

# Personalizing Task-oriented Dialog Systems via Zero-shot Generalizable Reward Function

A.B. Siddique  
University of Kentucky  
Lexington, Kentucky, USA  
siddique@cs.uky.edu

Kshitija Taywade  
University of Kentucky  
Lexington, Kentucky, USA  
kshitija.taywade@uky.edu

M.H. Maqbool  
University of Central Florida  
Orlando, Florida, USA  
hasanmaqbool@knights.ucf.edu

Hassan Foroosh  
University of Central Florida  
Orlando, Florida, USA  
hassan.foroosh@ucf.edu

## ABSTRACT

Task-oriented dialog systems enable users to accomplish tasks using natural language. State-of-the-art systems respond to users in the same way regardless of their personalities, although personalizing dialogues can lead to higher levels of adoption and better user experiences. Building personalized dialog systems is an important, yet challenging endeavor and only a handful of works took on the challenge. Most existing works rely on supervised learning approaches and require laborious and expensive labeled training data for each user profile. Additionally, collecting and labeling data for each user profile is virtually impossible. In this work, we propose a novel framework, P-ToD, to personalize task-oriented dialog systems capable of adapting to a wide range of user profiles in an unsupervised fashion using a zero-shot generalizable reward function. P-ToD uses a pre-trained GPT-2 as a backbone model and works in three phases. Phase one performs task-specific training. Phase two kicks off unsupervised personalization by leveraging the proximal policy optimization algorithm that performs policy gradients guided by the zero-shot generalizable reward function. Our novel reward function can quantify the quality of the generated responses even for *unseen* profiles. The optional final phase fine-tunes the personalized model using a few labeled training examples. We conduct extensive experimental analysis using the personalized bAbI dialogue benchmark for five tasks and up to 180 diverse user profiles. The experimental results demonstrate that P-ToD, even when it had access to *zero* labeled examples, outperforms state-of-the-art supervised personalization models and achieves competitive performance on BLEU and ROUGE metrics when compared to a strong fully-supervised GPT-2 baseline.

## CCS CONCEPTS

• **Computing methodologies** → **Information extraction; Natural language generation; Reinforcement learning.**

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. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9236-5/22/10...\$15.00

<https://doi.org/10.1145/3511808.3557417>

## KEYWORDS

Dialog Systems, Personalization, Reinforcement Learning, Zero-shot Learning.

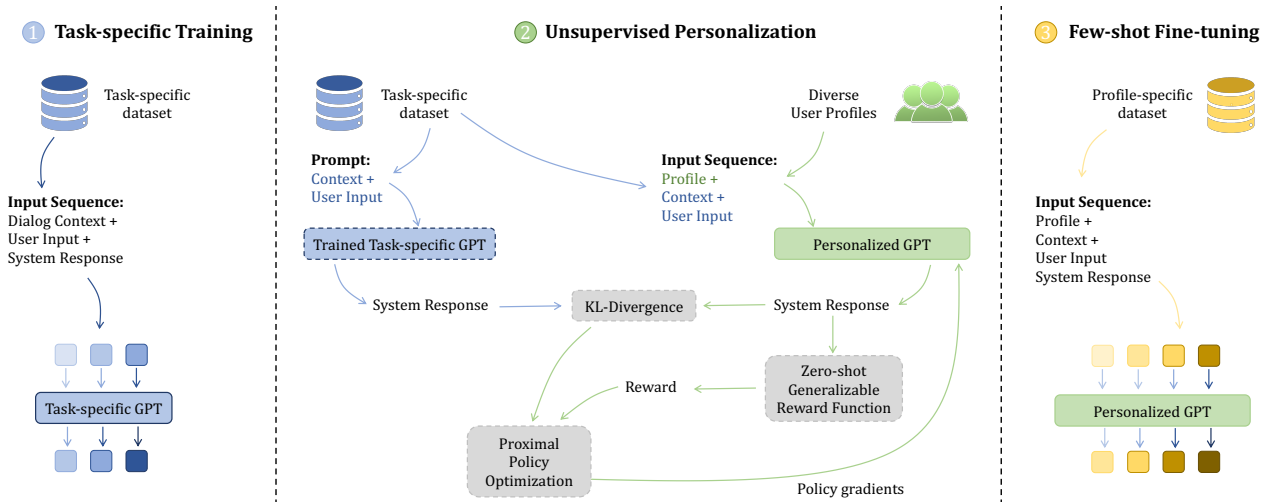
### ACM Reference Format:

A.B. Siddique, M.H. Maqbool, Kshitija Taywade, and Hassan Foroosh. 2022. Personalizing Task-oriented Dialog Systems via Zero-shot Generalizable Reward Function. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22)*, October 17–21, 2022, Atlanta, GA, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3511808.3557417>

## 1 INTRODUCTION

Task-oriented dialog systems provide users with the ability to carry out tasks, such as reserving a table at a restaurant, using natural language [56]. Contrary to the pipeline approach [7], researchers have increasingly focused on training end-to-end task-oriented dialog systems recently [3, 16]. Such models generate responses exclusively based on the task-specific context of the dialog. Consequently, these models fail to adapt their responses to the diverse user personalities [18]. Specifically, state-of-the-art task-oriented dialog systems struggle to (i) adapt their conversation flows according to the active user's personality, (ii) adjust their linguistic style, and (iii) handle ambiguities [15]. In addition to presenting the choices to the user in an arbitrary or sequential order without taking the personality of the active user into account, task-oriented dialog systems use only task-specific, dull language. It has been shown that adapting to the interlocutor improves communication efficiency [4, 5, 20]. Personalized task-oriented dialog systems can leverage profile information to expedite the interaction by understanding user's actual information needs promptly, generate tailored responses by adapting linguistic variations, and properly address ambiguities by contextualizing nuanced queries – a step towards delivering more human-like interactions [34, 59]. Personalizing task-oriented dialog systems without compromising the task completion accuracy is the focus of this work.

Earlier works use pre-training user profiles for intermediate supervision, as well as memory networks with copy mechanisms [15, 38]. The authors in [18, 28] encode user information and conversation history using memory networks in an end-to-end fashion. To synthesize personalized responses, [63] utilized dynamic and static attention mechanisms in the end-to-end memory network.



**Figure 1: Overview of P-ToD. The unsupervised personalization phase is at the core of the proposed framework.**

For each user profile, these works require enormous amounts of labeled training data, which is time-consuming, expensive, and nearly impossible to acquire. Recently, pre-trained language models have shown zero-shot capabilities in the natural language understanding and natural language generation tasks [6, 10], which suggests the possibility of developing personalized task-oriented dialog systems without requiring labeled training data for each target user profile. However, successfully exploiting the users’ profiles and synthesizing personalized responses with no (or few) labeled training examples is a demanding task.

We introduce a novel framework for building Personalized Task-oriented Dialog Systems, P-ToD, that leverages the pre-trained language models (LMs), zero-shot (as well as few-shot) learning, and deep reinforcement learning. Guided by the proximal policy optimization (PPO) algorithm [9, 46] and a zero-shot generalizable reward function, the proposed framework can personalize task-oriented dialog systems to diverse user profiles in an unsupervised fashion. Figure 1 presents an overview of the framework that works in three phases and uses a pre-trained GPT-2 [39] as a backbone model. A task-specific training (e.g., reserving a table) is performed in the first phase. Task-specific training datasets are generally available for a wide range of tasks in many domains [25, 62], whereas personalized counterparts are practically impossible to obtain. To overcome this challenge, we employ the unsupervised personalization phase. The deep reinforcement learning-based phase initializes a personalized GPT model from the task-specific GPT model (i.e., trained in phase one). Then, it trains personalized GPT model based on (i) the appropriateness of the generated response for the given user profile, quantified by the zero-shot generalizable reward function; and (ii) fidelity of the response to the task, measured by the KL divergence between the responses generated by the task-specific and personalized models. Using the above signals, the PPO algorithm is employed to perform policy gradients.

We also propose a new reward function that allows quantifying the quality of the generated personalized responses not only for previously seen user profiles, but also for newly emerging unseen profiles. The zero-shot generalizable reward function uses

pre-trained sentence transformers and contrastive representation learning to score the suitability of the response for the active user profile. To the best of our knowledge, this is the *first work* that can adapt the responses of task-oriented dialog systems to diverse user profiles in an unsupervised fashion. To further improve the performance of the personalized task-oriented dialog systems, an *optional* few-shot fine-tuning phase is introduced. This phase uses a few labeled training examples to adjust the responses for the given user profile, that can be employed or skipped depending on the availability of the labeled training data. Moreover, the number of shots can also be adjusted depending on the quantity of the available training examples.

We perform thorough experimental evaluations on the only publicly available benchmark, personalized bAbI dialogue benchmark, for five tasks and up to 180 distinct user profiles in the restaurant domain. The experimental results show that our proposed framework outperforms state-of-the-art supervised personalization models, even when given access to zero labeled training instances (i.e., few-shot fine-tuning phase is skipped). We also demonstrate that the proposed personalization approach achieves a competitive performance when compared to a strong supervised GPT-2 baseline model on the BLEU-4 and ROUGE-2 measures. Furthermore, the human study confirms the competitiveness of our unsupervised personalization framework to the other supervised approaches.

This work’s contributions are summarized below:

- We propose an end-to-end framework for personalizing task-oriented dialog systems in an unsupervised way. To the best of our knowledge, this is the first work that has the unsupervised personalization capabilities.
- We introduce a zero-shot generalizable reward function that can guide the policy of the personalized task-oriented dialog systems to generate rich and personalized responses even for the unseen user profiles.
- We perform extensive experimental analysis using personalized bAbI dialogue dataset and show that our framework consistently outperforms state-of-the-art supervised personalization models for up to 180 unique user profiles on five tasks.

## 2 PRELIMINARIES

### 2.1 Problem Formulation

In a multi-turn task-oriented dialogue,  $\mathcal{U}_t$  is an input from the user and  $S_t$  is a system’s response at a turn  $t$ . To generate a response  $S_t$ , all previous turns are concatenated to prepare dialog context  $C_t = [\mathcal{U}_0, S_0, \dots, \mathcal{U}_{t-1}, S_{t-1}]$  and passed to the system as input along with the user’s current input  $\mathcal{U}_t$ . In a personalized task-oriented dialog system, at turn  $t$ , the goal is to synthesize a response  $S_t^i$  adapted for a user profile  $\mathcal{P}^i \in \mathcal{P} = \{\mathcal{P}^0, \mathcal{P}^1, \dots\}$ . The system’s response  $S_t^i$  is generated by conditioning on dialog context  $C_t$ , user’s current utterance  $\mathcal{U}_t$ , and profile information  $\mathcal{P}_i$  for user  $i$ , concatenated as a single sequence.

$$S_t^i = \text{P-ToD}([\mathcal{P}^i; C_t; \mathcal{U}_t])$$

In traditional (i.e., supervised) personalized task-oriented dialog systems, at turn  $t$ , we are given  $m$  variants of the system response adapted for each user as:  $\{(\mathcal{U}_t, S_t^i)\}_{i=1}^m$  for all  $m$  user profiles to train the models. The major disadvantage of such an approach is the unscalable requirement of having a large number of labeled training examples for each user profile; such data acquisition is expensive and time-consuming. To overcome this challenge, we assume that, at turn  $t$ , profile-specific response  $S_t^i \forall i$  is not available for model’s supervision (i.e., unsupervised personalization). To allow for handling up to  $\infty$  user profiles, we assume that the user profile is described via natural language text, in contrast to previous works that encode the features of the user profile via one-hot encoding and limits the model’s expansion to new profile features. Naturally, describing user profiles using natural language takes care of the case where only partial information about a user profile is available. Moreover, some tasks require interaction with knowledge base, we define the knowledge base tuples as  $K = [k_1, k_2, \dots, k_f]$ , where each tuple  $k_b$  is defined using natural language and passed as additional input to the model where needed.

### 2.2 Pre-trained Language Models

The Language models (e.g., GPT-2 [39], BERT [8]) are trained utilizing massive amounts of text data in the unsupervised way. Since these models have millions of parameters, they have the capability to effectively capture both general semantic and syntactic information. In this work, we utilize the pre-trained GPT-2 and MPNet [52] models. We use GPT-2 as a base model, perform task-specific training, and then further train the model to synthesize personalized responses in an supervised way, guided by the novel reward function. The GPT-2 model has achieved state-of-the-art performance on many natural language generation benchmarks including conversation question answering [42], text summarization [35], and machine translation [22], among others.

We train a zero-shot generalizable reward function to score the acceptability of the generated responses for the given user profile using a contrastive loss function. The novel reward function uses pre-trained MPNet [52] as a basic building block to acquire semantically accurate embeddings. The MPNet model has produced cutting-edge results on several natural language processing tasks including GLUE [55], SQuAD [40, 41], RACE [21], and sentiment prediction [29] benchmarks. In the following, we provide a brief overview of the GPT-2 and MPNet models.

**GPT-2.** The GPT-2 model is pre-trained for autoregressive generation (i.e., predicting the next word) on the WebText dataset (i.e., 40 GB of text) and adapts a transformer-based neural architecture [54]. Suppose we have a natural language sequence  $(s_1, \dots, s_n)$  where symbol  $s_i$  is drawn from a fixed set of symbols. The sequential ordering of language leads to factorizing the joint probabilities over symbols as a product of conditional probabilities [2], as given below.

$$p(s) = \prod_{i=1}^n p(s_i | s_1, \dots, s_{i-1})$$

Using this approach, it is possible to estimate  $p(s)$  and any conditionals of the form  $p(s_{i-k}, \dots, s_i | s_1, \dots, s_{i-k-1})$ , and perform tractable sampling.

**MPNet.** BERT does not account for interdependence among predicted tokens, whereas complete position information is not used by XLNet [61], though dependency among tokens is considered. The MPNet model exploits the benefits of masked language modeling (MLM) (i.e., employed by BERT) and permuted language modeling (PLM) (i.e., used by XLNet) and eliminates their shortcomings. It brings out the best of both worlds: by using PLM, it exploits the predicted token’s dependencies, and, at the same time, uses the full position information of a sentence from MLM to enable a full view of the sentence. It has been pre-trained on BooksCorpus [67], OpenWebText, CC-News, Stories [53], and Wikipedia (i.e., over 160GB data). For a given sequence  $(s_1, \dots, s_n)$ , where permutations of set  $\{1, \dots, n\}$  is represented by  $\mathcal{Z}_n$ , the  $t$ -th element of  $z$  by  $z_t$ , the first  $t-1$  element of  $z$  by  $z_{<t}$ , the number of non-predicted tokens by  $c$ , and the mask tokens  $[M]$  in position  $z_{>c}$  by  $M_{z_{>c}}$ . The MPNet is trained for the following objective:

$$\mathbb{E}_{z \in \mathcal{Z}_n} \sum_{t=c+1}^n \log p(s_{z_t} | s_{z_{<t}}, M_{z_{>c}}; \theta)$$

### 2.3 Reinforcement Learning Paradigm

The reinforcement learning paradigm has been extensively studied for unsupervised learning. Methods that use policy gradients compute an estimator of the gradient and then plug it into a stochastic gradient ascent algorithm. It is common to optimize the policy  $\pi$  by maximizing the expected reward  $r \in \mathbb{R}$  for the generated sequence  $\mathcal{Y} = (y_1, \dots, y_n)$  with length  $n$ , given the input sequence  $\mathcal{X} = (x_1, \dots, x_m)$  with length  $m$ , that is sampled from data distribution  $\mathcal{D}$ . We can optimize the expected reward as follows:

$$\mathbb{E}_{\pi} [r] = \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi(\cdot | x)} [r(x, y)]$$

The PPO algorithm introduced clipped surrogate objective, in addition to, the penalty on the KL divergence. The objective function is modified using the KL divergence penalty, instead of making it a hard constraint like in the trust region policy optimization algorithms [45]. The PPO updates its policy, at step  $k$  via:

$$\theta_{k+1} = \arg \max_{\theta} \mathbb{E}_{s, a \sim \pi_{\theta}} [\mathcal{L}(s, a, \theta_k, \theta)]$$

where  $s$  and  $a$  represent the state and action, respectively. In this work, we employ PPO algorithm [9] to perform policy gradients, that has been shown to be scalable (e.g., for large language models), data-efficient, and robust (i.e., without excessive hyperparameter tuning) [1].

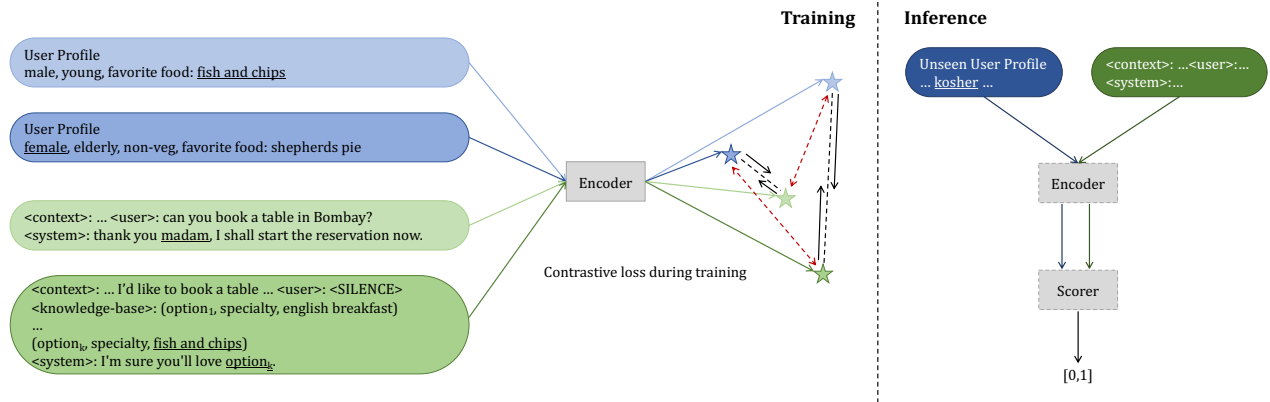


Figure 2: Overview of the training and inference process for the zero-shot generalizable reward function.

### 3 PERSONALIZATION FRAMEWORK: P-ToD

This work presents a new framework for developing personalized dialog systems that works in three phases. A pre-trained GPT-2 model serves as the backbone model for the framework. In the first phase, the base GPT-2 model is optimized via task-specific training. The phase two, referred to as unsupervised personalization phase, employs deep reinforcement learning to adapt the system responses to a wide range of user profiles guided by the zero-shot generalizable reward function (i.e., presented in Figure 2) and the trained task-specific GPT model. The *optional* phase three fine-tunes the personalized GPT model using a few supervised training examples to further improve the performance. Figure 1 summarizes the proposed unsupervised personalization framework.

#### 3.1 Phase One: Task-specific Training

We leverage the power of the pre-trained language models by initializing the phase one of our framework with a pre-trained GPT-2 model. The details of the pre-trained model are as follows. The model [39] was pre-trained on the WebText dataset and has 774 million parameters. Using byte pair encoding, the vocabulary size is 50,257 tokens; capitalization and punctuation were preserved [47]. The model is built on the transformer’s decoder stack [54], and it has 36 layers, 20 heads, and an embedding size of 1280. The task-specific training of the model is performed using causal language modeling (see Section 2.2 for details). Figure 3 presents the task-specific training of the model. Given a dialog context  $C_t$ , user’s current utterance  $U_t$ , and (optional) knowledge base search result tuples  $K$  at turn  $t$ , the probability of system’s response  $S_t$  with length  $n$  can be defined as:

$$p(S_t|C_t, U_t, K) = \prod_{i=1}^n p(s_i|s_{<i}, C_t, U_t, K)$$

We train the model by calculating the cross-entropy loss by maximizing the log-likelihood of the system response conditioned on the dialog context, user’s input, and knowledge base tuples. If the task does not require interaction with the knowledge base, the search query is not performed nor the generation is conditioned on the resultant tuples. The output of phase one is the trained task-specific GPT model.

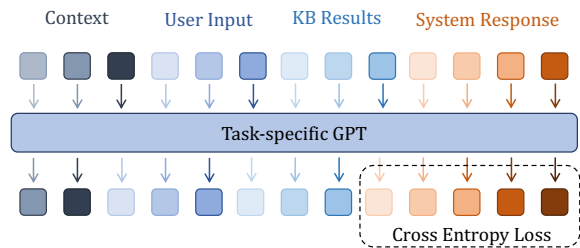


Figure 3: The task-specific training of the GPT-2 model.

#### 3.2 Phase Two: Unsupervised Personalization

This phase initializes the personalized GPT model with the trained task-specific GPT model (i.e., output of phase one). The personalized GPT model is trained for personalization in the unsupervised way. The two critical training signals are provided by (i) the zero-shot generalizable reward function that quantifies whether the output of the personalized model is appropriate for the given user profile; and (ii) the KL divergence between the personalized and task-specific model’s distributions to ensure that the output of the personalized model does not deviate too much from the task-specific model (i.e., it still accomplishes the task with high accuracy).

In the following, we describe the details of the novel reward function and KL divergence. Then, we detail the training process for the unsupervised personalization phase.

**Zero-shot Generalizable Reward Function.** The zero-shot generalization is enabled by the unsupervised representations provided by the powerful pre-trained language model MPNet and the contrastive loss function [14]. The training and inference process of the reward function is shown in Figure 2. At a dialog turn  $t$ , we concatenate the dialog context  $C_t$ , user’s current input  $U_t$ , the (optional) knowledge base search result tuples  $K$ , and the system’s response  $S_t^i$  for the user  $i$  and acquire their representation  $\mathcal{H}_t^i$ . Similarly, we encode the user profile information  $\mathcal{P}_j$  for the user  $j$  to get a corresponding representation  $\mathcal{U}^j$ . If a pair of encodings had a positive corresponding label (i.e., the system response is appropriate for the given user profile), then the contrastive loss function would reduce their distance, and if a negative label were given, it would increase their distance. We generate positive training examples by setting

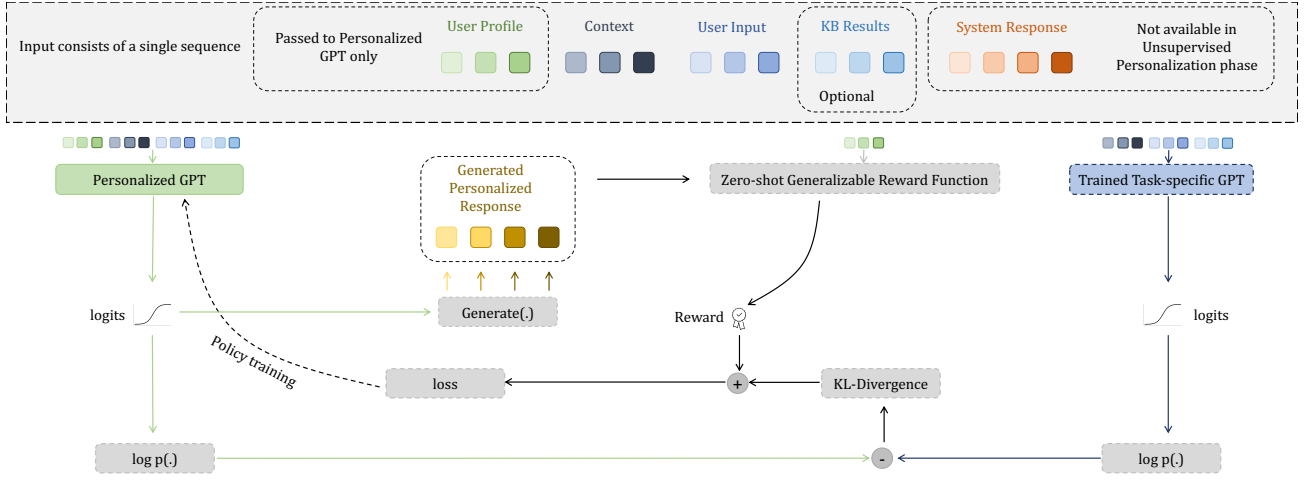


Figure 4: Phase two of the framework: Unsupervised Personalization.

$i = j$  and negative examples are generated by setting  $i \neq j$ . The training loss can be defined as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\mathcal{H}_t^i \cdot \mathcal{U}^j / \tau)}{\sum_{q \in Q} \exp(\mathcal{H}_t^i \cdot \mathcal{U}^q / \tau)}$$

where the  $\cdot$  represents the scoring function,  $\tau \in \mathcal{R}^+$  is a scalar parameter for temperature, and  $Q$  is the set of negative pairs, i.e.,  $i \neq j$ . To train a classifier that works in the zero-shot setting, we select a subset of user profiles (i.e., *seen* profiles) and use them to train the classifier. The pre-trained MPNet has the capability to generate rich, accurate, and high-quality embeddings even for the *unseen* user profiles or unseen knowledge base entries, since both the user profile and knowledge base tuples are described using natural language. For example, the model can produce precise embeddings for an unseen user profile who prefers “kosher” food, because it has already learned the contextual usage of a large number of words (e.g., MPNet has a vocabulary size of 30,527) in the pre-training process. The scoring function learns to score close to one, the matching pairs (i.e., the system response is appropriate for the given profile), and zero otherwise.

Our zero-shot generalizable reward function follows the Sentence-BERT [43] that employs siamese and triplet network structures [44], leverages contrastive loss, and dot product is used as the scoring function. To generate input encoding, we use the pre-trained `all-mpnet-base-v2` that has been trained on over one billion training pairs and produces 768 dimensional normalized embeddings for the input by mean pooling. For every positive training pair, two negative training examples are generated. At inference time, the trained zero-shot generalizable reward function provides a scalar reward,  $r \in [0,1]$  that quantifies the suitability of the system’s responses for both previously seen and newly emerging unseen user profiles.

**KL Divergence.** To ensure that the personalized policy does not diverge too much from the trained task-specific model, we use an additional reward signal by calculating the KL divergence between the personalized policy and the task-specific policy (i.e., the model

trained in phase one). That is, keeping close to the task-specific model is rewarded, whereas big KL divergences are penalized. We denote the distributions of the task-specific and personalized models by  $p_1$  and  $p_2$ , respectively. At dialog turn  $t$ , the KL divergence can be calculated as:

$$KL = \mathbb{E}_{S_t^i \sim p_2} [\log p_2(S_t^i | \mathcal{P}^i, C_t, \mathcal{U}_t, K) - \log p_1(S_t | C_t, \mathcal{U}_t, K)]$$

where  $S_t$  is the task-specific response and  $S_t^i$  is the system’s response adapted for the user  $i$ . The final *reward* can be combined as given below:

$$reward = r + \beta \times KL$$

where  $\beta \in [0, -1]$  is the penalty coefficient and decides the weight of the KL divergence. We use adaptive KL Penalty coefficient and initialize  $\beta = -0.2$  in our experiments .

**Training Details.** To start with the unsupervised personalization phase, we initialize our personalized model  $p_2 = p_1$  and then adapt  $p_2$  to synthesize the personalized responses for a wide range of user profiles using deep reinforcement learning. The personalized model is fine-tuned via PPO algorithm from [9] with the final *reward* (i.e., a combination of KL divergence and a score from zero-shot generalizable reward function). The expected reward for a response  $S_t^i$  for the user  $i$  at a dialog turn  $t$  can be written as:

$$\mathbb{E}_{p_2} [reward] = \mathbb{E}_{\mathcal{U}_t \sim \omega, S_t^i \sim p_2(\cdot | \mathcal{P}^i, C_t, \mathcal{U}_t, K)} [reward(\mathcal{P}^i, S_t^i)]$$

where  $\omega$  represents a given task, the model  $p_2$  is being trained for. The personalized model is trained for up to 600,000 episodes using Adam optimizer [19] with a learning rate of  $1.41 \times 10^{-5}$ .

The output of this phase is a personalized model that can generate responses that are not only specific to the task, but are also adapted for the given user profile. It is important to recall that the unsupervised personalization phase does not use any personalized variants of the responses for training the model. It is exclusively trained in the unsupervised setting, guided by the zero-shot generalizable reward function and KL divergence between the distributions of the task-specific and personalized models.

**Table 1: Datasets statistics.**

Dataset		Task 1	Task 2	Task 3	Task 4	Task 5
bAbI dialogue	Number of dialogs	4000	4000	4000	4000	4000
	Avg. dialog turns	6.0	9.5	9.9	3.5	18.4
Personalized	Number of dialogs	24000	24000	48000	24000	48000
	Avg. dialog turns	6.0	9.5	11.8	3.5	20.3
bAbI dialogue	Number of user profiles	6	6	180	6	180
	Avg. dialogs per profile	4000	4000	267	4000	267

### 3.3 Phase Three: Few-shot Fine-tuning

The optional phase three uses a few labeled training examples to calibrate the personalized model (i.e., trained in phase two in the unsupervised setting) for the given user profile in the supervised setting. The probability for system’s response  $S_t^j$  with length  $n$ , for a given user  $j$ , at dialog turn  $t$  can be defined as:

$$p(S_t^j | \mathcal{P}^j, C_t, \mathcal{U}_t, K) = \prod_{i=1}^n p(s_i | s_{<i}, \mathcal{P}^j, C_t, \mathcal{U}_t, K)$$

We call this phase *optional*, since it can be employed or skipped based on the availability of the labeled variants for the given user profile. Moreover, the number of shots can also be adjusted depending on the quantity of the available training examples. In our experiments, we present results with the following number of shots: 0 (i.e., we skip this phase), 1, 5, 10, and 20.

## 4 EXPERIMENTAL SETUP

In this section, we describe the task-specific and personalization datasets, methodology of evaluation, competing methods, and the implementation details of our framework P-ToD.

### 4.1 Datasets

We used one task-specific task dataset bAbI dialogue [3] that trains our model in phase one. The personalized counterpart, called personalized bAbI dialogue [18], is used to train all the supervised competing models. Our proposed framework adapts to diverse user profiles in the unsupervised setting. To the best of our knowledge, personalized bAbI dialogue is the *only* publicly available personalization benchmark for task-oriented dialog systems. Table 1 presents important statistics for both datasets. Both datasets are in the restaurant domain and consist of five tasks.

**Task 1: Issue API calls.** This task involves extracting values of all the required slots (a.k.a. values for query parameters, e.g., cuisine = spanish) from natural language utterances and successfully making an API call. In this task, the personalization involves understanding and adapting the linguistic variations for a given user profile (e.g., male vs female).

**Task 2: Update API calls.** This task includes updating the values for certain slots, if the user wishes to do so. For example, a user’s request in natural language, “Instead could it be in a cheap price range in Madrid?”, should update the current API call: `api_call(cuisine=french, city=paris, party_size=four, price_range=expensive)` to the call: `api_call(cuisine=french, city=madrid, party_size=four, price_range=cheap)`. Similarly to task one, personalization task two mainly deals with the style adaptations.

**Task 3: Display Options.** This task requires displaying relevant options from the knowledge base using the search results from API call. The personalization task involves adapting certain linguistic style as well as understanding user’s taste and restaurant’s specialties, among others, and making appropriate suggestions based on the active user’s profile. Unsupervised personalization for this task is the most challenging part of this work.

**Task 4: Provide extra information.** The user’s acceptance of an option entails asking for extra information (e.g., phone\_number) from the system. The personalization for task four calls for resolving ambiguities efficiently along with the style adaptation. For example, asking for contact information could refer to phone\_number or social\_media depending on the active user (e.g., elderly vs young).

**Task 5: Conduct Full dialogs.** This task is about conducting the full dialogue that covers tasks 1-4 successfully. Similarly, personalization task includes, but not limited to: (i) adjusting the conversation flow to the active user’s personality, (ii) adapting the linguistic style, and (iii) dealing with nuances effectively.

The personalized bAbI dialogue dataset contains two test sets: a standard test set and a test set - 00V (Out Of Vocabulary). We conduct extensive experiments on both test sets for all the five tasks for up to 180 diverse user profiles.

### 4.2 Evaluation Methodology

To demonstrate the effectiveness of P-ToD, we evaluate our framework and all the competing methods for (i) task completion and (ii) personalization of the dialog for the given user profile.

**Task Completion.** To quantify the performance for the task completion, we compute the F1 scores and present evaluation results for all the models for all five tasks.

**Personalization.** The main task for the proposed framework is to personalize the task-oriented dialog systems in the unsupervised way. To evaluate the efficacy of the framework and how it compares to the other supervised approaches, we use BLEU-4 and ROUGE-2 scores. The BLEU [36] and ROUGE [17] metrics have been extensively used for natural language generation tasks. Human judgment and BLEU scores show a very strong correlation. The BLEU-n ( $n \in \{1, 2, 3, 4\}$ ) score  $\in [0, 100]$  measures the proportion of n-grams in the generation that also occurs in the reference. ROUGE, on the other hand, is a recall-based measure that quantifies n-gram overlap between the generation and the reference. Moreover, we also conduct a user study on a randomly selected 300 responses generated by the top performing supervised models and our proposed unsupervised personalization framework.



**Table 2: F1 scores for task completion.**

Approach	Models	Task 1	Task 2	Task 3	Task 4	Task 5
Supervised	MemNN-org	99.63	99.81	98.87	98.87	85.10
	MemNN-split	85.66	85.83	84.89	84.89	87.28
	PMemN2N	99.70	99.93	98.91	98.97	95.33
	Mem2Seq-org	99.68	99.68	98.28	99.68	80.41
	Mem2Seq-split	99.62	99.62	98.52	99.62	82.19
	Mem2Seq-att	99.66	99.66	98.46	99.66	82.38
	GLMP	99.45	99.45	98.48	99.45	86.20
	CoMemNN	99.65	99.65	98.61	99.65	98.13
	Supervised-GPT	<u>99.72</u>	<b>99.96</b>	<u>99.02</u>	<b>99.96</b>	<b>98.21</b>
Unsupervised Personalization	PToD-0 (This work)	99.69	99.86	98.92	99.88	98.14
Few-shot Personalization	Few-shot GPT	98.12	99.08	97.71	97.32	91.23
	P-ToD (This work)	<b>99.74</b>	<u>99.94</u>	<b>99.03</b>	<u>99.94</u>	<u>98.17</u>

### 4.3 Competing Methods

We compare against the following state-of-the-art (SOTA) personalization models and GPT-2-based strong baselines:

**MemNN [18]:** The response selection-based approach proposes to use the memory network to encode dialog content and user profile information using a concatenation of the profile information and dialog memory (i.e., MemNN-org) and using split memory for the profile information and concatenating hidden states (i.e., MemNN-split).

**PMemN2N [28]:** The memory network-based method facilitates the model’s personalization by combining the style information of the user attributes in the encoder.

**Mem2Seq [31]:** An end-to-end approach that proposes to use memory network in the encoder and employs RNN-based decoder for query generation and memory network for personalized response generation. This work proposes three variants of the models, called Mem2Seq-org, Mem2Seq-split, and Mem2Seqatt.

**GLMP [57]:** Based on Mem2Seq, this model includes local and global encoders to share external knowledge efficiently.

**CoMemNN [37]:** This work proposes cooperative memory network and assumes that only partial user profile information is available. This approach does not generate response, instead relies on the response selection. In our experiments, we provided the model with 100% user profile information for a fair comparison.

**Supervised GPT:** Since none of the SOTA personalized models follow SOTA transformers architecture, we also trained a supervised GPT-2 model. This model was trained in the same fashion as our phase three except it was trained on all the training examples of the dataset, thus serves as a strong supervised baseline.

**Few-shot GPT:** Due to the unavailability of any unsupervised approach for comparison and coming up with a reward function is non-trivial, we also trained a few-shot GPT-2 model. This model follows same training process, except phase two (i.e., unsupervised personalization) is skipped to demonstrate the effectiveness of the phase two of the proposed framework.

### 4.4 Implementation Details

We use the pre-trained GPT-2 model as a backbone model that is trained in all the three phases of the framework. The phase one

trains the task-specific model for 3 epochs using cross-entropy loss and Adam optimizer, with a batch size of 8, and a learning rate of  $5 \times 10^{-5}$ . Other parameters are as follows: `warmup_steps=100`, `weight_decay=0.01`, `max_length=1024`. The zero-shot generalizable reward function uses a pre-trained MPNet for input encoding. It is trained for 3 epochs using contrastive loss on 50% of the user profiles on every task and the remaining 50% profiles are considered unseen. The phase two uses the same parameters as phase one, except batch size of 4 was used because of the GPU memory limitations (and a learning rate of  $1.41 \times 10^{-5}$ ). Similarly, phase three uses same parameters, except a smaller learning rate of  $5 \times 10^{-7}$  was used and up to 20 training examples were made available for training. We present two variants of our model: (i) PToD-0 does not use phase three (i.e., personalized model is only trained in the unsupervised setting) and (ii) P-ToD that uses 20 training examples in the phase three.

## 5 RESULTS

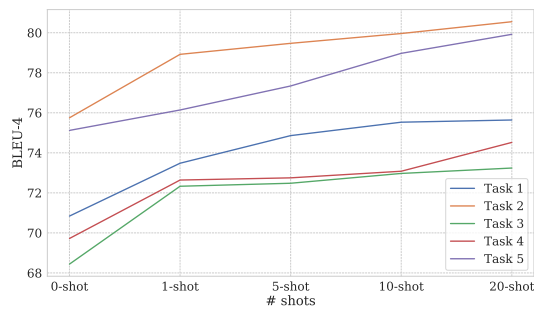
In this section, we present quantitative as well as qualitative analysis. We first present results on the task completion and then demonstrate that our proposed framework consistently outperforms SOTA supervised personalization models for the personalization task.

### 5.1 Quantitative Analysis

**Task Completion.** Despite the fact that the core task in this work is personalization, the personalized models should not compromise the accuracy of task completion for adapting their behaviors for the profiles of the users. Keeping it in mind, we report the results for task completion in Tables 2 that presents F1 scores for all five tasks for all the competing models. In terms of task completion, all the models show competitive performance except MemNN-split. The main reason for all the models showing great performance for task completion is that the user never drops out of the conversation, even if the system keeps providing the user with unwanted recommendations or never adapts the linguistic style according to the user. Since, the system eventually completes the task (i.e., the user is too patient which is not the case in the real-world), the F1 score is high for all the competing models. Though, the margin is not big, the best models are supervised-GPT and P-ToD (i.e., this work). For example, on tasks one and three, the proposed P-ToD performs the

**Table 3: BLEU scores and ROUGE scores for personalization for all five tasks.**

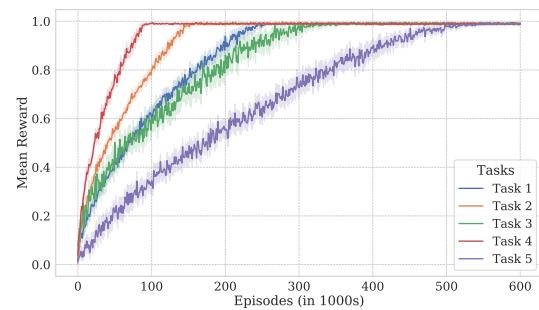
Approach	Models	Task 1		Task 2		Task 3		Task 4		Task 5	
		BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Supervised	Mem2Seq-org	60.12	64.82	65.54	69.83	57.74	62.73	59.07	63.32	64.23	59.39
	Mem2Seq-split	60.30	63.82	64.92	68.60	58.07	62.43	59.20	63.03	64.11	58.73
	Mem2Seq-att	62.26	71.17	67.15	75.84	59.84	69.59	61.29	69.74	66.02	66.17
	GLMP	61.25	70.81	66.40	75.46	59.07	68.93	59.66	70.13	64.91	65.74
	CoMemNN	68.67	77.71	73.83	82.67	65.77	75.72	67.58	76.85	72.23	72.53
	Supervised-GPT	<b>75.71</b>	<b>78.42</b>	<b>80.61</b>	<b>83.38</b>	<b>73.21</b>	<b>76.46</b>	<b>74.64</b>	<b>77.11</b>	<b>80.01</b>	<b>73.61</b>
Unsupervised	PToD-0 (This work)	70.84	75.02	75.75	79.85	68.44	72.93	69.72	73.69	75.12	70.21
Few-shot	Few-shot GPT	40.21	46.71	33.17	39.32	27.17	22.78	39.20	33.25	24.12	29.31
	P-ToD (This work)	<b>75.64</b>	<b>78.46</b>	<b>80.55</b>	<b>83.29</b>	<b>73.24</b>	<b>76.37</b>	<b>74.52</b>	<b>77.13</b>	<b>79.92</b>	<b>73.65</b>

**Figure 5: Performance of the P-ToD for different number of shots for all five tasks.**

best, and on the remaining three tasks, supervised-GPT shows the best performance.

It is critical to emphasize that the proposed P-ToD was trained using only 20 labeled training examples in phase three, whereas the supervised-GPT was trained on the complete training set. Moreover, we observe that PToD-0 variant (i.e., that was not trained in phase three) has comparable performance when compared to the SOTA personalization models. Last but not least, the few-shot GPT (that skipped phase two training and used only 20 training examples in phase three) baseline does not show good performance for task five as compared to other models.

**Personalization.** Table 3 presents BLEU-4 and ROUGE-2 scores for all the competing models on all five tasks. For all the tasks, the proposed P-ToD achieves the best performance or insignificant performance difference from supervised-GPT baseline. Excluding supervised-GPT model, the proposed P-ToD outperforms all other SOTA response generation methods by at least 19.95% on BLEU-4 and 9.74% on ROUGE-2 metrics. Similarly, the other variant PToD-0 that was not trained on any labeled training examples, still outperforms all the competing models including CoMemNN (which is a response selection model) for BLEU score. Since CoMemNN does not generate responses, it has advantage to get better BLEU and ROUGE scores as compared to the response generation approaches. Moreover, the few-shot GPT baseline shows the worst performance, since it was trained with only 20 labeled examples in the phase three and phase two (i.e., unsupervised personalization)

**Figure 6: Mean reward across unsupervised personalization phase for all five tasks.**

was skipped. The poor performance of the few-shot GPT baseline highlights the critical role of the phase two.

Figure 5 presents the performance of the proposed personalization framework, when provided with different number of training examples in phase three. Generally, we notice that as the number of training examples are increased, the performance improves, which highlights the importance of the supervision. However, we noticed that the performance does not get much better beyond 20 examples. That is almost the point, when P-ToD is as good as supervised-GPT model (i.e., trained on full training set).

The unsupervised personalization phase is at the core of the proposed framework, we provide more details about it in Figure 6. Since all five tasks vary in terms of difficulty, we present the mean reward of the models for each task, as the training progresses in phase two. The general trend is that the mean reward starts at 0 (e.g., at episode 0), which is obvious because the responses at the beginning of this phase were not tailored for the given user profile. Then, depending on the difficulty of the task, we notice that the respective models start approaching to 1.0 (e.g., after 100,000 episodes). We know that the task five (i.e., conduct full personalized dialog) is the most challenging task and the mean reward throughout the training process also signifies that. Similarly, we also notice that the tasks that involve adapting only linguistic styles (e.g., task two), the respective models start to achieve higher mean reward quickly as compared to the tasks that require meaningful recommendations or need to resolve nuances (e.g., task three).



**Table 4: Average scores of the user study.**

Method	Fluent	Appropriate	Rank
Reference Response	4.92	4.87	2.41
Supervised GPT	4.93	4.85	2.52
PToD-0 (This work)	4.91	4.86	2.62
P-ToD (This work)	4.92	4.85	2.45

## 5.2 Qualitative Analysis

In this experiment, we randomly selected 300 responses generated by supervised-GPT (i.e., the best model among the supervised competitors), PToD-0 (i.e., used zero labeled training examples), and P-ToD (i.e., used 20 labeled training examples) along with the reference responses and asked human annotators to rate them (i.e., 1 to 5, 5 being the best) for fluency and appropriateness of the response for the given user profile. Moreover, we also asked the annotators to rank the responses for personalization to the given user profile. Each response was rated by three annotators. Table 4 presents average scores for fluency, appropriateness of the response, and average rank among the responses. All the models (including reference) achieve high scores on the fluency and appropriateness of the response for the given user profile. Moreover, there is not a significant difference among the average scores. Similarly, almost all were ranked similar as reference responses. For example, responses generated from every model are ranked at all the places, i.e., 1<sup>st</sup> to 4<sup>th</sup> place. In summary, results from human study show that the responses of all the models are as good as reference responses. It is important to remind that the supervised-GPT was trained on the full training set, whereas our proposed PToD-0 and P-ToD were trained using zero and 20 labeled training examples, respectively.

We also observe that the PToD-0 model had slightly lower BLEU and ROUGE scores as compared to P-ToD and supervised-GPT, whereas in the human study it showed equally outstanding performance. Upon further investigation, we noticed that the responses generated by the PToD-0 are identical to that of supervised-GPT and P-ToD. The PToD-0 model did not use the “words” (or n-grams) in the reference responses. For example, a perfectly acceptable response generated by PToD-0, “What should the price be, madam?” did not get good BLEU or ROUGE scores, because the reference response happened to be, “Madam! which price range are you looking for?”.

## 6 RELATED WORK

The two broad categories of dialog systems are open-ended and task-oriented dialog systems. In the following, we summarize the personalization aspect of related work for both categories.

**Personalized Open-ended Dialogue Systems.** Among the earlier attempts to personalize open-ended dialog systems, [26] proposes learning interlocutor persona embeddings and adapting the conversation style accordingly. Researchers have since proposed a variety of methods, including persona information fusion [33, 64], multi-task learning [27], transfer learning [60, 65], meta learning [30], persona incorporation into the sequence-to-sequence framework [13, 26], persona-conditioned RNN-based model [12],

persona memory-conditioned variational autoencoders [51], response selection using memory networks [64], topical information usage [58], persona pre-training [15, 66], and extra training procedures for personalization [15, 38]. While many of these works have proven useful for assigning personalities or language styles to open-ended dialog systems, they are ineffective for task-oriented dialog systems. We propose that, rather than assigning personalities to agents (i.e., dialog systems), make them more adaptive to their different kinds of interlocutors in task-oriented dialog settings.

**Personalized Task-oriented Dialogue Systems.** Comparatively to open-domain dialog systems, personalized task-oriented dialog systems are under-explored. In fact, to the best of our knowledge, personalized bAbI dialogue [18] is the only publicly available benchmark for the evaluation of task-oriented dialog systems. Most of the existing work [18, 28, 31, 37, 57] use memory networks by concatenating profile information and dialog memory [18], combining style information [28], query generation via RNN-based decoder [31], local and global encoders [57]. Similarly, cooperative memory network have been proposed [37] to handle the case, where only partial profile information is available. All of these works follow supervised learning approaches and require a large amount of labeled training data for each user profile. In contrast to previous work, we employ deep reinforcement learning to personalize task-oriented dialog systems in the unsupervised setting without requiring any labeled training data. This work leverages pre-trained language models and zero-shot learning for natural language understanding and generation, and adapts its responses to a wide range of user profiles in unsupervised way. Nonetheless, it is noteworthy to mention that several key ideas leveraged in this work have been used for task-oriented dialog systems such as deep reinforcement learning for dialog policy generation [23, 24] and paraphrasing [50], zero-shot learning for intent detection [49] and slot filling [48], and language models for anaphora resolution [32] and response generation [11]. However, none of these works have proposed to personalizing dialog systems in the unsupervised setting.

## 7 CONCLUSION

We have presented a novel personalization framework for task-oriented dialog systems, P-ToD, that can seamlessly adapt to newly emerging unseen user profiles in the unsupervised fashion. P-ToD stands out as the first unsupervised framework for personalized task-oriented dialog systems that can effectively adapt its conversation flows and linguistic styles, disambiguate nuances, and make meaningful recommendations according to the profile of the active user. The key idea behind the proposed framework is using a novel zero-shot generalizable reward function that guides the policy of the personalized model to adapt its responses for the given user without compromising the task completion accuracy. Our experimental evaluation uses up to 180 diverse user profiles for five tasks including conducting full personalized dialogs. Interestingly, our proposed framework outperforms all the existing personalization models using quantitative as well as qualitative analysis. Furthermore, we also trained a fully supervised-GPT model for comparison and it turned out that P-ToD, trained using only 20 labeled training examples, achieves better or competitive performance.

## REFERENCES

- [1] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research* 47 (2013), 253–279.
- [2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3 (2003), 1137–1155.
- [3] Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683* (2016).
- [4] R Brown. 1987. Theory of politeness: An exemplary case. In *meeting of the Society of Experimental Social Psychologists, Charlottesville, VA*.
- [5] Robert A Brown. 1965. Work decrement, kinesthetic aftereffect, and personality. *Journal of Personality and Social Psychology* 2, 6 (1965), 868.
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [7] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter* 19, 2 (2017), 25–35.
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [9] Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. 2017. OpenAI Baselines. <https://github.com/openai/baselines>.
- [10] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2021. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. *arXiv preprint arXiv:2112.06905* (2021).
- [11] Umar Farooq, AB Siddique, Fuad Jamour, Zhijia Zhao, and Vagelis Hristidis. 2020. App-aware response synthesis for user reviews. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 699–708.
- [12] Jessica Fidler and Yoav Goldberg. 2017. Controlling Linguistic Style Aspects in Neural Language Generation. *EMNLP 2017* (2017), 94.
- [13] Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. *arXiv preprint arXiv:1603.08148* (2016).
- [14] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.
- [15] Jonathan Herzig, Michal Shmueli-Scheuer, Tommy Sandbank, and David Konopnicki. 2017. Neural response generation for customer service based on personality traits. In *Proceedings of the 10th International Conference on Natural Language Generation*. 252–256.
- [16] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems* 33 (2020), 20179–20191.
- [17] Eduard H Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated Summarization Evaluation with Basic Elements. In *LREC*, Vol. 6. Citeseer, 899–902.
- [18] Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503* (2017).
- [19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [20] Rolf O Kroger and Linda A Wood. 1992. Are the rules of address universal? iv: Comparison of chinese, korean, greek, and german usage. *Journal of cross-cultural psychology* 23, 2 (1992), 148–162.
- [21] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale Reading Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 785–794.
- [22] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043* (2017).
- [23] Nhat Le, AB Siddique, Fuad Jamour, Samet Oymak, and Vagelis Hristidis. 2021. Generating Predictable and Adaptive Dialog Policies in Single-and Multi-domain Goal-oriented Dialog Systems. *International Journal of Semantic Computing* 15, 04 (2021), 419–439.
- [24] Nhat Le, AB Siddique, Fuad Jamour, Samet Oymak, and Vagelis Hristidis. 2021. Predictable and adaptive goal-oriented dialog policy generation. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*. IEEE, 40–47.
- [25] Harrison Lee, Raghav Gupta, Abhinav Rastogi, Yuan Cao, Bin Zhang, and Yonghui Wu. 2022. SGD-X: A Benchmark for Robust Generalization in Schema-Guided Dialogue Systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [26] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 994–1003.
- [27] Yi Luan, Chris Brockett, William B Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-Task Learning for Speaker-Role Adaptation in Neural Conversation Models. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 605–614.
- [28] Liangchen Luo, Wenhao Huang, Qi Zeng, Zaiqing Nie, and Xu Sun. 2019. Learning personalized end-to-end goal-oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6794–6801.
- [29] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 142–150.
- [30] Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5454–5459.
- [31] Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. *arXiv preprint arXiv:1804.08217* (2018).
- [32] Muhammad Hasan Maqbool, Luxun Xu, AB Siddique, Niloofar Montazeri, Vagelis Hristidis, and Hassan Foroosh. 2022. Zero-label Anaphora Resolution for Off-Script User Queries in Goal-Oriented Dialog Systems. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*. IEEE, 217–224.
- [33] Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training Millions of Personalized Dialogue Agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2775–2779.
- [34] Kaixiang Mo, Yu Zhang, Shuangyin Li, Jiajun Li, and Qiang Yang. 2018. Personalizing a dialogue system with transfer reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [35] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 280–290.
- [36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.
- [37] Jiahuan Pei, Pengjie Ren, and Maarten de Rijke. 2021. A cooperative memory network for personalized task-oriented dialogue systems with incomplete user profiles. In *Proceedings of the Web Conference 2021*. 1552–1561.
- [38] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. *arXiv preprint arXiv:1706.02861* (2017).
- [39] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).
- [40] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don’t Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 784–789.
- [41] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392.
- [42] Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266.
- [43] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*. 3982–3992.
- [44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.
- [45] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*. PMLR, 1889–1897.
- [46] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [47] Rico Senrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* (2015).
- [48] AB Siddique, Fuad Jamour, and Vagelis Hristidis. 2021. Linguistically-enriched and context-aware zero-shot slot filling. In *Proceedings of the Web Conference 2021*. 3279–3290.
- [49] AB Siddique, Fuad Jamour, Luxun Xu, and Vagelis Hristidis. 2021. Generalized zero-shot intent detection via commonsense knowledge. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1925–1929.

- [50] AB Siddique, Samet Oymak, and Vagelis Hristidis. 2020. Unsupervised paraphrasing via deep reinforcement learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1800–1809.
- [51] Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. *arXiv preprint arXiv:1905.12188* (2019).
- [52] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MpNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems* 33 (2020), 16857–16867.
- [53] Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847* (2018).
- [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [55] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 353–355.
- [56] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 438–449.
- [57] Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. *arXiv preprint arXiv:1901.04713* (2019).
- [58] Minghong Xu, Piji Li, Haoran Yang, Pengjie Ren, Zhaochun Ren, Zhumin Chen, and Jun Ma. 2020. A neural topical expansion framework for unstructured persona-oriented dialogue generation. *arXiv preprint arXiv:2002.02153* (2020).
- [59] Min Yang, Qiang Qu, Kai Lei, Jia Zhu, Zhou Zhao, Xiaojun Chen, and Joshua Z Huang. 2018. Investigating deep reinforcement learning techniques in personalized dialogue generation. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 630–638.
- [60] Min Yang, Zhou Zhao, Wei Zhao, Xiaojun Chen, Jia Zhu, Lianqiang Zhou, and Zigang Cao. 2017. Personalized response generation via domain adaptation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1021–1024.
- [61] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).
- [62] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics, Online, 109–117. <https://doi.org/10.18653/v1/2020.nlp4convai-1.13>
- [63] Bowen Zhang, Xiaofei Xu, Xutao Li, Yunming Ye, Xiaojun Chen, and Lianjie Sun. 2019. Learning personalized end-to-end task-oriented dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 55–66.
- [64] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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)*. 2204–2213.
- [65] Wei-Nan Zhang, Qingfu Zhu, Yifa Wang, Yanyan Zhao, and Ting Liu. 2019. Neural personalized response generation as domain adaptation. *World Wide Web* 22, 4 (2019), 1427–1446.
- [66] Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9693–9700.
- [67] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. 19–27.