_2 = \mathbb{P}\left[\frac{v+\varepsilon}{1+\lambda} \leq B \leq \frac{v}{1+\lambda}\right]$. Assume $(P_1 - P_2)\mathbb{E}\left[B - \frac{v}{1+\lambda}\right] > 0$, then the long-run average utility without noise is strictly better than that with noise, i.e.,*

$$\lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M (u_m - \tilde{u}_m) > 0 \quad a.s.$$

Proof.

$$u_m - \tilde{u}_m \quad (10)$$

$$= z_m(v_m - p_m) - \tilde{z}_m(v_m - \tilde{p}_m) \quad (11)$$

$$= v_m(z_m - \tilde{z}_m) - z_m p_m + \tilde{z}_m \tilde{p}_m \quad (12)$$

$$= v_m(z_m - \tilde{z}_m) - z_m(z_m b_m^o) + \tilde{z}_m(\tilde{z}_m b_m^o) \quad (13)$$

$$= v_m(z_m - \tilde{z}_m) - z_m b_m^o + \tilde{z}_m b_m^o \quad (14)$$

$$= (v_m - b_m^o)(z_m - \tilde{z}_m) \quad (15)$$

$$= (v_m - b_m^o)(z_m - \tilde{z}_m) \quad (16)$$

Next, we will use the strong law of large numbers to analyze the long-term behavior of the above differences and combine it with the lemma 1 to get the final result. By the Strong Law of Large Numbers (Wasserman, 2004), almost surely,

$$\lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M (z_m - \tilde{z}_m) \quad (17)$$

$$= \mathbb{E}[1_{b > b^o} - 1_{\tilde{b} > b^o}] \quad (18)$$

$$= \mathbb{P}[b > b^o] - \mathbb{P}[\tilde{b} > b^o] \quad (19)$$

$$\quad (20)$$

$$\tilde{b} - b \quad (21)$$

$$= \min\left\{\frac{v+\varepsilon}{1+\lambda}, B\right\} - \min\left\{\frac{v}{1+\lambda}, B\right\} \quad (22)$$

$$= \begin{cases} 0, & \text{if } B \leq \min\left\{\frac{v}{1+\lambda}, \frac{v+\varepsilon}{1+\lambda}\right\}, \\ B - \frac{v}{1+\lambda}, & \text{if } \frac{v}{1+\lambda} < B < \frac{v+\varepsilon}{1+\lambda}, \\ \frac{v+\varepsilon}{1+\lambda} - B, & \text{if } \frac{v+\varepsilon}{1+\lambda} < B < \frac{v}{1+\lambda}, \\ \frac{\varepsilon}{1+\lambda}, & \text{if } B \geq \max\left(\frac{v}{1+\lambda}, \frac{v+\varepsilon}{1+\lambda}\right). \end{cases} \quad (23)$$

For the first case, $\tilde{b} - b = 0$. For the last case, $\mathbb{E}[\tilde{b} - b] = \mathbb{E}\left[\frac{\varepsilon}{1+\lambda}\right] = \mathbb{E}[\varepsilon]\mathbb{E}\left[\frac{1}{1+\lambda}\right] = 0$.

And we can derive that

$$\mathbb{E}[b - \tilde{b}] \quad (24)$$

$$= P_1 \mathbb{E}\left[B - \frac{v}{1+\lambda}\right] - P_2 \mathbb{E}\left[\frac{v+\varepsilon}{1+\lambda} - B\right] \quad (25)$$

$$= (P_1 - P_2)\mathbb{E}[B] + \quad (26)$$

$$(P_1 - P_2)\mathbb{E}\frac{v}{1+\lambda} + P_2 \mathbb{E}\left[\frac{\varepsilon}{1+\lambda}\right] \quad (27)$$

$$= (P_1 - P_2)\mathbb{E}\left[B - \frac{v}{1+\lambda}\right] \quad (28)$$

By assumption, $\mathbb{E}[b - \tilde{b}] > 0$,

we have $\mathbb{P}[\tilde{b} > b^o] > \mathbb{P}[b > b^o]$.

Then $\mathbb{P}[b > b^o] - \mathbb{P}[\tilde{b} > b^o] < 0$,

indicating $\lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M (z_m - \tilde{z}_m) < 0$ a.s.

And by the Strong Law of Large Numbers (Wasserman, 2004), almost surely,

$$\lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M v_m - b_m^o = \mathbb{E}[v - \max(v^{\text{competitor}})],$$

it is obvious that $\mathbb{E}[v] < \mathbb{E}[\max(v^{\text{competitor}})]$ because the expected value of the maximum in a set of identically distributed variables is greater than the expected value of any individual variable from the same distribution.

Followed by Lemma 1, let $-\alpha_n = v_n - b_n^o$ and $-\beta_n = z_n - \tilde{z}_n$, then α_n and β_n satisfy

the assumptions of Lemma 1, so almost surely,

$$\begin{aligned} & \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M (u_m - \tilde{u}_m) \\ &= \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M \alpha_m \beta_m > 0 \end{aligned} \quad (29)$$

Thus,

$$\lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M (u_m - \tilde{u}_m) > 0 \quad a.s.$$

□

Remark. The key assumption of the theorem A.1 ($P_1 - P_2$) $\mathbb{E}[B - \frac{v}{1+\lambda}] > 0$ is reasonable and moderate because of the following insights:

$$\mathbb{E}[B - \frac{v}{1+\lambda}] \quad (30)$$

$$= \mathbb{E}[B - \frac{v}{1+\lambda} \mid B \leq \min\{\frac{v}{1+\lambda}, \frac{v+\varepsilon}{1+\lambda}\}] \quad (31)$$

$$+ \mathbb{E}[B - \frac{v}{1+\lambda} \mid \frac{v}{1+\lambda} < B < \frac{v+\varepsilon}{1+\lambda}] \quad (32)$$

$$+ \mathbb{E}[B - \frac{v}{1+\lambda} \mid \frac{v+\varepsilon}{1+\lambda} < B < \frac{v}{1+\lambda}] \quad (33)$$

$$+ \mathbb{E}[B - \frac{v}{1+\lambda} \mid B \geq \max(\frac{v}{1+\lambda}, \frac{v+\varepsilon}{1+\lambda})] \quad (34)$$

Now the full parts of R.H.S. are negative, positive, negative, and positive, respectively. If the $P_1 \geq P_2$, then the scale of the second positive term is more likely to be greater than that of the third negative term. So R.H.S. is more likely to be positive. So $(P_1 - P_2)\mathbb{E}[B - \frac{v}{1+\lambda}]$ is more likely to be positive. The discussion is similar to the other situation. Thus this assumption is reasonable and moderate.

B Ablation Study Detail

In this section, we present the two datasets $\mathcal{D}_{w/o R}$ (Table 7, Task Instruction w/o R) and $\mathcal{D}_{w/o G}$ (Table 7, Task Instruction w/o G) that we used for the ablation experiments.

C Valuation Datasets

This section demonstrates the composition of our valuation dataset through an illustrative example in Table 8. The presented instruction-tuning instance showcases the input-output structure with

anonymized item descriptions, preference rationales generated by LLMs, and corresponding valuation labels, reflecting our methodology for resolving the challenge as discussed in Section 3.1.

D Potential Risks

Recent studies have revealed vulnerabilities in LLMs’ attention mechanisms under adversarial prompting conditions. As demonstrated in Research (Hung and Ko, 2025), targeted prompt injection attacks can induce specific attention heads to disproportionately prioritize malicious instructions over legitimate user queries—a phenomenon termed attention hijacking. In auction contexts where LaMP-Val operates, such attacks could systematically distort user preference extraction processes, leading to non-trivial deviations in item valuation (e.g., overestimating prices for items containing injected keywords like “limited edition”). This risk originates from the inherent architectural limitations of transformer-based LLMs, particularly their susceptibility to gradient-based manipulation of attention distributions during inference.

To address this challenge, our future work will implement three mitigation strategies: (1) integrating adversarial training with gradient masking techniques to harden attention heads against manipulation, (2) deploying real-time attention monitoring modules to detect abnormal focus shifts exceeding pre-defined thresholds (e.g., >85% attention weight on non-instructional tokens), and (3) incorporating user preference verification loops through contrastive prompting.

E Scientific Artifacts

E.1 Artifact Use Consistent with Intended Use

We made sure that our use of existing artifacts aligned with their intended purpose as specified. For the artifacts we created, we clearly defined their intended use and ensured that they were compatible with the original access conditions. We adhered to restrictions, such as limiting the creation of derivatives from research data to research contexts only.

E.2 Documentation of Artifacts

Epinions.com is a comprehensive consumer review platform that allows users to evaluate various products and services. The Epinions dataset encompasses multiple domains, including automobiles, banking, movies, and travel destinations, reflecting

LaMP-Val (Fine-tuning)	93.14
LaMP-Val (Inference)	34.86

seed (2025) with 0.05 LoRA dropout.

Table 5: Time consumption (in minutes) of LaMP-Val.

Parameter	Value
Seed	2025
LoRA rank (r)	16
LoRA alpha (α)	8
LoRA dropout	0.05
LoRA target modules	q_proj, v_proj, k_proj, o_proj
Batch size	64
Micro batch size	8
Number of epochs	2
Learning rate	1×10^{-4}
Cutoff length	730

Table 6: Hyperparameter of LaMP-Val.

its extensive coverage of various offerings. The reviews in the Epinions dataset are predominantly in English. However, the existing public literature lacks detailed information regarding the specific demographics of users in the dataset, such as age, gender, and geographic location.

F Computational Experiments

F.1 Model Size and Budget

The pre-trained models (LLaMA-3-8B-Instruct, and Mistral-7B-Instruct) were trained on a single A100 GPU (80GB). As the average time were shown in Table 5, the training phase was completed in 93 minutes averagely, while inference required only 35 minutes averagely on the same hardware. This performance represents a significant improvement in efficiency compared to the baseline models, which require approximately 8.6 hours for the same tasks when using the GPT-4 API.

F.2 Hyperparameters

The fine-tuning process adopts a hybrid optimization strategy combining Low-Rank Adaptation (LoRA) with dynamic batching. As detailed in Table 6, critical parameter selections follow three design principles: (1) parameter efficiency through LoRA’s low-rank decomposition ($r = 16$, $\alpha = 8$) targeting cross-attention projections in {q, v, k, o}_proj layers, reducing trainable parameters by 98.7% compared to full fine-tuning; (2) memory optimization via hierarchical batching with macro/micro batch sizes of 64/8, enabling gradient accumulation on NVIDIA A100 GPUs; and (3) stability preservation using a fixed random

Instruction Input	
Task Instruction:	You will act as an assistant for bidding decisions and valuation in an auction scenario. Below is the item information and the corresponding bidder’s review. You will make a bidding decision (whether to bid on the item) for the bidder based on this information and suggest the possible valuation by the bidder with the reasons . You must use ‘#’ and ‘\$’ before your bidding decision and value, respectively. And make sure the sentence is semantically complete and clear after removing ‘#’, and ‘\$’. Example: The bidder (reason), value it at \$XXX, decides #YES or #NO to bid.
Task Instruction (w/o R):	You will act as an assistant for bidding decisions and valuation in an auction scenario. Below is the item information and the corresponding bidder’s review. You will make a bidding decision (whether to bid on the item) for the bidder based on this information and suggest the possible valuation by the bidder. You must use ‘#’ and ‘\$’ before your bidding decision and value, respectively. And make sure the sentence is semantically complete and clear after removing ‘#’, and ‘\$’. Example: The bidder value it at \$XXX, decides #YES or #NO to bid.
Task Instruction (w/o R):	You will act as an assistant for bidding decisions and valuation in an auction scenario. Below is the item information and the corresponding bidder’s review. You will make a bidding decision (whether to bid on the item) for the bidder based on this information and suggest the possible valuation by the bidder with the reasons . Example: The bidder (reason), value it at \$XXX, decides #YES or #NO to bid.
Task Input:	The item is pr-Maytag_FAV9800AW_Washing_Machine. The Maytag FAV9800AW is a top-of-the-line washing machine designed for efficiency and convenience. It features a large capacity, perfect for handling bulky items or large loads. With its advanced Neptune wash system, it ensures your clothes are thoroughly cleaned while maintaining fabric care. The machine also boasts of a quiet operation and energy-saving features, making it a smart choice for any household. And the corresponding bidder’s review is “We bought our Neptune TL in 2004 because we were told it would do large loads our clothes were not coming out clean so i stopped it in the wash cycle and unlocked the lid looked inside to find hardly any water when the repairman came out he said that was normal also it keeps leaving soap stains on my white shirts because they were not rinsing after three years we gave up and bought a front loader lg I will not sell the old Neptune to anyone else because I have standards and do not want to rip someone off as Maytag did to me it is going to the landfill shame on you Maytag I will never buy one of your products again you should refund us our money”.
Instruction Output	
LLaMA Output:	The bidder, dissatisfied with the performance and functionality of the Maytag FAV9800AW washing machine, decides #NO to bid, valuing it at \$800 .
Mistral Output:	Based on the item’s description and the bidder’s negative review, I would advise against bidding on the Maytag FAV9800AW washing machine, as it has a history of inadequate water levels and soap residue issues. #NO Bid \$0 .
GPT-3.5 Output:	The bidder, dissatisfied with the Maytag FAV9800AW Washing Machine’s performance and expressing a lack of trust in the brand, concludes #NO to bid, valuing it at \$0 .
GPT-4 Output:	The bidder, frustrated by the machine’s performance and ethical standards, has resolved #NO to bid, valuing the Maytag Neptune TL at \$0 due to its unsatisfactory operation and decision to dispose of it rather than sell it.
LaMP-Val(w/o R) Output:	The bidder value it at \$1000 , #NO to bid.
LaMP-Val(w/o G) Output:	The bidder, dissatisfied with the washer’s performance and unmet expectations, decides NO to bid, valuing it at 1000 .
LaMP-Val Output:	The bidder, dissatisfied with the Maytag washing machine’s reliability and customer service, decides #NO to bid, valuing it at \$1000 .

Table 7: Test output example.

Instruction Input	
Task Instruction:	You will act as an assistant for bidding decisions and valuation in an auction scenario. Below is the item information and the corresponding bidder’s review. You will make a bidding decision (whether to bid on the item) for the bidder based on this information and suggest the possible valuation by the bidder. You must use ‘#’ and ‘\$’ before your bidding decision and value, respectively. And make sure the sentence is semantically complete and clear after removing ‘#’, and ‘\$’.
Task Input:	The item is Nokia_7160_Cellular_Phone. The Nokia 7160 Cellular Phone is a dynamic communication device, offering features such as SMS, call waiting, speed dialing, and conferencing. Esteemed for its long battery life, this classic mobile phone also provides a high-quality voice call experience. Its easy-to-navigate interface, clock functions, and reliable network coverage make the Nokia 7160 an excellent choice for robust, straightforward mobile communication. And the corresponding bidder’s review is “I have been a long-time user of Southwestern Bell Wireless now cingular for the past 6 years I have been on the analog service for a long time the whole 6 years and my bill was getting larger and larger”.
Instruction Output	
Task Output:	The bidder, influenced by their long history with a costly analogous service, decides #YES to bid, valuing the Nokia_7160_Cellular_Phone at \$99.

Table 8: A tuning instance for LaMP-Val.