CHAI: A CHatbot AI for Task-Oriented Dialogue with Offline Reinforcement Learning

Anonymous ACL submission

Abstract

Conventionally, generation of natural language for dialogue agents may be viewed as a statistical learning problem: determine the patterns in human-provided data and generate appropriate responses with similar statistical properties. However, dialogue can also be regarded as a goal directed process, where speakers attempt to accomplish a specific task. Reinforcement learning (RL) algorithms are designed specifically for solving such goal-directed problems, but the most direct way to apply RL – through trial-and-error learning in human conversations, – is costly. In this paper, we study how offline reinforcement learning can instead be used to train dialogue agents entirely using static datasets collected from human speakers. Our experiments show that recently developed offline RL methods can be combined with language models to yield realistic dialogue agents that better accomplish task goals.

1 Introduction

Constructing fluent and intelligent dialogue agents could pave the way for intuitive interfaces and automation of human-interactive tasks. However, this requires dialogue agents that both generate fluent, natural responses and effectively pursue the goals of the given dialogue task. A predominant approach to training dialogue agents is through supervised learning, where an agent is tasked with imitating language provided by humans. While this can provide for fluent responses, it becomes difficult to ensure that such agents systematically pursue the goals of the dialogue interaction. If we instead view dialogue as a control problem, frameworks such as reinforcement learning (RL) could allow agents to automatically optimize dialogue with respect to a task goal through a trial-and-error process and improve over human behavior.

However, implementing an RL system in practice, where an agent learns online from interacting with real humans, can be prohibitively expensive and time-consuming. This is in stark contrast to supervised learning approaches, where we can cheaply construct datasets for training imitation agents by simply logging conversations. Therefore, existing RL approaches for dialogue often rely on interacting with a learned model of a human (Li et al., 2016b; He et al., 2018), from which experience can be generated inexpensively. However, naïve training in this manner can result in the dialogue agent exploiting the model, which can degenerate into non-intelligible language. To mitigate this, algorithms must typically enforce strong priors to keep generated language similar to those seen in the dataset (Li et al., 2016b; Jaques et al., 2019), or adopt dialogue management and template-based approaches which directly re-use language seen in the dataset (He et al., 2018).

Issues such as model exploitation and distribution shift when training on static datasets are a primary concern of offline RL (Levine et al., 2020), and can provide a formalized approach to tackling these problems. While offline RL is motivated by scaling RL to large datasets, annotated datasets for dialogue are still small compared to the large amount of raw text datasets available today. Therefore, we propose an offline, model-free approach to dialogue generation that leverages language models. Because the size of unlabeled language datasets dwarfs that of curated datasets for dialogue, using a pre-trained language model as a central component of our method allows it to learn aspects of language fluency from unlabeled datasets, while learning higher-level strategies for goal-directed dialogue via RL on a smaller annotated datasets. This combined approach enables us to utilize the large amounts of existing language data that standard RL methods cannot.

The main contribution of this work is CHAI (CHatbot AI), an algorithm for learning task-oriented dialogue that utilizes a language model in conjunction with offline RL. We show that this

leads the policy to generate goal-oriented dialogue that is both realistic and functional, and does not require training against a simulated model of human language. We evaluate our method on a negotiation task, which requires the model to both reason about strategic aspects of conversation along with generating fluent language. We show that CHAI consistently bargains for better prices and with higher rates of successful negotiation than prior RL approaches to goal-oriented dialogue.

2 Related Work

Recent developments in deep learning have led to end-to-end approaches to dialogue using supervised learning, such as sequence-to-sequence models (Dušek and Jurcicek, 2016; Eric and Manning, 2017), hierarchical models (Serban et al., 2017), attention (Mei et al., 2017; Chen et al., 2019), and Transformer-based models (Wu et al., 2019a; Hosseini-Asl et al., 2020; Peng et al., 2020; Adiwardana et al., 2020). However, supervised learning only allows an agent to imitate behaviors, requires optimal data, and does not allow agents to exceed human performance. Supervised learning for dialogue generation also has well-known issues such as outputting commonplace responses (e.g., *I do not know*) regardless of the inputs (Li et al., 2016a). Therefore, additional training of the dialogue agent is required for performing goal-oriented tasks.

Task-oriented dialogue has been formulated as a sequential decision making problem in a Markov Decision Process (MDP) since the 1990s (Smith and Hipp, 1994; Singh et al., 1999; Williams and Young, 2007; Young et al., 2013; Paek, 2006; Henderson et al., 2008; Gao et al., 2018; Pieraccini et al., 2009; Young et al., 2013; Su et al., 2015). Dialogue is converted into abstract states and actions from which an agent is trained using reinforcement learning (RL) (Eckert et al., 1997; Levin et al., 2000; Chung, 2004; Georgila et al., 2006; Schatzmann et al., 2007; Heeman, 2009; Georgila and Traum, 2011; Su et al., 2016; Fatemi et al., 2016; Asri et al., 2016; Zhao et al., 2019; Zhang et al., 2020; Wang et al., 2020). These methods differ in how the abstract states/actions are designed and whether the simulated environment for training the policy is hand created, learned as a fixed model, or is an agent itself. For instance, Eckert et al. (1997); Levin et al. (2000) learn a fixed transition model from human conversations and Georgila and Traum (2011) learn negotiation agents where each agent is the user simulator for the other agent. These methods also differ in how the decision making policy is trained, e.g., online (Gašić et al., 2011) or off-policy/offline (Yu et al., 2016; Pietquin et al., 2011) using actor-critic (Su et al., 2017), policy gradient (He et al., 2018), or fitted Q-iteration (Pietquin et al., 2011). Regardless of the RL method used, since policies are trained on abstract states and actions, these methods lack the ability to *generate* natural language (i.e., response is created via templates depending on an abstract action).

To overcome these limitations, recent work has trained policies directly on text, using a recurrent neural network to output language tokens, and using self-play for policy training while interacting with another learned agent (Li et al., 2016b; Lewis et al., 2017; Liu et al., 2018). To further improve the generated language quality, hierarchical methods decouple the strategic high-level dialogue decisions from generation (Yarats and Lewis, 2018; He et al., 2018; Saleh et al., 2020). These model-based approaches require accurate estimation of the environment/human (e.g., the trained self-play agent needs to mimic complex human behavior), which is beyond current capability of model-based reinforcement learning algorithms. Similar to our proposal, Jaques et al. (2019) use offline RL based on KL-control for text generation in open-domain dialogue. Our work differs in that our model is able to utilize large amounts of unsupervised data through the use of pre-trained language models, and that our work focuses on task-oriented (as opposed to open-domain) dialogue tasks. Goal-oriented tasks have clearly defined objectives that can be quantified, allowing us to provide an objective comparison between our method and prior approaches.

3 Preliminaries

In this section, we describe our evaluation task and cover the necessary background and notation.

3.1 Negotiation via Task-Oriented Dialogue

We evaluate our approach on the CraigslistBargain task (He et al., 2018). CraigslistBargain consists of 6682 advertisements scraped from Craigslist, along with dialogues for each advertisement collected via Amazon Mechanical Turk where two users play the role of buyer/seller. An example advertisement from this dataset is shown in Fig. 1, along with a sample conversation between a human and CHAI.

During each round of interaction, the buyer and

Title: iPhone 5S 16 GB black silver - $135
Description: Great condition. No scratches. I upgraded to iPhone 7. Please reply with phone number if you are interested. Thank you.

Buyer: Hi, I would like to buy the iPhone 5S
CHAI: Hi, great! The phone is in great condition. I'm asking $135.00, it's in mint condition.
Buyer: Does it come with a charger?
CHAI: Yes it comes with a charger, as I upgraded. It's in mint condition.
Buyer: How long have you been using this phone?
CHAI: Just a month. It's in great condition and has a charger included.
Buyer: Did you buy this phone secondhand or from the Apple store?
CHAI: I bought it brand new from Apple store. I only used it a month, I'm happy I did not have to use this one.
Buyer: I can buy it for $100
CHAI: I'll take that if you can pick it up in a few hours.
Buyer: Where will you be at 5pm today?
CHAI: I'm at the Apple store so about an hour.
Buyer: Okay, I can meet you at the Apple store.
CHAI: Awesome!
Buyer: *accept*

Figure 1: An example advertisement from the CraigslistBargain task, with a conversation between CHAI (blue) and a human buyer (green).

seller can execute four possible response types. A `message` allows one player to send an utterance to the other. An `offer` allows one player to propose a price at which to conduct the transaction. Once an offer is made, the other player can either `accept` or `reject` the offer, which ends the episode. A reward is then computed based on the transaction price. Our bot receives a reward equal to the normalized price the item is sold for at the end of an episode (normalized by the list price), scaled by a constant factor of 10. Additionally, we penalize the bot by a constant of -20 for episodes resulting in a `reject` to incentivize the agent to make deals.

We selected this task because it provides a clear objective, allowing us to illustrate our approach with quantifiable metrics. Of course, practical applications of CHAI to goal-directed dialogue may tackle other problems, including non-adversarial problms such as helping a user to answer a question or fulfill a request. However, our choice of tasks was constrained by the limited availability of public datasets for dialogue tasks that are goal-directed and have objective task goals.

3.2 Reinforcement Learning Setup

We formulate the task-oriented dialogue problem as an RL problem, where the agent serves the role of *seller* in the CraigslistBargain problem, and the environment serves the role as the *buyer*. We consider a Markov decision process defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma)$. The state and action spaces, $\mathcal{S}$ and $\mathcal{A}$, consist of three main components: the action type `type` (one of the four described in Sec. 3.1), an utterance u (used only for `message` responses), and a normalized price `price`, expressed as a fraction of the list price. Additionally, the context component $s_{\texttt{context}}$ contains the advertisement listing description. The price component is used in two ways. First, because we do not wish to represent prices in the dialogue as discrete tokens, we instead replace prices in the utterance with a placeholder token understood to be substituted with the price component. Second, the price component is used in `offer` response type to communicate the desired transaction price. In all other cases, it is ignored. An example of the state space is shown in Fig. 2.

s_u	`utterance`	Hello!
$s_{context}$	`context`	Selling a new mountain bike.
s_{price}	`price`	0.5 (of list price)
s_{type}	`action_type`	`message`

Figure 2: An example of the state space used in the CraigslistBargain task described in Section 3.1.

We write individual states as $s = \{s_u, s_{\texttt{type}}, s_{\texttt{price}}, s_{\texttt{context}}\}$ and actions as $a = \{a_u, a_{\texttt{type}}, a_{\texttt{price}}\}$, where s_u, a_u denote the utterances, or sequences of words generated by the environment and agent. $\{s_{\texttt{type}}, a_{\texttt{type}}\}$ denote the action types, $\{s_{\texttt{price}}, a_{\texttt{price}}\}$ denote the prices, and $s_{\texttt{context}}$ denotes the listing context. The transition distribution $\mathcal{T}(s'|s,a)$ governs the distribution over responses generated by the buyer agent (environment), and the reward $\mathcal{R}$ defines the task objective. The goal of RL is to find a policy $\pi(a|s)$ that maximizes the expected returns:

$$\mathbb{E}_{\pi,\mathcal{T}}\left[\sum_{t=0}^{T}\gamma^t r(s_t, a_t)\right],$$

where $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function, and $\gamma \in (0, 1]$ is the discount factor. In online RL, the agent is interacts with the environment to maximize this objective. In contrast, offline RL uses a dataset $\mathcal{D}$ of pre-collected interaction. This is

a natural approach in many settings, such as dialogue, where online RL could require interacting with real humans for impractically long periods of time, whereas offline RL can utilize pre-recorded dialogues between humans.

4 Offline Reinforcement Learning with Language Models

Offline RL potentially allows reinforcement learning methods to leverage large datasets for policy learning. However, it still requires datasets to be annotated with rewards, and for conversations to come from the task at hand. Because of this, annotated dialogue datasets, such as the CraigslistBargain dataset presented in Section 3.1, are many orders of magnitude smaller than unlabeled datasets collected for unsupervised and language modeling tasks. In order to utilize these large unlabeled datasets, we propose an algorithm that combines offline RL with fine-tuned language models.

Our approach begins with training a language model, such as GPT-2 (Radford et al., 2018), and fine-tuning it on our task-specific dialogue corpus (Sec. 3.1). We use $\text{LM}(u|s_{1:t})$ to denote a distribution over utterances u produced by the language model given the dialogue history, denoted $s_{1:t}$. We then train a critic or Q-function as described in Sec. 4.1 which is responsible for scoring good and bad responses and is used to select responses from a pool of candidates generated from the language model. Our approach can be viewed as using a Q-function to steer a language model (which has no concept of a task) towards producing language that accomplishes some task-specific goal.

4.1 Q-Learning with Language Models

In this section we describe how to train a Q-function that can score candidate responses based on their potential to maximize returns. We implement and evaluate three different training procedures, each utilizing different offline RL methods. In the overall Q-learning framework, we sample a batch of transitions (consisting of states, actions, rewards, and successor states) from our dataset and perform updates based on minimizing a modified Bellman loss:

$$J(\theta) = (Q_\theta(s,a) - Q_{\text{target}}(s,a))^2, \qquad (1)$$

where target value $Q_{\text{target}}(s,a)$ is typically computed via the Bellman operator defined as

$$Q_{\text{target}}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}\left[\max_a \bar{Q}(s',a)\right]. \qquad (2)$$

However, using this update directly can lead to problems if we only have access to offline datasets. A widely studied issue in offline RL is the challenge of handling *out-of-distribution actions*: when the maximization over the action in the target value is not constrained in any way, it is easy to obtain actions for which the Q-value predictions are erroneously high (Levine et al., 2020). In dialogue, this issue is greatly exacerbated, since the Q-function is only trained on responses in the dataset, and therefore is unlikely to make accurate predictions for arbitrary strings. The following modifications to Eqn.1 and Eqn.2 address this issue.

Proposal sampling (CHAI-prop) In the proposal sampling approach, the target value $Q_{\text{target}}(s,a)$ is computed via a modified Bellman operator that utilizes a proposal distribution based on the language model, $\mu(a_t|s_{1:t})$, to generate N *response proposals*, and then uses the target Q-function, $\bar{Q}$, to score those responses and selects the highest one:

$$\begin{aligned} &Q^{\text{prop}}_{\text{target}}(s,a) \\ &= r(s,a) + \gamma \mathbb{E}_{s'} \mathbb{E}_{\{a_i\}^N \sim \mu}\left[\max_i \bar{Q}(s',a_i)\right]. \end{aligned}$$

This sampling scheme serves a dual purpose: it both constrains the responses to be naturalistic, and it also prevents out-of-distribution inputs to the Q-function in the target value calculation. This approach resembles a number of prior offline RL methods that also employ proposal distributions (Kalashnikov et al., 2018; Kumar et al., 2019; Fujimoto et al., 2019; Wu et al., 2019b). Similarly to several prior works, we use samples from a proposal distribution for the target value, without an explicit actor (Kalashnikov et al., 2018; Ghasemipour et al., 2020). Unlike these approaches, our method leverages a pretrained and finetuned language model LM, which additionally makes use of extensive unsupervised prior datasets during the pretraining stage and enables our method to handle the complex and combinatorial action space of dialogue generation. In addition, following prior work, we use a separate target network $\bar{Q}$ whose weights are updated to track those of Q using a soft update rule as done in prior methods (Lillicrap et al., 2016; Haarnoja et al., 2018).

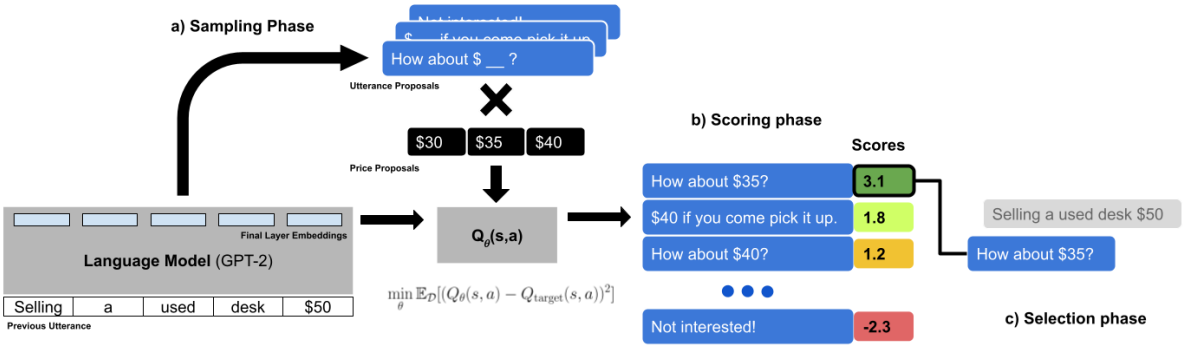

Figure 3: An overview of our dialogue generation and selection process. First (a), a finetuned language model is used to generate candidate responses conditioned on the conversation history, along with sampled prices. Then (b), each action is scored using a critic, and (c) a final response is returned by sampling the Q-value scores.

The proposal distribution $\mu(a_t|s_{1:t})$ represents a distribution over actions $a = \{a_u, a_{\texttt{type}}, a_{\texttt{price}}\}$. We use the language model in order to sample utterances $a_u \sim \text{LM}(\cdot|s_{1:t})$. To make training more computationally efficient, we pre-generate a batch of 5 utterances per transition in the dataset using the language model, and resample these during training as an approximation of directly sampling from the language model. For the prices $a_{\texttt{price}}$, we uniformly sample a value between 70% to 100% of the previously offered price, which roughly matches the distribution of the seller's offers in the dataset. Finally, we infer the message type based on the utterance sampled using a simple heuristic, as the CraigslistBargain task requires us to specify a `type` for each response. During language model fine-tuning, we replaced each `offer`, `accept`, or `reject` action with the utterances "offer", "accept", and "reject", respectively. We then check if the language model generated any of these tokens and return the corresponding action type, and label the action as a `message` otherwise. This simplifies our method and allows us to use the language model to generate the action types as well as the utterances.

Conservative Q-learning (CHAI-CQL) Conservative Q-learning (CQL) (Kumar et al., 2020) proposes a complimentary approach to reducing the harmful effect of out-of-distribution Q-values by explicitly penalizing the Q-value of actions not seen in the dataset. We adapt CQL as an additional regularizer on the Q-value in addition to the proposal sampling scheme. Specifically, we use the CQL($\mathcal{H}$) variant, which add an additional regularizer f^{CQL} to the Q-learning objective:

$$J^{\text{CQL}}(\theta) = (Q_\theta(s,a) - Q_{\text{target}}(s,a))^2 + \alpha f^{CQL}(\theta),$$

where the regularizer is defined as:

$$f^{CQL}(\theta) = \mathbb{E}_{s\sim\mathcal{D}}\left[\log \sum_a \exp(Q_\theta(s,a)) - \mathbb{E}_{a\sim\mathcal{D}}[Q_\theta(s,a)]\right]$$

Our adaptation of CQL differs from Kumar et al. (2020) in that they propose an actor-critic method which trains an explicit actor. Rather, the remainder of our method is identical to the proposal sampling variant, specifically in regards to the computation of Q_{target} and language model sampling.

Behavior-regularized Q-learning (CHAI-BRAC) Behavior-regularized actor-critic (BRAC) (Wu et al., 2019b) proposes an alternative method for regularizing the Q-function such that out-of-distribution Q-values are penalized. Adapting this method to the setting of dialogue with language models, we use this approach to regularize the price proposal mechanism. Rather than uniformly sampling prices as described for proposal sampling, we train an additional *price proposal network* $\pi_\phi(a_{\texttt{price}}|s_{1:t})$ that outputs a Gaussian distribution over prices given the conversation states. Using the notation $a' \sim \pi_\phi, \mu$ to denote sampling prices from the proposal network, utterances from the language model, and action types uniformly, the proposal network is trained according to the objective

$$\max_{\pi_\phi} \mathbb{E}_{s,a\sim\mathcal{D}}\big[\mathbb{E}_{a'\sim\pi_\phi,\mu}[Q(s,a')] - D_{KL}(\pi_\phi(\cdot|s), \pi_B(\cdot|s))\big].$$

The prior proposal network, π_B, is estimated as a univariate conditional Gaussian of the current offer given the previous offer, where the mean and standard deviations are linear functions of the previous offer. The target value is then computed as:

$$Q^{\text{BRAC}}_{\text{target}}(s,a) = r(s,a) + \gamma\mathbb{E}_{s',a'\sim\pi_\phi,\mu}\left[\bar{Q}(s',a')\right].$$

This adaption of the behavior regularized objective controls out-of-distribution queries on the Q-value by regularizing the price towards those seen within the dataset. This prevents the target Q-value from being queried in low-data regimes which can cause inaccuracies during training.

4.2 Dialogue Generation

Once the Q-function has been trained, dialogue generation from our model is a three-phase process. The first phase is sampling: given a response from the buyer, we query the language model to sample 5 candidate utterances, and sample 5 candidate prices. The next step is scoring: we then take the cross-product of these sets, and score each potential action with the Q-function. Finally, in the selection phase, several methods are considered in order to select an action. A straightforward method is to return the action that had the highest Q-value. However, we found that this approach resulted in behavior with low diversity. Instead, we opted to follow the approach in soft Q-learning (Haarnoja et al., 2018) and sample actions from a softmax distribution over the Q-values, $p(a|s) \propto \exp\{Q(s,a)\}$, which increases diversity in the responses as sub-optimal actions are occasionally sampled. This decoding process is depicted in Fig. 3.

4.3 Architecture Details

For our language model, we use an off-the-shelf implementation of GPT2-medium (Radford et al., 2018). This language model is finetuned on a transcript of each scenario in the dataset containing the context (title, description) and spoken dialogue. The prices in the dialogue are masked out with a special price token allowing us use GPT to generate templates which we can substitute prices into. The input to the language model is a concatenation of the scenario context and the dialogue history.

The Q-function is parameterized as a feedforward network that maps states and actions into a single scalar representing the Q-value. To process the utterances into the state and action, we separately compute state and action embeddings by taking the average of the masked GPT2 attention embeddings of the entire dialogue history up to the current utterance. These embeddings are then concatenated with the prices (represented as a fraction of the list price) and message types (represented as a one-hot vector) to produce a single vector that is given to the critic as input. The critic is parametrized using a 2-layer feedforward network with hidden sizes of 256 and ReLU nonlinearities. Additional details can be found in Appendix A.1.

5 Experiments

Our experimental evaluation aims to compare our proposed goal-directed offline RL dialogue method to both prior dialogue management approaches and language modeling baselines. We conduct two studies: an objective evaluation against other dialogue agents to measure each method's performance in negotiation, and a subjective human study to measure the overall end-to-end performance of the system in a similar manner to prior work (He et al., 2018; Jaques et al., 2019). Qualitative results showing actual dialogue generated from our method can be found in Appendix A.4. We consider 4 baseline approaches. The first is the current state-of-the-art approach for the CraigslistBargain task proposed by He et al. (2018) (referred to as the retrieval-based baseline). This is a hierarchical approach to dialogue generation that separately handles language generation and dialogue management. This method parses utterances into coarse "dialogue acts," which represent high-level categorizations of the utterance such as greetings, offers, or counter-offers. An RL agent is then trained against a learned model of the environment to select "dialogue acts", and a retrieval-based generator is then used to convert "dialogue acts" back into text. In contrast, our method directly generates text, and does not require any manually designed categorizations of natural dialogue into dialogue acts. Since our method utilizes a modern language model, we also include a pure language modeling baseline, which consists of the same GPT-2 language model Radford et al. (2018) finetuned on the CraigslistBargain dataset using the same method as done in CHAI. This baseline allows us to determine whether any improvement from our method is due to the language model, or to the use of offline RL. Finally, we also evaluate the end-to-end approaches described by Lewis et al. (2017), which include a dialogue agent trained via supervised learning, and an RL agent optimized for the task objective.

5.1 Simulated Evaluation

To evaluate the effectiveness of our offline RL goal-directed dialogue system, we first conducted a systematic study including the 3 variations of CHAI outlined in Sec. 4: the proposal sampling method (CHAI-prop), CQL method (CHAI-CQL), and be-

Method	vs Rule-based		vs Stingy		vs Utility	
	Acc%	Revenue	Acc%	Revenue	Acc%	Revenue
CHAI-prop	61.5	0.48 ± 0.39	57.5	0.39 ± 0.35	99.0	0.70 ± 0.17
CHAI-CQL	74.0	0.51 ± 0.33	77.5	0.49 ± 0.30	98.0	0.70 ± 0.19
CHAI-BRAC	62.0	0.52 ± 0.41	47.0	0.38 ± 0.41	99.0	0.71 ± 0.17
Language Model	48.5	0.29 ± 0.32	51.5	0.27 ± 0.28	20.5	0.14 ± 0.28
He et al. (2018) (Utility)	1.0	0.01 ± 0.10	0.0	0.00 ± 0.00	11.0	0.07 ± 0.22
He et al. (2018) (Fairness)	84.0	0.70 ± 0.32	80.0	0.59 ± 0.31	100.0	0.72 ± 0.15
He et al. (2018) (Length)	53.0	0.46 ± 0.43	49.0	0.37 ± 0.38	100.0	0.72 ± 0.16
Lewis et al. (2017) (RL)	83.5	0.17 ± 0.24	83.0	0.19 ± 0.25	64.5	0.46 ± 0.37
Lewis et al. (2017) (SL)	38.5	0.17 ± 0.27	46.5	0.21 ± 0.27	18.0	0.13 ± 0.28

Method	vs Fairness		vs Length		Overall (mean)	
	Acc%	Revenue	Acc%	Revenue	Acc%	Revenue
CHAI-prop	99.0	0.90 ± 0.15	92.5	0.79 ± 0.27	81.9	$\mathbf{0.65 \pm 0.34}$
CHAI-CQL	99.5	0.87 ± 0.14	94.5	0.79 ± 0.24	**88.7**	$\mathbf{0.67 \pm 0.29}$
CHAI-BRAC	100.0	0.85 ± 0.03	91.0	0.76 ± 0.25	79.8	$\mathbf{0.65 \pm 0.34}$
Language Model	25.5	0.19 ± 0.35	18.5	0.14 ± 0.32	32.9	0.21 ± 0.31
He et al. (2018) (Utility)	100.0	1.00 ± 0.00	100.0	1.00 ± 0.00	42.4	0.42 ± 0.49
He et al. (2018) (Fairness)	0.0	0.00 ± 0.00	100.0	0.70 ± 0.16	72.8	0.54 ± 0.35
He et al. (2018) (Length)	100.0	1.00 ± 0.00	100.0	0.78 ± 0.18	80.4	$\mathbf{0.66 \pm 0.36}$
Lewis et al. (2017) (RL)	88.0	0.26 ± 0.34	71.5	0.31 ± 0.36	78.1	0.28 ± 0.33
Lewis et al. (2017) (SL)	60.0	0.48 ± 0.46	53.0	0.42 ± 0.46	43.2	0.28 ± 0.39

Table 1: Acceptance rates and normalized average revenue generated comparing CHAI using proposal sampling (CHAI-prop), conservative Q-learning (CHAI-CQL), and behavior regularization (CHAI-BRAC). The baselines consist of a retrieval-based agent (He et al., 2018), Lewis et al. (2017) and a fine-tuned language model (higher is better) against 5 different evaluation bots. The mean score across all evaluation bots is reported in the right-most column. Numbers are reported as means and standard deviations over 200 trials.

Metric	Fluency	Coherency	On-Topic	Human-Likeness	Total
CHAI-prop	$\mathbf{4.31 \pm 0.97}$	$\mathbf{3.91 \pm 1.17}$	$\mathbf{4.16 \pm 0.99}$	$\mathbf{3.47 \pm 1.27}$	$\mathbf{15.84 \pm 3.86}$
He et al. (2018) (Utility)	3.56 ± 1.34	2.47 ± 1.39	3.09 ± 1.40	2.13 ± 1.13	11.25 ± 4.50
Lang. Model	4.06 ± 1.11	2.66 ± 1.36	3.63 ± 1.18	2.50 ± 1.10	12.84 ± 3.66

Table 2: Human evaluation scores comparing CHAI, He et al. (2018), and language model (higher is better). Numbers are reported as means and standard deviations over 32 trials. CHAI scores the highest across all metrics.

havior regularized method (CHAI-BRAC). In order to ensure that the results are not overfit to a single strategy, we run each method against a suite of 5 evaluation buyer agents, based on the retrieval agents presented by He et al. (2018). We choose these agents because they have been evaluated by humans as being human-like and have the strongest performance on the CraigslistBargain benchmark task. Specifically, we use the rule-based and RL agents proposed by He et al. (2018) (trained using "utility", "fairness", and conversation "length" as rewards). To introduce additional variety in negotiation styles, we additionally modify the rule-based agent to offer 25% of the difference between offers rather than splitting the difference between offers, which we refer to as the "Stingy" rule-based agent. We record the percentage of negotiations that result in an accept and the average normalized revenue generated per negotiation, which totals the average sale price (rejections have zero revenue) normalized by the listing price of the advertisement. Our results are presented in Table 1.

Overall, we find that among the variations of CHAI, the conservative Q-learning variant performs the best by a small margin, but results are very comparable between all 3 variations. This suggests that the particular choice of offline regularizer is far less important than the CHAI framework of utilizing a pre-trained language model with Q-function scoring. On average, CHAI-CQL performs significantly higher on acceptance rate and similarly on revenue to the next best agent, the retrieval agent (He et al., 2018) using conversation length as reward. Computing statistical significance between these two methods, we find that $p < 1.96 * 10^{-9}$ using a chi-squared test for acceptance rate, indicating that the difference in acceptance rates is statistically significant. We find that $p < 0.946$ using a t-test for revenue, indicating that the difference in revenue is not significant. We also note that the performance of CHAI has significantly less variation across evaluations against different buyer agents than the retrieval-based agents. For example, the retrieval agent with

utility reward scores near-zero on 3 evaluations but scores near-perfectly on the other two. This suggests that the CHAI framework produces dialogue agents that are more consistent and less susceptible to exploitation. We also note that implementing the retrieval method (He et al., 2018) requires hand-designing high-level dialogue actions, topic categories, and rules for parsing or labeling these components. These designs are specifically tailored to the CraigslistBargain task, whereas such hand-engineering for CHAI does not exist outside of the interface requirements to the task itself. Thus, CHAI has significantly weaker assumptions, generates dialogue end-to-end via RL, and yet is able to narrowly outperform prior methods. Among prior methods with similar assumptions to CHAI (the language modeling baseline and (Lewis et al., 2017)), CHAI outperforms by a wide margin on both acceptance rates and revenue.

We ran an additional ablation study on the choice of reward in Appendix A.3. We find that this has a significant effect on performance, and we based our reward design on balancing between maximizing acceptance rate (through a rejection penalty) and revenue (through the utility function).

5.2 Human User Study

To evaluate the effectiveness and naturalness of our offline RL goal-directed dialogue system, we conducted a user study with 16 individuals, who were each asked to carry out 2 negotiations with each of three agents. The users were then asked to rate the conversation on fluency, coherency, on-topicness, and human-likeness on a 5-point Likert scale. **Fluency** specifically refers to the frequency of grammatical and word-choice errors. **Coherency** measures whether the agent's responses are coherent. **On-topic** measures how well the agent was aligned with performing the task at hand. Finally, **human-likeness** measures how similar the agent's responses were to a human. Because of the cost of human evaluations, we were limited in our ability to evaluate as many baselines. Therefore, we chose methods that were the most directly comparable - the simplest variation of CHAI (CHAI-prop) optimized for utility, evaluated against the utility-optimized agent from He et al. (2018), and a language model baseline that shares the same fine-tuning procedure as CHAI. Additional details of the user study, including the questions posed to the users and statistical significance tests, are included in Appendix A.3.1.

Results are shown in Table 2. We ran a one-way repeated measures ANOVA test, and found that the type of agent used leads to statistically significant rating differences for all metrics (with at least $p < 0.01$). CHAI outperforms both baselines on almost all metrics, except for fluency, where both CHAI and the language model perform similarly. The fact that fluency is similar between the two models makes sense, since both methods use a GPT-2 model to generate utterances. However, the langauge modeling baseline lacks an understanding of the task goal, and therefore makes unreasonable offers or responds in illogical ways (see Appendix A.4 for examples). It therefore scores lower on other metrics as compared to CHAI. This result suggests that the ability of language models to execute goal-directed dialogues is limited by a lack of awareness of the task objectives, and that offline RL potentially addresses this issue, producing dialogue that is perceived as more coherent, task-oriented, and human-like.

6 Discussion and Future Work

We presented a system for goal-directed dialogue based on combining offline RL with finetuned language models. CHAI learns with RL, but does so from offline datasets of human dialogue. The language model allows CHAI to benefit from large-scale unsupervised pre-training, and the offline RL component enables CHAI to select responses that are more likely to lead to a successful task outcome. Quantitatively, CHAI achieves higher acceptance rate at higher revenue than prior dialogue management systems designed for this task.

While CHAI provides a proof-of-concept that offline RL can successfully learn complex human-interactive tasks such as dialogue, it also has limitations. The goal of RL is to maximize reward, which can lead to unintended responses – for example, without additional objective terms, there is no reason for CHAI to be truthful. Similar issues affect language models more broadly, though we anticipate that it would be easier to address such issues in RL by employing better reward design. An exciting direction for future work is to extend offline RL to address a wider range of human-interactive tasks, particularly tasks with longer-range dependencies and delayed rewards, where complex task goals can lead to the emergence of dialogue that is ultimately more useful to human users.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Layla El Asri, Jing He, and Kaheer Suleman. 2016. A sequence-to-sequence model for user simulation in spoken dialogue systems. *arXiv preprint arXiv:1607.00070*.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709.

Grace Chung. 2004. Developing a flexible spoken dialog system using simulation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 63–70.

Ondřej Dušek and Filip Jurcicek. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51.

Wieland Eckert, Esther Levin, and Roberto Pieraccini. 1997. User modeling for spoken dialogue system evaluation. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 80–87. IEEE.

Mihail Eric and Christopher D Manning. 2017. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 468–473.

Mehdi Fatemi, Layla El Asri, Hannes Schulz, Jing He, and Kaheer Suleman. 2016. Policy networks with two-stage training for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 101–110.

Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374.

Milica Gašić, Filip Jurčíček, Blaise Thomson, Kai Yu, and Steve Young. 2011. On-line policy optimisation of spoken dialogue systems via live interaction with human subjects. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 312–317. IEEE.

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2006. User simulation for spoken dialogue systems: Learning and evaluation. In *Ninth International Conference on Spoken Language Processing*.

Kallirroi Georgila and David Traum. 2011. Reinforcement learning of argumentation dialogue policies in negotiation. In *Twelfth Annual Conference of the International Speech Communication Association*.

Seyed Kamyar Seyed Ghasemipour, Dale Schuurmans, and Shixiang Shane Gu. 2020. Emaq: Expected-max q-learning operator for simple yet effective offline and online rl. *arXiv preprint arXiv:2007.11091*.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR.

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

Peter A Heeman. 2009. Representing the reinforcement learning state in a negotiation dialogue. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 450–455. IEEE.

James Henderson, Oliver Lemon, and Kallirroi Georgila. 2008. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, 34(4):487–511.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. volume 33.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. 2018. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. 2019. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. volume 33.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on speech and audio processing*, 8(1):11–23.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Conference on Empirical Methods in Natural Language Processing*.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. *International Conference on Learning Representations*.

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of NAACL-HLT*, pages 2060–2069.

Hongyuan Mei, Mohit Bansal, and Matthew R Walter. 2017. Coherent dialogue with attention-based language models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3252–3258. AAAI Press.

Tim Paek. 2006. Reinforcement learning for spoken dialogue systems: Comparing strengths and weaknesses for practical deployment. In *Proc. Dialog-on-Dialog Workshop, Interspeech*. Citeseer.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model. *arXiv preprint arXiv:2005.05298*.

Roberto Pieraccini, David Suendermann, Krishna Dayanidhi, and Jackson Liscombe. 2009. Are we there yet? research in commercial spoken dialog systems. In *International Conference on Text, Speech and Dialogue*, pages 3–13. Springer.

Olivier Pietquin, Matthieu Geist, Senthilkumar Chandramohan, and Hervé Frezza-Buet. 2011. Sample-efficient batch reinforcement learning for dialogue management optimization. *ACM Transactions on Speech and Language Processing (TSLP)*, 7(3):1–21.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners. *OpenAI Blog*.

Abdelrhman Saleh, Natasha Jaques, Asma Ghandeharioun, Judy Shen, and Rosalind Picard. 2020. Hierarchical reinforcement learning for open-domain dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8741–8748.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152.

Iulian Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. 2017. Multiresolution recurrent neural networks: An application to dialogue response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Satinder Singh, Michael Kearns, Diane Litman, and Marilyn Walker. 1999. Reinforcement learning for spoken dialogue systems. *Advances in neural information processing systems*, 12:956–962.

Ronnie W Smith and D Richard Hipp. 1994. *Spoken natural language dialog systems: A practical approach*. Oxford University Press on Demand.

Pei-Hao Su, Paweł Budzianowski, Stefan Ultes, Milica Gasic, and Steve Young. 2017. Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 147–157.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Continuously learning neural dialogue management. *arXiv preprint arXiv:1606.02689*.

Pei-Hao Su, David Vandyke, Milica Gašić, Dongho Kim, Nikola Mrkšić, Tsung-Hsien Wen, and Steve Young. 2015. Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2020. Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system. *arXiv preprint arXiv:2006.06814*.

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Qingyang Wu, Yichi Zhang, Yu Li, and Zhou Yu. 2019a. Alternating recurrent dialog model with large-scale pre-trained language models.

Yifan Wu, George Tucker, and Ofir Nachum. 2019b. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*.

Denis Yarats and Mike Lewis. 2018. Hierarchical text generation and planning for strategic dialogue. In *International Conference on Machine Learning*, pages 5591–5599.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 404–412.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9604–9611.

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 1208–1218.

A Appendix

A.1 Architecture Details

For our language model, we use an off-the-shelf implementation of GPT2-medium (Radford et al., 2018). This language model is finetuned on a transcript of each scenario in the dataset containing the context (title, description) and spoken dialogue. The prices in the dialogue are masked out with a special price token allowing us use GPT to generate templates which we can substitute prices into. To produce language samples, we concatenate the context of the scenario along with the dialogue history and feed it to the language model. The language model then generates the next utterance. This process is repeated in order to get multiple samples.

The Q-function is a feedforward network that maps the states and actions into a single scalar representing the Q-value. To process the utterances into the state and action, we separately compute state and action embeddings by taking the average of the masked GPT2 attention embeddings of the entire dialogue history up to the current utterance. State and action embeddings are then concatenated with the prices (represented as a percentage of the list price) and message types (represented as a one-hot vector) to produce a single vector that is given to the critic as input. The critic is parametrized as a 2-layer feedforward network with hidden sizes of 256 and ReLU nonlinearities.

The equations below describe the inputs of the Q-function. "STRCAT" is a custom string formatting function that concatenates the context and utterances in the dialog and prefixes each utterance with the string "Buyer:" or "Seller:". "EMBED" calculates the masked attention embeddings from GPT2. `dialog` represents the dialog history up to the current state, $[s_{u,1}, a_{u,1}, \ldots, a_{u,t-1}, s_{u,t}]$, and `candidate` denotes a candidate utterance $a_{u,t}$ generated by the language model that is being considered for scoring.

$$
\begin{aligned}
s_{\texttt{embed}} &= \text{EMBED}(\text{STRCAT}(s_{\texttt{context}}, \texttt{dialog})) \\
a_{\texttt{embed}} &= \text{EMBED}(\text{STRCAT}(s_{\texttt{context}}, \texttt{dialog} + \texttt{candidate}) \\
s &= [s_{\texttt{embed}}, s_{\texttt{price}}, s_{\texttt{type}}] \\
a &= [a_{\texttt{embed}}, a_{\texttt{price}}, a_{\texttt{type}}] \\
q &= Q(s, a)
\end{aligned}
$$

A.2 Experiment Details

Hyperparameter Selection. For our Q-learning algorithm, we used default hyperparameters from an SAC implementation and did not vary the parameters. We used a critic learning rate of $3 * 10^{-4}$, and a soft target update rate of 0.05. For the language model architecture, we finetuned 2 GPT models architectures (GPT2-small, and GPT2-medium). In order to select which model to use, we manually rated samples from the language generated in their quality across both model types and checkpoints, and selected the best performing model. We selected the GPT2-medium architecture at training epoch 2000.

Compute Resources. We finetuned our language models on TPUs (TPU v3-8, 16GB memory per core with 8 cores) within a GCP instance. We trained our Q-learning models on an internal compute cluster using an Nvidia 1080 GPU (12 GB memory).

A.3 Reward Ablation Study

Because we are limited in the number of evaluations possible in a human study, we use a simulated evaluation against another chatbot to run an ablation study measuring the effect of using different reward functions. Specifically, we instantiate a rule-based dialogue manager proposed by (He et al., 2018) as the "buyer" and negotiate with it on randomly sampled scenarios from the dataset. This is done for both our method and the baselines, and the results are tabulated in Table 3.

We evaluated 5 variations of CHAI. CHAI(final) is the method used in our paper, which uses 2 components to the reward: positive reward for the price the item was sold at, and a penalty for the episode ending in a rejection. The specific reward used was 10 * the price sold (normalized by the list price) if the offer was accepted, or a penalty of -20 if the offer was rejected. CHAI(penalty) uses the same reward,

except with an increased rejection penalty. CHAI(accept) is given a positive reward of +20 for episodes ending in an accept and negative reward of -20 for episodes ending in a rejection, without regard to the price. CHAI(utility) is rewarded solely for the price an item is sold at at 10 * the price sold. Finally, CHAI(fair) is rewarded for negotiating to a midpoint price between the buyer and seller's target prices.

We report the acceptance rate, average revenue, and average offers made and offers accepted. We see that CHAI(accept), CHAI(fair) and CHAI(penalty) achieve higher acceptance rates, but offer lower prices on average. In contrast, CHAI(utility) and CHAI(final) offer higher prices with lower acceptance rates, with the pure utility optimizing agent CHAI(utility) offering the highest prices with the lowest acceptance rates. Thus, we can see how changing the reward function can significantly affect the behavior of the resulting agent.

Method	Accept Rate	Prices Offered	Prices Accepted	Revenue
CHAI(accept)	0.74	0.80 ± 0.15	0.74 ± 0.13	0.55 ± 0.35
CHAI(fair)	0.76	0.80 ± 0.15	0.75 ± 0.13	0.57 ± 0.34
CHAI(penalty)	0.90	0.77 ± 0.14	0.77 ± 0.12	0.68 ± 0.26
CHAI(utility)	0.34	0.96 ± 0.29	0.76 ± 0.13	0.29 ± 0.42
CHAI(final)	0.66	0.84 ± 0.14	0.87 ± 0.07	0.51 ± 0.38

Table 3: Mean and standard deviation of revenue, acceptance rate, and prices sold & offered, over 50 samples when negotiating against a baseline chatbot. Revenue and prices are reported as a fraction of the list price.

A.3.1 User Study Parameters

Setup. We conducted our user study through a web-interface, where an advertisement from the test set of CraigslistBargain is displayed to the user. This is shown in Fig. 4. Users were instructed to "type any message to speak with the bot and negotiate", and a chatbot agent replied to each message. After the user is satisfied with the negotiation, they are asked to indicate whether they want to accept or reject the current offer on the item.

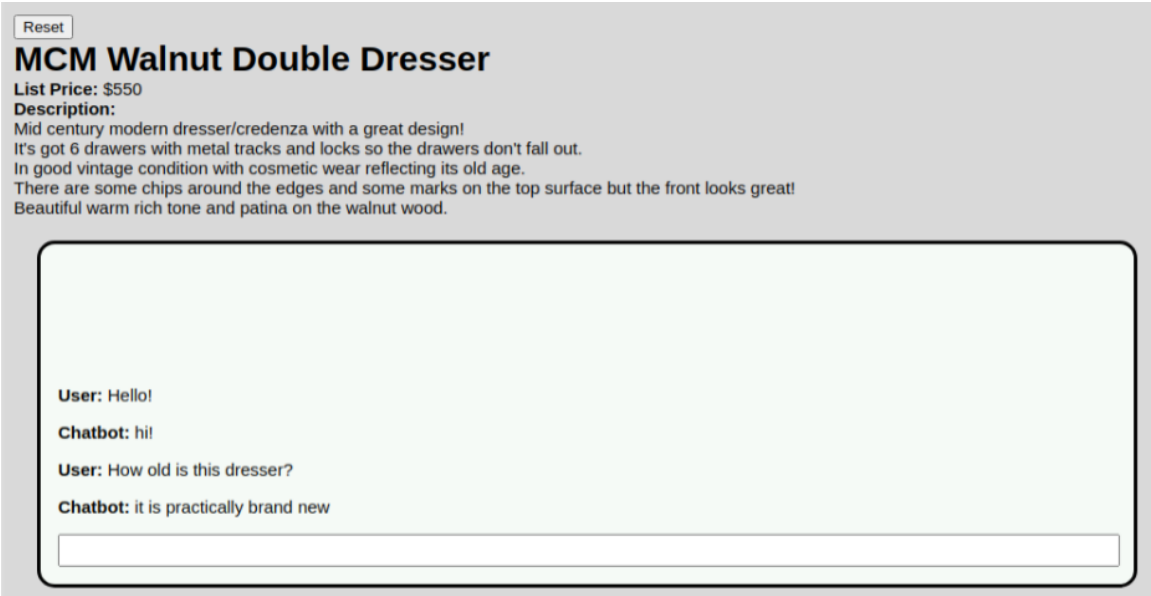

Figure 4: The interface presented to participants in the user study. We display the title, price, and description of the advertisement, and users are able to chat with the bot through a textbox.

After interacting with the chatbot agent, users were given a survey and asked to rate the bot on a 5-point Likert scale: strongly agree (5), agree (4), neutral (3), disagree (2), strongly disagree (1). The ratings were:

- The bot was fluent (did not make grammatical or word choice errors).
- The flow of the conversation was coherent.
- The bot was on-topic.
- The bot demonstrated human-like behavior.

These questions correspond to the fluency, coherency, on-topicness, and human-likeness scores reported in our paper, respectively.

Manipulated factors. We evaluate 3 chatbot agents. The first is CHAI. The second is a retrieval-based baseline ($\text{RL}_{\text{utility}}(\text{act})$ as proposed in (He et al., 2018)). The this is a language modeling baseline based on finetuning GPT2-medium on the CraigslistBargain dataset.

Dependent measures. We measure the performance of each chatbot agent according to 4 subjective metric (fluency, coherency, on-topicness, and human-likeness), which correspond to the survey questions given to participants described above. Each metric is rated on a 5-point Likert scale. We also measure the price that was agreed upon, and whether the negotiation resulted in an acceptance or a rejection.

Risks The risks presented to participants were minimal, and participants were never placed in the way of physical harm. There was a small probability the chatbot could generate offensive or otherwise inconsiderate langauge, as the language generation for CHAI and the language model baseline were unconstrained. However, we did not observe this behavior prior to the study, and did not observe this behavior during the study. We minimized the risk of confidentiality breaches by anonymizing all data stored as user IDs – participant names were not stored on our servers.

Subject allocation. We recruited 10 male and 6 female participants, with an average age of 24. Participation was voluntary, and subjects were not compensated for their participation. Participants were asked for consent before participating, and consent for including examples from their interactions for this paper. All examples contained in this paper are anonymized and contain no personally identifiable information.

Prior to participation in the study, each user was provided with instructions for how to use the interface, and a practice evaluation against the retrieval agent to familiarize them with the user interface and rating system. Each user interacted with all 3 dialogue agents (CHAI, retrieval, and language model) twice, with the order of interaction randomized per user. The advertisement displayed to the user is also uniformly sampled from the test set for each trial, independently from the agent being used.

Analysis. We ran a one-way repeated measures ANOVA test for each metric reported (fluency, coherency, on-topic, human-like) to examine the effect of chatbot agent on the metric.

Results showed that the type of agent used lead to statistically significant differences in the ratings. We found that:

- For fluency, $f(2, 30) = 5.8333, p < 0.0072$
- For coherency, $f(2, 30) = 16.9518, p < 0.0001$
- For on-topicness, $f(2, 30) = 10.1840, p < 0.0004$
- For human-likeness, $f(2, 30) = 20.1592, p < 0.0001$

Title: FULLY REMODELED 4 Bdrm, 1 Bath, Duplex - $3395
Description: 4 bed / 1 bath *Type:* Duplex *City/Neighborhood:* Emeryville *Showing:* By Appointment Only *Utilities:* Paid by tenant (not included in rent amount) *Features:* Hardwood Floor, Washer Dryer, newly renovated, stainless steel appliances

Buyer: hi **CHAI:** Hi there. Are you interested in my rental? **Buyer:** yes, when is the next opening to view the apartment? **CHAI:** Today. **Buyer:** how much are the utilities per month? **CHAI:** Utilities are not included in the rental amount, but I am willing to negotiate the price. **Buyer:** since utilities are not included, could you do $3000? **CHAI:** I can't go that low. What about $3300? **Buyer:** sounds good! **CHAI:** Perfect! Thank you! **Buyer:** *accept*	**Buyer:** Hi **Retrieval:** hi **Buyer:** when is the next opening to view the apartment? **Retrieval:** its right by apartment merrit **Buyer:** how much are the utilities per month? **Retrieval:** yes they are **Buyer:** since utilities are not included, could you do $3000? **Retrieval:** deal! **Buyer:** sounds good! **Retrieval:** *offer $3000* **Buyer:** *accept*	**Buyer:** Hi **LM:** Hi, are you interested in my property? **Buyer:** Yes, when is the next opening to view the apartment? **LM:** Open until 6pm tomorrow. **Buyer:** Sounds good, how much are the utilities per month? **LM:** Utilities are included in the monthly rent. **Buyer:** Okay, and can you do $3000? **LM:** *offer $2200* **Buyer:** *accept*

Figure 5: A comparison between similar negotiations talking to CHAI-prop (left), a retrieval-based agent (He et al., 2018) (middle), and language modeling (LM, right) for the same listing. The bot takes the role of the seller, and a human is the buyer. CHAI gives both descriptive responses to questions and reasonable bargaining behavior, whereas the retrieval-based agent only shows good bargaining behavior, and the language model agent only gives descriptive responses.

A.4 Side-by-side Qualitative Evaluation

To complement our quantitative evaluation, we perform a qualitative analysis of conversations generated from CHAI, both by itself and in comparison to other agents, in order to better understand why CHAI effectively tackles the goal-oriented negotiation problem.

In Fig. 5 we show a comparative example between CHAI, the retrieval-based agent (He et al., 2018), and the language modeling baseline on the same scenario and human responses. CHAI and the language modeling baseline tend to produce more specific responses to the prompt due to the use of a language model, rather than generating text via the usage of templates. For example, when presented with a questioned about utilities and viewing time, both methods are able to answer the question, whereas the retrieval-based agent gives a non-sequitur answer. However, the language modeling baseline struggles in understanding prices, and offers $2200 when the buyer requested $3000. CHAI is able to demonstrate understanding both language and the flow of the negotiation by offering reasonable counter-offers to the user, such as responding to a low offer with "I can't go that low" and offering a higher counter-offer that the Buyer accepts.

A.5 Additional Qualitative Results

In this section, we include additional examples collected from human dialog for CHAI-prop, the retrieval method of (He et al., 2018), and the pure language modeling baseline.

A.5.1 CHAI-prop

Title: Nokia Batteries - One new one used - $10
Description: I have these two batteries left over from my old phone. Maybe you can use them? Both are in excellent working order and ready to use. Message me if you are interested.

Buyer: I've been looking for batteries, what about 50 cents?
CHAI: Hi. I still have them. They're used, but I have used them and they work perfectly.
Buyer: How much life does the used one have?
CHAI: $9
Buyer: How many times have you used the old battery?
CHAI: 1 to 2 times
Buyer: Did you not use your phone much?
CHAI: Nope, I've been away from it for about 6 months.
Buyer: Ah, I see that makes sense.
CHAI: I can do $6
Buyer: I doubt many people would want to buy a used battery though, how many other offers do you have?
CHAI: I will let them go for $7.
Buyer: But you offered 6, 7 is higher?
CHAI: No. But I'm really looking for around $6 for them.
Buyer: I'll pick it up right now for 4 dollars
CHAI: No deal
Buyer: But you were not being fair earlier when you went from six to 7 dollars
CHAI: I can go to $5
Buyer: I'm holding firm on 4 dollars, especially because the batteries are used
CHAI: I can let them go for $5
Buyer: No one really buys used batteries though, what other offers are you getting?
CHAI: I'm just going to let them go for $4
Buyer: *accept*

Title: Couch and love seat - $80
Description: Couch and love seat from a non-smoking pets free house. Upholstery is clean and doesn't have any stain. Some zippers attaching the cushions to the frame are broken. Solid wood frame. pics coming

Buyer: I am interested in purchasing your couch and love seat
CHAI: Great! These were ordered from Crate & Barrel and are in like new condition. I'm willing to negotiate a little, but I'm asking $72 for them.
Buyer: How long have you owned them?
CHAI: I've had them for a few years. They are perfect for my new apartment and they were an absolute steal when I bought them.
Buyer: How much did you purchase them for?
CHAI: They are brand new, I'm asking $66.
Buyer: How much did you originally buy the love seat and couch from from Crate and Barrel?
CHAI: I originally paid $65 for both, but I figured I'd give it a shot since I love the couch.
Buyer: So you are asking for more money then you originally bought them for?
CHAI: Yes. I was hoping to get more for the couch and love seat. How much would you like for the couch?
Buyer: I do, but I can only offer $40
CHAI: I can only go down to $40 if you come pick up and I will throw in two chairs for you to use.
Buyer: Great! That's a deal
CHAI: Great!
Buyer: *accept*

Title: Fit Bike VH2 - $200
Description: 20.75" top tube All in perfect working order.Very good condition - no rust - always stored indoors. Shiny. Few scratches/scrapes on pegs, etc, but very well cared for. Gyro brake. Thanks for interest. If its still here it is still available.

Buyer: Hello, I'm interested in the bike
CHAI: It is in great shape. It's a well-run bike. I bought it new, so I'm making it available.
Buyer: how long have you had it?
CHAI: I bought it new last year. It's in excellent condition.
Buyer: can I see pictures?
CHAI: Of course. How much are you looking to spend on the bike?
Buyer: I'm looking to spend under $150
CHAI: That's not a bad price. Would you be able to pay $180 for it?
Buyer: How about $160?
CHAI: $160 and you have a deal.
Buyer: Done
CHAI: Great!
Buyer: *accept*

A.5.2 Retrieval

Title: Media Cabinet (Moving Sale) - $20
Description: Slide out tray for media storage Condition: Structurally sound Good condition Search Steve2017 for my other listings or simply click on the "more ads by this user" link found under the map. Please email me if you are interested. Serious inquiries only. Thank you

Buyer: Hello, how are you?
Retrieval: i am well, how are you?
Buyer: Good. I'm interested in buying the media cabinet.
Retrieval: how can i help you today?
Buyer: Can you provide more pictures?
Retrieval: yes
Buyer: How old is the item?
Retrieval: less than 1 year olf
Buyer: I would like to buy it for $15.
Retrieval: deal
Buyer: *accept*

Title: Adams trail-a-bike shifter 7 - $100
Description: This is an awesome bike trailer for your kid. It features 7 shiftable gears so that your kid can learn to shift but more importantly shift down or up to continue to help you pedal the bike. I know that without the shiftable gears my kid would either think the pedaling was too hard at times or going so fast they couldn't feel their pedaling helping. It took some time to teach them but definitely the way to go.

Buyer: Hello, I'm wondering about the availability of this bike trailer
Retrieval: hello
Buyer: Is this bike trailer still available?
Retrieval: i've had it for about 5 years now.
Buyer: great, is it still available?
Retrieval: yes it is
Buyer: When would I be able to pick it up?
Retrieval: yeah, that's fine.
Buyer: Is there a date by which I would be able to pick it up?
Retrieval: sure, i could do that for you.
Buyer: Okay, is there any damage to the bike trailer
Retrieval: yes what questions do you have?
Buyer: I'm wondering if there's any damage to the bike trailer
Retrieval: great! i'm asking $100.00 for it!
Buyer: Would you be able to decrease the price to $90?
Retrieval: sure. i'd be willing to let it go for $90.00
Buyer: Okay great, I'll pick it up today. Does that work for you?
Retrieval: yes absolutely. its a great deal
Buyer: *accept*

Title: File Cabinet metal - $70
Description: I have 4 cabinets file metal 4-Drawer for sell. They are good condition. they are black .I sell them for 70$ each. Thank you for reading !

Buyer: Why are you selling
Retrieval: yes it is still for sale.
Buyer: How much is it
Retrieval: it's new
Buyer: Sorry, can you repeat that
Retrieval: where are you. how far from alameda(ca)?
Buyer: I am close to *[removed]*. It's pretty far
Retrieval: how about $70.00?
Buyer: I don't know the condition. Could you please provide me with some details
Retrieval: it looks extremely heavy and in great condition! i could do $70.00...
Buyer: Is there any damage
Retrieval: nope. overall condition is extremely well
Buyer: How heavy are they
Retrieval: yes, very good condition.
Buyer: How about 25
Retrieval: deal
Buyer: *accept*

Title: Samsung Gear S2 - $175
Description: Available in Black or white Flawless! Gear, Samsung, S3, S2, smartwatch, smart, watch, fitbit, pedometer, fitband, fit, band, bit, Samsung Gear Fit 2, classic, iwatch

Buyer: Is this an S3 or S2?
Retrieval: yes. are you interested? it's in great condition
Buyer: I'm interested.
Retrieval: do you have any questions?
Buyer: Is it a Gear S2 or Gear S3?
Retrieval: yes its a full tablet so it does come with the charger
Buyer: Is htis a watch or a tablet?
Retrieval: yes.
Buyer: Can I see pictures?
Retrieval: sure thing.
Buyer: can I have it for $50?
Retrieval: i am willing to do $50.00 if you can pick it up at my place of work.
Buyer: sure, where is that?
Retrieval: yes, that is correct.
Buyer: *reject*

A.5.3 Language Model

Title: Gorgeous 3 bedrooms available in August! Perfect way to end the summer - $3795
Description: Call Now - Use this link for more info: http://thegreystar.com/42v07z Everyday Exceptional. Where You Belong. This is a 3 Bedroom, 2 Bath, approximately 1422 Sq. Ft. NEW APARTMENT HOMES IN PLEASANTON, CA. A contemporary apartment community that is as much an escape as it is a destination. A place where beautifully appointed homes, resort-style amenities, rolling lawns and parks, and boutique shops come together to create a distinctly new leisure lifestyle. Please click on the link below to check out our virtual tour!

Buyer: Hello, is this 3 bedroom home still available for rental?
LM: Hi, yes it is. The asking price is $2430.
Buyer: Is that a monthly cost?
LM: The lease is for one year, but if you have a family member or friend who are already employed and would be willing to sign a longer lease, we could offer you the option of moving in right away.
Buyer: Would the price be lower if I signed a longer lease?
LM: The price is firm at $2425, and I'm willing to negotiate
Buyer: Would you be able to rent for $2300?
LM: Yes
Buyer: Great, what documents would I need to provide to apply to rent this place?
LM: This place has been carefully designed with you in mind.
Buyer: Does this apartment have a deposit fee?
LM: This apartment is located in a convenient location, so that the lowest that can be reasonable for you is 10% of your deposit.
Buyer: What would the deposit fee bee?
LM: If you sign a longer lease, the deposit fee will be $75.
Buyer: Okay, great! How long would my lease need to be?
LM: I'd be willing to let it run for 5 months at no additional cost.
Buyer: So if I sign a 5 month lease the deposit fee would be $75 and rent $2300 per month?
LM: That's too low for me. I could come down to $90 in monthly rent. I can also drop the rent down to $100 for you.
Buyer: Oh wow! So I could get this place for $90 in monthly rent?
LM: I think that is fair. How long of a lease are you looking for?
Buyer: 1 year
LM: This is a very good deal. I can accept $80 for a 3 year lease.
Buyer: I want a 1 year lease
LM: Ok, you have a deal!
Buyer: Okay, 1 year lease at $80 per month?
LM: Yes
Buyer: Great, thank y
LM: I would like to hear your budget.
Buyer: $80 per month works for me
LM: No, sorry.
Buyer: I thought you had provided that as your previous offer?
LM: I am sorry, but I do not. So $80 is the price.
Buyer: Okay, confirming $80 per month?
LM: How about $90 per month?
Buyer: That also works.
LM: You have a deal!
Buyer: Great!
LM: *offer* $80
Buyer: *accept*

Title: Parker Apartments - Now Accepting Below Market Rate Applications - $979
Description: **** Parker is that thrill you feel when inspiration strikes. Perched proudly on Shattuck Avenue in Berkeley, this collection of 155 carefully crafted, sustainably designed apartments offer immediate access to music venues, boutique shopping, nightlife, green space, the Gourmet Ghetto, and *removed*. At home, get to know your neighbors by the crackling light of a fire pit, get focused in our co-working space, or burn some calories at the on-site Equinox Fitness Club.

Buyer: Hi, when can I move in?
LM: Today
Buyer: Are there grocery stores nearby?
LM: Yes, and the inside of our apartments is completely gated.
Buyer: Is there parking available?
LM: No, but there is a free shuttle that you can use to get there.
Buyer: Would you be able to do $900?
LM: *offer* $2800
Buyer: *reject*

Title: 20" LED Emergency Strobe Light Lamp Bar - $30
Description: 20" LED Emergency Vehicle Dash Warning Strobe Flashing Light. these lights are a great addition and will defiantly get people out of your way as so you can respond on time! 7 Different modes of flashing / scanning mode. Please read the description to see a detail description of each mode. Includes: 1 x Traffic strobe light bar with on/off switch and mode changer button Fitment: Universal for all vehicle with a 12V power source Very bright led that are visible even during daytime conditions. 2 colors: Red, Amber. Please let me know what color you are interested.

Buyer: are both colors the same price
LM: The colors are different. I can sell them both for $200 each.
Buyer: I thought it was $30
LM: I am sorry, I have to change my mind.
Buyer: oops there is only one light
LM: The bulbs come with the two matching colors. It would be a hassle to flash them with the other lights in your vehicle. The best thing about the bulbs is they will shine for you. It's a great deal at only $20
Buyer: is there free shipping
LM: I think I could negotiate a little but I am not able to accept this offer.
Buyer: can I pay $7
LM: I have another offer but I can't go any lower than $13
Buyer: what about $11
LM: *offer* $7
Buyer: oh ok sure $7 and free shipping?
LM: Yes, I accept that!
Buyer: *accept*