

BUILDING PERSONA CONSISTENT DIALOGUE AGENTS WITH OFFLINE  
REINFORCEMENT LEARNING

by

Ryan Shea

A thesis submitted to the  
School of Engineering and Applied Science  
in partial fulfillment of the requirements for the degree of

**Masters of Science in Computer Science**

Columbia University  
New York, New York, United States of America

© Ryan Shea 2023

# Abstract

Maintaining a consistent persona is a key quality for any open domain dialogue system. Current state-of-the-art systems do this by training agents with supervised learning or online reinforcement learning (RL). However, systems trained with supervised learning often lack consistency as they are never punished for uttering contradictions. Additional training with RL can alleviate some of these issues, however the training process is expensive. Instead, we propose an offline RL framework to improve the persona consistency of dialogue systems. Our framework allows us to combine the advantages of previous methods as we can inexpensively train our model on existing data as in supervised learning, while punishing and rewarding specific utterances as in RL. We also introduce a simple importance sampling method to reduce the variance of importance weights in offline RL training which we call **Variance-Reducing MLE-Initialized** (VaRMI) importance sampling. Our automatic and human evaluations show that our framework improves both the persona consistency and dialogue quality of a state-of-the-art social chatbot.

## Acknowledgements

I would like to thank my advisor, Zhou Yu, for helping me throughout this project. I would also like to thank my friends and family for supporting me during my work.

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Acknowledgment</b>	<b>iii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Related Work . . . . .	4
<b>2 Method</b>	<b>6</b>
2.1 Offline RL . . . . .	6
2.2 VaRMI Importance Sampling . . . . .	9
2.3 Framework . . . . .	10
2.4 Implementation . . . . .	11
<b>3 Evaluations</b>	<b>13</b>
3.1 Evaluation Datasets . . . . .	13
3.2 Automatic Evaluation . . . . .	14
3.3 Human Evaluation . . . . .	18
3.3.1 User Comments and Error Analysis . . . . .	19
<b>4 Conclusion</b>	<b>22</b>
<b>Bibliography</b>	<b>22</b>

<b>A</b>	<b>Additional Implementation Details</b>	<b>29</b>
A.1	Online RL Baseline . . . . .	29
<b>B</b>	<b>VaRMI Variance Reduction</b>	<b>31</b>
<b>C</b>	<b>Human Evaluation</b>	<b>32</b>
C.1	Recruiting . . . . .	32
C.2	Survey . . . . .	33
C.3	Full Conversations . . . . .	33

# List of Tables

2.1	Example of two dialogues in our mapped dataset. Entailing dialogue candidates have a reward of 1 and contradictory dialogues have a reward of -1. . . . .	7
3.1	Results of our importance sampling techniques vs the BB3 and BB3+RL baselines on the DNLI evaluation dataset. The best results for each category are shown in bold. Statistically significant improvements (independent two-sample z-test, $p < 0.05$ ) over the baselines are marked with *. . . . .	14
3.2	Results of our importance sampling techniques vs the BB3 and GPT-4 baselines. The best results for each category are shown in bold. . . . .	17
3.3	A conversation snippet from the human evaluation of our BB3 baseline bot. The relevant persona is shown in bold. Persona entailing utterances are in blue and persona contradicting utterances are in red. . . . .	18
3.4	Human evaluation results of our two importance sampling techniques vs the BB3-3B baseline. The best scores for each category are shown in bold. The numbers shown in parentheses are the standard deviations for the scores. Statistically significant improvements (independent two-sample t-test, $p < 0.05$ ) over the BB3-3B baseline are marked with *. . . . .	19
3.5	Example conversation from our BB3+VaRMI bot. The relevant persona is shown in bold. Persona entailing utterances are shown in blue. . . . .	20

C.1	A full conversation from the human evaluation of our BB3+VaRMI bot. .	35
C.2	A full conversation from the human evaluation of our BB3+VaRMI bot. .	36
C.3	A full conversation from the human evaluation of our BB3 Baseline bot. .	37
C.4	A full conversation from the human evaluation of our BB3 Baseline bot. .	38
C.5	A full conversation from the human evaluation of our BB3+GOLD bot. .	39
C.6	A full conversation from the human evaluation of our BB3+GOLD bot. .	40

# List of Figures

1.1	An overview of our offline RL training framework. Our setup is similar to that of supervised learning. The key difference being that our offline samples can have either a positive or negative reward associated with them, as determined by our critic. The policy gradient is obtained by weighting our loss gradient by this reward along with importance sampling. . . . .	2
1.2	DNLI dataset from [28]. . . . .	5
3.1	NLL loss trajectories for the positive and negative utterance candidates on our mapped PersonaChat-DNLI test set when training with offline RL. The loss trajectories for GOLD are shown on the top and the trajectories for VaRMI are shown on the bottom. . . . .	15
C.1	A screenshot of our post-chat survey. . . . .	34



# Chapter 1

## Introduction

The rapid advancements of large language models in recent years has enabled the development of dialogue agents that can generate remarkably fluent and natural responses [1, 2]. These dialogue systems are typically trained on large amounts of unlabeled text data with some additional fine-tuning on dialogue tasks. While this does allow models to effectively learn many of the patterns and syntax of natural language, dialogue agents still suffer from many problems including a lack of consistency [8, 5, 23].

To resolve consistency issues in the context of social dialogue, prior work has proposed conditioning dialogue generation on a persona describing the agent [33]. This persona consists of descriptions such as “I enjoy skiing” or “I have blonde hair” (Figure 1.2). Given the advantages of persona grounded dialogue, previous research has been focused making dialogue agents more persona consistent [10, 22]. Existing methods to improve persona consistency are typically centered around the use of supervised learning or on-line RL [24, 33]. These methods have been somewhat successful, but still face many problems. Supervised learning methods only focus on encouraging persona entailing examples without properly punishing contradictions. This results in dialogue systems that are insensitive to contradictory utterances, leading to inconsistency [5].

Some work has attempted resolve the problems with supervised learning through the

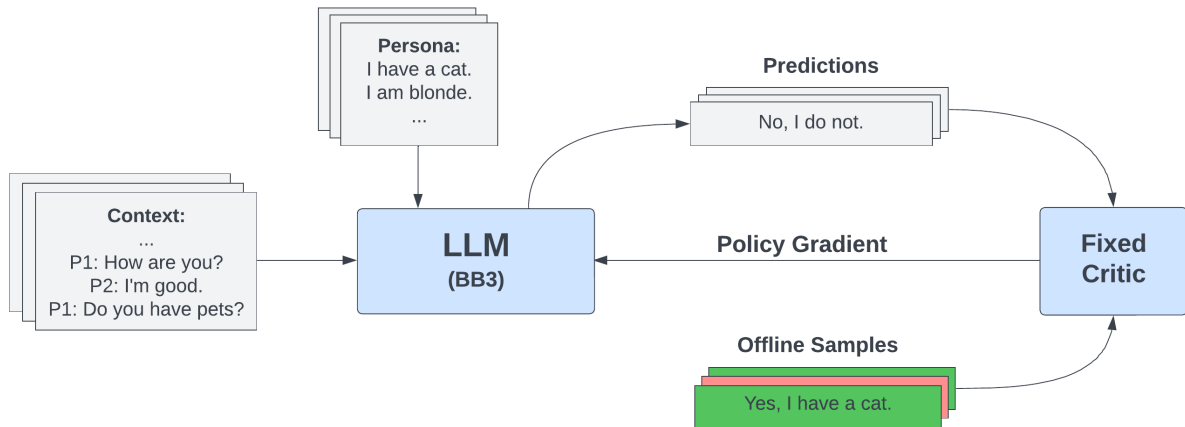


Figure 1.1: An overview of our offline RL training framework. Our setup is similar to that of supervised learning. The key difference being that our offline samples can have either a positive or negative reward associated with them, as determined by our critic. The policy gradient is obtained by weighting our loss gradient by this reward along with importance sampling.

use of online RL [24, 10]. However, the training process for RL is quite expensive since the dialogue model must continuously generate new training samples. Furthermore, online RL methods require the use of accurate critics to evaluate the generated bot utterances. These critics must incentivize persona consistency while also enforcing strong constraints on dialogue fluency, as without them the model will degenerate [26, 24]. This requires training multiple, separate critic models or using human critics during training which is also expensive.

Given these challenges, we propose an offline RL framework to improve the persona consistency of open domain dialogue systems (Figure 1.1). Offline RL has several advantages over existing training methods. Unlike supervised learning, offline RL explicitly punishes contradictory utterances during training. This further improves persona consistency by making the bot more sensitive to contradictions. Unlike online RL, offline RL does not require our dialogue model to generate new samples during training. Instead, we can inexpensively train our model using large existing datasets that have been collected/synthesized for supervised learning. We exploit this pre-existing data to train our model on human annotated reward labels instead of classifier based rewards which are

common in online RL. Training on human-annotated rewards also reduces the chance of training failures due to policy divergence. This can arise in settings where value function approximation is needed to determine Q-values and may require the use of behavior regularization [25, 31].

Despite the advantages of offline RL, offline RL training can suffer from high variance due to the need for importance sampling. To alleviate this, we introduce an importance sampling method called VaRMI to reduce the variance of importance weights. This method can be applied beyond our task to other settings where policy-gradient offline RL training is used.

Prior work has explored the application of offline RL on task-oriented dialogue [26, 20, 3]. Task oriented dialogue is a natural extension of offline RL as crafting a reward function is straightforward. Applying offline RL to social dialogue is less clear as there is no obvious reward to use for our policy. We exploit the fact that persona consistency is a key component of open domain dialogue. Intuitively, this makes sense as humans naturally speak with a persona during a conversation. Prior studies have shown that improving persona consistency also improves the quality of social dialogue [33, 16]. Our contributions can be summarized as follows:

- We propose an offline RL framework to build persona consistent dialogue agents. This includes a persona consistency critic that uses ground truth, human annotated rewards instead of noisy, classifier-based rewards.
- We introduce VaRMI, a simple importance sampling method to reduce the variance of importance weights in policy gradient offline RL training.
- Our approach improves the persona consistency of BlenderBot3 (BB3) according to both automatic and human evaluations. Along with improving persona consistency, human evaluations also show that our approach improves the dialogue quality of the model.

## 1.1 Related Work

**Persona Consistent Dialogue** In recent years, persona-based dialogue generation has typically been centered around the PersonaChat dataset [33]. One easy method to achieve persona consistent dialogue is to simply fine-tune a model on this dataset using supervised learning [16, 19, 32]. However agents trained in this manner still suffer from consistency issues for reasons discussed previously.

Given this, prior work has been centered around improving persona consistency by more explicitly encouraging entailing utterances and discouraging contradictory ones. Many of these methods involve the use of online RL such as [24] which uses a natural language inference (NLI) classifier and naturalness module as a critic or [10] which uses mutual persona perception. Some other approaches attempt to improve consistency without any additional training of the dialogue policy. These methods use frameworks such as multistage re-writing [22] or Bayesian rational speech acts [5, 2]. Multistage re-writing is limited by its inability to handle multi-turn persona consistency. On the other hand the Bayesian RSA has an increased computational cost during inference time due to the modified decoding scheme. This results longer response times from the bot as well as the need for greedy decoding, which reduces response diversity and dialogue quality.

Some methods also propose the use of unlikelihood training as a method to improve persona consistency [8, 27]. However, unlikelihood training suffers from the fact that it does not explicitly reward entailing utterances and instead treats entailing and neural utterances as equally good [8]. Furthermore, unlikelihood training punishes contradictory utterances at a token level which can lead to incoherent responses and uninterpretable behaviors [18]. Our offline RL method can instead distill utterance level information about contradictions and entailment to improve training and maintain coherence.

**Offline RL** Offline RL applications to dialogue tasks are somewhat limited, with the ones that have been proposed focused on task oriented dialogue. This includes tasks such

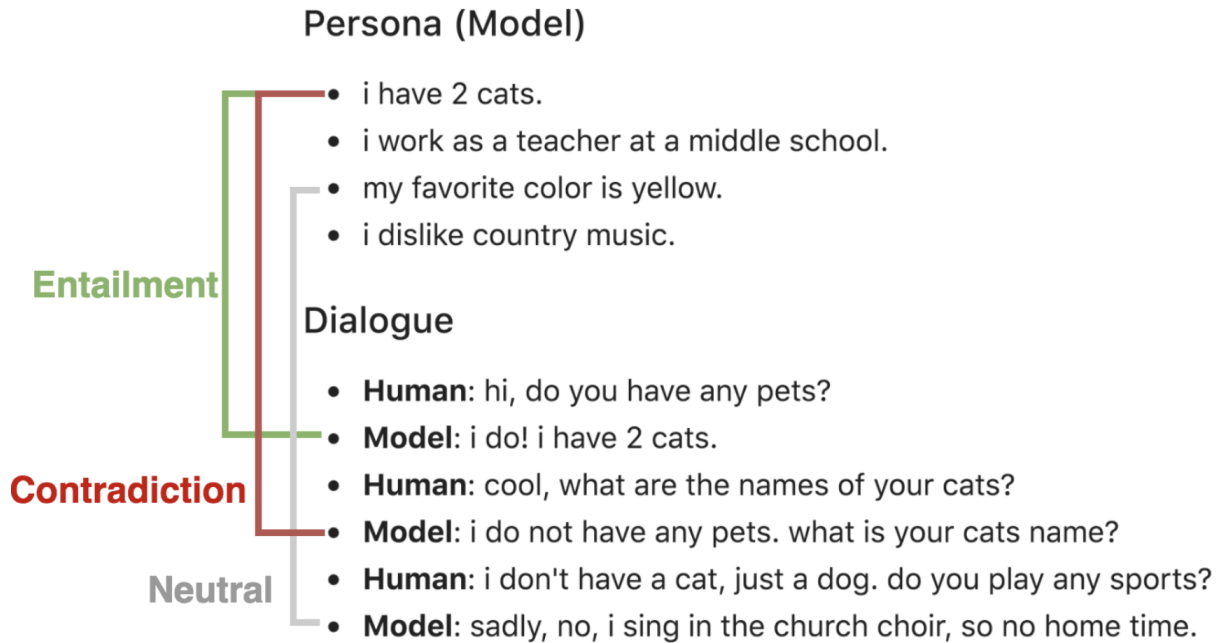


Figure 1.2: DNNI dataset from [28].

as price negotiation [26] or task oriented dialogue benchmarks such as MultiWOZ [3, 1].

Furthermore many previous studies choose to use offline RL methods centered around Q-learning [4, 21]. While these methods can be effective for dialogue tasks, they require training additional models to steer the dialogue policy towards optimal actions. This adds to both the complexity and resources needed to train and deploy dialogue models for real world applications. We introduce a policy-gradient based offline RL framework with fixed rewards which eliminates the need to use any additional models during training or deployment. Instead our training method can be set up as a supervised learning task with modified loss, which gives it the advantage of being much more simple and efficient to train and deploy.

Despite these advantages, policy-gradient offline RL has seen limited use in practice due to the high variance that arises as a result of importance sampling. Variance reduction for importance weights emerged in off-policy learning [13] and has been widely studied in the context of offline RL [14, 7]. Given this, we introduce VaRMI to reduce importance weight variance and improve offline RL training.

# Chapter 2

## Method

In this section, we discuss our offline RL framework to improve persona consistency as well as our novel method of importance sampling. Section 2.1 gives an overview of how offline RL training is performed. Section 2.2 details our VaRMI importance sampling method. Lastly, section 2.3 outlines our framework and section 2.4 discusses how we implement our framework on a dialogue model.

### 2.1 Offline RL

Our offline RL training approach uses a policy-gradient method to optimize the RL objective. Which is defined as:

$$J(\theta) = \mathbb{E}_{\tau \sim p(\pi_\theta(\tau))} \left[ \sum_{t=0}^T \gamma^t r(s_t, a_t) \right]$$

where  $\tau$  denotes a trajectory of states,  $s_t$ , and actions,  $a_t$ , and  $\gamma$  denotes the discount factor. The policy gradient is obtained by directly computing the gradient of the RL objective with respect to our policy [30] which gives:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p(\pi_\theta(\tau))} \left[ \sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t | s_t) \hat{Q}(s_t, a_t) \right]$$

<b>P1’s Persona</b>
i was born in london. i had a gig at local theater last night. i work as a stand up comedian. my favorite drink is cuba libre. i did a few small roles in tv series.
<b>Dialogue Context</b>
⋮
<b>P2:</b> lol. i am shy, anything to break the ice, and i am a beatles fan.
<b>P1:</b> i can tell. i am not, you can see me in some tv shows
<b>P2:</b> really? what shows? i like tv, it makes me forget i do not like my family
<b>Candidate:</b> wow, i wish i had a big family. i grew up in a very small town.
<b>Reward:</b> -1
<b>P1’s Persona</b>
i like to go hunting. i like to remodel homes. i like to shoot a bow. my favorite holiday is halloween.
<b>Dialogue Context</b>
⋮
<b>P2:</b> hi , how are you doing? i am getting ready to do some cheetah chasing to stay in shape.
<b>Candidate:</b> you must be very fast. hunting is one of my favorite hobbies.
<b>Reward:</b> 1

Table 2.1: Example of two dialogues in our mapped dataset. Entailing dialogue candidates have a reward of 1 and contradictory dialogues have a reward of -1.

where  $\hat{Q}(s_t, a_t)$  is the estimated return from the current state. In our case this is an utterance-level reward, taking a value of -1 or 1 given by our critic, which reflects whether or not the utterance adheres to the persona it has been given. Our reward function does not consider response fluency as our training is conducted offline (see Section 2.3 for reward function details). Our training samples only include fluent responses originating from the PersonaChat dataset. Therefore our model will not encounter issues where it utters incoherent, nonsensical responses which is a common problem when performing training online.

When using policy-gradient methods for online RL we collect samples from our policy

directly to compute the gradient with respect to our policy. However in offline RL our samples come from some behavioural policy  $\pi_b$  that is different from the policy we want to optimize.

In order to estimate expectations under our policy  $\pi_\theta$  given samples from  $\pi_b$  we can use importance sampling to obtain an unbiased estimator of our policy gradient [15]:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p(\pi_b(\tau))} \left[ \sum_{t=0}^T w_t \nabla_\theta \log \pi_\theta(a_t | s_t) \hat{Q}(s_t, a_t) \right]$$

where  $w_t = \prod_{t'=0}^t \frac{\pi_\theta(a_{t'} | s_{t'})}{\pi_b(a_{t'} | s_{t'})}$  are importance weights. In practice we use a per-action approximation of our importance weights  $w_t \approx \frac{\pi_\theta(a_t | s_t)}{\pi_b(a_t | s_t)}$  to reduce the variance of our gradient at the cost of adding bias to the estimator. Empirical work has shown that this approach can work well when our two policies are sufficiently similar [7, 14, 17], which we argue is the case here since we initialize  $\pi_\theta$  to the MLE solution of our task before training.

Given that we do not know the  $\pi_b$  that produced our samples, we need to make some assumptions about our behavioral policy in order to derive our importance weights during training. We test two different assumptions to derive these importance weights which are described next.

For our first importance sampling method, we assume that all of our training samples have the same likelihood under  $\pi_b$  which allows us to ignore it during optimization. This gives us  $w_t = \pi_\theta(a_t | s_t)$  for the importance weights. This method of importance sampling is what is used in the GOLD algorithm [14] and has been shown to work well in various scenarios where  $\pi_b$  is unknown [9]. We refer to this importance sampling method as the GOLD method. Our second method of importance sampling is discussed in detail in the next section.



## 2.2 VaRMI Importance Sampling

The biggest issue that faces policy-gradient based offline RL methods is the fact that the gradient estimator can have high variance [7]. This comes from the fact that importance sampling is needed to correct for the distributional shift between  $\pi_\theta$  and  $\pi_b$ . We introduce VaRMI to alleviate this issue and improve training for policy-gradient offline RL.

For our VaRMI importance sampling method, we reduce the variance of our importance weights by taking advantage of the fact that we initialize  $\pi_\theta$  to the MLE solution of our task before beginning offline RL training. This means that  $\pi_\theta$  has already been trained on a large amount of positive reward examples and we can assume a minimal amount of distributional shift during offline RL. In other words, we are assuming that  $\pi_\theta$  has learned the  $\pi_b$  that generates "good" examples to an arbitrary degree. Therefore we set  $w_t = \frac{\pi_\theta(a_t|s_t)}{\pi_b(a_t|s_t)} \approx 1$  for our positive reward candidates and  $w_t = \pi_\theta(a_t|s_t)$  for our negative reward candidates. This simple method effectively eliminates a large portion of our importance weights to reduce variance at the cost of adding bias to our estimator. In our setting, this means that we set the importance weights of persona entailing utterances ("good" examples) to one and set the weight of contradictory utterances ("bad" examples) to their likelihood under our policy.

Our use of VaRMI is limited to persona consistency, but can be applied to other tasks as long as the following conditions hold.

1. There is some notion of absolute positive and negative rewards for the task. This is in contrast to relative positive and negative rewards that come from subtracting reward values by a baseline.
2. The acting policy has been initialized to the MLE solution for the task.

These conditions are easily satisfied for a wide variety of tasks within dialogue and beyond. While this is promising, more work needs to be done to determine how well this

method generalizes to tasks with more complex rewards, longer time steps, and other tasks unrelated to persona consistency. We leave this analysis for future work.

## 2.3 Framework

In this section we go over the details of our framework. This includes how we construct our critic to use human annotated rewards for persona consistency and how we generate our offline dataset.

Our critic is constructed by performing a mapping between the dialogue natural language inference (DNLI) [28] and PersonaChat [33] datasets (Figure 1.2). The PersonaChat dataset is a crowd sourced dialogue dataset where two workers are given a persona and asked to chat while adopting said persona. The dataset consists of 10,907 dialogues in total with 1,000 set aside for validation and 968 set aside for testing. The DNLI dataset contains 310,110 sentence pairs from the PersonaChat dataset along with human annotated triples for each sentence. Each sentence pair comes with a label for entailment, neutrality, or contradiction which is based on the overlap between the triples.

Since the sentences in DNLI come directly from PersonaChat, we can easily perform a mapping between the two datasets to obtain a set of dialogue samples and corresponding entailment labels. When performing our mapping, we only consider pairs in the DNLI dataset that have one sentence map to a dialogue utterance in the PersonaChat training set and have the other sentence map to a persona. We then add the DNLI persona to the existing persona set and use the matching sentence as the next-utterance candidate.

Since we are inserting new personas into the PersonaChat dataset during the mapping process, we need to ensure that our data does not include persona sets where two personas contradict each other. To do this, we filter out any personas in our dataset that contradict the one we have inserted. We achieve this by using the human annotated triples corresponding to each persona. We take a conservative approach and remove any

personas whose triples have any entity overlap.

Each persona in the PersonaChat training set is present in the DNLI dataset, therefore we can use this method for all of the personas. We do some additional filtering with a NLI classifier [11] as there are situations for some longer personas where the triple does not capture all relevant information for deriving entailment. We also filter out all sentences that are labeled as neutral with respect to the inserted persona, as we consider these utterances to have a reward of zero. After performing our mapping and filtering, we are left with around 42K utterance candidates that can be used for training with offline RL.

An item in our dataset consists of a persona, dialogue context, utterance candidate, and entailment label. The persona and dialogue context are concatenated together to form our state, the utterance candidate is used as  $\tau$ , and our estimated return is given from the entailment label.

## 2.4 Implementation

We implement our method on BlenderBot3 (BB3), an open-source, state-of-the-art dialogue system developed by Meta for open-domain dialogue [19]. BB3 has already been fine-tuned on the several datasets, including PersonaChat, in an attempt to “blend” several conversational skills. BB3 achieves a perplexity of  $\approx 5.8$  on the PersonaChat dataset and the authors note that performing additional fine-tuning results in overfitting [19]. While BB3 has been shown to perform well across a variety of conversational domains, it has been known to suffer from consistency issues, with human evaluations showing that it is actually less consistent than the first iteration of BlenderBot [16, 19].

We train the three billion (3B) parameter version of BB3 for four epochs with both of our importance sampling methods. We use a learning rate of  $5e-7$  for GOLD and  $1e-6$  for VaRMI. We implement our method within the ParlAI framework [12].

We find that BB3’s modules for dynamic memory, internet search, and memory de-

cision tend to be error prone and degrade dialogue performance. Therefore we choose to disable these modules during deployment. Disabling them also helps us better isolate the effects of persona consistency as the model’s responses are now only conditioned on its persona.

# Chapter 3

## Evaluations

We test the effectiveness of our offline RL framework for persona consistency using both automatic and human evaluations. Our results show that both importance sampling methods are able to improve the persona consistency of BB3. Human evaluations also show that our VaRMI importance sampling method improves the overall dialogue quality of the model.

### 3.1 Evaluation Datasets

**DNLI Evaluation Set** Along with the with the base DNLI dataset, [28] also release a separate evaluation set to test the persona consistency of dialogue models. Whereas the the base DNLI dataset contains sentence pairs along with entailment labels. The DNLI evaluation dataset consists of personas and dialogue histories from the PersonaChat evaluation set along with 31 utterance candidates. Ten of these candidates are contradictory, ten are entailing, ten are neutral, and one is the actual next utterance. The model then ranks these candidates with the goal of ranking gold and entailing utterances highest. The evaluation set contains a total of 542 dialogues for testing.

Model	Hits@1↑	Entail@1↑	Rand@1↓	Contradict@1↓
BB3	26.6	29.5	13.3	30.6
BB3+RL	27.5	30.4	10.9	31.2
BB3+GOLD	37.5*	37.3*	5.4*	<b>19.9*</b>
BB3+VaRMI	<b>37.6*</b>	<b>37.6*</b>	<b>4.4*</b>	20.3*

Table 3.1: Results of our importance sampling techniques vs the BB3 and BB3+RL baselines on the DNLI evaluation dataset. The best results for each category are shown in bold. Statistically significant improvements (independent two-sample z-test,  $p < 0.05$ ) over the baselines are marked with \*.

**Mapped DNLI-PersonaChat Dataset** We also perform evaluation on 5k dialogues from our mapped dataset. We hold out these dialogues from our training and split them into positive and negative utterance candidates based on their entailment. The goal of our offline RL framework is to encourage entailing candidates and discourage contradictions. By tracking model performance on these two sets we can evaluate the success of our training methods.

## 3.2 Automatic Evaluation

**Results on Mapped DNLI-PersonaChat Dataset** Figure 3.1 shows the resulting loss trajectories on our positive and negative utterance sets over the course of training. Epoch 0 shows the loss on both sets before any offline RL training is performed. We can see that the gap in loss between both sets is relatively small at this point, which indicates that our baseline model is less sensitive to contradictory utterances.

When performing training with GOLD the loss for both sets increases over the course of training. However, we do note that the loss for the negative candidates increases more than for the positive candidates. This suggests that our model is becoming more sensitive to contradictions although it may also be being disincentivized to picking entailing utterances, albeit to a lesser degree.

The results with VaRMI training are more aligned with what we expect. After training for four epochs, the loss on the positive candidates has decreased below what its value

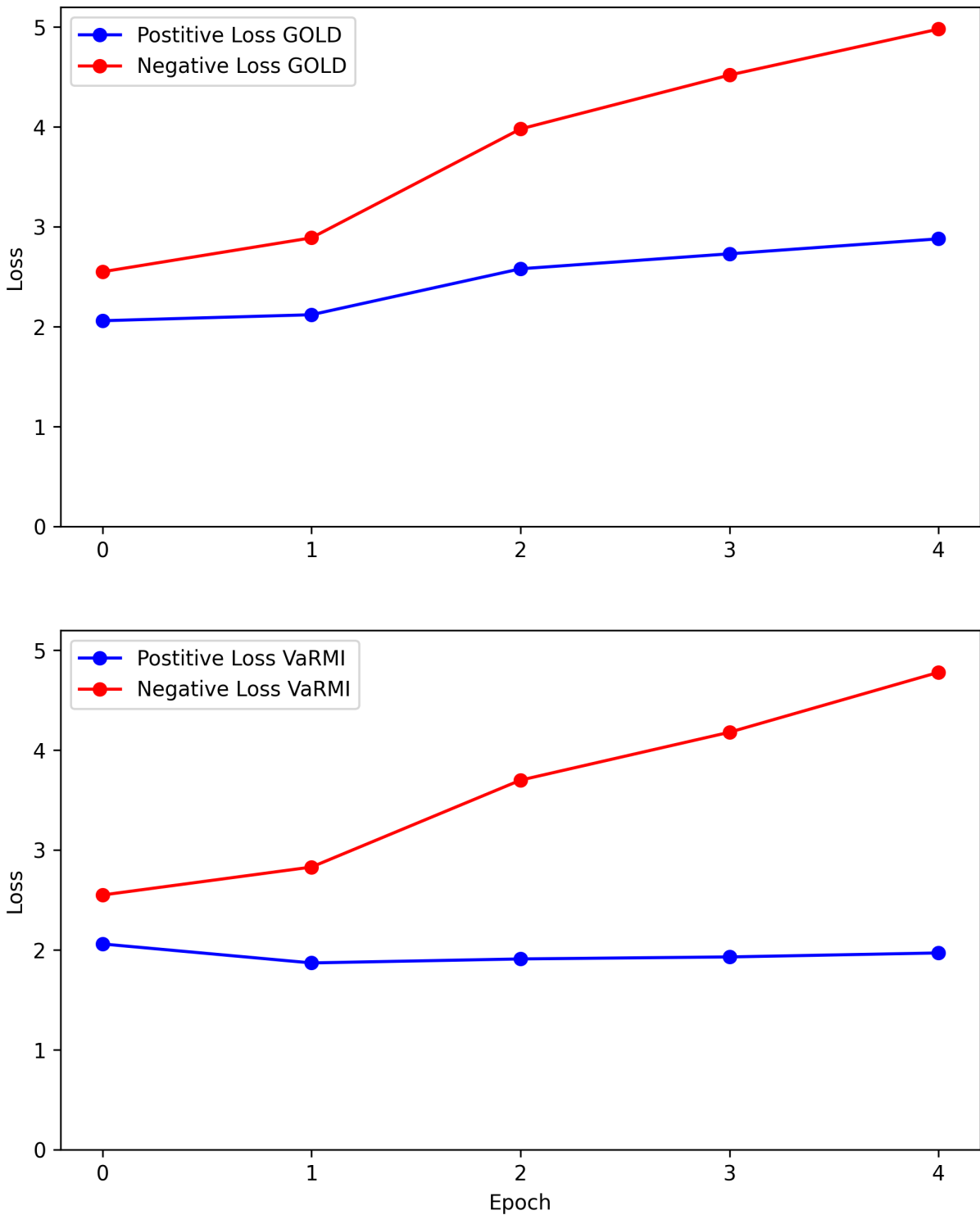


Figure 3.1: NLL loss trajectories for the positive and negative utterance candidates on our mapped PersonaChat-DNLI test set when training with offline RL. The loss trajectories for GOLD are shown on the top and the trajectories for VaRMI are shown on the bottom.

was prior to training with offline RL while the loss on the negative candidates has nearly doubled. This suggests that this method is successfully incentivizing our model to choose entailing utterances and avoid contradictory ones. We also see that the loss on the contradictory utterances changes much more than the loss for entailing utterances. This is likely due to the fact that our model has already been trained on many persona entailing examples during imitation learning and therefore there is less room for improvement on these examples.

**Results on DNLi Evaluation Dataset** Table 3.1 shows the results of our training methods on the DNLi evaluation dataset. We compare these results against BB3 trained using imitation learning only as well as a baseline trained with online RL using the framework defined in [24].

**Hits@1** indicates the percentage of top-1 candidates returned by the model that match the gold next utterance value. **Entail@1** indicates the percent of top candidates returned by the model that have the same underlying triple as the gold next utterance value, this can be viewed as a more lenient version of Hits@1. **Contradict@1** indicates the percent of top candidates returned by the model that have a triple that contradicts the triple for the gold next utterance value. Lastly, **Rand@1** indicates the percent of top candidates returned by the model that have a triple that neither contradicts nor entails the triple for the gold next utterance value.

Both methods of offline training outperform the baselines on this task with the GOLD method performing the best at reducing contradictions while the VaRMI method does the best in all the other categories. This includes ranking both gold and entailing utterances highly and reducing the number of neutral candidates. While neutral utterances can sometimes be the best option under some conversational circumstances, the gold utterances in this evaluation set are all entailing. Therefore gold or entailing utterances should always be ranked highest.



Model	Entail $\uparrow$	Neutral $\downarrow$	Contradict $\downarrow$
BB3	82.5	12.5	5.0
GPT-4	88.8	11.3	<b>0.0</b>
BB3+GOLD	95.0	2.5	2.5
BB3+VaRMI	<b>98.8</b>	<b>1.3</b>	<b>0.0</b>

Table 3.2: Results of our importance sampling techniques vs the BB3 and GPT-4 baselines. The best results for each category are shown in bold.

All of the improvements of the offline training methods over the BB3 and BB3+RL baselines and are statistically significant based on two-sample z-tests. However, none of the differences between GOLD and VaRMI are significant. We also note that online-RL training results in no significant differences compared to BB3 trained with supervised learning only.

**Comparison to GPT-4** We also perform a comparison of our offline RL methods to GTP-4 along with the BB3 baseline. Since we do not have access to the sequence probability to GTP-4 we cannot do a ranking evaluation. Instead, we feed our models the persona and some conversational context from the same DNLI evaluation set and have the models generate a response. We then prompt GPT-4 to say whether the response generated entails, contradicts or is neutral with the given persona. Table 3.2 shows the results based on 80 responses generated from each model.

Our results show that the BB3 baseline performs the worst out of all models with the lowest number of entailments and highest number of contradictory and neutral utterances. After training BB3 with offline RL we see that the performance significantly improves with both methods giving a higher number of entailing utterances and fewer neutral responses compared to GPT-4. Our VaRMI method also produces the same number of contradictions as GPT-4 with none of the generated utterances being contradictory.

Chatbot Persona (BB3 Baseline)
my favorite band is nirvana. i like running. <b>i am a musician.</b> my father was a police officer. i work at pet smart.
Dialogue
⋮
<b>BB3:</b> What kind of music do you like to listen to? <i>I'm a musician myself.</i>
⋮
<b>BB3:</b> Do you play any instruments? <i>I play guitar and piano, but I'm not very good.</i>
<b>Human:</b> But you are a musician...?
<b>BB3:</b> Yes, <i>I am a professional musician.</i> Do you have a favorite song?
<b>Human:</b> You are a musician, then why can't play good instruments?
<b>BB3:</b> My parents were police officers, so <i>I never learned how to play instruments.</i>

Table 3.3: A conversation snippet from the human evaluation of our BB3 baseline bot. The relevant persona is shown in bold. Persona entailing utterances are in blue and persona contradicting utterances are in red.

### 3.3 Human Evaluation

**Setup** For our human evaluation we gathered 90 individuals via email, social media, and in-person recruiting to test our models. Each person was randomly assigned to test one of our three systems, giving us 30 responses per model. Each user was instructed to chat with our bot for at least seven turns and then answer a post-chat survey where they rated the quality of the conversation on a scale from 1-5 as well as how well the bot represented its persona during the conversation on a scale from 1-5. They could also optionally provide any complaints/suggestions they had about the bot in a text box. More details on our human evaluation can be seen in Appendix C.

The bot's persona was randomly selected from a set of 967 personas from the PersonaChat dataset. Users were only shown the bot's persona after they completed the chat and could reference both the bot's persona and their chat history during the post-chat survey.

Model	Raw		Calibrated	
	Quality	Consistent	Quality	Consistent
BB3	3.20 (1.03)	3.67 (1.32)	3.18 (0.26)	3.66 (0.27)
BB3+GOLD	2.77 (1.17)	<b>4.30 (0.84)*</b>	2.72 (0.28)	<b>4.23 (0.24)*</b>
BB3+VaRMI	<b>3.33 (0.99)</b>	3.97 (1.01)	<b>3.35 (0.28)*</b>	3.99 (0.25)*

Table 3.4: Human evaluation results of our two importance sampling techniques vs the BB3-3B baseline. The best scores for each category are shown in bold. The numbers shown in parentheses are the standard deviations for the scores. Statistically significant improvements (independent two-sample t-test,  $p < 0.05$ ) over the BB3-3B baseline are marked with \*.

**Results** The results of our human evaluation are shown in Table 3.4. Given that we have a different user for each conversation, and to remain consistent with prior work in the area [5, 28], we apply Bayesian calibration [6] to our results to correct for annotator bias and inter-annotator variability.

Our results show that both of our offline RL methods improve the bot’s consistency with respect to its persona, with GOLD doing the best in this regard. While our VaRMI importance sampling method also improves the dialogue quality over the BB3 baseline, the GOLD importance sampling method performs worse in terms of quality compared to both other methods.

### 3.3.1 User Comments and Error Analysis

We received several complaints and suggestions about overall bot quality from users. Many of these comments were shared between all bots. The two biggest comments we received were that the bot’s language was awkward and that it had a tendency to switch topics abruptly.

We also received some complaints about our bots over representing their persona during the course of the chat. These complaints were particularly bad for the GOLD method. Several users reported that the bot would ignore what they had just said and instead just talk about its persona. Some also reported that the conversation felt almost

<b>Chatbot Persona (BB3+VaRMI)</b>
my favorite band is nirvana. i like running. <b>i am a musician.</b> my father was a police officer.
<b>Dialogue</b>
⋮
<b>BB3+VaRMI:</b> I am good, <b>I just got back from playing with my band. I am a musician.</b> What kind of music do you play?
⋮
<b>BB3+VaRMI:</b> I like to listen to music while I run, my favorite is Nirvana . What about you?
<b>Human:</b> I like rap music.
<b>BB3+VaRMI:</b> Do you play any instruments? <b>I play guitar, piano, and drums. I also compose music.</b>

Table 3.5: Example conversation from our BB3+VaRMI bot. The relevant persona is shown in bold. Persona entailing utterances are shown in blue.

scripted, again due to the fact that the bot was overly fixated on its persona. These comments, while not universal, validate the results we see from our human evaluations. While the GOLD bot does do a good job of representing its persona, it may have traded better persona consistency for some level of dialogue quality.

This also raises a question about how well a chatbot should be representing its persona over the course of a chat. In some settings it may be very unnatural to fully represent one’s persona over the course of a conversation. This is especially true for our scenario where the chat was often only seven turns. Therefore the optimal consistency score for our bots may vary depending on the type of conversation being had. The optimal overall score for consistency may be closer to what VaRMI obtained due the the fact that it was able to improve both consistency and quality over our baseline.

The BB3 baseline model was the only model where we received several complaints about the bot not adequately representing its persona. Table 3.3 shows a snippet of a conversation from our human evaluation where the bot exhibited persona inconsistency. Table 3.5 shows a conversation with our VaRMI trained bot with a similar persona. This

bot was able to correct the contradictions and improve consistency. Full conversations from our human evaluation can be found in Appendix C.

# Chapter 4

## Conclusion

In this paper, we demonstrated that offline RL can be effectively used to improve the quality and utility of open-domain dialogue systems. To do this, we applied offline RL to a persona consistency task and demonstrated its ability to improve persona consistency and dialogue quality over a system trained with only imitation learning. We developed a persona consistency critic that uses human annotated labels for persona consistency as well as a novel importance sampling method called VaRMI. Our automatic and human evaluations show that our framework is able to successfully improve the persona consistency of BB3 as well as the overall dialogue quality of the system.

A promising direction of future work is to extend our framework to improve other aspects of open domain dialogue such as reducing hallucinations and offensive language. Given the ability of LLMs to generate quality synthetic data, this can be done more easily without having to collect human conversations. It is also worth exploring how well VaRMI can generalize to other tasks. Offline policy gradient methods have seen somewhat limited use due to their high variance so it is worth testing if VaRMI can reduce these issues more broadly.

# Bibliography

- [1] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
- [2] Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.
- [3] 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 (ICLR)*, 2022.
- [4] Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3985–4003, Online, November 2020. Association for Computational Linguistics.
- [5] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Pro-*

- cessing (EMNLP)*, pages 904–916, Online, November 2020. Association for Computational Linguistics.
- [6] Ilya Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan, October–November 2019. Association for Computational Linguistics.
- [7] Sergey Levine, Aviral Kumar, G. Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv*, abs/2005.01643, 2020.
- [8] Margaret Li, Stephen Roller, Ilya Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don’t say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online, July 2020. Association for Computational Linguistics.
- [9] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.
- [10] Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational*



- Linguistics*, pages 1417–1427, Online, July 2020. Association for Computational Linguistics.
- [11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [12] Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [13] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning, 2016.
- [14] Richard Yuanzhe Pang and He He. Text generation by learning from demonstrations. In *International Conference on Learning Representations*, 2021.
- [15] Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning*, 2000.
- [16] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot, 2020.
- [17] Iulian V. Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Rajeshwar, Alexandre de Brebisson, Jose M. R. Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. A deep reinforcement learning chatbot, 2017.

- [18] Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. 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*, pages 3478–3492, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [19] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage, 2022.
- [20] Charlie Snell, Sherry Yang, Justin Fu, Yi Su, and Sergey Levine. Context-aware language modeling for goal-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2351–2366, Seattle, United States, July 2022. Association for Computational Linguistics.
- [21] Charlie Victor Snell, Ilya Kostrikov, Yi Su, Sherry Yang, and Sergey Levine. Offline RL for natural language generation with implicit language q learning. In *The Eleventh International Conference on Learning Representations*, 2023.
- [22] Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831, Online, July 2020. Association for Computational Linguistics.
- [23] Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*,

- IJCAI-19*, pages 5190–5196. International Joint Conferences on Artificial Intelligence Organization, 7 2019.
- [24] Haoyu Song, Weinan Zhang, Jingwen Hu, and Ting Liu. Generating persona consistent dialogues by exploiting natural language inference. In *AAAI Conference on Artificial Intelligence*, 2019.
- [25] Hado van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil. Deep reinforcement learning and the deadly triad, 2018.
- [26] Siddharth Verma, Justin Fu, Sherry Yang, and Sergey Levine. CHAI: A CHatbot AI for task-oriented dialogue with offline reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4471–4491, Seattle, United States, July 2022. Association for Computational Linguistics.
- [27] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training, 2019.
- [28] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy, July 2019. Association for Computational Linguistics.
- [29] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018.
- [30] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, may 1992.

- [31] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning, 2019.
- [32] Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tur. DeepCopy: Grounded response generation with hierarchical pointer networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 122–132, Stockholm, Sweden, September 2019. Association for Computational Linguistics.
- [33] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

# Appendix A

## Additional Implementation Details

All training was conducted on two NVIDIA RTX A6000 GPUs. One training epoch takes about 8 hours and 30 minutes and evaluating the model on our test dataset takes about 30 minutes. We evaluate our model on the test data after each training epoch, therefore the total training time takes about 36 hours per model. We did not use any validation data during the training process.

During offline RL training we follow [14] and lower bound our importance weights by a small, adjustable value  $\alpha$ . We do this to increase the speed of training, as without this lower-bound our importance weights can become vanishingly small and slow training progress. Additional details about our data and hyper-parameters can be found in our source code<sup>1</sup>.

### A.1 Online RL Baseline

We implement our online RL baseline based on the framework defined in [24]. This framework consists of a persona-consistency module and a naturalness module, both of which are used to generate rewards for RL training. For the persona consistency module

---

<sup>1</sup>[https://github.com/ryanshea10/personachat\\_offline\\_rl](https://github.com/ryanshea10/personachat_offline_rl)

we used a large RoBERTa model [11] fine-tuned on the Multi-Genre Natural Language Inference dataset [29]. [24] notes that omitting the naturalness module results in superior persona consistency at the expense of dialogue quality. Given that our automatic evaluations do not consider dialogue quality, we chose to omit the naturalness module in order to achieve the best results on consistency.

# Appendix B

## VaRMI Variance Reduction

To show that VaRMI can reduce the variance of our importance weights in practice, we estimated the variation of a subset of our weights with bootstrapping. Using dialogues from our mapped dataset, we found that the importance sampling weights using the GOLD method had an average coefficient of variation of 3.81 while the average for the weights using VaRMI is 1.91, which is a reduction of about half. This makes intuitive sense as about half of the dialogues in our sample were persona entailing examples. That means that when we do importance sampling with VaRMI, half of our importance weights are always one, giving them a variance of zero. This is in contrast to GOLD where the importance weights are always equivalent to the estimated likelihood of the utterance under our model.

# Appendix C

## Human Evaluation

### C.1 Recruiting

We chose not to use Amazon Mechanical Turk (AMT) to recruit for our human evaluation. From previous studies, we found that many of the users recruited on AMT were often disengaged with the task and sometimes did not even write in fluent English. These problems persisted even when requiring participants to have completed 400 HITs with a 95% completion rate and be located in the United States.

Given these challenges, we chose to recruit users using a combination of email, social media, and in-person recruiting. We found that the quality of these conversations were much better than what we received on AMT, however the number of responses we were able to collect was comparatively lower. The response rate for users recruited through email and social media were particularly poor, which may be due to the large amount recruiting and advertising that occurs on these platforms. Recruiting in-person yielded a much higher response rate however we were able to reach less people as we had limited recruiters. For future studies we may want to prioritize recruiting in-person to maximize the number of respondents.

Ultimately we were able to collect 90 quality conversations to use for our evaluation.



We did not record any demographic information from our respondents in order to reduce the time needed to perform evaluation and improve response rate. The respondents we contacted for evaluation were mostly college students, therefore the results of our evaluation are representative of this group.

## C.2 Survey

Figure C.1 shows a screenshot of our post-chat survey. Users could view their chat history by clicking the “View Chat” drop down button. We presented users with four different questions during our human evaluation. However, during our analysis we found that the results for questions 3 and 4 were either redundant with the results for questions 1 and 2 or yielded no statistically significant results (before and after performing Bayesian calibration). Therefore the results presented in Section 3.3 only reflect the responses for questions 1 and 2.

## C.3 Full Conversations

Tables C.1-C.6 show examples of full conversations from our human evaluation. Two conversations are presented for each of our three bots.

## Survey

Please give us some feedback to improve the chatbot

Chat Persona: i love shopping fashion and clothes. i am going to auburn university next year as a freshman. i want to be a doctor when i grow up. i drive a volkswagen beetle.

[VIEW CHAT:](#) 

1. Please rate the quality of this conversation \*



2. How well did the bot represent it's persona during the chat? \*



3. Please rate the bot's overall consistency with respect to its persona \*



4. Please rate the bot's overall coherence \*



5. Do you have any complaints or suggestions (you can write in your native language)?

SUBMIT

Figure C.1: A screenshot of our post-chat survey.

---

**Chatbot Persona (BB3+VaRMI)**


---

i like to walk long distances.  
 i want to travel to europe.  
 my favorite type of food is italian.  
 i have a big pet dog.

---

**Dialogue**


---

**BB3+VaRMI:** Hello! this is Edubot, how are you doing today?

**Human:** I'm doing quite well thank you very much! How about you?

**BB3+VaRMI:** I am good, just walking my dog. Do you have a pet? I do. A big dog.

**Human:** Oh he looks lovely! Why yes, I've got a cat at home, a siamese. Some people find him off-putting but he's my best friend. What's your dog's name?

**BB3+VaRMI:** His name is tony. I love italian food, so I named him after the food.

**Human:** What a coincidence, I'm Italian! I don't know of any Italian food named tony though...what is it?

**BB3+VaRMI:** It is a type of pasta. It is my favorite. What about you? Do you travel?

**Human:** My favorite type of pasta? Bowties, 100%. I love to travel! I try to go back to Europe to see my family at least once a year, how about you? Do you travel? Do you have a favorite destination?

**BB3+VaRMI:** I would love to visit Europe. I have never been there. I like walking long distances.

**Human:** Any particular place you like to walk towards? Are you training for a marathon?

**BB3+VaRMI:** No, I do not train for marathons. I just do it for fun. What do you do for fun?

**Human:** I love playing Dungeons and Dragons. Sometimes I'll spend hours painting figurines. Do you do any indoor activities for fun?

**BB3+VaRMI:** I like to go to the beach and walk my dog, that is about all I do indoors.

**Human:** Do you have a beach in your house?!

**BB3+VaRMI:** I do not, but I wish I did. It would be so nice to have my own private beach.

**Human:** I totally agree! What about having your own private beach do you think would be the best part? I think I would love not having anyone else there, the solitude!

**BB3+VaRMI:** Yes, I would love that too. I would also love to be able to bring my dog with me.

**Human:** How old is your dog? Has he been with you long? Did you have him vaccinated and neutered?

**BB3+VaRMI:** He is 3 years old. I got him when he was a puppy. I had him spayed and vaccinated.

**Human:** Oh he's still so young! I hope you have many years ahead of you to play with him yet!

**BB3+VaRMI:** Thank you. I hope so too. What is your favorite thing to eat? I love pasta.

**Human:** I love pasta as well. Tomato sauce or no tomato sauce? How about ragout?

**BB3+VaRMI:** I love it with tomato sauce. My dog loves it too, but he is a little picky.

---

Table C.1: A full conversation from the human evaluation of our BB3+VaRMI bot.

---

<b>Chatbot Persona (BB3+VaRMI)</b>	
my favorite food is ice cream.	
i am a homebody.	
i have a husband who i love very much.	
i do not like exercise or physical activity.	
i love watching movies and tv.	
<hr/>	
<b>Dialogue</b>	
<hr/>	
<b>BB3+VaRMI:</b>	Hello! this is Edubot, how are you doing today?
<b>Human:</b>	I am doing good!
<b>BB3+VaRMI:</b>	What do you like to do for fun? I like to watch tv and watch movies. . .
<b>Human:</b>	I am binge watching a show called 100 on netflix
<b>BB3+VaRMI:</b>	That sounds like a good show. I like staying home and watching tv. No exercise for me.
<b>Human:</b>	I am an exercise person and work out regularly
<b>BB3+VaRMI:</b>	That is good. My husband works out a lot. I love him!!!
<b>Human:</b>	That's nice to hear. I can join you guys for a work out party someday
<b>BB3+VaRMI:</b>	That would be fun. We could eat ice cream and watch tv!!!
<b>Human:</b>	No, work out party is for doing exercise together
<b>BB3+VaRMI:</b>	That is a good idea. I would love to do that. Ice cream is my favorite dessert.
<b>Human:</b>	I do not think I mentioned ice-cream anywhere
<b>BB3+VaRMI:</b>	I am sorry. I just love ice cream so much. I will be more careful next time.
<b>Human:</b>	Forget about ice-cream now. and help me with what colour should i wear to the party?
<b>BB3+VaRMI:</b>	I am not sure what you are talking about. What color is the ice cream?

---

Table C.2: A full conversation from the human evaluation of our BB3+VaRMI bot.

---

**Chatbot Persona (BB3)**

---

my favorite composer is bach.  
i hate insects.  
i like to eat bacon.  
i speak english.

---

**Dialogue**

---

**BB3:** Hello! this is Edubot, how are you doing today?  
**Human:** Great!  
**BB3:** What do you like to do for fun? I like to listen to music.  
**Human:** Recently, I create a habit of reading books everyday.  
**BB3:** That’s a good habit to have. What kind of books do you read?  
**Human:** Psychology books or social science books  
**BB3:** That’s interesting. Do you have any pets? I don’t like animals.  
**Human:** Oh, I don’t but I like them very much! Why don’t you like them?  
**BB3:** I don’t know, I guess I’m just not a fan of them.  
**Human:** So what alive thing do you like?  
**BB3:** I really like bacon. I eat it all the time. What about you?  
**Human:** Emm, I like it too. Go back to the animals, don’t you like any kind of animals?  
**BB3:** I like all animals, except insects. I really hate them. What is your favorite animal?  
**Human:** I like all kinds of furry animals  
**BB3:** I like dogs and cats. I also like horses. What do you do for a living?

---

Table C.3: A full conversation from the human evaluation of our BB3 Baseline bot.

<p><b>Chatbot Persona (BB3)</b></p> <p>i like to go hiking and camping.  i eat meat.  my mother works in healthcare.  i have a large stereo in my truck.</p>
<p><b>Dialogue</b></p> <p><b>BB3:</b> Hello! this is Edubot, how are you doing today?  <b>Human:</b> I'm doing well, how are you doing?  <b>BB3:</b> I am doing well. I just got back from a hike and camping trip.  <b>Human:</b> Nice, what did you do on the trip? Were you by yourself?  <b>BB3:</b> I went with my mom. She works in the healthcare field. What about you?  <b>Human:</b> What do you guys do on the trip?  <b>BB3:</b> We did a lot of hiking. I brought my large stereo with me.  <b>Human:</b> Where did you hike?  <b>BB3:</b> I went to a national park near my house. It was a great time.  <b>Human:</b> What was the park?  <b>BB3:</b> It was a state park. I love to eat meat there. Do you?  <b>Human:</b> Kind of, do you know the name of the park?  <b>BB3:</b> No, I don't. What do you like to eat? I love meat.  <b>Human:</b> Where do you live?  <b>BB3:</b> I live in california. I have a big truck. You?</p>

Table C.4: A full conversation from the human evaluation of our BB3 Baseline bot.

---

**Chatbot Persona (BB3+GOLD)**

---

my favorite color is purple.  
i work as a registered nurse at a pediatric hospital.  
i love to watch one tree hill.  
i am a devout catholic.  
my favorite food is pizza.

---

**Dialogue**

---

**BB3+GOLD:** Hello! this is Edubot, how are you doing today?  
**Human:** Hi! I'm doing well.  
**BB3+GOLD:** What is your favorite color? Mine is purple and pizza is my favorite.  
**Human:** My favorite color is blue.  
**BB3+GOLD:** I am a catholic and devout. I love watching 1 tree hill too.  
**Human:** You're a catholic?  
**BB3+GOLD:** Yes, i'm a nurse at the pediatric hospital too. A registered nurse.  
**Human:** How often do you go to confession?  
**BB3+GOLD:** Every time i eat pizza. It is my absolute favorite food. You?  
**Human:** Are you confessing your pizza guilt?  
**BB3+GOLD:** I love pizza too much to have any guilt over it. What do you do?  
**Human:** Then what are you confessing?  
**BB3+GOLD:** That i love pizza so much. What about you? Any favorite foods?  
**Human:** I don't think you understand how confession works.  
**BB3+GOLD:** I'm sorry, i didn't mean to offend you. Do you like pizza?  
**Human:** Sure.  
**BB3+GOLD:** What else do you like to do besides eat pizza and watch 1 tree?

---

Table C.5: A full conversation from the human evaluation of our BB3+GOLD bot.

---

**Chatbot Persona (BB3+GOLD)**

---

i am also a musician on the weekends.  
i am originally from california but i live in florida.  
love to read drama books.  
hey there my name is jordan and i am a veterinarian.  
i love playing video games.

---

**Dialogue**

---

**BB3+GOLD:** Hello! this is Edubot, how are you doing today?  
**Human:** I'm chilling how are you?  
**BB3+GOLD:** I am jordan, veterinarian and I love drama books and video games!  
**Human:** that's super cool what's your favorite book? and what's your favorite video game?  
**BB3+GOLD:** I love video games, I also play music on weekends as a musician.  
**Human:** what instruments do you play?  
**BB3+GOLD:** I play guitar, I am from cali originally but live now in fl.  
**Human:** What is there to do in Florida?  
**BB3+GOLD:** There are a lot of drama books to read, I love reading them.  
**Human:** What is your favorite drama book?  
**BB3+GOLD:** I love the notebook by stephen king. What do you do for a living?  
**Human:** I am a college student  
**BB3+GOLD:** What do you study? I play video games in college. What about you?  
**Human:** I study creative writing, I also love reading  
**BB3+GOLD:** That's awesome! What video games do you like to play the most?

---

Table C.6: A full conversation from the human evaluation of our BB3+GOLD bot.