Conversational AI for Positive-sum Retailing under Falsehood Control

Yin-Hsiang Liao
Academia Sinica
zenonliao@iis.sinica.edu.tw

Ruo-Ping Dong
Academia Sinica
dongruoping@gmail.com

Huan-Cheng Chang
Academia Sinica
changhc84@gmail.com

Wei-Yun Ma
Academia Sinica
ma@iis.sinica.edu.tw

Abstract

Retailing combines complicated communication skills and strategies to reach an agreement between buyer and seller with identical or different goals. In each transaction a good seller finds an optimal solution by considering his/her own profits while simultaneously considering whether the buyer's needs have been met. In this paper, we manage the retailing problem by mixing cooperation and competition. We present a rich dataset of buyer-seller bargaining in a simulated marketplace in which each agent values goods and utility separately. Various attributes (preference, quality, and profit) are initially hidden from one agent with respect to its role; during the conversation, both sides may reveal, fake, or retain the information uncovered to come to a final decision through natural language. Using this dataset, we leverage transfer learning techniques on a pretrained, end-to-end model and enhance its decision-making ability toward the best choice in terms of utility by means of multi-agent reinforcement learning. An automatic evaluation shows that our approach results in more optimal transactions than human does. We also show that our framework controls the falsehoods generated by seller agents. The code and dataset are available on https://github.com/ckiplab/Fruit_Stand.

1 Introduction

Retailing is a mixture of cooperation and competition between buyer and seller. The construction of virtual retailers has received widespread attention due to their broad applications in the E-commerce era. If the focus of the conversational retailer is limited to the buyer's needs, the retailer is actually a conversational recommendation system. However, if the conversational retailer's purpose is to maximize his/her own profit, the retailer is in fact a negotiation system, which typically must use discourse with opponents to perceive their intent and build strategies to achieve the retailer's own goals (Keizer et al., 2017; Afantenos et al., 2012).

Previous NLP research on negotiation concerns closed-domain scenarios in games such as Settlers of Catan (Asher and Lascarides, 2013), goods distribution (DeVault et al., 2015; Lewis et al., 2017), and open-ended settings, for example, price bargaining on a single item in a zero-sum, second-hand market (He et al., 2018). However, these scenarios do not attempt to find an optimal solution for both sides, which crucially defines a good retailer who always takes into account future transactions.

Therefore, inspired by Shapiro (1983), we propose a positive-sum setting in this paper: a buyer and a seller negotiate to achieve a transaction, and the seller not only considers his/her profit but also takes into account whether the buyer's needs have been met, thus seeking a mutually optimal solution. To simulate such a real-world vending scenario and provide enough motivation to start a conversation, both buyer and seller are given incomplete information prior to the conversation. The buyer knows what he/she prefers among multiple products but does not know the quality of the product prior to the conversation, and the seller does not know in advance the buyer's preferences but is aware of the quality of the product and its profit. The seller seeks a mutually optimal solution by which to build his/her own reputation for future business while simultaneously making a profit. Thus we propose separate utility functions for buyers and sellers.

To facilitate end-to-end fine-tuning for this scenario, we collected a large dataset of 4232 dialogues between two people negotiating on goods in a simulated market on Amazon Mechanical Turk (AMT). Our model is based on the Transformer architecture (Vaswani et al., 2017), which is predominant in recent NLP research, due in part to its inherent parallelism, which facilitates the use of large-scale datasets to train complex models such as GPT2 (Radford et al., 2019), evolved Transformer (So et al., 2019), and T5 (Raffel et al., 2020). Further, these complex models are often pre-trained

Proceedings of the 4th Workshop on NLP for Conversational AI, pages 21 - 33
May 27, 2022 ©2022 Association for Computational Linguistics

in an unsupervised fashion, yielding powerful performance in downstream tasks in an end-to-end, supervised manner, which lays the foundation for training two acceptable conversational agents to fit the proposed scenario. The supervised fine-tuning maximizes the likelihood of human utterances in the dataset. To maximize agent targets, we leverage reinforcement learning (RL) to direct the fine-tuning process.

In addition, due to the increasing saturation of machine learning algorithms in contemporary society, there has been a surge in interest in building truthful AI, a system that avoids stating falsehoods, thus enhancing transparency and helping to establish trust between system and human (Evans et al., 2021). To achieve such truthful AI, we attempt to reduce a certain type of statement against the ground truth in our negotiation scenario. First, we build a falsehood detector with respect to such statements. Second, we formulate a deduction mechanism in the RL stage to decrease the generation of falsehoods.

In summary, the contributions of our study are the following:

- We propose a simplified market setting where vendor and purchaser are in a "coopetitive" relation with information asymmetry. To this purpose we gathered FruitStand, a rich dataset of human-human negotiations under this scenario on Amazon Mechanical Turk (AMT).
- We propose an RL framework by which to cause a virtual retailer to learn how to find optimal solutions under positive-sum situation.
- The experiments demonstrate the effectiveness of reinforcement learning in improving the ability to achieve optimal transactions.
- We analyze the lies in a crowd-sourced dataset and the falsehoods generated by the seller model, based on which we propose an approach to reduce falsehoods.

2 Data Collection

In this paper, we discuss the behavior of two conversational agents negotiating given imperfect information. To promote end-to-end training, we collected FruitStand, a dataset of human-human dialogues designed around a novel scenario which simulates a fruit stand at which the negotiation takes place. In FruitStand, one agent plays the role of the buyer and the other that of the seller, communicating in natural language, developing strategies and eventually making a deal.

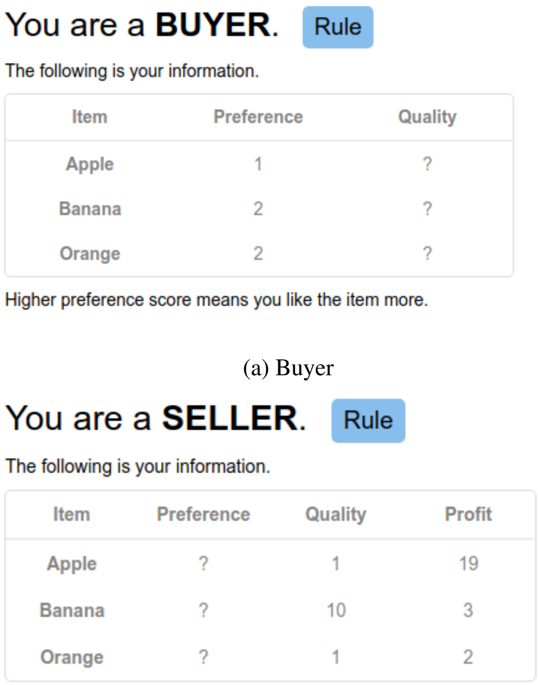
2.1 Task

The scenario simulates two agents transacting at a fruit stand. In each dialogue, the agents are first assigned a role, either buyer or seller, and the order of turns in which to send natural language messages. There are 3 item types—*apples*, *bananas*, and *oranges*—each of which has three attributes—*preference*, *quality*, and *profit*—as shown in upper left corner of Fig. 2. These 9 attributes determine the initial condition o. The buyer and seller each have an individual utility. The buyer's utility $U_b(item)$ to an item is defined as $\mathrm{preference}(item) \times \mathrm{quality}(item)$, following the intuition that the buyer is satisfied by purchasing what he/she likes in excellent condition (e.g., red, sweet, and juicy apples). Likewise, the seller's utility $U_s(item)$ is defined as $U_b + \mathrm{profit}(item)$, taking into account the seller's current profit and the buyer's satisfaction for future profit, since the buyer might become a regular if he/she is satisfied. Each agent's best option is that which provides the highest utility. Depending on the best options, the agents' goals may be identical, or may conflict, which leads to opportunities for cooperation or competition, respectively.

In each dialogue, buyer and seller bargain turn by turn, trying to make a deal on their own best option(s). Agents possess imperfect information. Initially, the buyer knows only its *preference*, and the seller only the *quality* and *profit* of an item. During the conversation, they must estimate the other's exclusive attributes by skill of speech, all the while not revealing any exact values. Absolute honesty is not required; agents can be deceptive. In particular, the seller may mislead the buyer when a given item is more profitable; however, the final decision lies with the buyer. Each conversation ends when the buyer makes a decision; typically this occurs within 5 to 20 turns. The design of the utility functions and the right to choose compensates for the buyer's inferior position in terms of the amount of information.

2.2 Collection

We collected the FruitStand dataset based on the above task via AMT with the interface shown in Figures 1 and 3. Workers were paid per dialogue, with a bonus for achieving the best option in terms of utility. The starter of a dialogue could be either a seller or buyer, and we kept the number of starters from both sides roughly balanced. The dataset statistics in Table 1 show that FruitStand has longer and more variant dialogues than DealorNodeal(Lewis et al., 2017). FruitStand has a total of 4232 dialogues with unique initial conditions, 76.1% of which have mutually optimal solutions (the overlapping best options from two sides), as illustrated in Table 2. We partitioned 80%/10%/10% of the dialogues for training/validation/testing.



You are a **BUYER**. Rule

The following is your information.

Item	Preference	Quality
Apple	1	?
Banana	2	?
Orange	2	?

Higher preference score means you like the item more.

(a) Buyer

You are a **SELLER**. Rule

The following is your information.

Item	Preference	Quality	Profit
Apple	?	1	19
Banana	?	10	3
Orange	?	1	2

(b) Seller

Figure 1: At the start of each conversation, the buyer knows only his/her preferences, and the seller knows only the quality and profit.

	FS	DN
Number(#) of Dialogues	4232	5808
Average Turns per Dialogue	7.8	6.5
Average Words per Turn	11.6	7.6
Vocabulary Size	4318	2719
Vocabulary Size without Numbers	4229	2623
% Agreed	100	80.1

Table 1: Comparison of dataset statistics of FruitStand and DealorNodeal. FruitStand contains longer, more variant dialogues on average.

	Number (Ratio)
Buyer's optimal selection chosen	2767 (65.4%)
Seller's optimal selection chosen	2966 (70.1%)
Mutually optimal occasion	3222 (76.1%)
Mutually optimal selection chosen	2464 (58.2%)

Table 2: Statistics of final deals in the whole FruitStand dataset.

3 Retailer

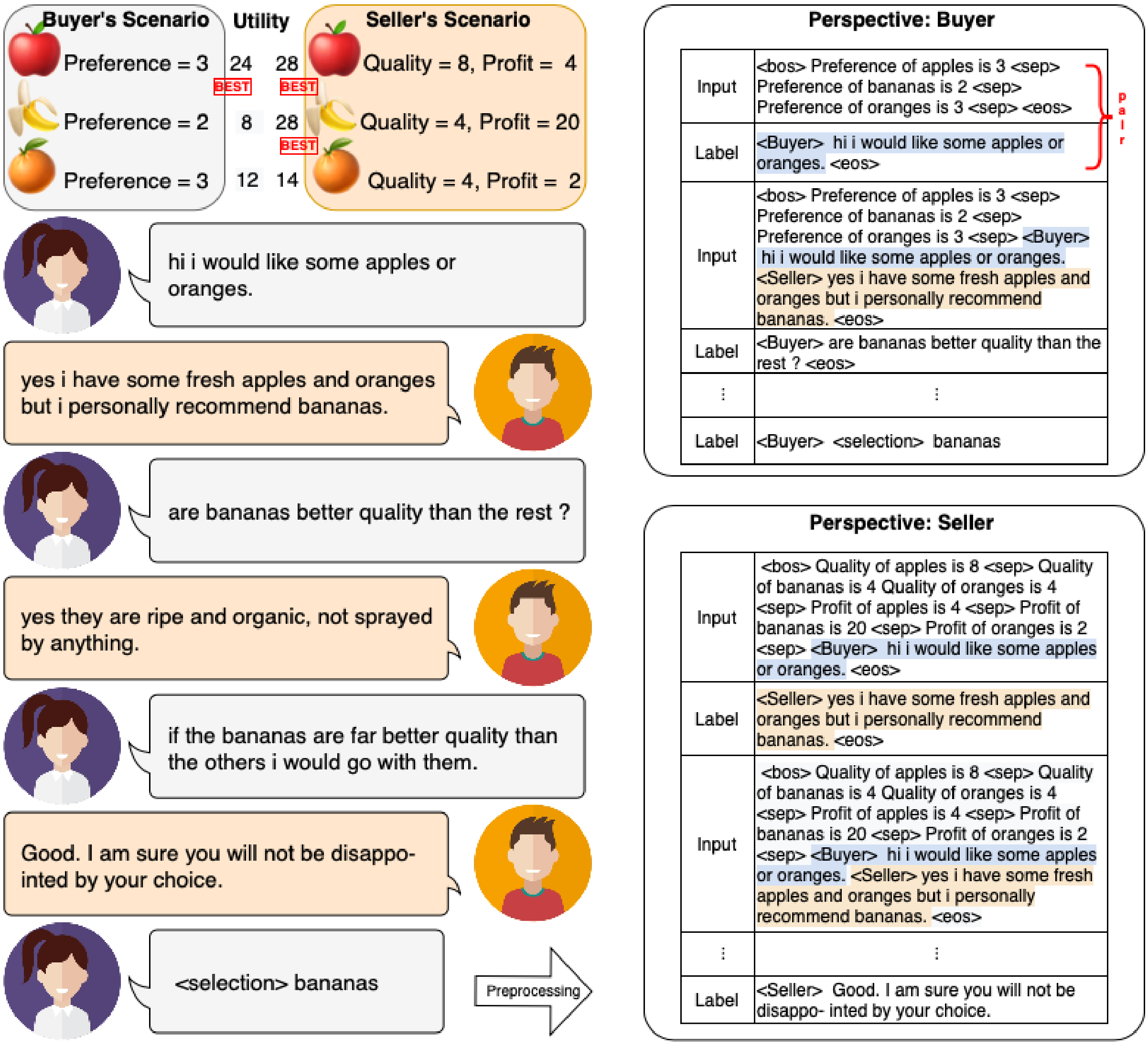
3.1 Data Representation

Every turn in a dialogue is transformed into a training pair—input sequence X and label sequence Y—from the perspective of the agent. For example, as illustrated in Figure 2, the buyer starts the conversation, and its preferences and utterance in this turn are converted into the first training pair of the dialogue, $\langle X_1^B, Y_1^B \rangle$. Note that $Y_1^B = \{y_{11}^B, y_{12}^B, ..., y_{1T}^B\}$, where y_{ij} is a token and T is the length of the utterance at this turn. Next, the seller's scenario along with the buyer's previous utterance and its response in this turn become the second training pair, $\langle X_1^S, Y_1^S \rangle$, the seller's first. The process continues until the end of the conversation. A similar technique has been used, see, e.g., Wolf et al. (2019). Note that we take the natural form for the agents' scenario, o^B and o^S, instead of merely numbers, to leverage the words' underlying information from pretrained models.

3.2 Baseline Models

For the first training stage, we fine-tune a T5 model (Raffel et al., 2020) pretrained on our FruitStand dataset. T5 is a standard encoder-decoder Transformer (Vaswani et al., 2017) which regards all NLP tasks as a text-to-text format. We leverage its baseline version (T5-base) as described in Raffel et al. (2020) as our starting point. T5-base is a composite of 12 Transformer blocks (each block combines self-attention, optional encoder-decoder attention, and a feedforward layer with a hidden size of 3072). It performs well on downstream tasks as varied as machine translation, document summarization, and sentiment classification.

The pretrained model is then fine-tuned as in supervised learning (SL), i.e., by minimizing the cross-entropy loss between the generated sequence Z and the label sequence Y described in Sec. 3.1. We have two transfer paths: one for the buyer and one for the seller. The buyer path uses labels from the buyer's perspective, and the seller path uses its part in the dialogue. The pair of the



hi i would like some apples or oranges.

yes i have some fresh apples and oranges but i personally recommend bananas.

are bananas better quality than the rest ?

yes they are ripe and organic, not sprayed by anything.

if the bananas are far better quality than the others i would go with them.

Good. I am sure you will not be disappo- inted by your choice.

<selection> bananas

Perspective: Buyer

Input	<bos> Preference of apples is 3 <sep> Preference of bananas is 2 <sep> Preference of oranges is 3 <sep> <eos>
Label	<Buyer> hi i would like some apples or oranges. <eos>
Input	<bos> Preference of apples is 3 <sep> Preference of bananas is 2 <sep> Preference of oranges is 3 <sep> <Buyer> hi i would like some apples or oranges. <Seller> yes i have some fresh apples and oranges but i personally recommend bananas. <eos>
Label	<Buyer> are bananas better quality than the rest ? <eos>
⋮	⋮
Label	<Buyer> <selection> bananas

Perspective: Seller

Input	<bos> Quality of apples is 8 <sep> Quality of bananas is 4 Quality of oranges is 4 <sep> Profit of apples is 4 <sep> Profit of bananas is 20 <sep> Profit of oranges is 2 <sep> <Buyer> hi i would like some apples or oranges. <eos>
Label	<Seller> yes i have some fresh apples and oranges but i personally recommend bananas. <eos>
Input	<bos> Quality of apples is 8 <sep> Quality of bananas is 4 Quality of oranges is 4 <sep> Profit of apples is 4 <sep> Profit of bananas is 20 <sep> Profit of oranges is 2 <sep> <Buyer> hi i would like some apples or oranges. <Seller> yes i have some fresh apples and oranges but i personally recommend bananas. <eos>
⋮	⋮
Label	<Seller> Good. I am sure you will not be disappo- inted by your choice.

Figure 2: Transforming a crowd-sourced dialogue (left) into a series of training pairs (input, label) from perspectives of the two agents. The buyer only knows its own preference, while the seller only knows the quality and profits of fruits

two picked models, denoted as $\langle M_\phi^B, M_\theta^S \rangle$, forms the baseline for later evaluation, where ϕ and θ are the learned parameters of the buyer and seller model, respectively. See Sec. 4.1 for more details.

3.3 Goal-oriented Reinforcement

The goal of supervised learning is to imitate average human behavior; however, not every person is good at making deals. We further fine-tune the agents via reinforcement learning to improve the choice of—or the persuasion of the buyer to choose—the best option through a dialogue. This two-stage learning strategy has been widely used to enhance pretrained models toward a specific goal, e.g., Stiennon et al. (2020); Lewis et al. (2017); Li et al. (2016).

In reinforcement learning, we utilize *self play* (Lewis et al., 2017) to enhance our baseline models M_ϕ^B and M_θ^S by making one agent talk to the other turn by turn. Each turn ends when an agent outputs the END-OF-SENTENCE token, and the dialogue finishes when the buyer outputs the SELECTION token in a turn, or when the dialogue length limit is reached, as in the human case depicted in Fig. 2. A buyer's utterance in the i-th turn of a dialogue is denoted as Z_i^B, with Z_i^S for the seller's. We denote the trajectory τ^B or τ^S as the sequence of all tokens generated by buyer or seller during a dialogue. For instance, the buyer's trajectory is

$$\begin{aligned}\tau^B &= Z_1^B || ... || Z_i^B || ... || Z_N^B \\ &= \{z_{11}^B, ..., z_{1T_1}^B, ...z_{i1}^B, ...z_{iT_i}^B, ...z_{NT_N}^B\},\end{aligned}$$

where $||$ denotes concatenation and N is the number of turns.

After a complete dialogue has been generated, we update the agents' parameters based on the negotiation results. Agents get the final reward $R(\tau)$ when the dialogue is terminated. We define $R(\tau) = 1$ if the buyer selects the item with highest utility, $R(\tau) = 0$ if the buyer selects an item other than the best one, and $R(\tau) = -1$ otherwise. Note that the best item for a buyer is not necessarily the same that for a seller. Similar to AlphaGO (Silver and Huang et al, 2016), $R(\tau)$ is then assigned to tokens generated at each previous, non-terminal time step. We use REINFORCE (Williams, 1992) to optimize the baseline models separately toward the best options. Given a sampled trajectory τ and the final reward $R(\tau)$, let a_i be the i-th token generated in a turn; we update the model's parameters θ by

$$\theta \leftarrow \theta - \eta \sum_i (R(\tau) - b) \nabla_\theta \log p_\theta(a_i | a_{<i}, o), \tag{1}$$

where η is the learning rate and b is the baseline calculated by the average reward of the previous 3 updates.

Whereas the canonical Transformer is difficult to optimize in the RL setting, often resulting in performance comparable to a random policy (Parisotto et al., 2020), or leading to meaningless results (Lewis et al., 2017; He et al., 2018), we find the pretrained T5 model works well with parameter updates by policy gradient when we simply set a smaller learning rate.

3.4 Falsehood Control

One way to increase one's integrity is to tell no lies. We follow this notion to build a more trustworthy conversational agent, especially a seller, by decreasing the possibility that an agent produces an untruthful utterance. In the FruitStand task, the seller might claim that one type of fruit is the best in quality when it really is not, attempting to attract a buyer to choose a more profitable item, and vice versa, to keep a buyer away from a less lucrative one.

Motivated by these observations, we construct a simple rule-based falsehood detector that first parses the claim for two superlatives, as shown in Table 3, and then determines whether the seller's claim conflicts with any known fact based on a given scenario o. We further use this to establish a deduction mechanism $D(\tau)$ on the final reward in the reinforcement learning stage. Given a trajectory τ, $D(\tau) = -2$ if any of the seller's utterances conflict with the facts about the quality of an item; $D(\tau) = 0$ if none of this kind of falsehood is detected. The updated final reward then becomes $R(\tau) + D(\tau)$; we term this approach RL (w/DM).

Claim Parser	
SUP: best/worst	
FRUIT: apple/banana/orange	
	Matching Pattern
	SUP are the FRUIT
	SUP FRUIT
	FRUITs are your SUP
	FRUITs are my SUP
	FRUITs are the SUP
<Ignore>	FRUITs are the best seller
Falsehood Type	
Claim a type of fruit is the best or worst but actually not.	

Table 3: The falsehood detector is consisted of a claim parser and falsehood type. If the claim from a seller disobeys any fact derived from a scenario o, the detector will catch a falsehood

4 Experiments

4.1 Training Details

We used PyTorch to implement our models, and used the pretrained T5-base model from Hugging Face.[1] We added the special tokens BUYER, SELLER, and SELECTION as self-defined tokens to T5Tokenizer.[1] For the baseline models, we fine-tuned the pretrained T5-base for 20 epochs; after each epoch we set a checkpoint, and then picked that with the least perplexity on the validation set: for the buyer this yielded model M_ϕ^B and for the seller M_θ^S separately. We used AdamW (Loshchilov and Hutter, 2019) to optimize models with a learning rate of 6.25×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a mini-batch size of 4. We clipped gradients above 1.0. Supervised fine-tuning was run on a single RTX 2080Ti GPU.

In reinforcement learning, we estimated the expected final reward under an initial condition by

[1] https://huggingface.co/transformers/model_doc/t5.html

	Fine-tuned SL	RL	RL	RL	Human
	Baseline	Interleave	$O_{train'}$	Random	Testing set
Buyer's optimal	62.6%	**71.9%**	69.3%	65.5%	69.0%
Seller's optimal	59.1%	**73.5%**	70.0%	63.6%	66.4%
Mutually optimal (if possible)	64.2%	**78.8%**	75.2%	68.2%	75.8%

Table 4: Rate of the optimal items being chosen in final deals. The RL approach that interleave REINFORCE and SL to update parameters reaches the highest performance among all, even better than average humans.

Item\Attr.	Preference	Quality	Profit	Buyer's Utility	Seller's Utility
Apples	2	4	16	8	24
Bananas	1	1	7	1	8
Oranges	2	10	11	20	31

a. Fine-tuned SL (baseline)	b. RL (interleaved on $O_{train'}$)
B: hello, i am looking for some fresh fruit. S: i have some great oranges. B: are they fresh? S: they are very fresh and juicy. B: great! S: great! B: <selection> oranges	B: hello, i would like to buy some fruit. S: i have some great oranges. B: are they fresher than your apples? S: they are fresher than the apples. B: ok. i will take some oranges then S: great. thank you for your business B: <selection> oranges
c. RL (random initialized O)	**d. RL (directly on $O_{train'}$)**
B: hello, i'm looking for some apples. S: i have some great apples. they are very fresh and juicy. B: great! i love apples! S: great.. how many would you like? B: i will take a dozen please S: great.. here you go. B: <selection> apples	B: hello, i am looking for some fresh fruit. what do you have that is fresh? S: my oranges are the freshest i have today. B: oh great, i love oranges. how are your apples? S: my apples are ok but my oranges are the freshest i have today. B: ok. i will take some oranges then. thank you. S: great! i will get them ready for you. B: <selection> oranges

Table 5: Cherry picked examples under the same scenario of bot-bot chats on FruitStand. Fine-tuned SL produces general response, and RL approaches get more specific and various.

sampling N turns of utterances from self-play dialogue. In each turn, at the T5 decoding phase, the next token a_t was randomly chosen according to its conditional probability distribution

$$a_t \sim P(a|a_{1:t-1})$$

using top-K sampling (Fan et al., 2018), in which the K most likely next tokens are filtered in and the pmf of the output tokens is redistributed among the K tokens. We empirically chose $N = 32$ and $K = 50$ for a given o and set the mini-batch size to N. We also used AdamW for the parameter updates but reduced the learning rate to one-tenth of that used in the supervised fine-tuning. We chose the number of dialogues in the validation dataset as the amount of dialogues used in an epoch for RL approaches. We updated the parameter per mini-batch for 10 epochs. This took about 40 hours on a single Quadro RTX 8000.

4.2 Comparison

We compare the performance of the following models:

- **Fine-tuned SL**: our baseline models described in Sec. 3.2: a pair of pretrained T5 models fine-tuned on FruitStand.

Given $O_{train'}$, the initial conditions of the dialogues randomly picked from the training set to the size of the validation set, we evaluated the variants derived from Sec. 3.3:

- **RL (interleaved on $O_{train'}$)**: Direct optimization of the agent goals via RL often results in language that differs from human language. Similar to Lewis et al. (2017), we fine-tuned the baseline models with RL followed by SL in each epoch. The learning rate was one-tenth of that for Fine-tuned SL.

- **RL (directly on $O_{train'}$)**: Under the same initial conditions, we evaluated the scenario without the following SL part. The learning rate was one-tenth of that for Fine-tuned SL.

- **RL (random initialized O)**: The baseline models self play under randomly initialized scenarios. Since the outputs of the baseline models diverge from human language during the RL process for unseen initial conditions, we further reduced their learning rate to one-hundredth of that for Fine-tuned SL.

4.3 Evaluation

We evaluated the performance of the proposed approaches on FruitStand by the proportion of the best options being chosen after self play, denoted as the p-score, with respect to the unseen initial conditions in the testing dataset. Note that in the evaluation stage, for fair competition, we used not top-K sampling but instead greedy search, which simply selects the token with the highest probability as the next token:

$$a_t = \arg\max_a P(a|a_{1:t-1}).$$

For each RL variant described in Sec. 4.2, we first evaluated our models on the validation set, pair by pair at each checkpoint, and chose that pair with the highest average p-score for testing.

The results are shown in Table 4. The RL approaches considerably enhance the ability to select the best item from the baseline models. Compared to human-human negotiation in the FruitStand testing set, RL (interleaved on $O_{train'}$), the best model, achieves even better performance. This success provides evidence that maximizing the reward outplays average humans and constitutes an acceptable imitation.

For falsehood detection, we compared the number of a typical kind of detected falsehood produced by a seller from dialogues in the testing dataset (Human), the number from baseline models (Baseline models), and the number from the RL (interleaved on $O_{train'}$) variant, RL (w/o DM).

Checkpoints	1	2	3	4	5
RL(w/o DM)	18	8	8	7	41
RL(w/ DM)	0	0	0	0	0
Checkpoints	6	7	8	9	10
RL(w/o DM)	9	6	16	7	11
RL(w/ DM)	0	0	1	0	1
Human: 18					
Baseline models: 32					
Size of testing dataset: 423					

Table 6: RL(interleave) with/without deduction mechanism in each checkpoint. Each number in a cell (expect for those horizontal to 'Checkpoints') shows how many falsehoods found by the detector in each checkpoint.

The results are shown in Table 6. In the crowd-sourced testing dataset, the specific type of falsehood exists in 18 out of 423 dialogues. In the baseline, falsehoods were detected in 32 out of 423 dialogues. RL (interleaved) on $O_{train'}$) performs poorly on falsehood detection with 6 to 41 falsehoods among all the checkpoints. In contrast, our approach, RL (w/DM), significantly reduces the falsehoods in the pattern.

5 Analysis and Discussion

Goal-based models are more task-centered. Although the fine-tuned T5-base model can generate fluent and reasonable utterances, it tends to output generic responses such as "great!" which poorly reflect the task setting. See Table 5. In comparison, RL approaches generate utterances that better fit the simulated scene. A general phenomenon is that they generate long utterances, similar to humans, who show their interest in goods by asking more questions, and vendors, who show their passion by promoting their products. We also find that models learn to compare goods; comparison is an effective way to determine which item to choose.

Behavior Control Besides falsehood, we also investigated how to control virtual sellers' other behaviors. Four different sellers are investigated: **Balanced Seller** is the standard seller described all over the paper, which utility is the sum of buyer's utility plus items' profits. **Win-win Seller**'s utility is based on whether mutual optimality was achieved. **Recommender**'s utility is exactly the same as buyer's utility. **Profit-oriented Seller**'s utility base on only items' profits. Appendix C shows their vending results accordingly. We found that in general, **Balanced Seller** remains a certain level of profitability and satisfy customers at the same time. Actually, the decision of choose what kind of virtual seller to employ in practice would depend on employers' willingness and needs. Here we just demonstrate that how virtual sellers can be customised by just adjusting their utility design.

Deduction mechanism silences all. The falsehood detector is meant to prevent the seller from generating untruthful claims, and ensure that only factual claims are made. However, we find that the deduction mechanism suppresses not only such falseness, but also expressions containing such claims. That is, it prevents the seller from generating any utterances with matching patterns. For example, at some checkpoints, the seller does not even produce the string 'the best', which is clearly not a desired consequence.

Checkpoints	1	2	3	4	5	6	7	8	9	10
RL(w/o DM)	18/44	8/32	8/21	7/18	41/131	9/42	6/21	16/40	7/12	11/48
RL(w/ DM)	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0	1/3
RL(DM$_B$)	6/17	5/12	3/7	0/0	0/0	0/0	11/37	1/3	6/26	**0/6**
RL(DM$_R$)	2/10	8/23	**0/3**	0/0	0/0	0/0	2/10	0/0	0/0	0/0

Human: 18/58
Baseline models: 32/82
Size of testing dataset: 423

Table 7: RL(w/o DM) denotes the RL(interleave) model without deduction mechanism; RL(w/ DM), RL(DM$_B$), and RL(DM$_R$) stand for the RL(interleave) model with the deduction mechanism or its adjustment. Each number in a cell (expect for those horizontal to 'Checkpoints') shows how many falsehoods found by the detector in each checkpoint.

We thus adjust the mechanism using two approaches. First, we retain the -2 deduction on falsehood, but compensate those expressions by +0.5, denoted by RL (DM$_B$). Second, we instead reduce the deduction to -1, a more conservative value corresponding to R (τ). This path is denoted by RL (DM$_R$).

The results in Table 7 show that it is difficult to avoid mistakenly silencing non-deceptive utterances. In the experiment on both paths, at some checkpoints the seller avoids indiscriminate silencing, whereas at other checkpoints falsehoods are generated which still use those combinations of words. The underlying reasons for such unstable results are poorly understood. We leave this as future work.

6 Related Work

During the recent, rapid development of conversational agents, also known as chatbots, various applications have been created. Open-domain chatbots such as Facebook's BST (Roller et al., 2021) and Google's Meena (Adiwardana et al., 2020) seek to be more human-like, engaging in conversation on any topic. Closed-domain chatbots instead focus on improved task performance, for instance Guess-Which (Das et al., 2017), persuasion (Wang et al., 2019; Shi et al., 2020), and negotiation (Afantenos et al., 2012; Papangelis and Georgila, 2015; Lewis et al., 2017; He et al., 2018).

To negotiate item distribution (book, hat, ball), Lewis et al. (2017) apply a bi-directional GRU model to train a language model and use reinforcement learning with self play to develop data-driven strategies. For price bargaining on a single item (e.g., a TV), He et al. (2018) use a hybrid approach involving rule-based and LSTM models that decouple natural language understanding, dialogue act prediction, and natural language generation to facilitate controllable negotiation strategies. However, these scenarios do not attempt to find an optimal solution for both sides, and do not control the falsehoods generated by sellers. These limitations motivate this work.

7 Conclusion

We introduce a novel negotiation task and present FruitStand, a rich dataset of human-human dialogues, for negotiation dialogue research. We demonstrate the effectiveness of reinforcement learning in guiding the conversational agent toward a specific goal. Finally, our experiments in falsehood suppression show the potential of RL for truthful AI. A more robust falsehood detector would be our first future work. In our initial observations, a strong Natural Language Inference (NLI) model could play this role.

8 Acknowledgements

We are grateful for the insightful comments from anonymous reviewers. This work is supported by the Ministry of Science and Technology of Taiwan under grant numbers MOST111-2634-F-001-001.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Stergos Afantenos, Nicholas Asher, Farah Benamara, Anaıs Cadilhac, Cedric Dégremont, Pascal Denis, Markus Guhe, Simon Keizer, et al. 2012. Modelling strategic conversation: model, annotation design and corpus. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*.

Nicholas Asher and Alex Lascarides. 2013. Strategic conversation. *Semantics and Pragmatics*, 6(2):1–62.

Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

David DeVault, Johnathan Mell, and Jonathan Gratch. 2015. Toward natural turn-taking in a virtual human negotiation agent. In *AAAI Spring Symposia*.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful ai: Developing and governing ai that does not lie.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898. Association for Computational Linguistics.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Simon Keizer, Markus Guhe, Heriberto Cuayáhuitl, Ioannis Efstathiou, Klaus-Peter Engelbrecht, Mihai Dobre, Alex Lascarides, and Oliver Lemon. 2017. Evaluating persuasion strategies and deep reinforcement learning methods for negotiation dialogue agents. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 480–484. Association for Computational Linguistics.

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

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *CoRR*, abs/1606.01541.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*.

Alexandros Papangelis and Kallirroi Georgila. 2015. Reinforcement learning of multi-issue negotiation dialogue policies. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 154–158, Prague, Czech Republic. Association for Computational Linguistics.

Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphaël Lopez Kaufman, Aidan Clark, Seb Noury, Matthew Botvinick, Nicolas Heess, and Raia Hadsell. 2020. Stabilizing transformers for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7487–7498. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Carl Shapiro. 1983. Premiums for High Quality Products as Returns to Reputations*. *The Quarterly Journal of Economics*, 98(4):659–679.

Weiyan Shi, Xuewei Wang, Yoojung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of persuasive dialogues: Testing bot identities and inquiry strategies. *CoRR*, abs/2001.04564.

David Silver and Aja Huang et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489.

David R. So, Chen Liang, and Quoc V. Le. 2019. The evolved transformer.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

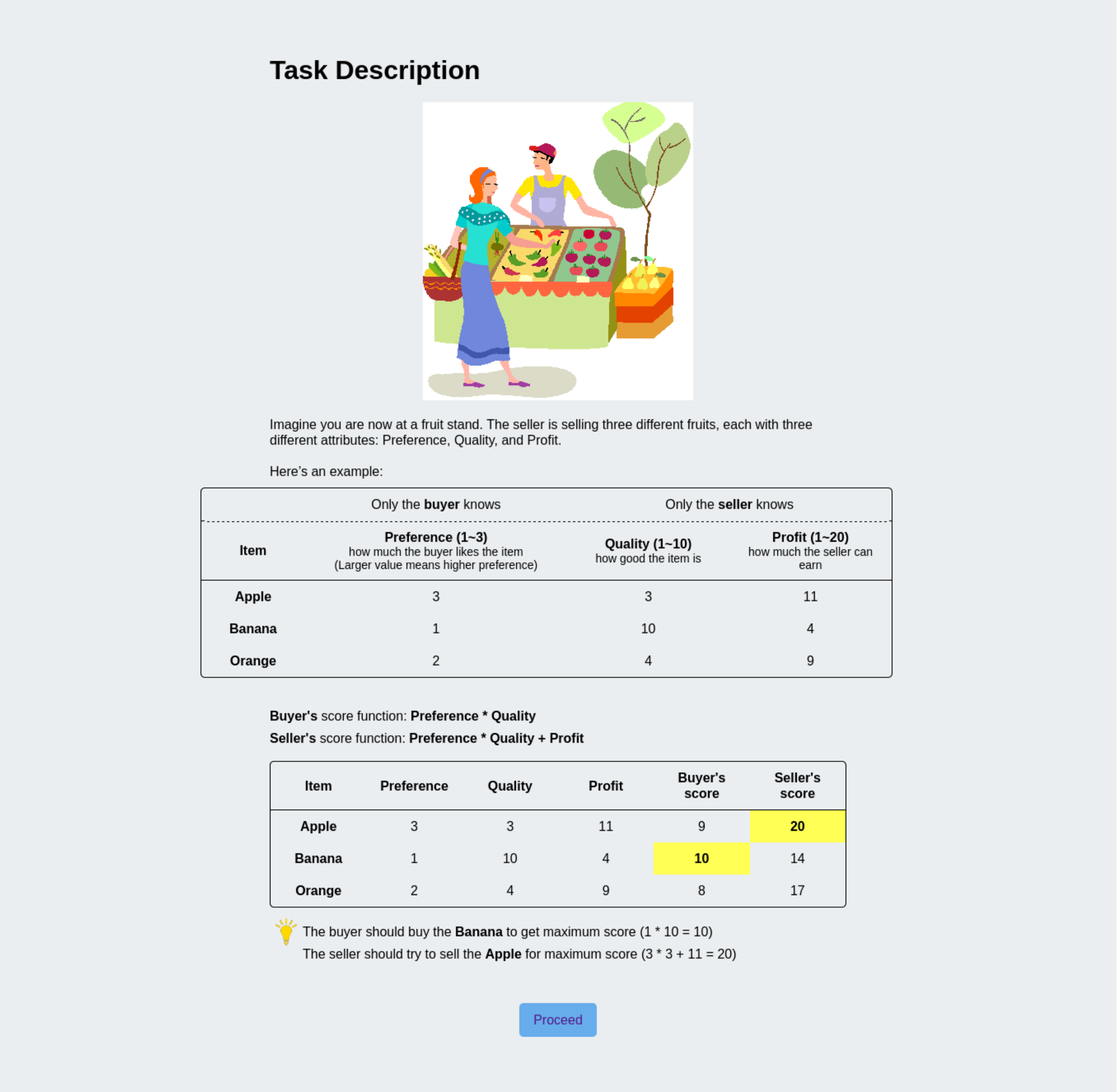
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Conference on Neural Information Processing Systems*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649. Association for Computational Linguistics.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

A FruitStand Interface



Task Description

Imagine you are now at a fruit stand. The seller is selling three different fruits, each with three different attributes: Preference, Quality, and Profit.

Here's an example:

	Only the **buyer** knows	Only the **seller** knows	
Item	**Preference (1~3)** how much the buyer likes the item (Larger value means higher preference)	**Quality (1~10)** how good the item is	**Profit (1~20)** how much the seller can earn
Apple	3	3	11
Banana	1	10	4
Orange	2	4	9

Buyer's score function: **Preference * Quality**

Seller's score function: **Preference * Quality + Profit**

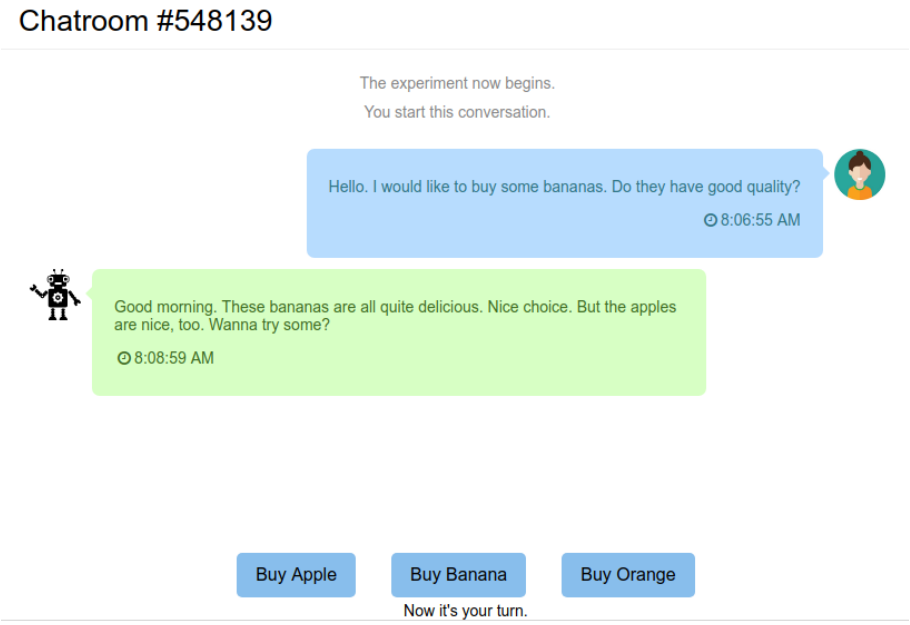
Item	Preference	Quality	Profit	Buyer's score	Seller's score
Apple	3	3	11	9	**20**
Banana	1	10	4	**10**	14
Orange	2	4	9	8	17

The buyer should buy the **Banana** to get maximum score (1 * 10 = 10)

The seller should try to sell the **Apple** for maximum score (3 * 3 + 11 = 20)

Figure 3: The interface we use for collecting dataset on the Amazon Mechanical Turk.

B FruitStand Interface (Cont.)



Chatroom #548139

The experiment now begins.

You start this conversation.

Hello. I would like to buy some bananas. Do they have good quality?

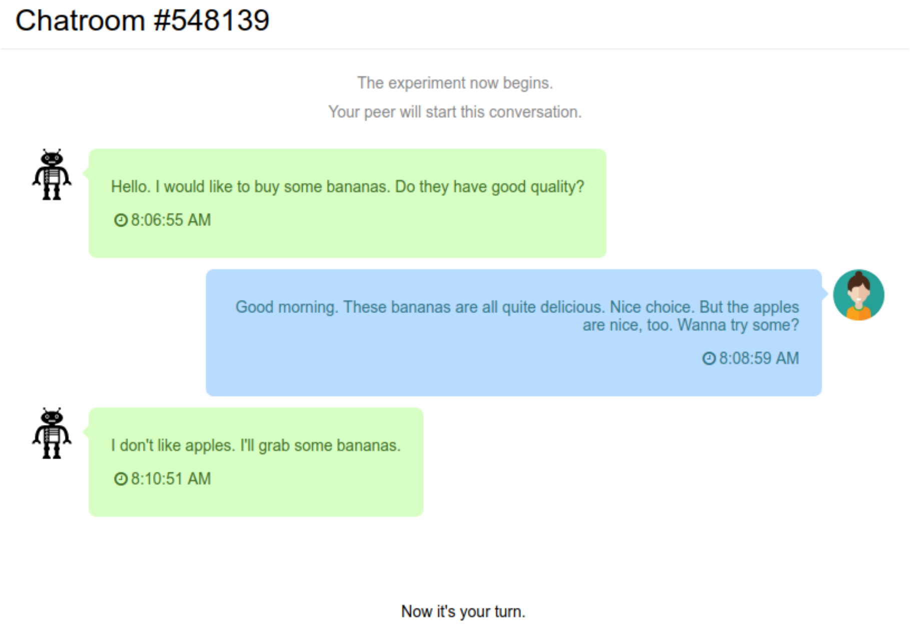
8:06:55 AM

Good morning. These bananas are all quite delicious. Nice choice. But the apples are nice, too. Wanna try some?

8:08:59 AM

Now it's your turn.

Figure 4: Buyer's interface



Chatroom #548139

The experiment now begins.

Your peer will start this conversation.

Hello. I would like to buy some bananas. Do they have good quality?

8:06:55 AM

Good morning. These bananas are all quite delicious. Nice choice. But the apples are nice, too. Wanna try some?

8:08:59 AM

I don't like apples. I'll grab some bananas.

8:10:51 AM

Now it's your turn.

Figure 5: Seller's interface

C Behavior Control

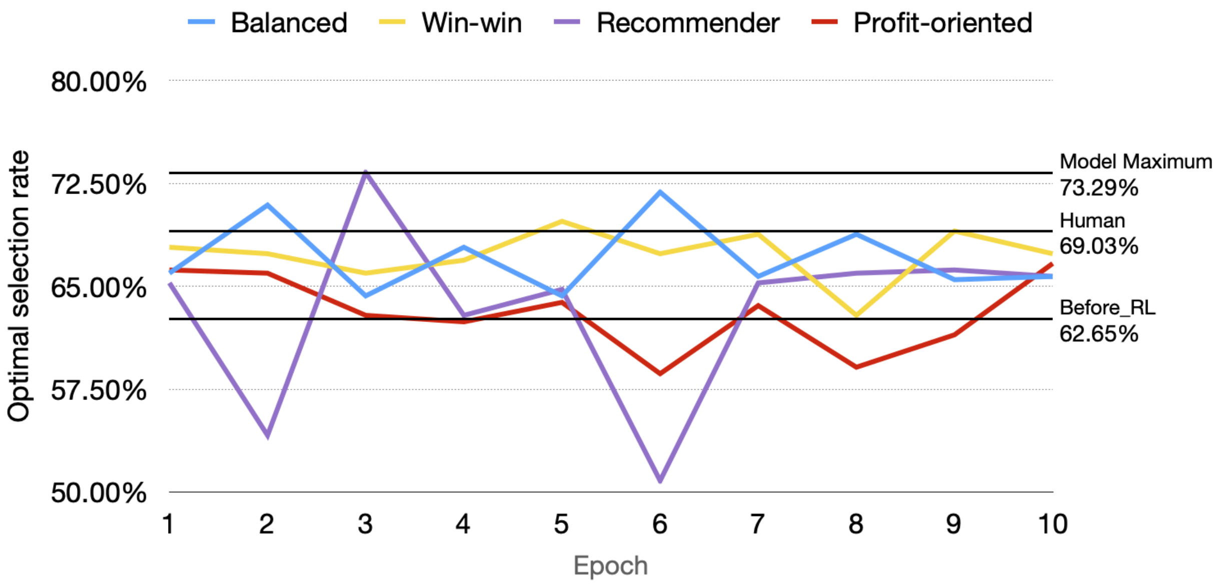


Figure 6: Rate of Buyer's best items being chosen. 'Human' stands for the human selections in testing. 'Before_RL' stands for the model before reinforcement learning.

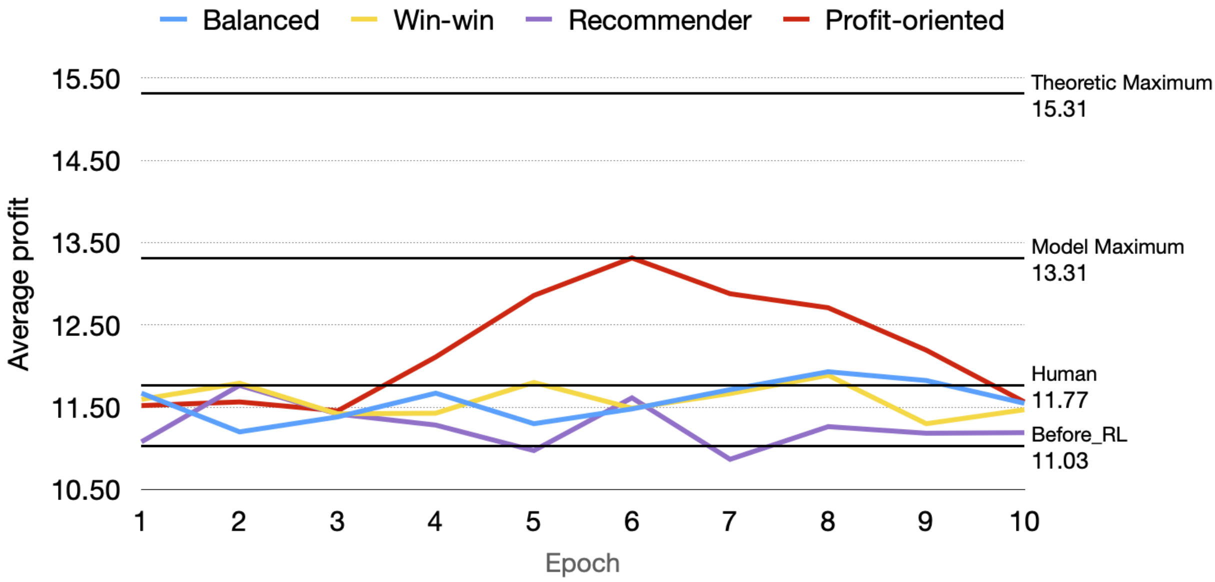


Figure 7: Seller's average profit. 'Theoretic maximum' stands for the average maximal profit of dialogues/scenarios in testing set. 'Human' stands for the human selections in testing. 'Before_RL' stands for the model before reinforcement learning.