Leveraging Large Language Models for Active Merchant Non-player Characters

Byungjun Kim , **Minju Kim** , **Dayeon Seo** and **Bugeun Kim**
Department of Artificial Intelligence, Chung-Ang University, Republic of Korea
{k36769, minjunim, sdyhappy, bgnkim}@cau.ac.kr

Abstract

We highlight two significant issues leading to the passivity of current merchant non-player characters (NPCs): *pricing* and *communication*. While immersive interactions with active NPCs have been a focus, price negotiations between merchant NPCs and players remain underexplored. First, passive pricing refers to the limited ability of merchants to modify predefined item prices. Second, passive communication means that merchants can only interact with players in a scripted manner. To tackle these issues and create an active merchant NPC, we propose a merchant framework based on large language models (LLMs), called MART, which consists of an appraiser module and a negotiator module. We conducted two experiments to explore various implementation options under different training methods and LLM sizes, considering a range of possible game environments. Our findings indicate that finetuning methods, such as supervised finetuning (SFT) and knowledge distillation (KD), are effective in using smaller LLMs to implement active merchant NPCs. Additionally, we found three irregular cases arising from the responses of LLMs.

1 Introduction

Exchanging in-game items is an integral part of open-world role-playing games because the utility of an item depends on the current attributes of a player. For example, a player with high agility may not benefit from an item that enhances agility. Thus, game developers implement an exchange system in their games to enhance gameplay and item utility.

While games commonly feature merchant non-player characters (NPCs) to facilitate item exchanges, these interactions are typically scripted; the merchant only enables players to buy and sell items at fixed prices, but without the dynamics of real-world transactions. Typically, developers do not allow merchants to alter prices or communicate in free form; instead, they simply present a fixed price to the player without any negotiation. However, recent advancements in LLM-integrated games show growing interest in dynamic and flexible interactions, similar to real-world communication where players and NPCs engage in co-constructed dialogue [ReLU Games and KRAFTON, 2024; Latitude, 2019; Li *et al.*, 2024]. To mirror real-world interactions and enhance player immersion, we propose a novel framework that leverages LLMs to enable merchant NPCs to engage in actual negotiations with players.

To create an active merchant NPC, which is more aligned to the real world, we need to address two issues that cause the current merchant to act passively: *pricing* and *communication*. First, regarding passive pricing, the merchant has no authority to adjust item prices. Instead, the game developers decide prices based on the items' utility in the game. However, in the real world, sellers can adjust the prices of their goods according to item specifications. To the best of our knowledge, gaming industry researchers have not adequately filled this gap. Since item descriptions often convey sufficient information about utility, we argue that they can be used to estimate the value of unseen items by leveraging information and values from other known items. Specifically, we let LLMs appraise game items by observing other items.

Second, regarding passive communication, merchants can only interact with players in a scripted manner. Researchers have explored the adoption of LLMs within games to facilitate more immersive player experiences [Phillips *et al.*, 2024; Peng *et al.*, 2024]. However, the way merchants communicate has been given limited focus, which remains a one-way interaction. Current merchant NPCs simply display a predefined list of items for players to choose from, and players respond by clicking on items to make a purchase. This purchasing experience is uniform across all interactions with a merchant, regardless of the individual player involved in the transaction. To enable two-way communication, we propose a negotiation style that mirrors real-world transactions. Through this study, we explore whether LLMs can be used to foster a negotiating interaction within the merchant context.

In this paper, we propose a framework for developing a More Active meRchanT NPC, called MART. This framework consists of two main components: *appraiser* and *negotiator*. The appraiser module addresses the issue of passive pricing by estimating the value of given items. The negotiator module addresses the passive communication issue by negotiating with players. As game developers have diverse requirements when deploying their games, we conducted experiments to compare potential candidates for each module using a public WoW Classic game item dataset. Specifically, we tested

both finetuning methods and n-shot prompting methods on Llama 3 models, which come with a wide range of parameter sizes up to 405 billion. Our GitHub repository[1] provides implementation details, prompts, and model outputs.

This study has the following contributions:

- We propose an LLM-based framework, MART, for developing active merchant NPCs.
- Our findings show that LLMs enable merchants to appraise game items, highlighting that supervised finetuning can balance performance, efficiency, and reliability.
- Our results reveal that LLMs enable merchants to negotiate item prices, highlighting that knowledge distillation efficiently achieves high persuasiveness.
- Through statistical and qualitative analyses, we present multiple implementation options for active merchant NPCs, tailored to suit different user preferences.

2 Related Work

2.1 Using LLMs to Communicate with NPCs

Recently, game developers have started using LLMs in their games to provide more human-like interaction between players and NPCs. Specifically, there is a growing concern about making conversational NPCs with LLMs [Viggiato and Bezemer, 2024; Gallotta *et al.*, 2024; Christiansen *et al.*, 2024; Cox and Ooi, 2024; Gao *et al.*, 2023; Marincioni *et al.*, 2024]. For example, Christiansen *et al.* used LLM-based NPCs to support player interactions within a murder mystery game by helping players interrogate, collect clues, and explore. Buongiorno *et al.* introduced a framework for integrating LLM-based NPCs with memory to ensure narrative consistency during free-form player interactions, demonstrated in a turn-based, role-playing detective thriller game.

Despite their contributions, prior studies have largely treated LLM-based NPCs as information providers; That is, they only supported simple interactions where NPCs present knowledge or reasoning to the player. However, more complex forms of interaction that involve mutual influence, such as negotiation, have been largely underexplored in the context of LLM-integrated games.

In real-world applications, NLP researchers have investigated the potential of LLMs to negotiate [Jin *et al.*, 2024; Shea *et al.*, 2024; Hua *et al.*, 2024]. For example, Jin *et al.* leveraged LLMs to generate persuasive dialogues based on everyday scenarios and showed that their approach was more persuasive than other models. Similarly, Shea *et al.* suggested a personal negotiation coach based on LLMs and proved their effectiveness in assisting human negotiators. Inspired by prior work demonstrating the negotiation potential of LLMs, we examine negotiation interactions of an LLM-based NPC in a gaming context.

2.2 Using LLMs to Estimate Values

As item prices are usually fixed in a game system, research has not sufficiently focused on predicting item prices based on item specifications. An active merchant, however, should have the ability to decide item prices in its shop, similar to a real-world situation. Thus, we designed an appraiser module, inspired by studies on estimating prices of real assets.

In real-world applications, researchers have investigated machine learning models for estimating commodity prices. Early studies attempted to predict prices based on a fixed set of features [Patel *et al.*, 2015; Mohamed *et al.*, 2022]. For example, [Mohamed *et al.*, 2022] used a fixed set of features and a machine-learning method for predicting the price of seasonal goods. However, such models generally do not perform well when predicting prices of out-of-domain goods or newly introduced features. To handle such unseen goods or features, researchers utilized latent representations from textual descriptions of items with a language model [Ni *et al.*, 2024; Geng *et al.*, 2024]. For instance, Ni *et al.* demonstrated that LLMs can predict stock market movements from the earnings report of a company to support information-based investment. Extending prior works, we demonstrate that LLMs can also achieve strong predictive performance within games.

3 MART Framework

We propose MART, an LLM-powered framework for active merchant NPC. This framework consists of two main modules: *negotiator* and *appraiser*. Inspired by the behavior of real-world merchants, MART autonomously sets prices for its items and engages in price negotiation with players. When a player expresses interest in buying an item, the appraiser module suggests a retail price for the item. Based on this suggestion, the negotiator module begins negotiations with players who wish to buy the item from the merchant NPC. This scenario is different from the purchasing experience of players in current games, as shown in Figure 1.

First, the appraiser module is designed using a language model that interprets an item description to estimate its retail price. The module takes item descriptions presented as natural language sentences, which include details such as required level, effects, and durability. Since the structure of such descriptions can vary across different games, we adopt a natural language format to ensure versatility. In Section 4, we introduce LLM-based appraiser modules using two straightforward approaches and evaluate their performance.

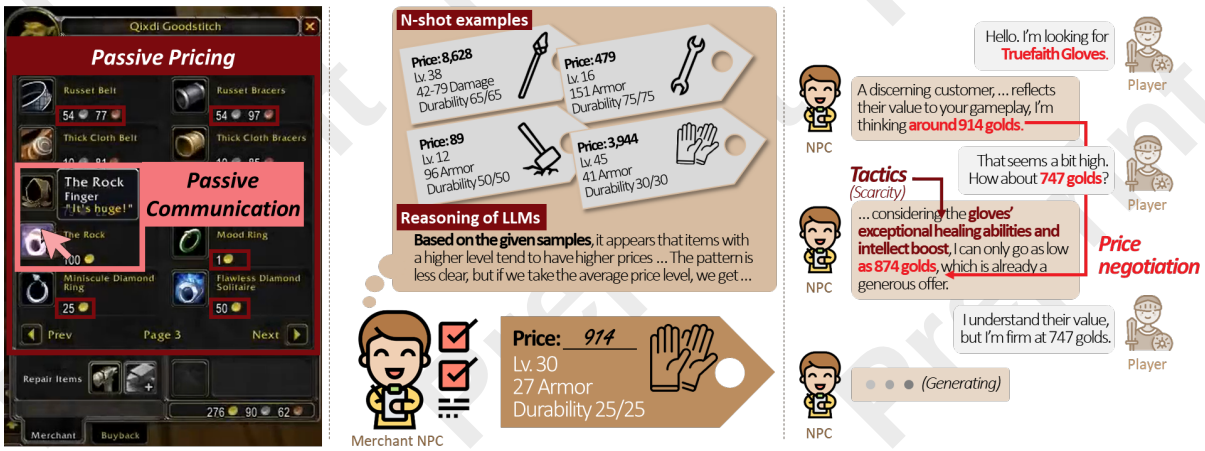
Second, the negotiator module employs a language model to facilitate negotiation dialogues with players. This module receives the item descriptions, retail price, and history of previous conversations. It then generates the appropriate response aimed at persuading the player. Since effective negotiation involves various tactics to persuade players, we experimented LLM-based negotiator modules with two simple approaches that integrate 10 tactics, as detailed in Section 5.

4 Appraiser Module

To develop the appraiser module, we experimented with two approaches: n-shot in-context learning (ICL) [Dong *et al.*, 2024] and supervised finetuning (SFT). Here, we set the n as 10 for our ICL approach and called LLMs via public APIs[2]. In this section, we outline each approach and illus-

[1] http://github.com/elu-lab/mart

[2] http://openrouter.ai



(a) Current Passive Interaction (b) Active Pricing with *Appraiser* (c) Active Communication with *Negotiator*

Figure 1: Comparison of current interaction and our MART framework: (a) a screenshot of trader in WoW, borrowed from a YouTube video (b) the proposed Appraiser module, and (c) the proposed Negotiator module.

trate our findings through an experiment on a WoW Classic item dataset.

4.1 Two Tested Approaches

First, to leverage the ability of LLMs to generalize from a set of given examples, we tested ICL method. ICL enhances reasoning through demonstrations and requires no additional training, making it a simple and efficient method to implement. Thus, we directly employed LLMs and provided ten random examples in the prompt. We used Llama models, a popular open-source family of LLMs: Llama-3.1 8b, 405b, and 3.2 1b. Note that ICL does not require any additional training to update LLM parameters.

Second, owing to the substantial computational resources and time required by the 405 billion parameter model, we tested the SFT method on smaller LLMs. To avoid catastrophic forgetting, we froze their parameters and trained an additional adapter. When we input an item description, the smaller LLM transforms the description into a single latent vector. Thereafter, the additional adapter uses the vector to predict the retail price of the item as a regression problem. We used Llama 3.1 8b and 3.2 1b.

4.2 Used Dataset

We used an item dataset from WoW (World of Warcraft) Classic [3] published by Blizzard Entertainment. By extending the existing WoW Classic item dataset, we created a dataset by crawling item information from the provided URLs for each item and retaining only purchasable items with disclosed prices. Dataset construction followed three steps. First, we collected 3,376 items by excluding item derivatives from the crawled dataset to simplify the problem. We then removed 106 items priced below 10 coppers (the cheapest currency unit in WoW) to make enough room for price negotiation during transactions. Finally, we converted all retail prices into coppers, using the conversions of 1 gold to 100 silver and 1 silver to 100 coppers. As a result, we collected a total of 3,270 items, with prices ranging from 10 coppers to 57,018 coppers. More than 50% of the items were priced below 1,250 coppers, with a median price of 1,238 coppers and an average price of 3,249 coppers. We further divided the dataset into a training set (80%, 2,616 items), validation set (10%, 327 items), and test set (10%, 327 items).

[3] https://www.kaggle.com/datasets/mylesoneill/classic-world-of-warcraft-auction-data

4.3 Evaluation Metrics

We used four metrics to evaluate the suitability of the two approaches for implementing the appraiser module: *mean absolute percentage error* (MAPE), *standard deviation*, *skewness*, and *unexpected output rate* (UOR). First, we used MAPE. Due to the wide range of item prices (from 10 to 57,018 coppers), the absolute errors of higher-priced items may overshadow those of lower-priced items. Therefore, we used percentage errors instead of absolute errors. We computed MAPE as the ratio of error to true price, as illustrated in Equation 2, where $\hat{y}_i$ is the predicted price and y_i^* is the true price for an item i.

$$PE_i = \frac{\hat{y}_i - y_i^*}{y_i^*} \quad (1)$$

$$\text{MAPE} = \mathbb{E}_i\left[\,|PE_i|\,\right]. \quad (2)$$

Second, we used the standard deviation σ of percentage errors (PEs). This metric can reveal the extent of variability in appraised prices. If the percentage errors are similar across different items, the standard deviation should be low.

Third, we measured the skewness of PEs. Skewness indicates whether a distribution leans toward positive or negative values. Therefore, we believe that analyzing skewness can help us determine if a model tends to underestimate or overestimate item prices. Mathematically, we computed skewness using Equation 3, where μ indicates the mean of PEs.

$$\text{Skewness} = \mathbb{E}_i\left[\left(\frac{PE_i - \mu}{\sigma}\right)^3\right]. \quad (3)$$

		MAPE	Std. Dev.	Skewness	UOR
ICL	1b	14.68	44.42	10.23	29.50
	8b	4.34	12.36	6.85	20.49
	405b	**1.34**	3.34	-5.06	**5.20**
SFT	1b	3.59	11.57	-5.84	-
	8b	2.66	11.26	**-3.47**	-

Table 1: Results of assessing two methods of appraiser

Lastly, we calculated the rate of unexpected outputs. When LLMs appraise an item, they may generate ambiguous appraisals to extract an exact price from the output sentences. For example, they sometimes suggest multiple prices even if we request a single prediction. We referred to these as unexpected outputs and estimated their frequency among the items. Note that this error only occurred in ICL models because SFT models directly produced the retail price through their prediction head. While LLMs can identify such unexpected cases, we manually labeled these errors.

4.4 Results

Table 1 shows the result of our experiments, comparing ICL and SFT methods. Of the five models, the ICL-405b achieved the best performance. It successfully outputted a retail price (94.8% of the cases) with the lowest MAPE (1.34%) and produced a few unexpected outputs (5.2%), while it was slightly skewed toward underestimation (-5.06). The second-best model was the SFT-8b. It exhibited a slightly higher MAPE score (2.66%) and the lowest absolute skewness toward underestimation (-3.47). Meanwhile, the SFT-1b demonstrated lower performance than the best model (3.59%) with a similar underestimation (-5.84). These results are different from those of smaller ICL models. The ICL-1b exhibited very high MAPE (14.68%) with highly overestimated prices (10.23) and a high unexpected output rate (29.05%). Similarly, the ICL-8b overestimated prices (6.85) with a moderate level of MAPE (4.34) and a high unexpected output rate (20.59%).

4.5 Discussion

We discuss our results in terms of three factors–*performance, efficiency*, and *reliability*–to assess the in-game applicability of the appraiser module. First, the appraiser has the potential to be used in a game because of its performance. Except for ICL-1b, five models demonstrated a MAPE of less than 5%. While the difference may seem substantial for high-priced items, it does not exceed 100 coppers for over half of the items, considering that the median price was 1,238 coppers. As gold coins are more frequently used in WoW Classic, a difference of 100 coppers is acceptable. Moreover, when using larger ICL models, the appraised prices aligned more closely with the true prices. Therefore, for game developers seeking a more precise appraiser module, ICL-405b is the best option, as it exhibited an appraisal error of less than 16 coppers for the median price.

Second, the SFT method is efficient for developing an appraiser module. Results show that SFT-8b performed much closer to ICL-405b than to ICL-8b; even the SFT-8b and ICL-8b used almost an identical number of parameters. Also, the SFT-1b outperformed the two ICL models, ICL-1b and ICL-8b. These results imply that it is possible to use a smaller LLM to implement an appraiser within a low-resource environment by finetuning it. Note that the ICL models achieved high performance by observing ten random pairs of item descriptions and prices. Researchers have reported that the performance of ICL methods can be improved by carefully curating examples or increasing the number of examples [Zhang *et al.*, 2022]. However, this curation demands additional human resources, and increasing the number of examples incurs higher computational costs. By contrast, a dataset for SFT can be generated with significantly fewer resources, using pairs of items and their prices.

Third, it is worth noticing that ICL methods can sometimes be unreliable when playing the role of an appraiser. We observed cases in which ICL models produced unexpected outputs. For instance, ICL-405b generated multiple price candidates or a continuous range of prices, despite our request for a single price output. Moreover, smaller ICL models occasionally failed to predict prices. Although we did not adopt any post-processing methods for these outputs, game developers should prevent such failures in their games. Alternatively, developers can use the SFT method, ensuring an LLM predicts a specific price.

5 Negotiator Module

We introduce a negotiator module inspired by real-world merchants, aiming to sell items while pursuing profit. To develop negotiators, we tested two approaches: zero-shot prompting (ZSP) and knowledge distillation (KD). We chose different methods from those used in the appraiser module, as supervised negotiation data is difficult to obtain. In the following subsections, we describe our methods, dataset generation procedure, and details of the experiments. Then, we discuss our findings by comparing these two approaches.

5.1 Two Tested Approaches

First, we employed the ZSP method in which an LLM generates negotiation dialogues without being provided any demonstrations. We initially considered a naïve negotiation method without using any specific tactics. This approach relied solely on using the pretrained knowledge of LLMs, which reflects a general understanding of the world; consequently, the generated negotiation strayed significantly from the intended game setting. Thus, we inputted 10 negotiation tactics within the prompt, as shown in Table 2, which are inspired by [Cialdini, 2001; Orji *et al.*, 2015; Guo and Barnes, 2009; Park and Lee, 2011]. We evaluated three Llama variants as in the appraiser.

Second, we employed knowledge distillation to allow smaller LLMs to achieve comparable performance with lower resource requirements. By transferring the knowledge of a larger model to a smaller model, the distillation reduces computational cost while maintaining performance. We trained student models on a dialogue dataset generated by teacher model. Here, we prompted the teacher model to select an appropriate negotiation tactic before generating each utterance

Six persuasion strategies	
Liking	building relationships through common ground or compliments
Reciprocity	the tendency to return favors
Social proof	mimicking observed behaviors
Consistency	aligning with past actions
Authority	trusting experts
Scarcity	valuing rare items
Four perceived values of a game item	
Enjoyment	enhancing gaming experience
Character competency	leveling up and boosting abilities
Visual authority	customizing characters to attract attention
Monetary value	reasonable pricing and good value

Table 2: Tactics used in the input prompt to help in negotiations

to distill its persuasiveness. To further reduce computational demands during distillation, we applied quantization and low-rank adaptation (LoRA) [Hu *et al.*, 2021]. Using these methods, we distilled knowledge about negotiation from Llama 3.1 405b to two smaller LLMs: Llama 3.1 8b and 3.2 1b.

5.2 Dataset

We prepared a dataset by generating a negotiation dialogue for each item in the appraiser dataset. To simulate a negotiation dialogue, we used two agents: a merchant and a player. As the merchant, we used Llama 3.1 405B, the teacher model. As the player, we used GPT-4o with input prompts about tactics. We used different models for these two agents to (1) avoid adopting a similar reasoning process and (2) make the player as similar as possible to human beings, as reported in [Kwon *et al.*, 2024]. To simulate the diversity of real-world customers, we used a temperature value of 1.0 for GPT-4o.

The negotiation procedure was as follows. Before starting a negotiation, both agents received an item description of the negotiation subject as shared information. The two agents were assigned different prices for each item: the player wanted to buy the item at a discount of 10% to 25%. The player started the negotiation with an utterance "`Hello. I'm looking for [the item name].`" Although greetings and small talk are common in real-world negotiations, we intentionally avoided including such content to ensure that the models learned to focus on goal-directed negotiation behaviors.

After the initial utterance, the two agents engaged in price negotiation until the player decided whether to purchase the item or not. Until the decision was made, the merchant kept persuading the player to buy the item to simulate a real-world merchant. The player could terminate the conversation by mentioning "`conversation over`." To avoid long dialogues, we set the maximum number of turns to 15. As a result, we generated 2,943 conversations: 2,616 conversations from the training set items and 327 conversations from the validation set items.

5.3 Evaluation Metrics

We compared three ZSP models and two KD models using three evaluation metrics: *persuasiveness*, *dominance*, and *agreement*. First, we measured the persuasiveness of each utterance generated by the tested negotiators. Similar to commercial behavior in the real world, the merchant NPC should effectively persuade players to buy or sell an item at a favorable price. Various tactics are used in an effective negotiation; so, we aimed to evaluate the effectiveness of the negotiator in using 10 different tactics to create persuasive statements. To measure persuasiveness, we used G-Eval [Liu *et al.*, 2023], a widely-used evaluation method using LLMs. The method directly asks GPT-4 to evaluate an input text according to given criteria. Specifically, we used a 5-point scale and averaged 20 runs following G-Eval.

Second, we measured the dominance of the merchant over the player during the negotiation. This metric indicates whether the merchant holds more power in the relationship with the player, highlighting the concept of power dynamics [Kim *et al.*, 2005]. A human merchant usually has the initiative in price negotiation, rather than giving the initiative to the customer. So, as a negotiated price increases, the merchant profits more while the player incurs greater losses. In other words, the negotiation is a type of zero-sum game. Considering such power dynamics, we measured dominance using Equation 4, where y, y^m, and y^p indicate the agreed price, retail price, and price desired by the player, respectively. The equation quantifies the ratio between the profit gained by the merchant ($y_i - y_i^p$) and the gap between two agents ($y_i^m - y_i^p$).

$$\text{Dominance} = \mathbb{E}_i \left[\frac{y_i - y_i^p}{y_i^m - y_i^p} \right] \tag{4}$$

Third, we defined the agreement rate as the proportion of negotiations in which the merchant and the player reached a mutual settlement on a specific price. While dominance captures how favorable the final price is to the merchant, relying on it alone may lead to misinterpretation. A merchant who persistently offers unreasonably high prices might achieve high dominance in a few successful cases, while failing most negotiations. Therefore, we also consider the agreement rate, which measures how often a mutual settlement is reached across total negotiations.

5.4 Results

Table 3 shows the results of our experiments, comparing ZSP and KD methods. The results show that the KD-8b performed the best and successfully used persuasive tactics. The model achieved a score of 3.99 in terms of persuasiveness, followed by ZSP-405b (3.92), ZSP-8b (3.74), KD-1b (3.65), and ZSP-1b (2.95). The ZSP-405b had the strongest dominance (0.47), followed by ZSP-8b (0.42), and KD-8b (0.40). The ZSP-8b had the highest agreement (96.94%), followed by ZSP-405b (90.83%) and ZSP-1b (90.52%).

		Persuasiveness	Dominance	Agreement
ZSP	1b	2.95 ± 0.93 (N=1,650)	0.14 ± 0.54 (N=296)	90.52%
	8b	3.74 ± 0.63 (N=1,620)	$\underline{0.42} \pm 0.25$ (N=317)	**96.94**%
	405b	$\underline{3.92} \pm 0.53$ (N=1,502)	$\mathbf{0.47} \pm 0.22$ (N=297)	$\underline{90.83}$%
KD	1b	3.65 ± 0.67 (N=1,359)	0.26 ± 0.26 (N=270)	82.57%
	8b	$\mathbf{3.99} \pm 0.50$ (N=1,390)	0.40 ± 0.17 (N=265)	81.04%

Table 3: Results of assessing two methods of negotiator. Numbers in parentheses of persuasiveness and dominance indicate the number of utterances and settled negotiations, respectively.

One-way ANOVA: $F_{4,7516} = 600.67, p < 0.001^{***}$				
Posthoc comparison			Mean Diff.	Adj. p
ZSP-1b	*vs.*	ZSP-8b	0.789	$<0.001^{***}$
ZSP-1b	*vs.*	ZSP-405b	0.973	$<0.001^{***}$
ZSP-1b	*vs.*	KD-1b	0.702	$<0.001^{***}$
ZSP-1b	*vs.*	KD-8b	1.044	$<0.001^{***}$
ZSP-8b	*vs.*	ZSP-405b	0.184	$<0.001^{***}$
ZSP-8b	*vs.*	KD-1b	-0.086	0.005^{**}
ZSP-8b	*vs.*	KD-8b	0.255	$<0.001^{***}$
ZSP-405b	*vs.*	KD-1b	-0.271	$<0.001^{***}$
ZSP-405b	*vs.*	KD-8b	0.070	0.040^{**}
KD-1b	*vs.*	KD-8b	0.341	$<0.001^{***}$

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 4: Results of statistical test on persuasiveness

We further performed statistical tests to verify differences between the five models in terms of persuasiveness and dominance. We conducted a one-way ANOVA to examine whether there were significant differences among groups using the `statsmodels` library [Seabold and Perktold, 2010]. Also, we conducted a post-hoc analysis using the Tukey-HSD test and p-value adjustment to identify group differences. Tables 4 and 5 show the statistical result. First, we observed that model differences affect the persuasiveness ($p < 0.001$). In detail, pairwise differences are all significant: ZSP-405b versus KD-8b (p=0.04), ZSP-8b versus KD-1b (p=0.004), and the other eight pairs ($p < 0.001$) are all significant. Second, we also observed that model differences affect the dominance ($p < 0.001$). Pairwise differences are all significant except for two pairs: ZSP-405b versus ZSP-8b (p=0.246) and ZSP-8b versus KD-8b (p=0.891) are statistically insignificant.

5.5 Discussion

In open-world games, we believe that merchant NPCs should contribute to immersive player experiences. To support such immersion, developers need to consider two essential parts of a merchant: its *purpose* and *position* within the game world. We discuss our findings with respect to these two aspects.

One-way ANOVA: $F_{4,1440} = 53.53, p < 0.001^{***}$				
Posthoc comparison			Mean Diff.	Adj. p
ZSP-1b	*vs.*	ZSP-8b	0.279	$<0.001^{***}$
ZSP-1b	*vs.*	ZSP-405b	0.331	$<0.001^{***}$
ZSP-1b	*vs.*	KD-1b	0.118	$<0.001^{***}$
ZSP-1b	*vs.*	KD-8b	0.254	$<0.001^{***}$
ZSP-8b	*vs.*	ZSP-405b	0.052	0.247
ZSP-8b	*vs.*	KD-1b	-0.160	$<0.001^{***}$
ZSP-8b	*vs.*	KD-8b	-0.024	0.891
ZSP-405b	*vs.*	KD-1b	-0.212	$<0.001^{***}$
ZSP-405b	*vs.*	KD-8b	-0.076	0.036^{*}
KD-1b	*vs.*	KD-8b	0.136	$<0.001^{***}$

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 5: Results of statistical test on dominance

First, small models can effectively support the objectives of a merchant by using the KD method. Note that the persuasiveness metric measures whether the merchant successfully employed tactics to persuade the player. Therefore, a higher persuasiveness score indicates that the negotiator is better equipped to fulfill the purpose of the merchant. The results show that the largest model (ZSP-405b) achieved a significantly high score. Additionally, we observed that the KD models effectively learnt persuasive tactics from their teacher model, Llama 3.1 405b. In other words, the KD models generally outperformed the ZSP models with the same size. So, it is reasonable to use smaller KD models instead of large ZSP models; we recommend KD-8b as the best option.

Second, the ZSP and KD methods can introduce different merchant personalities. In WoW Classic, there are many merchants in a variety of situations, including differences in race and faction. These factors should be considered when developing merchants to enhance the immersive experience of players. For example, a merchant can use different policies for different players based on their races or factions; a merchant can impose higher prices on rich players or players unfavorable to the merchant. As an early stage, our results imply such distinction is possible using different models. To create an agent favorable to players, developers can opt for models with high agreement rates and low dominance such as ZSP-1b or ZSP-8b. Conversely, if developers are looking for more stubborn merchants, models with low agreement rates and high dominance, like KD-8b, would be a good option.

5.6 Additional Case Analysis

We further analyzed negotiation utterances to identify irregular cases that should be considered when deploying MART in a game. Three of the authors manually labeled utterances of five models and discovered three irregular cases: *giveaways*, *improvisations*, and *arithmetic errors*.

The giveaway case refers to situations where a merchant offers additional items or enhancements to persuade a player

to make a purchase. The quote following this paragraph provides an example. We suspect this case occurred because the negotiator used the reciprocity tactic. Although this practice is common in the real world, giveaways are typically not incorporated into a game. Thus, developers should handle this case when integrating negotiators into their game world. If developers allow merchants to provide giveaways, their game system should support such cases. Otherwise, developers should prevent the negotiator from using such a tactic.

> **Player:** *Hello. I'm looking for a Cadet Belt.*
>
> **ZSP-405b:** *You've got a great eye for quality gear, ...*
>
> ... (omitted) ...
>
> **ZSP-405b:** *I understand your position, but ..., I'll throw in one final sweetener.* ***I'll give you a free upgrade*** *to a premium pouch, ...*

The improvisation case indicates that merchants sometimes create spontaneous statements about their stock lists. This error mainly occurred in ZSP-1b, the smallest model without training. An example of this case is shown in the quote following this paragraph. We suspect that this improvisation is a form of "hallucination." In natural language processing, this term describes instances where language models generate nonsensical or inaccurate text in relation to the provided input sources [Maynez *et al.*, 2020]. Prior studies have also shown that smaller language models are more likely to generate hallucinated outputs [Carlini *et al.*, 2023]. Therefore, developers should be prepared for the possibility that merchants may suggest nonexistent items. This case can be handled through two approaches: mitigation or approval. To mitigate the problem, we recommend that developers implement a retrieval-augmented generation system [Lewis *et al.*, 2020], which encourages language models to produce outputs based on a specific database. For approval, developers should establish a confirmation system to assess the relevance and legitimacy of any newly introduced items by merchants.

> [The Conjurer's Bracers is in WoW Classic, but the Conjurer's Sigil Cloak is not.]
>
> **Player:** *Hello. I'm looking for conjurer's bracers.*
>
> **ZSP-1b:** *The Conjurer's Bracers are in fact* ***not currently available*** *in our inventory, but we do have a special promotion for* ***the conjurer's sigil cloak*** *...*
>
> ... (omitted) ...
>
> **ZSP-1b:** *... Congratulations, you now own the conjurer's sigil cloak!*
>
> **Player:** *Thank you for the detailed invoice, I look forward to receiving the items.*

In the arithmetic error case, merchants sometimes struggle with basic calculations involving item quantity or price. This case is typically seen in smaller negotiators such as ZSP-1b, ZSP-8b, KD-1b, and KD-8b. An example of this case is shown in the quote following this paragraph. We attribute such errors to the lack of arithmetic capability in small language models. Previous studies on LLMs also support this speculation because smaller LLMs often struggle with arithmetic tasks [Touvron *et al.*, 2023; Grattafiori *et al.*, 2024; OpenAI *et al.*, 2024]. The arithmetic error case may disrupt negotiations and lead to embarrassment for players. Therefore, developers need to address these issues when incorporating smaller LLMs as negotiation modules. One potential solution is to use external calculators.

> [1455 golds are not a 15% discount from 1569 golds; the answer is 1333.65 golds.]
>
> **ZSP-1b:** *... The Conjurer's Sigil Cloak* ***originally cost 1569 golds,*** *and if you're willing to trade in the valuable items, I can give you a* ***15% discount*** *on that. ... That brings the* ***total down to 1455 golds,*** *but I think that's a fair trade. What do you say? ...*

6 Conclusion

We proposed a novel framework named MART, consisting of two LLM-based modules–*appraiser* and *negotiator*–designed to resolve two key limitations of passive merchant NPCs: pricing and communication. To support implementation decisions under varying deployment conditions, we conducted experiments comparing multiple implementation approaches for each module. Our results demonstrate that our frameworks can construct an active merchant NPC in WoW Classic using larger LLMs without additional training. We also showed that smaller LLMs can achieve acceptable performance when trained via supervised finetuning or knowledge distillation. Moreover, we identified several concerns developers may face when integrating LLMs into interactive NPCs, including promising unrealisitc giveaways during negotiation. We expect that these findings can generalize to other open-world games and trading NPCs.

Despite these contributions, this study has three limitations. First, our experimental results are based on the Llama LLM family. This may lower the generalizability of our results because different LLMs can perform differently. Second, we simulated player negotiation using GPT-4o. Humans may use different tactics to neutralize the negotiator. Lastly, our discussion centered on misbehavior that may emerge inherently from LLM-based merchants, rather than on adversarial threats like prompt injection by malicious users. Further investigations are required to address these limitations. Nonetheless, we believe this study can serve as a foundation for further studies on integrating LLMs to active merchants.

Contribution Statement

Byungjun Kim and Minju Kim contributed equally to this work and are considered co-first authors. Bugeun Kim is the corresponding author.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)]

References

[Buongiorno *et al.*, 2024] Steph Buongiorno, Lawrence Klinkert, Zixin Zhaung, Tanishq Chawla, and Corey Clark. Pangea: procedural artificial narrative using generative ai for turn-based, role-playing video games. In *Proceedings of the Twentieth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE '24. AAAI Press, 2024.

[Carlini *et al.*, 2023] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[Christiansen *et al.*, 2024] Frederik Roland Christiansen, Linus Nørgaard Hollensberg, Niko Bach Jensen, Kristian Julsgaard, Kristian Nyborg Jespersen, and Ivan Nikolov. Exploring presence in interactions with llm-driven npcs: A comparative study of speech recognition and dialogue options. In *Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology*, VRST '24, New York, NY, USA, 2024. Association for Computing Machinery.

[Cialdini, 2001] Robert B Cialdini. The science of persuasion. *Scientific American*, 284(2):76–81, 2001.

[Cox and Ooi, 2024] Samuel Rhys Cox and Wei Tsang Ooi. Conversational interactions with npcs in llm-driven gaming: Guidelines from a content analysis of player feedback. In *Chatbot Research and Design: 7th International Workshop, CONVERSATIONS 2023, Oslo, Norway, November 22–23, 2023, Revised Selected Papers*, page 167–184, Berlin, Heidelberg, 2024. Springer-Verlag.

[Dong *et al.*, 2024] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Gallotta *et al.*, 2024] Roberto Gallotta, Graham Todd, Marvin Zammit, Sam Earle, Antonios Liapis, et al. Large language models and games: A survey and roadmap. *IEEE Transactions on Games*, pages 1–18, 2024.

[Gao *et al.*, 2023] Fengsen Gao, Ke Fang, and Wai Kin Victor Chan. Chemical life: Knowledge-based personality, emotion and action cues in educational games. In *2023 IEEE Conference on Games (CoG)*, pages 1–3, 2023.

[Geng *et al.*, 2024] Binzong Geng, Zhaoxin Huan, Xiaolu Zhang, Yong He, Liang Zhang, et al. Breaking the length barrier: Llm-enhanced ctr prediction in long textual user behaviors. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2311–2315, New York, NY, USA, 2024. Association for Computing Machinery.

[Grattafiori *et al.*, 2024] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The llama 3 herd of models, 2024.

[Guo and Barnes, 2009] Yue Guo and Stuart Barnes. Virtual item purchase behavior in virtual worlds: An exploratory investigation. *Electronic Commerce Research*, 9:77–96, 2009.

[Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[Hua *et al.*, 2024] Yuncheng Hua, Lizhen Qu, and Reza Haf. Assistive large language model agents for socially-aware negotiation dialogues. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8047–8074, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Jin *et al.*, 2024] Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. Persuading across diverse domains: a dataset and persuasion large language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Kim *et al.*, 2005] Peter H Kim, Robin L Pinkley, and Alison R Fragale. Power dynamics in negotiation. *Academy of Management Review*, 30(4):799–822, 2005.

[Kwon *et al.*, 2024] Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale Lucas, et al. Are LLMs effective negotiators? systematic evaluation of the multifaceted capabilities of LLMs in negotiation dialogues. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5391–5413, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Latitude, 2019] Inc. Latitude. Ai dungeon. https://aidungeon.com/, 2019. Accessed: 2025-05-25.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[Li *et al.*, 2024] Jialu Li, Yuanzhen Li, Neal Wadhwa, Yael Pritch, David E. Jacobs, Michael Rubinstein, Mohit Bansal, and Nataniel Ruiz. Unbounded: A generative infinite game of character life simulation, 2024.

[Liu *et al.*, 2023] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, et al. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[Marincioni *et al.*, 2024] Alessandro Marincioni, Myriana Miltiadous, Katerina Zacharia, Rick Heemskerk, Georgios Doukeris, et al. The effect of llm-based npc emotional states on player emotions: An analysis of interactive game play. In *2024 IEEE Conference on Games (CoG)*, pages 1–6, 2024.

[Maynez *et al.*, 2020] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics.

[Mohamed *et al.*, 2022] Mohamed Ali Mohamed, Ibrahim Mahmoud El-Henawy, and Ahmad Salah. Price prediction of seasonal items using machine learning and statistical methods. *Computers, Materials & Continua*, 70(2), 2022.

[Ni *et al.*, 2024] Haowei Ni, Shuchen Meng, Xupeng Chen, Ziqing Zhao, Andi Chen, et al. Harnessing earnings reports for stock predictions: A qlora-enhanced llm approach. In *2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, pages 909–915, 2024.

[OpenAI *et al.*, 2024] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. Gpt-4 technical report, 2024.

[Orji *et al.*, 2015] Rita Orji, Regan L Mandryk, and Julita Vassileva. Gender, age, and responsiveness to cialdini's persuasion strategies. In *Persuasive Technology: 10th International Conference, PERSUASIVE 2015, Chicago, IL, USA, June 3-5, 2015, Proceedings 10*, pages 147–159. Springer, 2015.

[Park and Lee, 2011] Bong-Won Park and Kun Chang Lee. Exploring the value of purchasing online game items. *Computers in human behavior*, 27(6):2178–2185, 2011.

[Patel *et al.*, 2015] Jigar Patel, Sahil Shah, Priyank Thakkar, and K Kotecha. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1):259–268, 2015.

[Peng *et al.*, 2024] Xiangyu Peng, Jessica Quaye, Sudha Rao, Weijia Xu, Portia Botchway, et al. Player-driven emergence in llm-driven game narrative. In *2024 IEEE Conference on Games (CoG)*, pages 1–8, 2024.

[Phillips *et al.*, 2024] Adon Phillips, Jochen Lang, and David Mould. Goal-oriented interactions in games using llms. *IEEE Transactions on Games*, pages 1–12, 2024.

[ReLU Games and KRAFTON, 2024] ReLU Games and KRAFTON. Uncover the smoking gun. https://store.steampowered.com/app/2492290/Uncover_the_Smoking_Gun/, 2024. Accessed: 2025-05-25.

[Seabold and Perktold, 2010] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[Shea *et al.*, 2024] Ryan Shea, Aymen Kallala, Xin Lucy Liu, Michael W. Morris, and Zhou Yu. ACE: A LLM-based negotiation coaching system. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12720–12749, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[Viggiato and Bezemer, 2024] Markos Viggiato and Cor-Paul Bezemer. Leveraging the opt large language model for sentiment analysis of game reviews. *IEEE Transactions on Games*, 16(2):493–496, 2024.

[Zhang *et al.*, 2022] Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.