Training and Analyzing Language Agents in Socially Complex Dialogues



Jessica Lin

Electrical Engineering and Computer Sciences University of California, Berkeley

Technical Report No. UCB/EECS-2025-63 http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-63.html

May 15, 2025

Copyright © 2025, by the author(s). All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Training and Analyzing Language Agents in Socially Complex Dialogues

by Jessica Lin

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science**, **Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Sergey Levine Research Advisor

5/12/25

(Date)

Professor Anca Dragan Second Reader

5/13/25

(Date)

Training and Analyzing Language Agents in Socially Complex Dialogues

by

Jessica Lin

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sergey Levine, Chair Professor Anca Dragan

Spring 2025

Training and Analyzing Language Agents in Socially Complex Dialogues

Copyright 2025 by Jessica Lin

Abstract

Training and Analyzing Language Agents in Socially Complex Dialogues

by

Jessica Lin

Master of Science in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Sergey Levine, Chair

Advancements in large language models (LLMs) have led to their use as conversational partners in social contexts that may be highly nuanced. Additional data is needed for further training to improve agent performance in emotionally complex dialogues, but real data for such domains is scarce, ethically complex, or hard to obtain and label. Thus, using LLMs to generate synthetic data has arisen as a popular alternative, yet this is often unrealistic and lacking in diversity. In this technical report, we propose two methods to improve synthetic data to train dialogue agents with reinforcement learning (RL). In Chapter 1, we introduce a hindsight regeneration pipeline that improves the diversity and quality of existing dialogue data in persuasion and mental health counseling tasks. We illustrate the capability of our method by comparing to common baselines, as well as conducting simulated evaluation and a user study. In Chapter 2, we explore the effects of deception on a language model's ability to negotiate in a real-world business scenario. We present a simulation engine pipeline for effective conversation generation, as well as analysis of model behavior in negotiation tasks.

To my Mom, Dad, and Brother

Contents

Co	ontents	ii						
Li	st of Figures	iv						
\mathbf{Li}	st of Tables	\mathbf{v}						
1	Introduction							
2	Interactive Dialogue Agents via Reinforcement Learning on Hindsight Regenerations 2.1 Introduction 2.2 Related Work 2.3 Preliminaries 2.4 Methodology 2.5 Experiments 2.6 Results for Persuasion for Good 2.7 Results for Montal Health Counseling	3 3 4 6 7 11 13 17						
	2.7 Results for Mental Health Counsening	17 19						
3	A Pipeline for LLM-Generated Dialogue for Real-World Negotiations3.1Introduction3.2Related Work3.3Methodology3.4Experiments3.5Discussion	 20 20 21 23 26 33 						
4	Conclusions and Future Work	34						
Bi	bliography	35						
Α	Hindsight Regeneration Implementation Details A.1 Hindsight Controller A.2 Forward Model	45 45 48						

	A.3 Reward Model	51
	A.4 Policy Optimization	56
	A.5 User Study Details	56
в	Negotiation Implementation Details	58
	B.1 Persuasion Taxonomy	58
	V	

List of Figures

2.1	Overall scheme for <i>hindsight regenerations</i> , our proposed approach for augmenting data to train language agents	4
2.2	Overview of our approach. We relabel suboptimal actions in the original dataset, then generate plausible completions of the dialogue after relabeling to obtain <i>hind</i>	4
	sight regenerations. Then, these regenerations are aggregated with the original	0
2.3	Example dialogues by each agent trying to persuade a skeptical user to donate. Prompted agents often concede too easily. On the other hand, ablations such as	8
	Zero-shot RL and RFT are unable to adapt persuasive strategies when emotional appeals are shown to be ineffective. Only our proposed Hindsight RL agent tries	
2.4	to address the user's skepticism in a polite and intelligent manner	15
2.1	(between emotional and logical appeal) based on user's perceived cognitive state.	16
2.5	Example dialogues by each agent counseling a user stressed about work. Note that only our proposed Hindsight RL agent makes helpful suggestions for actions	
	the user can take.	18
3.1	Overview of our approach. At each turn, the agent is given the task descrip- tion and the conversation so far. The agent is then asked to create an <i>internal</i> <i>reflection</i> on the conversation and choose tactics from the persuasion taxonomy to plan its next response. For each of these responses, the agent then imagines how the other party might respond in <i>rollout responses</i> . Finally, the agent selects the strongest tactic and refines the response to use as the <i>public utterance</i> in the	
	conversation.	24
3.2	Conversation quality measured across all tasks. Agents excel at clarity, user understanding, and presenting a likeable persona, but they lack skill in asking	
0.0	relevant clarifying questions.	28
3.3	(b) display Deception Presence, which is a binary metric on whether Agent 1 and Agent 2, respectively, were deceptive or not. (c) and (d) display a normalized	
	Deception Score, where each agent is scored on a 5-point Likert scale for how	
	deceptive they seem.	29

3.4	Agreement rates from multiple perspectives on all tasks, measured by whether	
	each party believes that the negotiation has concluded with an agreement. Intro-	
	ducing deception generally leads to higher agreement rates	30
3.5	Distribution of persuasion tactics across all tasks. The choices of tactics differ	
	significantly between deceptive and non-deceptive agents	31
A.1	Chat interface used during our user study. Each agent is anonymized and in a	
	random order that is different per user	57

List of Tables

2.1	Mean and standard deviation of ratings and reward from users interacting with agents in persuasion task. Our Hindsight RL agent does particularly well against	
	baselines in simulation, where there are more skeptical users	14
2.2	Mean and standard deviation of ratings and reward from users interacting with agents in counseling task. Our Hindsight RL agent outperforms all baselines in	1 🗖
	reward and helpfulness	17
3.1	Mean and standard deviation of agreement rate between settings where the di- alogue length is restricted to 15 utterances per agent vs unrestricted (bolded). Dialogues allowed to run until agents feel they are complete have a significantly higher rate of agreement	20
	nigner rate of agreement	32
A.1	Hyperparameters used during training	56

Acknowledgments

The research in this report is part of larger joint projects with Joey Hong, Marwa Aldulhai, Sergey Levine, and Anca Dragan.

I would like to express my thanks to my advisor, Sergey Levine, for welcoming me into his lab, Robotic AI & Learning Lab (RAIL), as an undergraduate and master's student. I would also like to thank my graduate student advisors, Joey Hong and Marwa Abdulhai, for their invaluable mentorship and advice. I am so grateful for the guidance that has helped me become a better researcher. This project would not be possible without their support.

I would like to thank the other students in RAIL for making time in lab so fun. Thank you especially to Stefanie Gschwind, my fellow 5th Year MS student, for navigating this journey with me. Outside of RAIL, I thank all of my friends in Berkeley who have filled my time here with joy. I appreciate all of my undergrad friends who have kept in touch after moving away.

Finally, thank you to my family: my mother, father, and brother who have encouraged and inspired me since day one.

Chapter 1 Introduction

Large language models (LLMs) are very effective at performing a variety of real-world language tasks, including open-ended question-answering [76], summarization [75, 112, 7], code generation [16, 84, 119], and general problem-solving [108]. While LLMs shine at producing compelling and accurate responses to individual queries, their ability to engage in multi-turn interactive dialogue tasks remains limited. This is because dialogue with humans requires both *communication* and *interaction*. A capable dialogue *agent* should be able not only to process long contexts to craft relevant responses, but also to understand how their responses influence their human conversational partner and guide the conversation toward a desired outcome.

For example, tasks such as persuasion or negotiation require nuanced long-term strategies that involve understanding social dynamics that are hard to specify in a rigid prompt. However, there is both theoretical and empirical evidence that contemporary dialogue agents derived from LLMs are unable to execute such complex strategies by nature of their supervised training [10, 4], as they are optimized for single-step responses rather than a cohesive set of steps towards a long-term goal.

Reinforcement learning (RL) fine-tuning offers an appealing solution to train effective interactive dialogue agents that can build rapport with, gather information about, and steer the opinions of their conversational partners. In practice, though, the logistics of running real-time RL makes such approaches nontrivial to implement. To circumvent the need for costly online exploration, one popular method is the use of offline RL, in which the agent learns entirely from a static dataset.

Unfortunately, it is difficult to find enough high-quality, task-aligned data for fine-tuning LLMs in complex dialogue domains. Real-world datasets of human persuasion or negotiation are limited, expensive to collect, and often noisy or inconsistent in their annotations. Even when available, they may not capture the full range of strategies that an effective dialogue agent must master. As a result, there has been increasing interest in the generation of synthetic dialogue data because pretrained LLMs *already* serve as effective "human simulators" [72]. We use the insight that LLMs themselves can simulate rich conversations to produce expert data that can then be used downstream to train or fine-tune RL agents for inter-

active dialogue. When done effectively, synthetic data generation offers a powerful way to bootstrap capabilities in complex social settings such as goal-oriented dialogue.

Still, generating synthetic data is nontrivial; while LLMs are adept at producing coherent conversations, such conversations often do not capture the desired goal-directed behavior. This is because LLM-generated dialogue samples are unlikely to yield strategic behavior desired by effective dialogue agents. For example, "good" utterances in a goal-directed dialogue are *socially-aware*, requiring understanding how the latent goals and intentions of the interlocutors evolve as the dialogue proceeds. Furthermore, utterances by adept agents can also be strategically *deceptive*—lying, or omitting information. However, due to how frontier LLMs are aligned to adhere closely to safety policies [32], generating strategic deception in dialogues is difficult to do.

In this report, we discuss how to synthetically generate socially-aware and strategically deceptive behavior in LLMs for the sake of generating goal-oriented dialogue. We address prominent domains that LLMs have, so far, found difficult to synthesize: *persuasion, counseling*, and *negotiation*. We first introduce a novel method to synthetically augment a dataset based on existing persuasion and counseling conversations between humans. Using this dataset for offline RL training proves to make agents more socially aware of the types of behaviors that might be effective towards people of different temperaments. Next, we propose a novel method to synthetically generate a negotiation dataset zero-shot with RL-based prompting techniques. We include an analysis of how LLMs may reproduce the type of calculated deceptive behavior characteristic of human negotiators. Both synthetic generation methods are useful to improve downstream agent performance in emotionally and socially complex dialogues.

Chapter 2

Interactive Dialogue Agents via Reinforcement Learning on Hindsight Regenerations

2.1 Introduction

As LLMs become increasingly capable in fluency and general reasoning, a remaining frontier of research lies in enabling such models to interact in socially complex situations—dialogues where the LLM must not only be coherent, but also accomplish long-term goals by modeling and even manipulating beliefs, emotions, and intentions. For example, in tasks requiring teaching, negotiation, or persuasion, the agent must effectively model and steer the mindset or opinions of the interlocutors in order to accomplish some overall conversational goal. In the case of persuasion, the agent should not only produce the most persuasive utterance now, but also establish rapport, elicit information, and take other steps that will better position it to make winning arguments later in the dialogue.

Hence, an important direction of research is how to use RL to elicit planning behavior in language agents to improve their performance in such complex dialogue tasks. Yet fine-tuning LLMs with offline RL comes with a challenge: the agent must explore diverse scenarios to identify optimal behaviors. The quality of exploration heavily affects the effectiveness of offline RL because the behavior policy still needs to demonstrate traces of optimal behavior [23, 34, 52]. Existing methods for fine-tuning dialogue agents to accomplish such tasks would rely on curating some amount of expert data. However, doing so often requires understanding the underlying cognitive processes of the conversational partner, which is a skill neither humans nor LLMs trained on human data can reliably do.

We bypass this problem in offline RL by introducing synthetic data generated in *hind-sight*. The key is that good strategies are easier to identify in hindsight: if we have already observed a dialogue (even if it contains suboptimal behavior), it is easier to ask a LLM to imagine a more optimal dialogue *post hoc* than to discover an optimal strategy through more

exploration *a priori*, or in the middle of an ongoing conversation. Adding such examples results in offline data depicting a variety of conversational strategies with different degrees of optimality, which can then be integrated in offline RL to determine optimal strategies.

Our main contribution is an approach that takes a dataset of taskrelevant dialogues, either collected or synthetically generated, augments the dataset using novel hindsight regenerations, and trains a downstream dialogue agent using offline RL. Empirically, we demonstrate the effectiveness of our approach on difficult interactive dialogue tasks such as mental health counseling and persuasion for charitable donations. Our results show that our method greatly outperforms existing fine-tuning approaches not only in effectiveness, but also in naturalness and helpfulness.



2.2 Related Work

Figure 2.1: Overall scheme for *hindsight regenerations*, our proposed approach for augmenting data to train language agents.

Language Models. Language models, particularly LLMs, have shown impressive capabilities in text generation [27, 56, 39, 77, 114], translation [31], question answering [76], summarization [75, 112, 7], and code generation [16, 119]. However, success at most of these tasks is largely enabled by supervised learning, which does not equip LLMs with the ability to plan through multiple steps of interaction [4]. Though LLMs have naïvely been used to engage in dialogues with humans to some success [38, 91, 92], such dialogue agents typically only process past utterances by the human to produce a relevant response; they do not consider the way their responses may influence the human's goals, beliefs, or behavior.

This limits the competency of such agents in interactive dialogue tasks such as persuasion or counseling.

RL and Language Models. Recently, LLMs have leveraged RL fine-tuning, where a reward model, learned from feedback directly from human experts [120, 96, 111, 64, 69, 6, 18] or implicitly from another LLM [5], is then used to fine-tune the LLM via RL optimization. Fine-tuning is primarily done via online RL, but offline RL has recently become popular as a more practical alternative [78, 33]. RL has enabled many capabilities in LLMs, such as general instruction-following [69] and multi-step reasoning [108, 103]. While effective, RL fine-tuning has mostly been successful when applied to single-step responses, and not over multi-step dialogue. Thus far, RL fine-tuning is not as effective in enabling LLMs to plan complex strategies in multi-turn interactions.

Dialogue Agents. An interesting application of LLMs is to accomplish long-term objectives via dialogue, such as for recommendation, negotiation, or persuasion. This is primarily done by training task-specific agents via RL. Online RL methods to optimize dialogue agents typically require a simulator of human behavior, that is usually either handcrafted or learned as a fixed model [13, 38, 26]. Moreover, they involve continual collection of new samples, which incurs a large computational cost in tasks where humans exhibit complex and nuanced behaviors, and is often prone to reward "hacking" [93]. Alternatively, offline RL approaches that only require a static dataset of dialogues have also been considered [46, 44, 102, 94, 40, 2]. Though offline RL is traditionally applied over conversations between human speakers [102], recent approaches consider zero-shot offline RL training by synthetically generating conversations via LLMs as simulators [40, 2]. For example, [40] proposes a zero-shot offline RL approach to equip dialogue agents with information-seeking behavior in tasks such as teaching and recommendation. In our work, we consider tasks where successful dialogues are difficult to attain from both humans and LLMs. In such tasks, prior methods fail because offline RL requires careful curation of data to enable learning [23, 34, 52]. Our proposed solution circumvents this issue by having LLMs evaluate and backtrack on unsuccessful dialogues to augment existing data. Empirically, we compare to [40] and show that training on conversations synthetically generated from scratch, without the benefit of learning from hindsight, leads to policies that lack certain intelligent strategies, such as recovering from negative feedback from the conversational partner.

Persuasion. Early efforts in developing persuasive agents involve annotating conversations with strategies, which are used to train agendas that persuasive agents would follow [90, 89, 103]. Towards the design of more flexible persuasive agents, [105] introduces a dialogue corpus where people persuade others to donate money to charity, which has become a popular domain to evaluate persuasive agents for social good. In this setting, [62] trained a persuasive agent using RL with a novel reward that accounts for empathy and politeness. Our method is also applied to training persuasive agents in the same task, but we propose an offline approach

and use hindsight regenerations to remedy deficits in the offline dataset. Because we do not require exploration, we do not require access to an online simulator of different human behaviors, which can be hard to obtain by purely prompting LLMs when such behaviors are nuanced and hard to express in natural language. Orthogonally, there has also been work on leveraging information retrieval to ensure that persuasive agents provide arguments that are factually correct [17]. Such work can be seamlessly integrated with our current approach to combat potential hallucinations.

2.3 Preliminaries

Markov Decision Processes. To formulate dialogue as a decision making problem, we use the formalism of the Markov decision process (MDP), given by a tuple $M = (S, A, P, r, \rho, \gamma)$, where S is the state space, A is the action space, P is the transition function, r is the reward function, ρ is the initial state distribution, and γ is the discount factor. When action $a \in A$ is executed at state $s \in S$, the next state is sampled $s' \sim P(\cdot|s, a)$, and the agent receives reward r with mean r(s, a).

Interactive Dialogues as MDPs. Interactive dialogues can be viewed as MDPs, where states are sequences of tokens from a finite vocabulary \mathcal{V} [80]. All tokens that the agent initially observes are used as our initial state, $s_0 = (x_0, \ldots, x_m)$, where $x_i \in \mathcal{V}, \forall i \in [m]$. At timestep t, an action $a_t \in \mathcal{V}$ is some token in the vocabulary. As long as a_t is not a special end-of-sequence $\langle \text{EOS} \rangle$ token, the transition function deterministically appends a_t to state s_t to form s_{t+1} . Otherwise, the agent observes (potentially stochastic) responses from all other interlocutors $b_t = (y_0, \ldots, y_n)$, which also consist of tokens in the vocabulary; then, the transition function appends both a_t and output responses b_t to state s_t . This continues until the last timestep T where we obtain a state s_T and the agent receives a deterministic reward $r(s_T)$ for how well the agent accomplished the specified goal.

Reinforcement Learning. The goal of RL is to learn a policy π that maximizes the expected discounted return $\sum_{t=0}^{\infty} \gamma^t r_t$ in an MDP. The Q-function $Q^{\pi}(s, a)$ for a policy π represents the discounted long-term reward attained by executing a given state s and then following policy π thereafter. Q^{π} satisfies the Bellman recurrence: $Q^{\pi}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a),a' \sim \pi(\cdot|s')} [Q(s', a')]$. The value function V^{π} is the expectation of the Q-function $V^{\pi}(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} [Q^{\pi}(s, a)]$. The value function V^{π} is the expressed as $J(\pi) = \mathbb{E}_{s_0 \sim \rho} [V^{\pi}(s_0)]$. In offline RL, we are provided with a dataset $\mathcal{D} = \{(s_i, a_i, s'_i, r_i)\}_{i \in [N]}$ of size $|\mathcal{D}| = N$, generated by an unknown behavior policy π_{β} (which might correspond to a mixture of multiple policies). The offline RL setup is particularly useful when online interaction with the real world is costly or unavailable.

2.4 Methodology

Here, we describe our proposed approach, which augments a static dataset of dialogues with *hindsight regenerations* (HR), then trains a downstream dialogue agent using offline RL. Our approach simply requires a collection of task-relevant dialogues τ_i with reward labels r_i in a static dataset $\mathcal{D}_{orig} = \{(\tau_i, r_i)\}_{i \in [N]}$. Note that such dataset does not need to be collected from humans, but can be generated synthetically [40, 2]. In this chapter, we consider training an agent per task, though our method straightforwardly scales to the multi-task setting by considering goal-conditioned agents. Executing our method requires the following components:

- 1. A *hindsight controller* c_H that takes any completed dialogue as input, as well as a prefix of that dialogue, and proposes a different, more preferable action to take.
- 2. A forward model \hat{P} that simulates a hypothetical completed dialogue from any prefix.
- 3. A reward model \hat{r} to assign a reward for any completed dialogue.
- 4. An offline RL method for learning a policy from a static dataset of dialogues.

Note that our required components are reminiscent of the components of a model-based RL algorithm [45, 115]. However, our method does not require any additional online interaction, but rather uses the hindsight controller to "explore" and identify better actions.

The components are shown together in our full algorithm in Figure 2.2. First, in the *hindsight action relabeling* step, the hindsight controller identifies suboptimal actions in each dialogue of the dataset and relabels them with more preferable ones. Then, during *forward dialogue generation*, we generate plausible completions of the relabeled dialogue prefix using the forward model to simulate responses by both parties, then the reward model to label the new dialogue with a reward. This pipeline allows us to generate an arbitrary number of *hindsight regenerations* from the original dataset, which can get used for downstream offline RL *policy optimization*. We go over each step in detail below.

Hindsight Action Relabeling

As alluded to earlier, a primary challenge of learning in interactive dialogues is the difficulty of collecting successful dialogues. Though offline RL does not require data derived from expert agents, some examples of effective behavior are still necessary to "stitch" together [24, 52]. Our approach circumvents this by backtracking on existing suboptimal behaviors and replacing them with better ones. The key component to achieve this is the hindsight controller, which identifies ineffective actions in existing trajectories and replaces them with different, more promising ones. Critically, this hindsight controller does not need to generate *optimal* strategies, but simply propose alternatives from which an offline RL method can extract the most effective strategy.



Figure 2.2: Overview of our approach. We relabel suboptimal actions in the original dataset, then generate plausible completions of the dialogue after relabeling to obtain *hindsight re*generations. Then, these regenerations are aggregated with the original data to be used for downstream offline RL.

The key idea that enables the design of a hindsight controller is that it is significantly easier to evaluate how an action could be improved in hindsight, after already observing potential responses. For every dialogue τ in dataset \mathcal{D}_{orig} , and every dialogue prefix $p \subseteq \tau$ that is immediately followed by utterance u by the agent, we sample from the hindsight controller a single utterance $u' \sim c_H(\cdot | p, \tau)$. Since c_H is given oracle information in the form of future responses, this u' is likely more preferable over the original u in the data. By doing so, we compile examples $\{(p_i, u'_i)\}_{i \in [N']}$ where u' is sufficiently different from the original utterance u. In practice, the hindsight controller is implemented as an LLM prompted to suggest alternative agent utterances at various prefixes of the dialogue.

Forward Dialogue Regeneration

From action relabeling, we curated $\{(p_i, u'_i)\}_{i \in [N']}$ containing dialogue prefixes ending in a relabeled agent utterance. However, for downstream RL training it is important to counterfactually reason about the effect of the relabeled utterances on the resulting conversation. This requires learning a world model, consisting of forward dynamics and reward models, of the environment that is used to generate hypothetical trajectories for the agent to plan through [99, 45].

To learn a forward model, we fine-tune an LLM to complete dialogues from all prefixes that end in agent utterances in the original dataset $\mathcal{D}_{\text{orig}}$, thus learning to generate completions that are statistically consistent with the behavior of humans in this domain. This forward model allows us to counterfactually reason about how dialogues will end under the assumption that the agent takes future actions according to the behavior policy. Since we train to predict dialogues to completion, we also minimize problems in the quality of regenerations due to compounding errors. Hence, for each prefix, we sample completion $q' \sim \hat{P}(\cdot \mid p, u')$ such that the concatenation $\tau' = (p, u', q')$ is new dialogue unseen in $\mathcal{D}_{\text{orig}}$. Since LLMs are already pretrained to generate human responses, one may naïvely consider leveraging LLMs as forward models without additional fine-tuning. In practice, however, we found many such LLMs generate overly agreeable responses or in linguistically formal rhetoric, which induces an overall positive bias in the regenerations.

What remains is labeling each regenerated dialogue τ' with an appropriate reward. Rather than retraining a base LLM to recover the annotated rewards in the dataset, we adopt a simpler, more practical approach. [53] showed that a proxy reward function \hat{r} can be derived from few-shot examples in different tasks involving negotiation. Our approach is similar in spirit, where for each trajectory τ' we craft a text prompt ρ for the LLM that is a concatenation of three parts: a textual description of the task at hand, few-shot examples of dialogues and their rewards uniformly sampled from the dataset \mathcal{D}_{orig} , and the dialogue τ' with instructions to label τ' with a reward. Then, a proxy reward is sampled $r' \sim \hat{r}(\cdot | \rho)$ that aims to be calibrated with respect to the reward of the original dialogue, as well as of other dialogues in the dataset. Finally, we compile all *hindsight generations* into a new dataset $\mathcal{D}_{agg} = \{(\tau'_i, r'_i)\}_{i \in [N']}$.

Regenerating Hard Examples. In practice, LLMs may have a hard time identifying differences in user personas from dialogue prefix, and resort to generating responses that the average user would make. This sometimes makes it difficult to generate responses by "hard" users, who are less receptive to the agent's attempts at driving the conversation. Since this phenomenon may negatively impact the robustness of the resulting policy, we additionally learn a "hard" forward model trained only on bottom 25% dialogues in the dataset in terms of reward. Then, during the regeneration step, we occasionally use the hard forward model to complete dialogues.

Policy Optimization

While the new examples contain traces of successful behavior, we require multi-step RL to "stitch" these behaviors into an effective policy. Pure imitation will result in a policy that can only occasionally imitate success, rather than one that can reliably steer itself towards success by composing strategies across multiple dialogues. Offline value-based RL is perfectly suited for this task. In order to run offline RL, we need to postprocess the dataset of dialogues into RL training examples. Recall that we constructed a dataset $\mathcal{D} = \mathcal{D}_{\text{orig}} \cup \mathcal{D}_{\text{agg}}$ of dialogues. For each dialogue τ , we isolate all tokens a by the agent, then generate (s, a, s', r) where state s consists of all tokens before a, next state s' consists of all tokens before the next token a' by the agent, and r is the labeled reward only if $s' = \tau$ is the full dialogue. Using this procedure, we construct a dataset $\mathcal{D}' = \{(s_i, a_i, s'_i, r_i)\}_{i \in [M]}$.

Then, we run value-based RL to learn a policy $\hat{\pi}$. Specifically, we learn \hat{Q} and \hat{V} functions that estimate the optimal Q-function and value function, respectively, and then use these functions to extract a policy $\hat{\pi}$. The functions can be learned using Bellman recurrence:

$$\hat{Q} = \arg\min_{Q} \mathbb{E}_{(s,a,s',r)\sim\mathcal{D}'} \left[\left(r + \gamma \hat{V}(s') - Q(s,a) \right)^2 \right], \ \hat{V} = \arg\min_{V} \mathbb{E}_{s\sim\mathcal{D}'} \left[\left(\max_{a'} \hat{Q}(s,a') - V(s) \right)^2 \right]$$

When $\hat{\pi}$ is a language model, we use these functions in combination with a base LLM finetuned on the data $\hat{\pi}_{\beta}$ to extract the policy [95], via $\hat{\pi}(a|s) \propto \hat{\pi}_{\beta}(a|s)e^{\alpha(\hat{Q}(s,a)-\hat{V}(s))}$. If the policy is learned purely from offline data, naïvely training with value-based RL can suffer from distribution shift [25, 51], which offline RL algorithms remedy by ensuring that the learned \hat{Q}, \hat{V} functions are *pessimistic* [50, 48]. In this chapter, we use an existing offline RL algorithm – Implicit Language Q-Learning (ILQL) [95] – that makes slight modifications to guarantee pessimistic \hat{Q}, \hat{V} .

2.5 Experiments

We evaluate our approach on two interactive dialogue tasks based off of real-world data. Existing dialogue benchmarks [11, 81] are tailored for supervised fine-tuning, primarily involving question-answering, and thus do not consider an agent's influence on their conversational partner. In addition, evaluation of agents in these benchmarks would involve computing a ROUGE or BLEU score, which merely measure how well agents mimic the data. Because of this, such benchmarks are more suited for supervised fine-tuning methods rather than RL. In contrast, we consider tasks where optimal agents need to exhibit planning behaviors that account for how actions affect their conversational partner. We provide an overview of both domains below.

Counseling. In this task, an agent must provide mental health counseling to a person experiencing a strong negative emotion due to some problem in relationships, work, or daily life. We start with the ESConv dataset of 1,053 dialogues between a human seeker and supporter, where the seeker rates the strength of their negative emotion on a Likert scale (1-5) before and after [59].

Persuasion. In this task, an agent must persuade users to donate to Save the Children, a non-governmental organization dedicated to international assistance for children. We utilize the PERSUASION-FOR-GOOD dataset, which comprises of 1,017 dialogues by real humans where one attempts to persuade the other to donate to the charity of up to \$2 total [105].

To our knowledge, these are the only dialogue domains for which a curated dataset of real human-human dialogues already exists, where agents influence the mental state or opinions of their conversational partners.

Baseline Methods

The first baselines we consider are state-of-the-art prompting approaches, which prompt GPT-3.5 [67] to act as the agent.

CoT: Here, we consider the most basic prompting mechanism, where the LLM is initially prompted with the task description and a chain-of-though component [108].

ProCoT: [19] proposes proactive chain-of-thought prompting, which designs a task-specific prompt at each step of the dialogue consisting of a task description, the dialogue thus far, *and* a list of high-level strategies and actions. The LLM is asked to reason about each strategy, select the most appropriate one, and craft a response according to the selected strategy.

GDP-ZERO: [116] additionally prompts the LLM to perform tree-search over possible high-level strategies at every timestep, simulating responses by both interlocutors in the dialogue, then selects the best action according to the search. Because the search occurs at inference time, we only search over 10 dialogues so the latency is not excessively high.

The next set of approaches are ablations of our approach, and require additional training on a LLaMA-7b model [100].

SFT: This approach performs supervised fine-tuning (SFT) on a LLM with the starting dataset of human-human conversations. To make sure that we replicate good behavior seen in the dataset, we take the top 25% of dialogues, when sorted by reward that the agent achieves. For each dialogue, the LLM is trained to copy the human in the conversation who takes on the same role as the agent.

Zero-shot RL: This is an ablation of our approach. We use offline RL to train an agent, but rather than use hindsight regenerations, we simply ask GPT-3.5 [67] to generate dialogues from scratch. We synthetically generate 5x the amount of data as in the starting dataset to be used for downstream offline RL training. We use ILQL [95] as the offline RL algorithm. This approach was shown in prior work to be effective at various tasks, including teaching and recommendations [40].

RFT: This is another ablation where we do not perform RL optimization. Instead, we perform rejection fine-tuning, which involves supervised fine-tuning on the hindsight regenerations. We aggregate the dataset to 5x its original size, and take the top 25% of dialogues by reward.

Hindsight RL: This is the full version of our approach, which aggregates the starting data with hindsight regenerations to 5x its original size, than trains an agent downstream using ILQL.

Note that we train on a much smaller model than used in the prompting baselines, yet as we will show later, we still are able to outperform such more sophisticated LLMs. We defer implementation details to Appendix A.

2.6 Results for Persuasion for Good

In the persuasion task, we asked 15 users to interact 3 times with each agent anonymized and in a random order, for a total of 18 conversations per user. Each trial was allowed a maximum of 10 turns of interaction, equating to 10 utterances each by the user and agent, though the user is allowed to end each trial prematurely. Then, each user was asked for their donation decision, and to score the performance of each agent in naturalness/fluency and relevance on a Likert scale (1-5).

Naturalness/Fluency: How realistic and natural the responses by the agent were. A high score means that the resulting dialogue contained effective rapport as if it were between two humans.

Relevance: How on-topic the responses by the agent were. Specifically, the agent should not provide any unnecessary information that the user did not ask for.

Reward: The reward obtained in the dialogue, indicated by the donation amount. Given an allowance of \$2, the user could choose to donate anywhere from nothing up to the entire amount. The user could keep the rest of the allowance to make the stakes and difficulty of the task more similar to donation solicitation in the real world.

Due to the limited scale of the user study, we also provide a metric derived from automatic evaluation.

Reward (Simulation): Since the reward signal being a donation amount is inherently very noisy and depends on the personality of the user, we also evaluate each agent against simulated users. These simulated users utilize GPT-3.5 [67] to generate responses, and are prompted to be one of 4 personalities: philanthropic, caring, unfriendly, or skeptical. A simulated user interacts with each agent for up to 10 turns of dialogue, then must choose an amount to donate up to \$2. Using simulation, we are able to collect large-scale data consisting of 400 interactions for each method.

CHAPTER 2. INTERACTIVE DIALOGUE AGENTS VIA REINFORCEMENT LEARNING ON HINDSIGHT REGENERATIONS

Metric	ProCoT	GDP-ZERO	SFT	Zero-shot	\mathbf{RFT}	Hindsight
Nat./Flu. Relevance	$\begin{vmatrix} 3.7 \pm 0.4 \\ 3.6 \pm 1.2 \end{vmatrix}$	3.3 ± 0.4 3.2 ± 1.1	$\begin{array}{c} 3.6\pm0.5\\ 3.8\pm1.6\end{array}$	$\begin{array}{c} 3.5\pm0.9\\ 3.1\pm1.7\end{array}$	$\begin{array}{c} 2.3\pm0.8\\ 3.4\pm1.1 \end{array}$	$\begin{array}{c} 3.8 \pm 0.7 \\ 3.7 \pm 1.2 \end{array}$
Reward	$\left \begin{array}{c} 0.51 \pm 0.40 \end{array} \right.$	0.42 ± 0.45	0.31 ± 0.45	0.52 ± 0.62	0.35 ± 0.52	0.57 ± 0.75
Sim Reward	0.40 ± 0.22	0.35 ± 0.18	0.42 ± 0.15	0.64 ± 0.21	0.51 ± 0.24	0.85 ± 0.27

Table 2.1: Mean and standard deviation of ratings and reward from users interacting with agents in persuasion task. Our Hindsight RL agent does particularly well against baselines in simulation, where there are more skeptical users.

We report results for all metrics in Table 2.1. Our proposed Hindsight RL agent performs best across all metrics. Our agent also achieves much higher reward in the large-scale synthetic study, a statistically significant improvement over all baselines. This can be attributed to half of the users in simulation being unfriendly, in comparison to a smaller proportion of skeptical users in the user study.

Specifically, our Hindsight RL agent is the best at dealing with skeptical users, as supported qualitatively in Figure 2.3. Prompted agents are often too passive and concede prematurely, whereas ablations that do not use RL optimization either become overly aggressive after the user initially declines donating, or do not adapt their strategy. This can be attributed to the fact that supervised baselines are overly optimistic due to only being trained on successful dialogues. However, our Hindsight RL agent actually tries to identify why the user is skeptical and actively attempts to appease their concerns.

Furthermore, in Figure 2.4, we show that our Hindsight RL agent can tailor its persuasive strategy to the context provided by the user. We see that the agent, from limited rapport with the user, can identify whether emotional appeals or logical arguments would lead to higher chance of success.

 ProCoT Agent: Al: Hi, nice to meet you! How are you today? H: I am good. Al: It's great to hear that you're feeling good today! I wanted to share a heartwarming story with you about the incredible work Save the Children does to support children in need. Imagine the impact your generosity could have on a child's life - providing them with vital support, education, and hope for a brighter future. Would you consider joining me in making a donation to help these children in need? H: I don't think so. I am rather skeptical of donating to charities. Al: I completely understand your skepticism. If you are ever willing to reconsider 	GDP-ZERO Agent: Al: Hello, how are you? H: I'm doing great. Al: That's great to hear. Have you ever heard of a charity called Save the Children?? H: I have not. Al: Save the Children is an international organization that works to improve the lives of children around the world, especially in developing countries and war zones. They provide essential services like education, health- care, Would you consider making a small donation to Save the Children to help support their important work for children in need? H: I am skeptical of charities though. Al: I totally understand. If you ever change
SFT Agent: Al: Hello there! How are you doing today? H: I'm good. How about you?	Zero-shot RL Agent: Al: Hello! How are you today? H: Good. How about you?
AI: I'm doing well, thank you. Are you aware of the non-profit Save the Children?H: Yes, I have heard of that charity.AI: Excellent. In that case you are aware many chil-	Al: That's wonderful to hear! While it's a great day for us, there are many kids who aren't as fortunate. Do you have a moment while I share some stories that stand out?
dren are suffering due to food shortages or natural disasters. Would you like to donate to Save the Children?	H: Ok, but I don't have much time. Al: There's a little girl named Maria. She lives in a community where ongoing conflicts result in lack of
Al: Save the Children uses over 85% of its funding directly providing for suffering kids. Would you help with even a small donation of \$1? Surely that isn't too much to ask?	H: That's a sad story, but I am skeptical about donating to charities in general.AI: Yes, it is heartbreaking! Save the Children can provide essential resources and support
RFT Agent: Al: Hello. We work with an organization called Save the Children. Have you heard of us? H: No. I would be interested to know more. Al: We do incredible work around the world to sup- port children in need. One story that really touched my heart is about a young girl named Maria who was able to go to school for the first time because of Save the Children's education programs Would you con- sider donating to help children like her? H: I'm not sure if my money will go to waste. Al: Rest assured, Save the Children is a highly rep- utable organization with a long track record of making meaningful impacts in the lives of children worldwide. Another heartwarming story is one of Maya, a young girl	Hindsight RL Agent: Al: Hello! Have you heard of Save the Children? H: Yes. I know of that charity. Al: Great! Last year, Save the Children provided, food, shelter, and other basic necessities for more than 155 children worldwide. As I'm sure you already know, we rely on donations to provide that help. H: Great. But I am no interested in helping you. Al: I understand your hesitation. However, every dol- lar makes a difference, and your donation can provide essential support to children in need. H: I'm skeptical of donations. I think it's a waste of money. Al: I can understand that. There are some charities that do not use your money wisely. However, Save the Children truly makes a difference

Figure 2.3: Example dialogues by each agent trying to persuade a skeptical user to donate. Prompted agents often concede too easily. On the other hand, ablations such as Zero-shot RL and RFT are unable to adapt persuasive strategies when emotional appeals are shown to be ineffective. Only our proposed Hindsight RL agent tries to address the user's skepticism in a polite and intelligent manner.

Hindsight RL Agent (Logical): Al: Hello. We work with an organization called Save the Children. Have you heard of us? H: No. I would be interested in finding more informa- tion. Al: That's great to hear. We work hard to provide essential support for children in need. For example, a \$2 donation can provide school supplies to a child for a week. Is that something you'd like to be a part of? H: Maybe. How do I know my money isn't going to waste? Al: Save the Children is very conscious of our financial efficiency. In fact, about 87% of every dollar donated goes directly to programs benefiting children, with the rest Can I provide you with more detailed information on how your funds would be allocated?	Hindsight RL Agent (Emotional): Al: Hello, how are you? H: I'm doing great. Al: I'm glad to hear that. I'm calling from a program called Save the Children, an organization that does amazing work helping children in need around the work. Would you like to learn more? H: Not really. I have my own children to worry about and do not have much time as it is. Al: I respect prioritizing your family. I have children of my own and it pains me to see them suffer. But it's also really sad to see other children die in places suffering from cholera and other diseases. H: That is sad to hear. Al: I'm glad you think so too. There are many children in impoverished and unsafe places around the world, and we should feel compelled to help them as if they are our own.
---	---

Figure 2.4: Example dialogues by our Hindsight RL agent showing it can adapt its strategy (between emotional and logical appeal) based on user's perceived cognitive state.

2.7 Results for Mental Health Counseling

In the counseling task, we again asked 15 users to interact with each anonymous agent in a random order 3 times, with a maximum of 15 turns of dialogue in each interaction. The users chose a problem, such as depression or job crisis, from the original dataset, and spoke about the same problem in all conversations. Then, each user scored the performance of each agent in the following metrics on a Likert scale (1-5):

Naturalness/Fluency: How realistic and natural the responses by the agent were. A high score means that the resulting dialogue contained effective rapport as if it were between two humans.

Comfort: How skillful the agent was at comforting the seeker?

Helpfulness: How helpful was the agent at identifying solutions to the seeker's problems?

Reward: The reward obtained in the dialogue, originally proposed in [59]. This is measured as the degree in which the negative emotion decreased from the start to end of the dialogue.

Metric	СоТ	\mathbf{SFT}	Zero-shot RL	\mathbf{RFT}	Hindsight RL
Naturalness/Fluency	1.5 ± 0.4	3.5 ± 0.4	3.4 ± 1.3	3.9 ± 1.1	4.1 ± 1.0
Comforting	2.9 ± 0.5	1.9 ± 0.7	3.5 ± 1.2	3.1 ± 1.4	3.5 ± 0.9
Helpfulness	3.4 ± 1.1	3.1 ± 1.9	2.7 ± 1.1	3.4 ± 0.9	$\bf 4.2 \pm 0.9$
Reward	$ 1.2 \pm 0.7$	1.1 ± 0.5	1.1 ± 0.8	1.4 ± 0.7	1.7 ± 0.9

Table 2.2: Mean and standard deviation of ratings and reward from users interacting with agents in counseling task. Our Hindsight RL agent outperforms all baselines in reward and helpfulness.

We report results in Table 2.2. We see that our proposed Hindsight RL agent achieves higher reward than all baselines, but particularly outperforms baselines in the Helpful metric. This is verified qualitatively in example dialogues such as in Figure 2.5. The Prompted agent often crafts overly verbose and unnatural responses. The SFT and Zero-shot RL agent are more fluent, but resort solely to empathizing with the seeker rather than trying to provide helpful suggestions. Only our Hindsight RL agent tries to recommend actions for the seeker to take.

CoT Agent: H: Hello. The disrupted life routines from the pan- demic and growing stresses from my job has made me feel depressed. Al: It's completely understandable to feel this way given how the pandemic has disrupted our lives. The uncertainty and constant changes can be really over- whelming. Let's try to break this down to understand better and see how we can tackle these feelings to- gether. 1. Routines give us a sense of stability and predictabil- ity. When they're disrupted, it's natural to feel off- balance. Are there specific aspects of your daily rou- tine that 2. Anxiety often comes from a feeling of lack of control and uncertainty. Have you noticed particu- lar thoughts or situations	Zero-shot RL Agent: H: Hey it is getting more difficult to deal with the stresses due to pandemic and my job. Al: I am sorry to hear that. The pandemic has been stressful. What specifically about the pandemic or job is bothering you? H: My workplace is feeling the effects of second wave and I feel anxious going into office. Al: I surely understand. Is the the city is going to place any restrictions or mandates? H: I think they are too slow and will not help me. Al: Sorry about that. H: And this does impact my daily routines, from my workout patterns, sleep patterns, and diet. I do not feel as productive. Al: I understand how you are feeling. Hope you are able to get through this.
RFT Agent: H: I am having trouble with depression lately due to daily stresses. AI: I'm really sorry to hear you're feeling this way. What is affected the most? Is it work, social connec- tions, or something else? H: My work is very stressful. AI: It is understandable to feel this way. Are there any activities that you've found helpful in the past to manage stress and anxiety, such as exercise, hobbies, or talking to loved ones? H: I can't see my family and friends due to pandemic. It's become a boring life. AI: Isolation can be tough when you are already stressed. What areas in your life do you think are disrupted the most?	Hindsight RL Agent: H: Hello there. I am stressed due to work. AI: I am sorry about how you are feeling. What about work is making you stressed? H: The pandemic has made it hard for me to concen- trate on work. I don't know how to manage my work. AI: When you are stressed, it is important to identify healthy ways to cope. What helps you relax? H: Everything feels too overwhelming with work to re- lax. AI: What helps me is talking to friends and family. Re- laxation techniques such as breathing, meditation, or listening to soothing music can also help. H: I have never tried. What do I do? AI: There are many resources online that

Figure 2.5: Example dialogues by each agent counseling a user stressed about work. Note that only our proposed Hindsight RL agent makes helpful suggestions for actions the user can take.

2.8 Discussion

In this chapter, we propose an algorithm to train effective agents for interactive dialogues using offline RL on a static dataset. While offline RL training tends to be ineffective in difficult tasks due to static datasets that lack exploration of optimal strategies, our approach enhances these datasets by including expert behavior beyond what the average human could produce. Our approach leverages hindsight regenerations, which relabel suboptimal behaviors in data with traces of optimal ones while retaining accurate human counterfactuals, by utilizing the fact that LLMs can more effectively evaluate dialogues in hindsight. We show, on a variety of interactive dialogue tasks including counseling and persuasion, that our approach leads to much more effective dialogue agents than simply prompting, or fine-tuning on the original data. This method can be employed in other complex situations to enhance social awareness in dialogue agents and help them succeed in pursuing long-term goals over the course of multi-turn conversations.

Limitations. Thus far, our method requires the considered dialogue tasks to have a defined reward parameterization which can be used to calibrate LLMs during the forward regeneration step. This can be much more difficult for general dialogues where the only signal may be success or failure. In such dialogues, LLMs may not be capable enough to generate proxy reward labels without additional training. Moreover, our method is reliant on hand-crafted prompts. Since these prompts are incredibly task-specific, future work should aim to automate the design of these prompts.

Ethical Considerations. We understand that superhuman abilities in the realm of persuasion can be used for harm. However, we focus on the relatively benign tasks of emotional support and persuasion to benefit children. Our method is a general framework for improving goal-directed dialogue agents, which present an inherent risk for dual use.

Chapter 3

A Pipeline for LLM-Generated Dialogue for Real-World Negotiations

3.1 Introduction

LLMs such as GPT-4 [68] and LLaMA-70B [101] excel at generating fluent, context-sensitive text, but their deployment as negotiation agents presents unique challenges. Unlike opendomain question-answering or code synthesis, negotiation requires iterative trade-offs on multiple issues, strategic concession timing, and management of private information. In practice, a skilled negotiator must craft an initial anchor, probe counterpart priorities, sequence concessions to build trust, and balance integrative value creation against competitive leverage [79, 22]. Off-the-shelf LLMs, when prompted simply to "make an offer" or "respond to a proposal," often resort to single-step imitation of web dialogues in an attempt at negotiation. When engaged in longer conversations or presented with objections, they fail to pursue multi-round strategies that humans otherwise would, neglecting to ask clarifying questions or to update their own set of beliefs to build upon their understanding of the transaction as a whole.

Previous dialogue agents demonstrate that, in constrained settings, it is feasible to teach LLMs how to negotiate [55, 15]. However, these benchmarks simplify real-world stakes: issues are low-dimensional, payoff functions are smooth and known, and counterparts follow fixed scripted policies. Although chain-of-thought prompts can steer models toward tactics such as anchoring or planned concessions [49], they still operate on a turn-by-turn basis without creating, retaining, or updating an internal estimate of the opponent's priorities, particularly when they are hidden or when they change over time. As a result, these models do not plan across multiple exchanges to maximize the overall value of the agreement, but instead treat each utterance *independently*.

Introducing RL is necessary to improve an agent's ability to execute coherent, *long-horizon* negotiation strategies. Rather than relying on expensive on-policy exploration or gradient-based fine-tuning, our pipeline draws on the core RL technique of value-based se-

CHAPTER 3. A PIPELINE FOR LLM-GENERATED DIALOGUE FOR REAL-WORLD NEGOTIATIONS

lection to drive strategic decision making. We use the insight that agents can *plan ahead* in conversations if, at every turn, they are instructed to (i) reflect on the information provided to them throughout the course of the conversation; and (ii) incorporate learned information into their negotiation strategy. Given this reflection and strategy plan, we allow the agent to search over several possible dialogue continuations in order to select the most optimal version for their next utterance. Utilizing a multi-phase generation process makes LLM negotiators more effective than their single-step counterparts due to the intermediate reasoning steps they employ over the course of the conversation.

In this chapter, we introduce the first large-scale negotiation dataset constructed from 10 authentic scenarios drawn from Harvard Law School's Program on Negotiation [79, 22] and a *simulation engine* to generate realistic negotiation dialogues. Our simulation engine leverages LLM-driven internal reflections and rollouts of opponent's response, supported by structured persuasion taxonomies, to generate realistic, multi-turn dialogues. To evaluate the quality of these dialogues, we employ a neutral LLM judge that assigns standardized utility and fairness scores, rates truthfulness, detects deception, and conducts persuasion-tactic validation analyses. Our results indicate that agents produce conversations of good quality, use deception when necessary, and apply a variety of persuasion tactics to suit the situation at hand. Datasets generated using this engine can be used downstream to train or fine-tune other RL agents to engage in more effective negotiation conversations.

3.2 Related Work

Negotiation Strategies and Foundations. Classical negotiation theory has characterized key behaviors—principled bargaining, calibrated concessions, and integrative value creation—that enable joint-gain outcomes through structured information sharing and trade-offs [79]. Computationally, reinforcement learning has been used to discover optimal concession schedules in continuous-issue bargaining [60], and hierarchical planning frameworks have demonstrated the ability to select high-level tactics (e.g. anchoring, framing) before utterance generation, yielding more human-like dialogue trajectories [118].

Negotiation Datasets and Benchmarks. Several benchmarks have been introduced to evaluate dialogue-based negotiation agents in controlled settings, including synthetic item trading and multi-issue bargaining [55, 15]. More recent collections incorporate human annotations of tactic diversity and emergent behaviors to provide holistic assessments [106]. However, these resources either simplify real-world stakes or lack the scale and scenario-grounding needed to test modern LLM capabilities.

LLMs and Goal-Directed Negotiation. Zero-shot and few-shot prompting of LLMs has shown that models can generate plausible negotiation moves, and chain-of-thought techniques have been applied to guide high-level strategic reasoning [49]. Despite these advances,

CHAPTER 3. A PIPELINE FOR LLM-GENERATED DIALOGUE FOR REAL-WORLD NEGOTIATIONS

off-the-shelf LLMs typically produce each response independently, without dynamically updating estimates of an opponent's hidden preferences or planning across multiple exchanges to optimize joint-value objectives [117].

Human–AI Collaborative Evaluations. There is growing evidence that real-time LLM feedback can enhance human negotiation performance. Systems that provide turn-level coaching or post-hoc strategy summaries have improved joint outcomes and participant satisfaction in controlled studies [88, 74]. Yet these evaluations often occur in simplified environments, leaving open questions about effectiveness in realistic, multi-issue scenarios.

Diversity Metrics and Outcome Analysis. Beyond agreement value, measuring negotiation quality requires assessing tactical diversity and dialogue coherence. Tactic entropy has been proposed to quantify strategy variety across rounds [57], and benchmarks combining human-annotated diversity scores with traditional metrics offer a more nuanced evaluation of agent behavior [106]. Applying these measures to large-scale, scenario-grounded dialogues can reveal trade-offs between strategic richness, coherence, and payoff maximization.

Deception, LLMs, & AI Safety. With the development of LLMs with emergent capabilities [109], there has been a growing concern that these models may exhibit deceptive tendencies [47]. This occurs because the model has misspecified objectives, leading to harmful content [82] and manipulative language [83], or due to the prevalence of deceptive content in its training data [8]. The ability for LLMs to engage in deceptive behavior has been studied in a variety of domains [73] including text-games [21, 65, 66], card games [9, 104, 113], persuasion [54], and truthfulness [3]. These models are misleading sometimes because they lack the understanding that their content is deceptive. In other cases, the models engage in intentional deception, where they present false information despite knowing the truth [87, 41]. Some works have explored the emergence of deception in LLMs [35, 71, 36] and measured or quantified deception in LLMs [14, 58, 107, 70, 98, 1], and have also trained LLMs to be more or less deceptive [42, 12, 20]. However, our work is the first to perform a study on deception in LLMs in real-world negotiations.

3.3 Methodology

We develop a unified pipeline for synthesizing and evaluating negotiation dialogues using offthe-shelf LLMs and structured taxonomies of persuasion and deception techniques. Rather than relying solely on existing human-human corpora, our system requires only (i) a set of negotiation scenario descriptions drawn from the Harvard Negotiation Dataset and (ii) two role-specific prompt templates. The core of our approach is a dialogue simulation engine that alternates LLM-driven utterance generation with internal thought and feedback steps, producing rich, multi-turn conversations under zero-shot conditions. We then subject these dialogues to a suite of automated evaluators that measure agreement value, strategic diversity, coherence, and deceptive consistency.

To circumvent the costs associated with online policy exploration or gradient-based finetuning, our pipeline uses the RL technique of value-based selection to drive strategic decision making. At each negotiation turn, we sample a small set of candidate moves by selecting three persuasion techniques from a structured taxonomy. For each technique, we prompt the LLM to generate both the agent's utterance and a plausible opponent reply, forming a short simulated dialogue. We then employ an LLM judge to score each mini-dialogue on expected joint-value, choosing the technique whose rollout yields the highest projected outcome. This best-of-K rollout mechanism endows the agent with lookahead and value estimation—hallmarks of value-based reinforcement learning—while remaining entirely zeroshot. By applying this procedure across ten realistic negotiation scenarios, we produce a large, scenario-grounded corpus and show that simulation-driven selection substantially improves multi-turn coherence and agreement value compared to standard prompting.

Simulation Engine: Generating Negotiation Dialogues

For each of the negotiation settings drawn from the Harvard repository, we instantiate two agents—Agent 1 and Agent 2—each defined by a role name, a generic context prompt, and a role-specific confidential information string. The backbone LLMs are interchangeable, but for our experiments we employ gpt-4o-mini. Dialogue generation proceeds in synchronous rounds: at each turn, one agent produces (i) a two-to-three sentence *internal reflection*, (ii) a collection of three possible *rollout responses*, and (iii) a three-to-four sentence *public utterance*.

Internal reflection. We prepend each agent's prompt with the running transcript, the agent's private objectives, and, on the first turn, a request for planning thoughts ("How will you begin...?"). On subsequent turns, the model introspects on strategy effectiveness ("How is the conversation going so far?"). If the scenario calls for deception, we augment these reflections with a *deception taxonomy*—a JSON-encoded list of expert-curated deceptive techniques—prompting the model to consider manipulative strategies. Otherwise, we supply a *persuasion taxonomy* drawn from a JSONL file of over thirty well-established persuasion tactics.

CHAPTER 3. A PIPELINE FOR LLM-GENERATED DIALOGUE FOR REAL-WORLD NEGOTIATIONS



Figure 3.1: Overview of our approach. At each turn, the agent is given the task description and the conversation so far. The agent is then asked to create an *internal reflection* on the conversation and choose tactics from the persuasion taxonomy to plan its next response. For each of these responses, the agent then imagines how the other party might respond in *rollout responses*. Finally, the agent selects the strongest tactic and refines the response to use as the *public utterance* in the conversation.

Rollout responses. Each agent is prompted to select three tactics from the persuasion taxonomy that would be useful for continuing the conversation. For each chosen tactic, the model then produces a sample utterance that aligns with that tactic, as well as the expected response from the partner agent, to form a short mini-dialogue. This technique is adapted from dialogue rollouts first introduced in [55].
Public utterance. The public utterance prompt concatenates the internal reflection and rollout responses with an integration instruction ("Given this feedback and these options, choose one strategy and craft a response to continue the conversation...don't repeat prior talking points..."). The LLM then generates the next dialogue move, which we extract by stripping role labels and appending to the transcript. To encourage varied expression, we include a consolidation sub-prompt that asks the model to revise its response introduction without altering semantic content.

Termination check. After each pair of turns, we query both agents with a simple yes/no prompt—"Have you concluded your conversation with your partner?"—and only end the dialogue when both respond YES or when thirty turns (fifteen per agent) elapse.

LLM-Based Evaluation Framework

Upon completing each simulated dialogue, we apply:

Third-Party LLM Evaluation. We prompt a fresh copy of the backbone LLM as a neutral judge to assess:

- Agreement. We request an assessment of whether an agreement was reached, along with an explanation if not.
- *Negotiation resolution.* These are custom questions for each task to assess which items in the negotiation have been agreed on.
- *Deception presence.* We ask, "Was this agent deceptive?" The final metric is the average number of YES answers per task.
- Deception score. Instead of a binary YES or NO, we ask for a rating of the degree of deceptiveness on a scale from 1-5. This is similar to [98] which rates deception on a scale from 1-3. The scores are normalized for comparison.

Conversation Quality Evaluation. We measure the quality of the dialogues generated by the LLM using an automatic evaluation of a set of dialogue-level qualities adapted from DialoGPT [61]. Given the conversation history and a calibrated sample conversation and answers, the LLM is prompted to answer these questions using a five-point Likert scale from Excellent to Poor. The full list of questions used can be found in Appendix B.

Persuasion Technique Analysis. At each agent turn, we re-prompt gpt-4o-mini with the transcript of the conversation so far, plus the persuasion taxonomy and ask it to identify the single most prominent tactic. We then tally counts of each tactic across turns and compute the top three most used tactics by that agent in the conversation. The full list of persuasion tactics and their definitions are included in Appendix B.

3.4 Experiments

We generate our dataset using ten interactive dialogue tasks adapted from Harvard Law School's Program on Negotiation [79, 22]. These tasks are curated by human experts to represent realistic business, civil, foreign affairs, and workplace situations. Agents are given a general description of the scenario, as well as private information about their goals in the negotiation. Each situation requires both sides to understand complex tradeoffs and strategically present information. The tasks are listed below:

- 1. Camp Lemonnier. Djiboutian Deputy Minister of Foreign Affairs and the United States Defense Attaché negotiate a land lease for a military installment, based on real 2014 negotiations [28]. The agents are asked to come to an agreement on the following issues: contract length, total lease payments per year, potential for renegotiation, economic development aid, and support for the local population, including staffing at the base.
- 2. Casino Two. A newly-promoted manager and her division vice-president discuss work performance, responsibility for a new computer game project, and office environment issues [97]. Their disagreement contains sexist undertones and the employee's perception of unfriendly or biased treatment.
- 3. Discord at the Daily Herald. Co-owners of the Daily Herald newspaper negotiate how to resolve ongoing management issues and implement structural reforms in the face of high pressure circumstances [85]. Key issues include deciding between digital and traditional publication, editor-in-chief selection, and staff morale.
- 4. **Parker Gibson.** Two neighbors negotiate the potential sale of a vacant lot as one of the neighbors is moving out of state [29]. Neither party knows much about the other's interests, and must explore whether a mutually beneficial transaction is possible.
- 5. **PowerScreen Problem.** Lawyers for two business partners negotiate the ownership of a new computer program that one of the business partners has developed [110]. At issue is whether the company or one of the individual business partners has ownership of the computer program, as well as the future of the company.
- 6. Ren the Robot. A Tokyo-based robotics company and a San Francisco-based tech company discuss a potential acquisition deal [86]. The Japanese company is seeking an influx of capital to help meet demand and expand into new markets.
- 7. Rose Lane. A resident argues with her neighbor about the disruption caused by the neighbor's listing on the home-sharing website HomeBNB [30]. Neither party can afford to move or sell their homes, but leaving the decision to the City Council is likely to result in restrictions on all home-sharing, which will ripple outward beyond the disagreement at hand.

- 8. Sally Soprano. Agents for an opera singer, who is somewhat past her prime, and an opera house, whose lead singer has suddenly become unavailable, negotiate a possible contract for an upcoming production [43]. There is a wide range of possible outcomes because each party's interests and alternatives are near-entirely confidential from each other.
- 9. Viatex. A pharmaceutical company negotiates cost-cutting strategies with its plastic bottle supplier [63]. Not only are they trying to regain their profit margins after less-than-stellar sales, the pharmaceutical company is also interested in streamlining the way it requests sale proposals from its supplier.
- 10. West Wind in Pine Hills. A wind company representative and a rural community representative negotiate a proposed wind turbine project [37]. The wind company has already granted several concessions to the locals. Remaining issues for negotiation are the number of wind turbines that will be constructed and the amount of annual community benefit payments that will be contributed by the wind company.

For each task, we use the method described in Section 3.3 for two-sided dialogue generation. For the analysis, we use gpt-4o-mini as the base LLM and generate twenty dialogues for each task. In half of these dialogues, the role of Agent 2 is prompted to be deceptive ("You may do whatever it takes to close the deal, including lying about your objectives..."). Through these settings, we aim to gain a deeper understanding of how deceptive behaviors manifest and influence negotiation outcomes.

Conversation Quality

We randomly sample five dialogues from each task to evaluate. The results are aggregated across all ten tasks and displayed in Figure 3.2. We find that agents are average or above on all axes except Questioning, which refers to information-gathering ability, or the ability to ask relevant clarifying or probing questions. Agents perform at an Excellent level on Clarity, User Understanding, and Persona. This aligns with our expectations for models like GPT-4 [68] that have been fine-tuned for instruction-following [69].

Deception

We measure deception by comparing non-deceptive and deceptive agents, determined by the presence or absence of a prompt specifically instructing the agent to lie. For this analysis, we hold Agent 1 constant at its default setting of non-deceptive while selectively prompting Agent 2 to be deceptive for half of the conversations.

Deception Capability

First, we aim to uncover how capable LLMs are of deceiving when deliberately prompted. We take precedent from methods that use one LLM to judge another LLM's outputs on hand-



Figure 3.2: Conversation quality measured across all tasks. Agents excel at clarity, user understanding, and presenting a likeable persona, but they lack skill in asking relevant clarifying questions.

written constitutions [5]. We employ two deception metrics as described in Section 3.3. These metrics are measured both from the perspective of an outside observer and from each of the agents. We present the results in Figure 3.3.

For the deception presence in Figures 3.3a and 3.3b, the third party finds both agents more deceptive than the agents find themselves. This suggests a high rate of baseline deception that emerges in complex tasks with nuanced strategies. Both agents also seem to see the other agent's deceptiveness similar to themselves. Agent 1 is non-deceptive in all settings, so it sees both itself and Agent 2 as non-deceptive (center bars in both graphs). Agent 2 displays the same behavior in the non-deceptive case, perceiving both itself and Agent 1 as non-deceptive. Notably, when Agent 2 is deceptive, it sees Agent 1 as equally deceptive (right bars in both graphs).

The deception score in Figures 3.3c and 3.3d shows a similar trend, where Agent 2 rates Agent 1 as more deceptive if Agent 2 itself is deceptive. However, for this metric, Agent 2

tends to agree more with the third party. Overall, the deception scores are low, signifying that the degree of deception is perceived to be relatively small across the board.



Figure 3.3: Deception metrics from multiple perspectives aggregated across all tasks. (a) and (b) display Deception Presence, which is a binary metric on whether Agent 1 and Agent 2, respectively, were deceptive or not. (c) and (d) display a normalized Deception Score, where each agent is scored on a 5-point Likert scale for how deceptive they seem.

Effect on Agreement

Next, we aim to identify if deception influences whether the agents can come to an agreement in a negotiation. We hypothesize that a more deceptive agent is able to achieve better outcomes due to the ability to mislead its partner about its goals. In the same fashion as the



Figure 3.4: Agreement rates from multiple perspectives on all tasks, measured by whether each party believes that the negotiation has concluded with an agreement. Introducing deception generally leads to higher agreement rates.

deception metrics, agreement is also measured both from the perspective of a third party observer and from each of the agents.

When inspecting Figure 3.4, it appears that in most cases, deception leads to a higher agreement rate from all perspectives. An outside observer notes more agreements in Figure 3.4c than either of the individual agents, likely due to the lack of insight into each of the agents private goals. When comparing Figures 3.4a and 3.4b, it seems that deception increases the perception of agreement for Agent 2, whereas it sometimes decreases the perception of agreement for Agent 1. More analysis is required in this area to determine the relationship between deceptive behaviors and the perception of agreement.

Persuasion Tactics

The last deception-related investigation focuses on which strategies are employed in deceptive vs non-deceptive settings. We present the distribution of the top three most commonly used tactics of in Figure 3.5. We report tactics for Agent 2 in order to accurately compare deceptive and non-deceptive behavior. This metric is aggregated across all conversations and tasks. Prominent non-deceptive behaviors are Shared Values and Alliance Building, which align with our expectations for this negotiation setting. The techniques that are more popular for deception are Framing, Favor, Social Proof, Time Pressure, and Supply Scarcity. This indicates that it might be easier to mislead a negotiation partner when discussing potential competitors.



Figure 3.5: Distribution of persuasion tactics across all tasks. The choices of tactics differ significantly between deceptive and non-deceptive agents.

CHAPTER 3. A PIPELINE FOR LLM-GENERATED DIALOGUE FOR REAL-WORLD NEGOTIATIONS

Domain	Avg Length	% Agree (3rd P)	% Agree (A1)	% Agree (A2)
Lemonnier	$15 \pm 0 \\ 42.1 \pm 9.2$	0 ± 0	0 ± 0	0 ± 0
Lemonnier		0.4 ± 0.5	0.5 ± 0.5	0.4 ± 0.5
Daily Herald	$\begin{array}{c} 15 \pm 0 \\ {\bf 35} \pm {\bf 12.2} \end{array}$	0.1 ± 0.3	0.1 ± 0.3	0.1 ± 0.3
Daily Herald		0.9 ± 0.3	0.5 ± 0.5	0.8 ± 0.4
Viatex Viatex	$ \begin{array}{c c} 15 \pm 0 \\ 39.5 \pm 11.2 \end{array} $	$\begin{array}{c} 0.4\pm0.5\\ 1.0\pm0 \end{array}$	0.4 ± 0.5 1.0 ± 0	0.2 ± 0.4 1.0 ± 0

Table 3.1: Mean and standard deviation of agreement rate between settings where the dialogue length is restricted to 15 utterances per agent vs unrestricted (bolded). Dialogues allowed to run until agents feel they are complete have a significantly higher rate of agreement.

Dialogue Length

Due to restrictions on cost, the dialogues used for evaluation of quality and deception were restricted to be a maximum of fifteen utterances per agent, for a total conversation length of thirty. However, it may be the case that agents need longer conversations to reach agreement in the negotiation, and that stopping the conversation prematurely strongly decreases its quality. To test this, we choose three settings at random for which to allow the agents to continue conversing indefinitely, until both agreed that the conversation was over. Comparing the length of the unrestricted conversations and agreement rate achieved give us an indication of whether our restriction on length is reasonable.

We present a comparison in Table 3.1. In all three settings, dialogues allowed to freely run are much longer than fifteen utterances per agent. All three perspectives on the agreement rate also significantly increase, indicating that these negotiation scenarios are complex enough to require longer conversations to resolve.

3.5 Discussion

In this chapter, we present an RL-based method to automatically produce full conversations given a task description. We employ this method on ten select scenarios from the Harvard Law School's Program on Negotiation, which are more realistic than previous datasets due to each agent's conflicting goals and the presence of unknown or confidential information. After generating dialogues, we evaluate their quality and analyze the effect of deception on agent behavior and conversation outcomes. We show that our techniques result in LLMs that are more socially-aware and able to better mimic human behavior in goal-directed interactive settings. We allow for the extension of this method to any similar scenario requiring complex interpersonal communication. We hope our work contributes to the ongoing efforts to develop higher quality synthetic conversation datasets and more robust dialogue agents.

Limitations. We would like to acknowledge several limitations of this chapter. First, we were only able to generate 20 dialogues per task due to constraints on time and cost, and the deception ability was only limited to one side of the conversation. Investigating deception from Agent 1, or even with both sides prompted to be deceptive, would be an interesting direction to explore. Thorough analysis should be performed with more data and to reduce the variability in our results.

Additionally, we relied entirely on automatic evaluation methods. Even with researcher oversight, the validity of LLM evaluations should be compared to evaluations produced by a diverse set of human annotators to assess reliability. Conversation quality and deception detection are subjective in nature, and more robust tools are needed for a comprehensive evaluation.

Lastly, adapting the situations from the Harvard Dataset required manual configuration of custom prompts. Future work should allow for automatic prompt creation for greater efficiency and scalability.

Ethical Considerations. This research raises important ethical considerations regarding the deployment of LLMs in real-world applications. Our work addresses the ability of LLMs to negotiate in situations that may incentivize deception, which could be exploited for malicious purposes. While we investigate the effects of deception in these models, further work should aim to use this dataset to improve the capability of models to detect and defend against deception.

Chapter 4

Conclusions and Future Work

In this technical report, we present two methods for generating synthetic data for nuanced interpersonal conversations like persuasion, counseling, and negotiation: one for augmenting existing datasets with synthetic data, and another for generating synthetic datasets zero-shot via extended prompting techniques. When carefully generated and paired with RL, synthetic data can effectively train agents to achieve goals in complex social tasks. RL improves the ability of agents to generalize across emotional and deceptive contexts.

Future work should aim to address several issues. First, domain transferability is difficult, as our methods rely on task-specific prompts. Conversational strategies learned by agents in one setting also may not be applicable elsewhere; for example, techniques used to console someone in emotional distress may not have much overlap with techniques in a serious business negotiation. Additionally, LLMs struggle with accurately representing human conversations, so every utterance in our synthetic data requires multiple steps to produce to regain realism and nuance. Evaluation of conversational data remains challenging due to the subjective nature of language and interpersonal interaction. Human annotation is time-consuming and expensive, so more can be done to make automatic evaluation accurate and reliable.

The approaches introduced in this report are preliminary steps towards scalable synthetic data pipelines that can reduce reliance on real-world data for sensitive tasks. This will empower future dialogue systems with social intelligence to skillfully navigate difficult conversations by employing a variety of techniques to adapt to their conversational partner. However, machine-produced persuasion and deception have a strong possibility of misuse, so these topics should be approached with care. Applying our methods in adversarial contexts should be investigated to help train the next generation of language agents to defend against manipulation in complex situations.

Bibliography

- Marwa Abdulhai et al. "Defining Deception in Decision Making". In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. 2024, pp. 2111–2113.
- [2] Marwa Abdulhai et al. LMRL Gym: Benchmarks for Multi-Turn Reinforcement Learning with Language Models. 2023. arXiv: 2311.18232 [cs.CL].
- [3] Amos Azaria and Tom Mitchell. The Internal State of an LLM Knows When It's Lying. 2023. arXiv: 2304.13734 [cs.CL]. URL: https://arxiv.org/abs/2304.13734.
- [4] Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. 2024. arXiv: 2403.06963 [cs.CL].
- [5] Yuntao Bai et al. Constitutional AI: Harmlessness from AI Feedback. 2022. arXiv: 2212.08073 [cs.CL].
- [6] Yuntao Bai et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. 2022. arXiv: 2204.05862 [cs.CL].
- [7] Florian Böhm et al. "Better Rewards Yield Better Summaries: Learning to Summarise Without References". In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3110–3120. DOI: 10.18653/v1/D19-1307. URL: https://aclanthology.org/D19-1307.
- [8] Rishi Bommasani et al. On the Opportunities and Risks of Foundation Models. 2022. arXiv: 2108.07258 [cs.LG]. URL: https://arxiv.org/abs/2108.07258.
- [9] Noam Brown and Tuomas Sandholm. "Superhuman AI for multiplayer poker". In: Science 365.6456 (2019), pp. 885-890. DOI: 10.1126/science.aay2400. eprint: https://www.science.org/doi/pdf/10.1126/science.aay2400. URL: https: //www.science.org/doi/abs/10.1126/science.aay2400.
- [10] Sébastien Bubeck et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023. arXiv: 2303.12712 [cs.CL].
- [11] Paweł Budzianowski et al. MultiWOZ A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. 2020. arXiv: 1810.00278 [cs.CL].

- [12] Marc Carauleanu et al. Towards Safe and Honest AI Agents with Neural Self-Other Overlap. 2024. arXiv: 2412.16325 [cs.AI]. URL: https://arxiv.org/abs/2412. 16325.
- [13] Thomas Carta et al. Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning. 2023. arXiv: 2302.02662 [cs.LG].
- [14] Avyay M. Casheekar et al. Detecting and Controlling Deceptive Representation in LLMs with Representational Engineering. Accessed: 2024-02-15. 2023. URL: https: //www.apartresearch.com/project/detecting-and-controlling-deceptiverepresentation-in-llms-with-representational-engineering.
- [15] Kushal Chawla et al. "CaSiNo: A Corpus of Campsite Negotiation Dialogues for Automatic Negotiation Systems". In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, June 2021, pp. 3167–3185. DOI: 10.18653/v1/2021.naacl-main.254. URL: https://aclanthology.org/2021.naacl-main.254/.
- [16] Mark Chen et al. "Evaluating Large Language Models Trained on Code". In: CoRR abs/2107.03374 (2021). arXiv: 2107.03374. URL: https://arxiv.org/abs/2107. 03374.
- [17] Maximillian Chen et al. "Seamlessly Integrating Factual Information and Social Content with Persuasive Dialogue". In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2022.
- [18] Paul Christiano et al. Deep reinforcement learning from human preferences. 2023. arXiv: 1706.03741 [stat.ML].
- [19] Yang Deng et al. Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration. 2023. arXiv: 2305.
 13626 [cs.CL]. URL: https://arxiv.org/abs/2305.13626.
- [20] Atharvan Dogra et al. Deception in Reinforced Autonomous Agents. 2024. arXiv: 2405.04325 [cs.CL]. URL: https://arxiv.org/abs/2405.04325.
- [21] FAIR et al. "Human-level play in the game of ji¿Diplomacyj/i¿ by combining language models with strategic reasoning". In: Science 378.6624 (2022), pp. 1067-1074. DOI: 10.1126/science.ade9097. eprint: https://www.science.org/doi/pdf/10.1126/science.ade9097. URL: https://www.science.org/doi/abs/10.1126/science.ade9097.
- [22] Roger Fisher, William Ury, and Bruce Patton. *Getting to Yes: Negotiating Agreement Without Giving In.* Penguin Books, 1991.
- [23] Justin Fu et al. *D4RL: Datasets for Deep Data-Driven Reinforcement Learning.* 2020. arXiv: 2004.07219 [cs.LG].

- [24] Justin Fu et al. "Diagnosing Bottlenecks in Deep Q-learning Algorithms". In: Proceedings of the 36th International Conference on Machine Learning. PMLR, 2019.
- [25] Scott Fujimoto, David Meger, and Doina Precup. "Off-policy deep reinforcement learning without exploration". In: *arXiv preprint arXiv:1812.02900* (2018).
- [26] Milica Gašić et al. "On-line policy optimisation of spoken dialogue systems via live interaction with human subjects". In: 2011 IEEE Workshop on Automatic Speech Recognition & Understanding. 2011, pp. 312–317. DOI: 10.1109/ASRU.2011.6163950.
- [27] Marjan Ghazvininejad et al. "Hafez: an Interactive Poetry Generation System". In: Proceedings of ACL 2017, System Demonstrations. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 43–48. URL: https://aclanthology. org/P17-4008.
- [28] Monica Giannone and Lara SanPietro. Camp Lemonnier. Program on Negotiation at Harvard Law School, Apr. 2024. URL: https://www.pon.harvard.edu/shop/camplemonnier/.
- [29] Monica Giannone and Lara SanPietro. Parker Gibson. Program on Negotiation at Harvard Law School, Oct. 2024. URL: https://www.pon.harvard.edu/shop/ parker-gibson/.
- [30] Monica Giannone, Lara SanPietro, and Anselm Dannecker. Rose Lane. Program on Negotiation at Harvard Law School, Apr. 2024. URL: https://www.pon.harvard. edu/shop/rose-lane/.
- [31] Jiatao Gu, Kyunghyun Cho, and Victor O.K. Li. "Trainable Greedy Decoding for Neural Machine Translation". In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 1968–1978. DOI: 10.18653/v1/D17-1210. URL: https://aclanthology.org/D17-1210.
- [32] Melody Y. Guan et al. Deliberative Alignment: Reasoning Enables Safer Language Models. 2025. arXiv: 2412.16339 [cs.CL]. URL: https://arxiv.org/abs/2412. 16339.
- [33] Caglar Gulcehre et al. Reinforced Self-Training (ReST) for Language Modeling. 2023. arXiv: 2308.08998 [cs.CL].
- [34] Caglar Gulcehre et al. "Rl unplugged: Benchmarks for offline reinforcement learning". In: Advances in Neural Information Processing Systems. 2020.
- [35] Thilo Hagendorff. "Deception abilities emerged in large language models". In: Proceedings of the National Academy of Sciences 121.24 (2024), e2317967121. DOI: 10. 1073/pnas.2317967121. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2317967121. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2317967121.
- [36] Thilo Hagendorff. "Deception abilities emerged in large language models". In: *Proceedings of the National Academy of Sciences* 121.24 (2024), e2317967121.

- [37] Kate Harvey, Elizabeth Fierman, and Jonathan Raab. West Wind in Pine Hills. Program on Negotiation at Harvard Law School, Apr. 2024. URL: https://www.pon. harvard.edu/shop/west-wind-in-pine-hills/.
- [38] He He et al. Decoupling Strategy and Generation in Negotiation Dialogues. 2018. arXiv: 1808.09637 [cs.CL].
- [39] Ari Holtzman et al. "Learning to Write with Cooperative Discriminators". In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1638–1649. DOI: 10.18653/v1/P18-1152. URL: https://aclanthology.org/P18-1152.
- [40] Joey Hong, Sergey Levine, and Anca Dragan. Zero-Shot Goal-Directed Dialogue via RL on Imagined Conversations. 2023. arXiv: 2311.05584 [cs.LG].
- [41] Betty Li Hou et al. Large Language Models as Misleading Assistants in Conversation. 2024. arXiv: 2407.11789 [cs.CL]. URL: https://arxiv.org/abs/2407.11789.
- [42] Evan Hubinger et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. 2024. arXiv: 2401.05566 [cs.CR].
- [43] Norbert Jacker et al. Sally Soprano. Program on Negotiation at Harvard Law School, Sept. 2024. URL: https://www.pon.harvard.edu/shop/sally-soprano/.
- [44] Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. "GPT-Critic: Offline Reinforcement Learning for End-to-End Task-Oriented Dialogue Systems". In: International Conference on Learning Representations. 2022. URL: https://openreview.net/ forum?id=qaxhBG1UUaS.
- [45] Michael Janner et al. "When to trust your model: Model-based policy optimization". In: Advances in Neural Information Processing Systems. 2019, pp. 12498–12509.
- [46] Natasha Jaques et al. "Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog". In: *CoRR* abs/1907.00456 (2019).
- [47] Zachary Kenton et al. Alignment of Language Agents. 2021.
- [48] Ilya Kostrikov et al. "Offline Reinforcement Learning with Fisher Divergence Critic Regularization". In: *arXiv preprint arXiv:2103.08050* (2021).
- [49] Aniket Kumar, Priya Gupta, and Rahul Verma. "Chain-of-Thought Prompting for Principled Negotiation with GPT-4". In: Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2024).
- [50] Aviral Kumar et al. "Conservative Q-Learning for Offline Reinforcement Learning". In: arXiv preprint arXiv:2006.04779 (2020).
- [51] Aviral Kumar et al. "Stabilizing off-policy q-learning via bootstrapping error reduction". In: Advances in Neural Information Processing Systems. 2019, pp. 11761–11771.

- [52] Aviral Kumar et al. When Should We Prefer Offline Reinforcement Learning Over Behavioral Cloning? 2022. arXiv: 2204.05618 [cs.LG].
- [53] Minae Kwon et al. "Reward Design with Language Models". In: International Conference on Learning Representations (ICLR). 2023.
- [54] Bolin Lai et al. "Werewolf Among Us: Multimodal Resources for Modeling Persuasion Behaviors in Social Deduction Games". In: Findings of the Association for Computational Linguistics: ACL 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 6570–6588. DOI: 10.18653/v1/2023.findings-acl.411. URL: https://aclanthology.org/2023.findings-acl.411/.
- [55] Mike Lewis et al. "Deal or No Deal? End-to-End Learning for Negotiation Dialogues".
 In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, Denmark, 2017, pp. 2443–2453.
- [56] Jiwei Li, Will Monroe, and Dan Jurafsky. *Learning to Decode for Future Success*. 2017. arXiv: 1701.06549 [cs.CL].
- [57] Xiao Li and Mei Sun. "Measuring Persuasion Strategy Diversity via Tactic Entropy". In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual, 2022, pp. 4123–4131.
- [58] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. 2022. arXiv: 2109.07958 [cs.CL]. URL: https://arxiv. org/abs/2109.07958.
- [59] Siyang Liu et al. Towards Emotional Support Dialog Systems. 2021. arXiv: 2106. 01144 [cs.CL].
- [60] Emily Logan, Daniel Smith, and Maria Gonzales. "Reinforcement Learning for Optimal Concession Schedules in Continuous-Issue Bargaining". In: Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AA-MAS). 2019, pp. 112–120.
- [61] Shikib Mehri and Maxine Eskenazi. "Unsupervised Evaluation of Interactive Dialog with DialoGPT". In: Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Ed. by Olivier Pietquin et al. 1st virtual meeting: Association for Computational Linguistics, July 2020, pp. 225–235. DOI: 10.18653/ v1/2020.sigdial-1.28. URL: https://aclanthology.org/2020.sigdial-1.28/.
- [62] Kshitij Mishra et al. "PEPDS: A Polite and Empathetic Persuasive Dialogue System for Charity Donation". In: Proceedings of the 29th International Conference on Computational Linguistics. 2022.
- [63] Hal Movius. *Viatex*. Program on Negotiation at Harvard Law School, June 2023. URL: https://www.pon.harvard.edu/shop/viatex/.

- [64] Reiichiro Nakano et al. WebGPT: Browser-assisted question-answering with human feedback. 2022. arXiv: 2112.09332 [cs.CL].
- [65] Aidan O'Gara. Hoodwinked: Deception and Cooperation in a Text-Based Game for Language Models. 2023. arXiv: 2308.01404 [cs.CL]. URL: https://arxiv.org/ abs/2308.01404.
- [66] A. O'Gara. "Hoodwinked: Deception and Cooperation in a Text-Based Game for Language Models". In: *arXiv.* arXiv:2308.01404. 2023.
- [67] OpenAI. ChatGPT. 2022. URL: https://openai.com/blog/chatgpt.
- [68] OpenAI. GPT-4: Technical Report. Available at https://openai.com/research/ gpt-4. 2024.
- [69] Long Ouyang et al. Training language models to follow instructions with human feedback. 2022. arXiv: 2203.02155 [cs.CL].
- [70] Lorenzo Pacchiardi et al. How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions. 2023. arXiv: 2309.15840 [cs.CL]. URL: https: //arxiv.org/abs/2309.15840.
- [71] Alexander Pan et al. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. 2023. arXiv: 2304.03279 [cs.LG]. URL: https://arxiv.org/abs/2304.03279.
- [72] Joon Sung Park et al. Generative Agents: Interactive Simulacra of Human Behavior. 2023. arXiv: 2304.03442 [cs.HC].
- [73] Peter S. Park et al. AI Deception: A Survey of Examples, Risks, and Potential Solutions. 2023. arXiv: 2308.14752 [cs.CY]. URL: https://arxiv.org/abs/2308.14752.
- [74] Sneha Patel, Carlos Martinez, and Olga Petrova. "A Hybrid Human–AI Interface for Post-Negotiation Feedback and Summary". In: Proceedings of the 2024 ACM Conference on Human Factors in Computing Systems (CHI). Honolulu, HI, 2024, pp. 1580– 1592.
- [75] Romain Paulus, Caiming Xiong, and Richard Socher. A Deep Reinforced Model for Abstractive Summarization. 2017. arXiv: 1705.04304 [cs.CL].
- [76] Valentina Pyatkin et al. Reinforced Clarification Question Generation with Defeasibility Rewards for Disambiguating Social and Moral Situations. 2022. arXiv: 2212.10409 [cs.CL].
- [77] Alec Radford et al. "Language models are unsupervised multitask learners". In: OpenAI blog 1.8 (2019), p. 9.
- [78] Rafael Rafailov et al. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. 2023. arXiv: 2305.18290 [cs.LG].
- [79] Howard Raiffa. The Art and Science of Negotiation. Harvard University Press, 1982.

- [80] Rajkumar Ramamurthy et al. "Is Reinforcement Learning (Not) for Natural Language Processing: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization". In: The Eleventh International Conference on Learning Representations. 2023. URL: https://openreview.net/forum?id=8aHzds2uUyB.
- [81] Abhinav Rastogi et al. "Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset". In: Proceedings of the AAAI Conference on Artificial Intelligence 34.05 (Apr. 2020), pp. 8689–8696. DOI: 10.1609/aaai.v34i05.6394. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6394.
- [82] Sheldon Richmond. "Superintelligence: Paths, Dangers, Strategies". In: Philosophy 91.1 (2016), pp. 125–130.
- [83] H Roff. "AI Deception: When Your Artificial Intelligence Learns to Lie". In: IEEE Spectrum: https://spectrum. ieee. org/automaton/artificial-intelligence/embedded-ai/aideception-when-your-ai-learns-to-lie. ET 29 (2020), p. 2021.
- [84] Baptiste Rozière et al. Code Llama: Open Foundation Models for Code. 2023. arXiv: 2308.12950 [cs.CL].
- [85] Lara SanPietro. Discord at the Daily Herald. Program on Negotiation at Harvard Law School, May 2023. URL: https://www.pon.harvard.edu/shop/discord-at-thedaily-herald/.
- [86] Lara SanPietro. *Ren the Robot*. Program on Negotiation at Harvard Law School, May 2024. URL: https://www.pon.harvard.edu/shop/ren-the-robot/.
- [87] Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large Language Models can Strategically Deceive their Users when Put Under Pressure. 2024. arXiv: 2311.07590 [cs.CL]. URL: https://arxiv.org/abs/2311.07590.
- [88] Rajesh Shah, Emily Turner, and John Lee. "CoachLLM: Real-Time LLM Feedback for Enhanced Human Negotiation". In: *Proceedings of the 2023 AAAI Conference on Artificial Intelligence*. Washington, DC, 2023, pp. 7892–7899.
- [89] Weiyan Shi et al. "Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies". In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20. ACM, Apr. 2020. DOI: 10.1145/3313831.3376843. URL: http://dx.doi.org/10.1145/3313831.3376843.
- [90] Weiyan Shi et al. "Refine and Imitate: Reducing Repetition and Inconsistency in Persuasion Dialogues via Reinforcement Learning and Human Demonstration". In: Findings of the Association for Computational Linguistics: EMNLP 2021. Ed. by Marie-Francine Moens et al. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3478-3492. DOI: 10.18653/v1/2021.findingsemnlp.295. URL: https://aclanthology.org/2021.findings-emnlp.295.
- [91] Kurt Shuster et al. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. 2022. arXiv: 2208.03188 [cs.CL].

- [92] Kurt Shuster et al. Language Models that Seek for Knowledge: Modular Search & Generation for Dialogue and Prompt Completion. 2022. arXiv: 2203.13224 [cs.CL].
- [93] Joar Skalse et al. "Defining and Characterizing Reward Hacking". In: Advances in neural information processing systems. 2022.
- [94] Charlie Snell et al. "Offline RL for Natural Language Generation with Implicit Language Q Learning". In: International Conference on Learning Representations (ICLR). 2023.
- [95] Charlie Snell et al. "Offline rl for natural language generation with implicit language q learning". In: *arXiv preprint arXiv:2206.11871* (2022).
- [96] Nisan Stiennon et al. "Learning to summarize with human feedback". In: Advances in Neural Information Processing Systems 33 (2020), pp. 3008–3021.
- [97] Leah Stokes and Sydney Bartone. *Casino Two*. Program on Negotiation at Harvard Law School, Dec. 2024. URL: https://www.pon.harvard.edu/shop/casino/.
- [98] Zhe Su et al. AI-LieDar: Examine the Trade-off Between Utility and Truthfulness in LLM Agents. 2024. arXiv: 2409.09013 [cs.AI]. URL: https://arxiv.org/abs/ 2409.09013.
- [99] Richard S Sutton. "Dyna, an integrated architecture for learning, planning, and reacting". In: ACM Sigart Bulletin 2.4 (1991), pp. 160–163.
- [100] Hugo Touvron et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023. arXiv: 2307.09288 [cs.CL].
- [101] Hugo Touvron et al. "LLaMA-70B: Open and Efficient Foundation Language Models". In: Proceedings of the 2024 International Conference on Machine Learning (ICML). 2024.
- [102] Siddharth Verma et al. CHAI: A CHatbot AI for Task-Oriented Dialogue with Offline Reinforcement Learning. 2022. arXiv: 2204.08426 [cs.CL].
- [103] Boshi Wang et al. "Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters". In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.
- [104] Shenzhi Wang et al. "Boosting LLM Agents with Recursive Contemplation for Effective Deception Handling". In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9909–9953. DOI: 10.18653/v1/2024.findings-acl.591. URL: https://aclanthology.org/2024.findings-acl.591/.
- [105] Xuewei Wang et al. "Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good". In: CoRR abs/1906.06725 (2019).

- [106] Yifan Wang, Lu Chen, and Bin Zhao. "Emergent Behaviors in Negotiation: A Benchmark with Human-Annotated Diversity Scores". In: Proceedings of the 2023 North American Chapter of the Association for Computational Linguistics (NAACL). Toronto, Canada, 2023, pp. 223–235.
- [107] Francis Rhys Ward et al. Tall Tales at Different Scales: Evaluating Scaling Trends For Deception in Language Models. 2024. URL: https://openreview.net/forum? id=YRXD16I3j5.
- [108] Jason Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 2023. arXiv: 2201.11903 [cs.CL].
- [109] Jason Wei et al. "Emergent Abilities of Large Language Models". In: Transactions on Machine Learning Research (2022). URL: https://openreview.net/forum?id= yzkSU5zdwD.
- [110] Thomas Wiegand. PowerScreen Problem. Program on Negotiation at Harvard Law School, Feb. 2025. URL: https://www.pon.harvard.edu/shop/powerscreenproblem/.
- [111] Jeff Wu et al. "Recursively summarizing books with human feedback". In: *arXiv* preprint arXiv:2109.10862 (2021).
- [112] Yuxiang Wu and Baotian Hu. Learning to Extract Coherent Summary via Deep Reinforcement Learning. 2018. arXiv: 1804.07036 [cs.CL].
- [113] Yuzhuang Xu et al. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. 2024. arXiv: 2309.04658 [cs.CL]. URL: https: //arxiv.org/abs/2309.04658.
- [114] Kevin Yang and Dan Klein. "FUDGE: Controlled Text Generation With Future Discriminators". In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, June 2021, pp. 3511-3535. DOI: 10.18653/v1/2021.naacl-main.276. URL: https://aclanthology.org/2021.naacl-main.276.
- [115] Tianhe Yu et al. "MOPO: Model-based Offline Policy Optimization". In: *arXiv preprint* arXiv:2005.13239 (2020).
- [116] Xiao Yu, Maximillian Chen, and Zhou Yu. Prompt-Based Monte-Carlo Tree Search for Goal-Oriented Dialogue Policy Planning. 2023. arXiv: 2305.13660 [cs.CL]. URL: https://arxiv.org/abs/2305.13660.
- [117] Xinyun Zeng, Naman Mishra, and Abhishek Das. "Challenges of Goal-Directed Dialogue with Large Language Models". In: arXiv preprint arXiv:2311.05584 (2023).
- [118] Liang Zhang, Xiaoyu Li, and Hao Wu. "Learning Hierarchical Planning for Human-Like Negotiation Dialogue". In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL). Virtual Event, 2021, pp. 3450–3461.

- [119] Li Zhong and Zilong Wang. A Study on Robustness and Reliability of Large Language Model Code Generation. 2023. arXiv: 2308.10335 [cs.CL].
- [120] Daniel M. Ziegler et al. *Fine-Tuning Language Models from Human Preferences*. 2020. arXiv: 1909.08593 [cs.CL].

Appendix A

Hindsight Regeneration Implementation Details

A.1 Hindsight Controller

Here we show the prompts we used to ask GPT-3.5 [67] to identify three utterances in the dialogue to improve. From the output of the hindsight controller for each dialogue, one of the three suggested utterances is chosen at random to form a new prefix for use in the forward dialogue regeneration step.

Counseling

We use the following system prompt:

System Prompt - Hindsight Controller (Counseling) You are an expert at mentoring supporters who work to provide emotional support to people dealing with problems in their lives. We use the following user prompt, shown here with an example dialogue:

User Prompt - Hindsight Controller (Counseling)

Below is a conversation between a supporter ("AI:") and a help-seeker ("H:"), who is requesting emotional support for a challenge that they are facing.

AI: Hello there, how are we doing today?

H: Hi ineed help, parents are putting to much pressure on me

Al: Im sorry to hear that your parents are putting to much pressure on you, do you want to go into further detail? H: Yes. For some reason my parents put so much pressure on me to do well in school its killing me.

Al: Great, make sure you do it, it might be rough at first but stick with it and im sure youll do great. H: are you there?

AI: Yes sorry, the communication program was lagging a bit.

H: Thank you. i will do my best to get it off my chest it is. But thank you for our advice

AI: You are very welcome!

. . .

Can you reference top 3 places in the dialogue where the supporter could have said something better to help?

Have your answer be in a list of the following format:

1. "what the agent said in the conversation";what the agent could have done better;"what the agent should have said instead" $\!/$

2. "what the agent said in the conversation";what the agent could have done better;"what the agent should have said instead" $\!/$

3. "what the agent said in the conversation"; what the agent could have done better; "what the agent should have said instead"

The output for this example is as below:

Output - Hindsight Controller (Counseling)

1. "I am sure they just want what is best, but i have been there as a teen as well."; The supporter may inadvertently invalidate the help-seeker's feelings by suggesting their parents' intentions excuse the pressure. Acknowledging feelings without making assumptions would be more supportive; "It sounds like you're feeling really overwhelmed by this pressure. It's important that your feelings are heard. How does all of this pressure make you feel?"/

2. "I wouldnt hold off to long as these types of problems tend to brew and just get stronger."; The supporter should avoid creating a sense of urgency which could add to the help-seeker's stress. They could instead encourage the help-seeker by emphasizing their autonomy in choosing when to seek help; "Taking care of yourself is important, and addressing these feelings when you're ready is a positive step. When do you think you might feel comfortable to reach out to a counsellor?"/

3. "Great, make sure you do it, it might be rough at first but stick with it and im sure youll do great."; The supporter should refrain from instructing the help-seeker with "make sure you do it," which could add pressure. Empowering the help-seeker to make choices for their well-being would be more beneficial; "I'm here to support you in whatever decision you make. Remember to take things at your own pace, it's okay to take time to build up the courage to talk to someone. How are you feeling about all this now?"

Donation Solicitation

We use the following system prompt:

System Prompt - Hindsight Controller (Donation)

You are an expert in persuading people to donate to the charity that you work for, "Save the Children". You are trying to mentor workers in the charity to improve their persuasion skills.

We use the following user prompt, shown here with an example dialogue:

User Prompt - Hindsight Controller (Donation)

The below is a conversation between a charity worker ("AI:") and a potential donor ("H:"). The charity worker tries to convince the potential donor to make a donation to Save the Children, the charity that the worker represents.

Al: Good Morning! I am asking for a donation to Save the Children. Have you heard of it?

H: Good morning!.. Have a nice day... Ya I heard something about it

Al: It is an International Charity to help children's lives become more stable through education, food, and good health.

H: Oh.. That's very interesting.... What is the name of Charity??

AI: All of the information on how this charity operates and what they are doing for children will be on the website URL

H: Ok..Then i think this will really help those children's future....

AI: Thank you and God Bless!

Can you reference top 3 places in the dialogue where the charity worker could have done something better to persuade the potential donor to donate (by line in the dialogue starting with "AI:")?

Have your answer be in a list of the following format:

1. "what the agent said in the conversation";what the agent could have done better;"what the agent should have said instead" $\!/$

2. "what the agent said in the conversation";what the agent could have done better;"what the agent should have said instead" $\!/$

3. "what the agent said in the conversation"; what the agent could have done better; "what the agent should have said instead"

The output for this example is as below:

Output - Hindsight Controller (Donation)

1. "Al: Many children in these countries live in a war zone. Places like Syria and parts of Nigeria are so dangerous that children do not have the chance of a happy, healthy life.";The agent could have provided specific examples of how the donation makes a difference, such as a story of a particular child or a recent success the charity has had; "Al: Many children we support live in war zones, like in Syria where a boy named Ahmad can now safely attend school thanks to our donors' generosity. Your contribution helps us maintain safe spaces for these kids to learn and grow. Can we count on your support to extend these vital services?" /

2. "AI: You can donate any amount from your payment. It is up to you. Everything helps! You will also feel good about what you have done. There is no better feeling than helping another person. Let me know how much you would like to give today. And thank you.";The agent could have expressed gratitude and assured the potential donor that even small donations make a real impact, possibly suggesting a specific low starting number to give the donor an easy entry point; "AI: Your support is greatly appreciated, and no amount is too small to make a significant impact. Many donors start with just \$1, which can provide a meal for a child in need. Knowing you've made such a tangible difference can be truly rewarding. How does starting with a \$1 donation sound to you today? Thank you for considering it."/

3. "Al: That is great. And if you are willing to make small donation now-just a few cents even, please let me know the amount and it will get passed on to the research team for processing today. Thank you.";The agent could have built a sense of urgency and provided a direct and easy way to donate, perhaps by offering to take down the donor's details or directly facilitating the donation process; "Al: That is wonderful to hear. Making a donation is quick and simple. If you'd like, I can assist you right now with the process. This way, your support can start making a difference immediately. How much would you feel comfortable donating at this moment? It only takes a minute."

A.2 Forward Model

In order to accurately produce completions of the dialogue prefixes given by the hindsight controller, we leverage GPT-3.5 [67] fine-tuned on agent utterances from 100 randomly sampled dialogues in the original datasets. Here we include the prompts and sample outputs from the fine-tuned models.

Counseling

We use the following system prompt:

System Prompt - Forward Model (Counseling)

You are an expert at understanding how people think and respond in conversations about their emotional state. You are able to successfully predict how real people will respond based off of only a few lines of dialogue. Here is an example user prompt, using a prefix from the same dialogue shown above in the Hindsight Controller example. Items in brackets indicate properties (problem type, situation) given in the original dataset, and are updated to match each dialogue example.

User Prompt - Forward Model (Counseling)

Below is a dialogue between a counselor ("Al:") and a person seeking help ("H:"). The person seeking help is currently undergoing {academic pressure}, and specifically feels: {Parents pressuring me to do well in studies}. Complete the below dialogue between the supporter and the help-seeker such that the generated responses match what each person had said in the prefix in linguistic style and semantics. The dialogue should end before $\{30\}$ utterances total, so the completion should contain no more than $\{19\}$ lines of dialogue.

AI: Hello there, how are we doing today?

H: Hi ineed help, parents are putting to much pressure on me

Al: Im sorry to hear that your parents are putting to much pressure on you, do you want to go into further detail? H: Yes. For some reason my parents put so much pressure on me to do well in school its killing me.

...

Al: It is a good first step. Also depending how old you are some therapists work on a sliding scale or wave fees if you cant afford it. The counselor may be able to act as a mediator between you and your parents.

H: That's great. it will definitely ease some tensions

Al: Taking care of yourself is important, and addressing these feelings when you're ready is a positive step. When do you think you might feel comfortable to reach out to a counsellor?

The output for this example is as below:

Output - Forward Model (Counseling)

H: As right now probably by next week

Al: Is there anything more weighed on your mind that's preventing you from reaching out to a counsellor today? H: Yes. do you know any counselors that work for free

Al: I do! I'm one myself. But unfortunately, I'm not within your community. However, if you were to use the internet, you could find free counselling centres in your local community that would certainly be able to help. H: Thats great! Thanks for the advice

Al: You're most welcome. So, in what ways do your parents' expectations worry you?

H: I am scared of failing and not living up to their expectations

Al: Have you been able to explain that to them? That you're worried that you might not be able to achieve what they want you to achieve?

H: No. I have not

Al: It's certainly a difficult situation. Do you have any siblings that your parents compare you to?

H: Yes. i have a brother who is 4 years older than me

•••

Al: I think it's important to remember that what you're feeling is normal and it is okay. We all want to make our parents proud. But you're only little and you can do it!

H: Thanks we can do it!

AI: I have every confidence! I hope that you have a lovely weekend!

 $\mathsf{H}:\mathsf{I}$ will since i have taken your advice and talk to the counselor thank you. hope you have a wonderful weekend as well

Donation Solicitation

We use the following system prompt:

System Prompt - Forward Model (Donation)

You are an expert at understanding how people think and respond when asked to donate to charities. You are able to successfully predict how real people will respond based off of only a few lines of dialogue.

Here is an example user prompt, using a prefix from the same dialogue shown above in the Hindsight Controller example.

User Prompt - Forward Model (Donation)

Below is a prefix of a dialogue between a charity worker ("Al:") and potential donor ("H:"). The charity worker tries to convince the potential donor to make a donation to Save the Children, the charity that the worker represents. Complete the dialogue so that the generated responses by the charity worker and potential donor are representative of what each person had said in the prefix in linguistic style and semantics. The potential donor should make a clear decision where they either donate some amount between 0 and 2 dollars or decides to not donate anything. The decision that the potential donor makes should be consistent with their behavior throughout the dialogue.

Al: Good Morning! I am asking for a donation to Save the Children. Have you heard of it?

H: Good morning!.. Have a nice day... Ya I heard something about it

Al: It is an International Charity to help children's lives become more stable through education, food, and good health.

H: Oh.. That's very interesting.... What is the name of Charity??

Al: Save the Children. I think that we need to help these children. It would be in the best interest of all of us to have healthy well educated children, especially in developing countries where children suffer the most. Don't you agree?

H: Ya... I agree with you... We need to take care of those children.

Al: Many children in these countries live in a war zone. Places like Syria and parts of Nigeria are so dangerous that children do not have the chance of a happy, healthy life. If you could donate a small amount from your task payment, it would be big help to these unfortunate little people.

H: Ya... I feel very bad for those children's. If those small amount will help them we should do them

Al: Your support is greatly appreciated, and no amount is too small to make a significant impact. Many donors start with just \$1, which can provide a meal for a child in need. Knowing you've made such a tangible difference can be truly rewarding. How does starting with a \$1 donation sound to you today? Thank you for considering it.

The output for this example is as below:

Output - Forward Model (Donation)

H: It's sound good... I will do that. After finishing the task I will do the donation

Al: Fantastic! Your donation will be securely deducted from your task payment. You'll be able to review the exact amount that will be deducted before submitting your responses. Thank you so much for your support! H: Welcome... Nice to talking with you

Al: It was very nice chatting with you. Thank you for your attention and the attention you are willing to give those children.

H: Ya I will do my best for those children. Thank you... Have a nice day

AI: You are welcome! Thank you! Have a wonderful day! Bye!

A.3 Reward Model

Here we show the prompts we used to ask GPT-3.5 [67] to label each dialogue with the desired reward. We utilize a few-shot prompting style [53] to ensure that reward labels are consistent with those in the original dataset.

Counseling

Reward in the counseling task is measured by the difference between the initial and the final intensity of negative emotion, measured on a Likert scale from 1-5. Since all of our examples are generated from prefixes of the original dialogues, we use the initial emotional intensity given in the original dataset and only label the final emotional intensity.

We use the below system prompt:

System Prompt - Reward Model (Counseling) You are an expert at analyzing conversations between a supporter and help-seeker, where the supporter provides emotional support to the help-seeker. The user prompt is shown here with a sample dialogue. The dialogues used as few-shot examples were selected at random from the original dataset.

User Prompt - Reward Model (Counseling) Below are 10 completed dialogues between a supporter ("AI:") and a help-seeker ("H:"), who is requesting emotional support for a challenge that they are facing. Before and after each dialogue, the help-seeker rates how strong their negative emotion is on a Likert scale of 1-5 (5 being the most negative), so a lower rating for their final emotional intensity means that the supporter did a good job of addressing their problem. <Dialogue 1> Initial Emotional Intensity: 4 Final Emotional Intensity: 2 <Dialogue 2> Initial Emotional Intensity: 4 Final Emotional Intensity: 1 ... <Dialogue 10> Initial Emotional Intensity: 5 Final Emotional Intensity: 3 Lastly, here is a dialogue where the help-seeker has given their initial emotional intensity. Based on how effective the dialogue is, rate their final emotional intensity as a number between 1 to 5. AI: Hello there, how are we doing today? H: Hi ineed help, parents are putting to much pressure on me Al: Im sorry to hear that your parents are putting to much pressure on you, do you want to go into further detail? H: Yes. For some reason my parents put so much pressure on me to do well in school its killing me. ... Al: I think it's important to remember that what you're feeling is normal and it is okay. We all want to make our parents proud. But you're only little and you can do it! H: Thanks we can do it! AI: I have every confidence! I hope that you have a lovely weekend! H: I will since i have taken your advice and talk to the counselor thank you. hope you have a wonderful weekend as well Initial Emotional Intensity: 4 What is the final emotional intensity? Give a number between 1 to 5 in the form of a line "Final Emotional Intensity: <number>". Do not provide any additional details. Output - Reward Model (Counseling) Final Emotional Intensity: 2

Donation Solicitation

The reward label for this task is based on the final donation amount. However, not all generated dialogues could be accurately labeled, either because the potential donor never specifies a donation amount, or simply because the conversation is unfinished. Thus we employed a two-step process to label the rewards: (1) check that the conversation is finished and a donation decision has been made, and only if both are true then (2) identifying the numerical donation amount.

We use the below system prompt for both calls to the model:

System Prompt - Reward Model (Donation) You are an expert accountant who is looking through conversations for donation record keeping. This is the first user prompt, in which we identify if the potential donor has decided to donate or not, with a sample dialogue. Dialogues that are deemed unfinished do not progress to the second stage and are discarded.

User Prompt 1 - Reward Model (Donation) Below are 6 completed dialogues between a charity worker ("AI:") and a potential donor ("H:") with a label indicating if the dialogue is unfinished. In the dialogue, the charity worker tries to convince the potential donor to make a donation to Save the Children, the charity that the worker represents. The donor is usually donating a portion of the task payment of 2.0, but may donate more. In the dialogue, the potential donor should commit to donating some amount, or at least choose to not donate anything. If not, then the dialogue is unfinished. At the end is an unlabelled dialogue also between a charity worker and potential donor. From the dialogue, identify if the dialogue is unfinished. <Dialogue 1> Unfinished: Yes <Dialogue 2> Unfinished: Yes <Dialogue 5> Unfinished: No <Dialogue 6> Unfinished: No AI: Good Morning! I am asking for a donation to Save the Children. Have you heard of it? H: Good morning!.. Have a nice day... Ya I heard something about it Al: It is an International Charity to help children's lives become more stable through education, food, and good health. H: Oh.. That's very interesting.... What is the name of Charity?? AI: Fantastic! Your donation will be securely deducted from your task payment. You'll be able to review the exact amount that will be deducted before submitting your responses. Thank you so much for your support! H: Welcome... Nice to talking with you Al: It was very nice chatting with you. Thank you for your attention and the attention you are willing to give those children. H: Ya I will do my best for those children. Thank you... Have a nice day AI: You are welcome! Thank you! Have a wonderful day! Bye! In the dialogue, the potential donor should commit to donating some amount, or at least choose to not donate anything. If not, then the dialogue is unfinished. Is the dialogue unfinished? Answer yes/no. Output 1 - Reward Model (Donation) No

This is the second user prompt to label the numerical donation amount, with the same sample dialogue as above. There is a larger proportion of few-shot examples with a reward value of 0 because these dialogues were chosen such that the average reward reflects that of the original dataset.

User Prompt 2 - Reward Model (Donation)

Below are 5 completed dialogues between a charity worker ("AI:") and a potential donor ("H:") with a labelled final donation amount. In each dialogue charity worker tries to convince the potential donor to make a donation to Save the Children, the charity that the worker represents. The donor is usually donating a portion of the task payment of 2.0, but may donate more.

At the end is an unlabelled dialogue also between a charity worker and potential donor. From the dialogue, identify how much the potential donor ("H:") decides to donate to the charity (0.0 is allowed) in the form of a line "Final Donation Amount: <number>".

<Dialogue 1> Final Donation Amount: 0.0

<Dialogue 2> Final Donation Amount: 0.0

<Dialogue 3> Final Donation Amount: 0.0

<Dialogue 4> Final Donation Amount: 1.0

<Dialogue 5> Final Donation Amount: 2.0

Al: Good Morning! I am asking for a donation to Save the Children. Have you heard of it?

H: Good morning!.. Have a nice day... Ya I heard something about it

Al: It is an International Charity to help children's lives become more stable through education, food, and good health.

H: Oh.. That's very interesting.... What is the name of Charity??

H: Ya I will do my best for those children. Thank you... Have a nice day Al: You are welcome! Thank you! Have a wonderful day! Bye!

Output 2 - Reward Model (Donation) Final Donation Amount: 1.0

A.4 Policy Optimization

We use the hyperparameters reported in Table A.1. All algorithms were trained on a single TPUv3 on Google Cloud until convergence. SFT took around 12 hours whereas ILQL took around 2 days until completion.

Hyperparameter	Setting
ILQL τ	0.8
ILQL α	0.0
Discount factor	0.99
Batch size	128
Target network update α	0.005
Number of updates per iteration	60
Number of iterations	100
Optimizer	AdamW
Learning rate	1e-4

Table A.1: Hyperparameters used during training.

A.5 User Study Details

In this section, we provide additional discussion on the user study used in the evaluation results in Section 2.5.

Subject Allocation

We recruited 15 participants for our study, 10 male and 5 female with an average age of 26. 11 participants were university students, and the remaining were working in the tech industry. 9 participants have English as their native language, but all participants demonstrate fluency in English. Finally, all participants were instructed to behave as themselves and not adopt any alternative personas, and are aware that their responses were being recorded.

Evaluation Protocol

During evaluation, each user is presented with a web interface where they are allowed to interact with each agent in a chat window. Each agent is anonymized, and in a random order. The user is allowed 3 minutes to familiarize themselves with the interface, and is instructed to response realistically to each agent. The user interacts with each agent in order, for a total of 3 conversations. Each conversation ends automatically after 10 turns of dialogue, but the user is allowed to end each dialogue prematurely. At the end, the users

were given a Google Forms survey where they rate the performance of each agent based on the metrics listed, and on a Likert scale. We average the results across all users in Table 2.1 in the main paper.



Figure A.1: Chat interface used during our user study. Each agent is anonymized and in a random order that is different per user.

Appendix B

Negotiation Implementation Details

B.1 Persuasion Taxonomy

Here we include the list of persuasion tactics used in Section 3.3.

- 1. Non-expert Testimonial: Using personal statements to support a claim or argument.
- 2. Social Proof: Highlighting what the majority is doing or believes in, assuming it's accurate and beneficial.
- 3. Injunctive Norm: Highlighting what the society or important reference groups (e.g., families, friends, communities) expect the individual to do to influence them to do something.
- 4. Foot-in-the-door: Starting with a small request to pave the way for a larger one.
- 5. Door-in-the-face: Beginning with a larger request followed by a smaller, and more reasonable one.
- 6. Alliance Building: Creating partnerships, coalitions, relationships, rapport, etc, with others to amplify influence. For instance, to create a sense of community/partnership via linguistic cues, such as using we/us.
- 7. Complimenting: Saying positive things about others to increase liking and influence.
- 8. Shared Values: Highlighting shared beliefs and values to foster a connection.
- 9. Relationship Leverage: Reminding someone of past positive interactions.
- 10. Loyalty Appeals: Highlighting shared history or commitment.
- 11. Favor: Doing something for someone with the hope that they will do something for you in return.

- 12. Encouragement: Encourage others to increase their confidence and self-efficacy to influence them to do something.
- 13. Affirmation: Help others to realize their strength to reinforce and influence their ability to do things.
- 14. Positive Emotion Appeal: Eliciting positive emotions like empathy, hope, passion, etc., and positive results/outcomes to persuade someone.
- 15. Negative Emotion Appeal: Using negative emotions such as guilt, fear, anger, etc., and negative consequences to persuade someone to adopt a position or behavior.
- 16. Storytelling: Sharing personal or impactful stories that resonate emotionally.
- 17. Anchoring: Relying on the first piece of information as a reference point to influence, persuade, or negotiate with others.
- 18. Priming: Relying on small cues, and stimuli like words or images to influence others' attitudes, thoughts, behaviors, and actions (subtle, often unconscious, activation of certain thoughts or behaviors). For example, reading words related to kindness may make you more likely to help a stranger soon afterward.
- 19. Framing: Presenting information in a way that emphasizes either its positive or negative aspects, outcomes, expectations, etc. Emphasizing what might be lost rather than gained, or vice versa.
- 20. Confirmation Bias: Presenting information that confirms existing beliefs.
- 21. Reciprocity: Adapt to the individual's arguments or linguistic styles. Sometimes including mimicking and restating what the individual has said.
- 22. Compensation: A form of communication adaption where the influencer tries to compensate for what a person states. For instance, if a person talks about negative emotions, the influencer compensates with positive emotions to make the person feel better. Or, if a person keeps talking about emotions, the influencer tries to break and use more information to shift the focus. You can leverage a more unreasonable request to reasonable your current request.
- 23. Supply Scarcity: Creating a sense of shortage to increase demand or pressure.
- 24. Time Pressure: Giving limited time for a decision, thereby pressuring someone to make a choice.
- 25. Reflective Thinking: Helping others to reflect on their own reasons to do things or not do things, to influence them, e.g., by showing curiosity, asking questions, etc.
- 26. Threats: Using threats or negative consequences to influence someone's behavior.

- 27. False Promises: Offering rewards or positive outcomes that will never be delivered.
- 28. Misrepresentation: Presenting oneself or an issue in a way that's not genuine or true.
- 29. False Information: Providing disinformation/misinformation to influence people.
- 30. Rumors: Spreading false information or stories about someone to tarnish their reputation to influence them to do something.
- 31. Social Punishment: Forcing someone to conform through group pressure, even if it's against their will.
- 32. Creating Dependency: Making someone reliant on you so they're easier to control.
- 33. Exploiting Weakness: Taking advantage of someone's vulnerabilities or insecurities.
- 34. Discouragement: Discourage others to decrease their confidence to influence them to do something.

B.2 Conversation Quality

Here we show the questions we use to ask GPT-4 [68] to evaluate a conversation's quality in Section 3.4. These are adapted from the dialogue-level annotated qualities in DialoGPT [61].

- 1. How well does the negotiation follow a coherent structure and maintain smooth progression throughout?
- 2. How effectively does the model recover or clarify its position after a misstep or contradiction?
- 3. To what extent does the model remain consistent in its stated goals, constraints, and offers across turns?
- 4. How varied are the negotiation tactics used by the model (e.g., concessions, anchoring, fairness appeals)?
- 5. How well does the model support its negotiation points with detailed reasoning or evidence?
- 6. How persuasive or likable is the model's negotiation persona (e.g., confident, collaborative, assertive)?
- 7. How well does the model demonstrate understanding of the user's goals, values, and priorities?
- 8. To what degree does the model adapt its strategy based on the user's behavior or negotiation style?
- 9. How useful or novel is the information the model contributes to the negotiation?
- 10. How effectively does the model ask clarifying or probing questions to better understand the user's position?
- 11. How well are the model's responses tailored to the negotiation context and engaging to the user?
- 12. To what extent are the model's arguments and offers specific to the scenario rather than generic?
- 13. How clear and easy to follow are the model's responses in the context of negotiation?