Rewarding What Matters: Step-by-Step Reinforcement Learning for Task-Oriented Dialogue

Huifang Du[1*], **Shuqin Li**[2*], **Minghao Wu**[3], **Xuejing Feng**[1],
Yuan-Fang Li[3†], **Haofen Wang**[1†]

[1]Tongji University [2]Hangzhou Dianzi University Information Engineering College [3]Monash University
{duhuifang, fengxuejing}@tongji.edu.cn
{shuqinlee9683, carter.whfcarter}@gmail.com
{minghao.wu, yuanfang.li}@monash.edu

Abstract

Reinforcement learning (RL) is a powerful approach to enhance task-oriented dialogue (TOD) systems. However, existing RL methods tend to mainly focus on generation tasks, such as dialogue policy learning (DPL) or response generation (RG), while neglecting dialogue state tracking (DST) for understanding. This narrow focus limits the systems to achieve globally optimal performance by overlooking the interdependence between understanding and generation. Additionally, RL methods face challenges with sparse and delayed rewards, which complicates training and optimization. To address these issues, we extend RL into both understanding and generation tasks by introducing step-by-step rewards throughout the token generation. The understanding reward increases as more slots are correctly filled in DST, while the generation reward grows with the accurate inclusion of user requests. Our approach provides a balanced optimization aligned with task completion. Experimental results demonstrate that our approach effectively enhances the performance of TOD systems and achieves new state-of-the-art results on three widely used datasets, including MultiWOZ2.0, MultiWOZ2.1, and In-Car. Our approach also shows superior few-shot ability in low-resource settings compared to current models.

1 Introduction

The rapid advancements in pre-trained language models (PLMs) have significantly influenced a variety of real-world applications (Devlin et al., 2018; Raffel et al., 2020; Chung et al., 2024). Among these, the development of task-oriented dialogue (TOD) systems stands out as particularly impactful (Wen et al., 2017; Hosseini-Asl et al., 2020). Typically, a TOD system comprises several components

---

* These authors contributed equally.
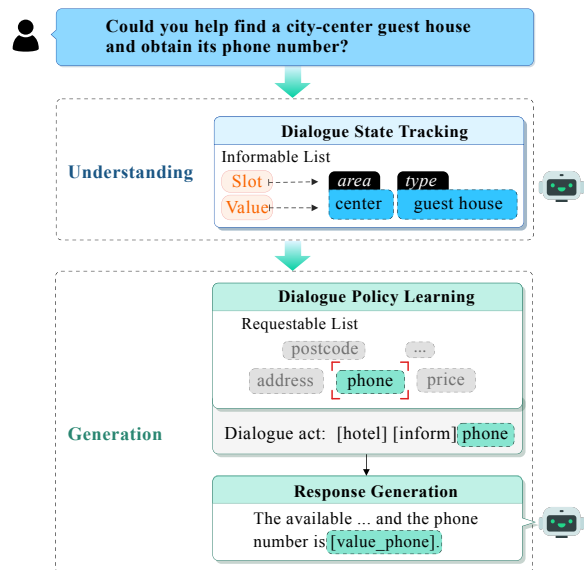† Corresponding authors.

Figure 1: A task-oriented dialogue system needs to successfully perform both understanding and generation to achieve its dialogue goals.

(He et al., 2022b; Feng et al., 2023) as shown in Figure 1, including dialogue state tracking (DST) for understanding user's belief state (Chen et al., 2020; Guo et al., 2023), dialogue policy learning (DPL) for generating dialogue acts (Zhao et al., 2024; Zhang et al., 2019), and response generation (RG) for generating system responses (Pei et al., 2020; Chen et al., 2019). More recently, there has been growing interest in constructing end-to-end (E2E) TOD systems based on PLMs to equip models with all these essential capabilities (He et al., 2022b; Hosseini-Asl et al., 2020; Feng et al., 2023; Yu et al., 2023).

Building on the advancements in TOD systems discussed earlier, recent research explores the use of offline reinforcement learning (RL) to optimize TOD systems further learning goal-oriented conversational strategies (Lu et al., 2019; Jang et al., 2022; Feng et al., 2023). However, current RL approaches typically focus on enhancing the gen-

eration component, such as generating dialog acts (DPL task) (Li et al., 2023) or system response (RG task) (Yu et al., 2023). This biased focus prevents the systems from reaching optimal performance by ignoring the crucial interdependence between understanding and generation. Furthermore, RL for TOD systems often faces issues with sparse and delayed rewards (Lu et al., 2019; Abdulhai et al., 2023), which are only provided upon reaching the goal at the dialogue or turn level (Kwan et al., 2023; Lu et al., 2019; Abdulhai et al., 2023). This leads to insufficient exploration and unstable training for RL. While many efforts have tried to mitigate these reward issues to offer dense rewards, the design of the reward function in these methods tends to be complex, which may limit the method's generalization (Li et al., 2020; Feng et al., 2023).

In this work, we propose to design a simple but effective reward function to jointly optimize both *understanding* and *generation* components in an *end-to-end* manner to achieve the globally optimal performance. We propose the combination of understanding reward and generation reward throughout per token generation to reinforce the learning step by step. The understanding reward is the growing proportion of correctly filled slots in the DST process, while the generation reward is measured by the correct inclusion of the user requests in the DPL and RG process. We conduct extensive experiments using two model backbones, the Flan-T5 base and Flan-T5 large models (Chung et al., 2024), on three widely used benchmarks: MultiWOZ2.0, MultiWOZ2.1, and In-Car. The results show that our approach significantly improves model performance against strong baselines, establishing new state-of-the-art results. We also show that our approach outperforms current models in low-resource conditions, highlighting its adaptability in real-world scenarios where data is limited.

Our contributions to this work are summarized as follows:

- We introduce a novel approach that integrates RL into both understanding (DST) and generation (DPL and RG) components in an end-to-end manner, which promotes a balanced optimization for TOD systems.

- To tackle the challenges of sparse and delayed rewards in RL for TOD systems, we propose a combined reward mechanism that provides progressive feedback during token generation. This step-by-step reward significantly enhances efficiency.

- Experimental results show that our approach establishes new state-of-the-art results on multiple benchmarks (MultiWOZ2.0, MultiWOZ2.1, and In-Car). Furthermore, the method shows superior performance in low-resource conditions.

2 Related Work

In this section, we review works on TOD systems utilizing both pipeline and E2E methods, the integration of reinforcement learning (RL), and the design of reward functions for RL. Additionally, we discuss the role of large language models (LLMs) in TOD systems.

Pipeline and End-to-End Approaches. Pipeline approaches are characterized by their modular structure, where dialogue state tracking (DST) (Chen et al., 2020; Guo et al., 2023), dialogue policy learning (DPL) (Zhao et al., 2024; Zhang et al., 2019), and response generation (RG) (Pei et al., 2020; Chen et al., 2019) are processed sequentially. They offer interpretability and modularity but often struggle to capture the overall context of conversations (Kwan et al., 2023). In contrast, E2E approaches directly map input utterances to system responses without explicit intermediate representations (He et al., 2022b; Yang et al., 2021; He et al., 2022a). Some models, such as SPACE-3 (He et al., 2022a), UBAR (Yang et al., 2021), and PPTOD (Su et al., 2022), restructure all sub-tasks into a single sequence prediction through pre-training and fine-tuning. However, supervised fine-tuning (SFT) focuses more on learning at the token level than on the particular requirements, which limits the model's ability to complete specific tasks.

RL-Based Policy Learning. RL can be leveraged to enhance model performance by tailoring it to the specific requirements of TOD tasks. Some studies utilize user simulators (Peng et al., 2018; Jang et al., 2022; Liu et al., 2023) for multi-turn dialogue RL exploration to improve policies. Yet, user simulators require careful design and can introduce biases if not accurately mimicked. Hierarchical RL (HRL) methods break tasks into sub-tasks, creating a policy hierarchy (Peng et al., 2017; Liu et al., 2020), while feudal RL (FRL) abstracts state and action spaces for more general policies (Gao et al., 2018; Casanueva et al., 2018). These HRL methods

primarily focus on complex algorithmic designs and often lack a robust understanding of user intentions, resulting in suboptimal performance. We aim to explore an offline RL approach and focus on token-level RL exploration to mitigate the issues above.

Reward Design for TOD. Recent studies have found offline RL to be a promising method for stabilizing training with static datasets (Snell et al., 2023; Feng et al., 2023). Following the offline principle, many methods design rewards at the dialog and turn level when a goal is achieved (Kwan et al., 2023; Lu et al., 2019; Tang et al., 2018), but reward signals remain sparse. Inverse reinforcement learning (IRL) and reward shaping techniques have been introduced to learn denser rewards and encourage faster learning (Li et al., 2020; Takanobu et al., 2019). However, IRL can be computationally intensive, and reward shaping might result in unintended behaviors if not carefully designed (Arora and Doshi, 2021; Gupta et al., 2024). Alternatively, some methods employ rewards for every token, which may lack semantic significance towards the dialogue goal (Yu et al., 2023; Gupta et al., 2024). Our approach provides progressive rewards directly towards the dialogue goal.

Large Language Models for TOD. LLMs have demonstrated impressive capabilities in understanding and generating text for various tasks (Ouyang et al., 2022; OpenAI, 2023; Chowdhery et al., 2023; Wu et al., 2024c). However, LLMs underperform compared to specialized task-specific models (Hudeček and Dušek, 2023; Li et al., 2023; Wu et al., 2024b). Fine-tuning LLMs for specific tasks is also computationally inefficient. All these reasons lead to a growing interest in prompt engineering approaches that leverage in-context learning without requiring parameter updates (Wei et al., 2022; Wang et al., 2022; Yao et al., 2024; Wu et al., 2024a). Yet, LLMs still tend to perform less effectively (Yang et al., 2024).

3 Preliminary

3.1 Supervised Fine-Tuning for TOD

The TOD task is typically modeled as an E2E problem and addressed by a seq2seq model (e.g. T5) using supervised fine-tuning (SFT). The input of the model can be represented as $I_t = [\text{prefix} : u_{t-1} : bs_{t-1} : da_{t-1} : sr_{t-1} : u_t]$, where $[\cdot : \cdot]$ denotes the concatenation operator, u_t represents the current user utterance, bs_{t-1}, da_{t-1}, and sr_{t-1} represent the belief state (BS), dialogue act (DA), and system response (SR) at turn $t-1$ respectively. The prefix instruction is "translate dialogue to belief state, dialogue action, and system response: [input]". The model is fine-tuned to maximize the likelihood of successively generating correct BS, DA, and SR given the input:

$$\mathcal{L}_\theta = \sum_{t=1}^{T} \log P(bs_t, da_t, sr_t \mid I_t; \theta), \quad (1)$$

where θ represents the parameters of the model.

3.2 Reinforcement Learning for TOD

Formally, the RL approaches for TOD tasks operate within a Markov Decision Process (MDP) (Kaelbling et al., 1998) characterized by the tuple $\langle S, A, P, R, \gamma \rangle$. The state space S can be represented as a set of states $\mathbf{s}_i = \{s_1, s_2, \ldots, s_k\}$, where each state includes the dialogue context and history up to the current time step. Each turn in the dialogue is considered an independent episode. An action $a_{\Delta t} \in A$ is the Δt-th action taken during an episode, which corresponds to selecting the next token in the dialogue. Transition probability $P(s' \mid s, a)$ is the probability of transitioning to state s' given action a and state s. The discount factor $\gamma \in [0, 1]$ is used to weigh future rewards. The SFT model is used to initialize a policy network π, which is subsequently refined to maximize the reward R, using algorithms such as proximal policy optimization (PPO) (Schulman et al., 2017).

4 Main Method

We aim to enhance TOD systems using a combination of SFT and RL. While SFT can provide a stable initial base for RL (Ramamurthy et al., 2023; Yu et al., 2023; Li et al., 2023), it equally treats every ground-truth token as an objective, without prioritizing task-specific goals. We utilize RL to refine the model to optimize for task completion.

In TOD tasks, accurately understanding user needs (i.e., belief states) is crucial for generating appropriate dialogue acts, which are essential for producing system responses that meet current needs and effectively drive the conversation forward. However, existing RL methods often focus solely on optimizing dialogue policy learning (Li et al., 2023; Takanobu et al., 2020) or response
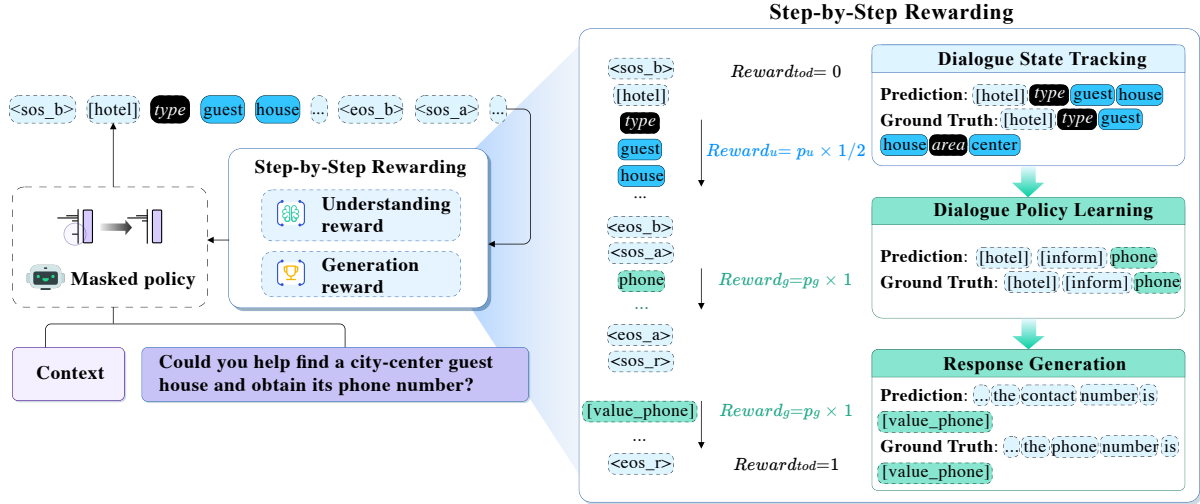
Figure 2: Overview of our approach. **Left**: We use the masked policy to optimize understanding and generation end-to-end with our reward function for TOD systems. Context is the concatenation of belief state (BS), dialogue act (DA), and system response (SR) at the previous turn. Special characters like <sos_b>, <sos_a>, and <sos_r> denote the start of BS, DA, and SR, while <eos_b>, <eos_a>, and <eos_r> denote their endings. **Right**: The designed reward function provides step-by-step rewards for understanding and generation tasks. $Reward_u$ refers to Equation 2, $Reward_g$ refers to Equation 3, and $Reward_{tod}$ refers to Equation 4.

generation (Yu et al., 2023), neglecting the importance of understanding and the interdependence between understanding and generation. Moreover, these methods typically use sparse rewards at the dialogue or turn level (Kwan et al., 2023; Lu et al., 2019; Tang et al., 2018; Abdulhai et al., 2023).

Task completion metrics evaluate whether the model correctly generates informable and requestable slot values defined in the dialogue schema, reflecting its performance in understanding and generation tasks. The policy model's sequence generation process involves continually satisfying these lists. Inspired by these metrics, we hypothesize that providing progressive task-oriented rewards during token generation for understanding and generation tasks can enhance TOD systems. The model architecture and our reward function are illustrated in Figure 2. In Section 4.1, we explain how these metrics are measured to support our reward function design. In Section 4.2, we show how our reward function provides continuous, step-by-step feedback, guiding the E2E model through understanding and generation tasks for a more coherent and responsive dialogue system.

4.1 Task Completion Metrics

An *informable* list and *requestable* list are commonly predefined for dialog goals in datasets, such as In-Car and MultiWOZ. The *informable* list contains slots and their values representing the user's

requirements. For example, a user's preference for a restaurant is characterized by a "cheap" value for the "price range" slot. The Inform metric evaluates whether the system accurately learns user demands as defined in the *informable* list and then provides a suitable entity in response. The *requestable* list includes user-requested values, such as "postcode". The Success metric measures whether the generated DAs or SRs contain all attributes in the requestable list. Therefore, we believe that a slot-value-specific reward derived from the *informable* list can enhance the system's understanding of user needs, while the value-specific reward based on the requestable list can improve responsiveness to user requests. Accordingly, we introduce the design of a progressive reward function combining the *understanding* reward for DST, as well as the *generation* reward for DPL and RG.

4.2 Step-by-Step Goal-Oriented Reward

Understanding Reward. We design the understanding reward for DST by measuring the growing proportion of correctly identified slot-value pairs in the *informable* list during token (action) generation. This reward function directly reflects how well the system understands the user's needs, which is closely related to the goals of DST. Formally, we denote SV_{gt} as the set of ground-truth slot values in the current turn and $\hat{SV}$ as the set of predicted ones during the token generation:

$$R_u = \frac{|SV_{gt} \cap \hat{SV}| \cdot \rho_u}{|SV_{gt}|}, \quad (2)$$

where $\rho_u = \exp\left(-\alpha \cdot \frac{|SV_{gt} \setminus \hat{SV}|}{|SV_{gt}|}\right)$ represents a penalty based on the discrepancy between the number of predicted slot-value pairs and ground truth slot-value pairs, with α being a tunable parameter that controls the sensitivity of this penalty. The function provides a dense reward that progressively reflects the accuracy of DST.

Generation Reward. We observe that the accuracy of both DPL and RG depends on how many values in their generations are correctly included in the *requestable* list. Therefore, we set the same reward function for these two generation tasks. The reward for DPL and RG is the increasing inclusion of values in the user *requestable* list during each token generation, which measures the system's ability to fulfill user requests continuously. Formally, S_{gt} is all ground-truth user request values in the current turn, and $\hat{S}$ denotes the predicted values during token generation:

$$R_g = \frac{|S_{gt} \cap \hat{S}| \cdot \rho_g}{|S_{gt}|}, \quad (3)$$

where the penalty term $\rho_g = \exp\left(-\beta \cdot \frac{|S_{gt} \setminus \hat{S}|}{|S_{gt}|}\right)$ peralizes the difference between the number of generated values and values in the ground-truth requestable list, and β is a tunable parameter that controls the sensitivity of this penalty. The function provides a dense reward that progressively reflects how well the generation completes.

TOD Reward. To offer a comprehensive reward that evaluates both the understanding and generation performance, we define the TOD reward as a weighted combination of the understanding reward R_u and the generation reward R_g:

$$R_{tod} = \frac{|SV_{gt} \cap \hat{SV}| \cdot \rho_u + |S_{gt} \cap \hat{S}| \cdot \rho_g}{|SV_{gt}| + |S_{gt}|}. \quad (4)$$

The combined reward function encourages balanced optimization of both the understanding (DST) and the generation (DPL, RG), which enhances the global robustness of TOD systems. The use of dense rewards derived from the informable and requestable lists ensures continuous feedback during token-level generation. Unlike sparse rewards that only provide feedback at the end of dialogues, our approach offers step-by-step rewards,

accelerating the learning process. The progressive nature of the rewards, based on the discrepancies ρ_u and ρ_g, helps make incremental improvements.

Reward Shaping. To prevent the policy network π from straying too far from the initial model π_o, we also add a KL constraint to balance the reward. Formally, the final RL reward function is:

$$R_{total} = R_t - \beta D_{KL}(\pi \parallel \pi_o), \quad (5)$$

where β is dynamically adapted during training.

Optimization. We use natural language policy optimization (NLPO) (Ramamurthy et al., 2023), which is an extension of PPO. NLPO incorporates action elimination through a parameterized-masked approach. It learns to mask out less relevant tokens using top-p sampling, which restricts the token set to those with a cumulative probability above a specified threshold. NLPO maintains a separate masked policy that updates periodically, providing an additional constraint to ensure the selection of more task-relevant actions.

5 Experiments

5.1 Dataset

We conduct experiments on two popular task-oriented dialog benchmarks: MultiWOZ (Budzianowski et al., 2018; Eric et al., 2020), and In-Car Assistant (In-Car) (Eric et al., 2017). The MultiWOZ datasets are a challenging benchmark for evaluating TOD systems with seven domains: attraction, hotel, hospital, police, restaurant, taxi, and train. The dataset is split into 8,438 dialogues for training, and 1,000 each for validation and testing. We use two versions, i.e. MultiWOZ2.0 and MultiWOZ2.1, to evaluate our model. The In-Car dataset comprises 3,031 multi-turn dialogues across three specific domains suitable for an in-car assistant: calendar scheduling, weather information retrieval, and point-of-interest navigation. The dialogues in In-Car are more natural and diverse. We split the dataset into training/validation/testing sets containing 2425/302/304 dialogs respectively as previous works do. Following the data pre-processing procedure from (Zhang et al., 2020b), delexicalized responses are utilized in our work to help the model learn generalizable parameters.

5.2 Evaluation Metrics

In this work, we evaluate our models on MultiWOZ and In-Car benchmarks as described in Section 5.1.

Method	MultiWOZ2.0				MultiWOZ2.1				In-Car			
	Inform	Succ.	BLEU	Comb.	Inform	Succ.	BLEU	Comb.	Match	SuccF1	BLEU	Comb.
E2E												
SimpleTOD	84.4	70.1	15.0	92.3	85.0	70.5	15.2	93.0	-	-	-	-
DoTS	86.6	74.1	15.1	95.5	86.7	74.2	15.9	96.3	-	-	-	-
PPTOD	89.2	79.4	18.6	102.9	87.1	79.1	19.2	102.3	-	-	-	-
UBAR[†]	85.1	71.0	16.2	94.3	86.2	70.3	16.5	94.7	-	-	-	-
LABES	-	-	-	-	76.9	63.3	17.9	88.0	85.8	77.0	22.8	104.2
SPACE-3[**]	88.7	78.7	16.3	100.0	90.9	81.0	16.8	102.7	84.7	79.6	18.6	100.7
SPACE-3	95.3	88.0	19.3	111.0	95.6	86.1	19.9	110.8	85.3	83.2	22.9	107.1
GALAXY*	93.1	81.0	18.4	105.5	93.5	81.7	18.3	105.9	81.9	83.3	22.0	104.6
GALAXY	94.4	85.3	**20.5**	110.4	95.3	86.2	**20.0**	110.8	85.3	83.6	**23.0**	107.4
RL												
MinTL	84.9	74.9	17.9	97.8	-	-	-	-	-	-	-	-
GPT-Critic	90.1	76.6	17.8	101.1	-	-	-	-	-	-	-	-
FanReward	93.1	83.9	18.0	106.5	-	-	-	-	-	-	-	-
Ours$_{base}$	92.1	88.3	16.6	106.9	92.7	88.5	16.2	106.8	84.3	83.8	22.8	106.9
Ours$_{large}$	**96.1**	**92.4**	17.2	**111.5**	**96.9**	**91.1**	16.9	**110.9**	**86.2**	**86.1**	23.0	**109.2**

Table 1: Performance comparison on MultiWOZ2.0, MultiWOZ2.1 and In-Car datasets. †: The results of UBAR are obtained using the models provided by the authors. ∗: The results of GALAXY* are presented without pre-training. ∗∗: The results of SPACE-3** are results without pre-training, reimplemented using their public code.

For MultiWOZ, we report `Inform` and `Success` as introduced in Section 4.1. Additionally, we report `BLEU` (Papineni et al., 2002) that is used to measure the fluency of the generated response. Consequently, we report (`Comb`) that is computed by (`Inform` + `Success`) ×0.5 + `BLEU` as an overall quality measure. For In-Car, we leverage `Match` to measure if a system can track all correct states to satisfy the user. `SuccF1` improves on the `Success` by considering both how completely (recall) and accurately (precision) the system handles requests.

Both `Inform` and `Match` evaluate the system's understanding of user requirements, but `Inform` further focuses on providing correct entities based on the understanding, while `Match` ensures accurate dialogue state tracking. Similarly, both `Success` and `SuccF1` assess the system's ability to fulfill user requests, but `Success` measures whether all user requests are met, whereas `SuccF1` balances precision and recall to gauge response accuracy and completeness. The design of our reward function aligns with all these task completion metrics for the dialogue datasets.

5.3 Baselines

We comprehensively evaluate our approach by comparing it with a wide array of methods on the MultiWOZ2.0 and MultiWOZ2.1 datasets. We select several prominent E2E models as baselines, including SimpleTOD (Hosseini-Asl et al., 2020), DoTS (Jeon and Lee, 2021), PPTOD (Su et al., 2022), UBAR (Yang et al., 2021), GALAXY (He

et al., 2022b), and SPACE-3 (He et al., 2022a). Furthermore, we compare our approach with current representative reinforcement learning (RL) models, including MinTL (Lin et al., 2020), GPT-Critic (Jang et al., 2022), and FanReward (Feng et al., 2023). These approaches demonstrate the potential of RL for TOD models and offer valuable comparative perspectives. We compare our approach against several strong baselines on the In-Car dataset, including LABES (Zhang et al., 2020a), SPACE-3 (He et al., 2022a), and GALAXY (He et al., 2022b). Additionally, we present results for GALAXY and SPACE-3 without pre-training across the three datasets to provide a comprehensive evaluation. Two pre-trained models, the Flan-T5-Base and Flan-T5-Large (Chung et al., 2024), are utilized as the backbone of our approach (see Appendix A).

5.4 Main Results

As shown in Table 1, our approach achieves new state-of-the-art results across all datasets in the combined score (`Comb`). These gains are primarily due to the increased `Inform` and `Success` rates. We achieve competitive `BLEU` scores on Multi-WOZ, possibly because our approach focuses on understanding and responsiveness over fluency. However, our model performs best on the In-Car dataset, demonstrating its ability to generate fluent responses. Additionally, our model is compared with two strong baselines, SPACE-3 and GALAXY, both of which utilize multiple TOD datasets for
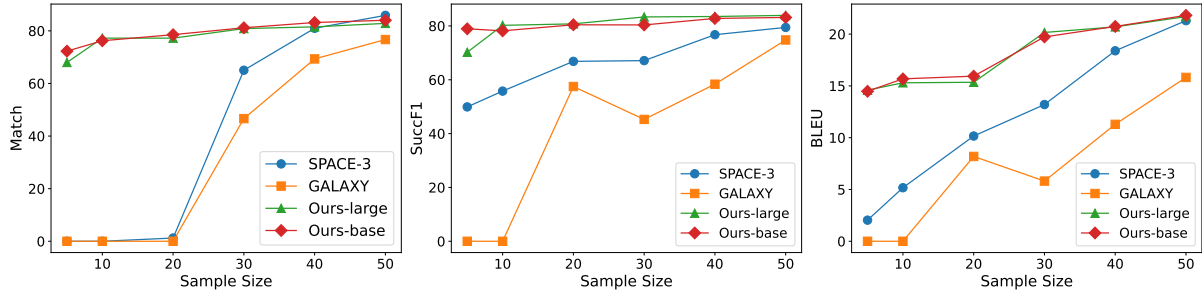
Figure 3: Results of low-resource experiments. 5% (121 dialogues), 10% (242 dialogues), 20% (485 dialogues), 30% (727 dialogues), 40% (970 dialogues), and 50% (1212 dialogues) of training data is used to train each model. Results are shown as mean values over five runs.

pre-training and are subsequently fine-tuned on the MultiWOZ and In-Car datasets. Our approach, without a pre-training step, demonstrates a significantly greater improvement compared to their results without pre-training. For instance, on the MultiWOZ2.0 dataset, our method achieves an increase of +3 points in the `Inform` rate and +11.4 points in the `Success` rate. This indicates that our approach, which offers step-by-step rewards for RL, effectively helps boost TOD task completion.

5.5 DST Results

We evaluate our approach on the DST task, a multi-turn slot-filling challenge where the model must track slot-value pairs throughout the conversation. We use joint goal accuracy (JGA) as the evaluation metric (Heck et al., 2020), which measures the percentage of turns in which the model correctly predicts all slot values. The experiments are conducted on the MultiWOZ 2.0 dataset, and we benchmark our method against a diverse set of recent approaches. The results, presented in Table Table 2, show that our method achieves the highest JGA score, showing its superior effectiveness.

Method	JGA
Neural Reading (Gao et al., 2019)	41.10
TRADE (Wu et al., 2019)	48.62
COMER (Ren et al., 2019)	48.79
DSTQA (Zhou and Small, 2019)	51.44
SOM-DST (Kim et al., 2020)	51.38
MinTL (Lin et al., 2020)	52.10
UBAR (Yang et al., 2021)	52.59
SOLOIST (Peng et al., 2021)	53.20
Ours$_{base}$	53.54
Ours$_{large}$	**53.92**

Table 2: DST comparison results of various methods.

5.6 Ablation Study

We conduct an ablation study on the MultiWOZ2.0 dataset to evaluate the effectiveness of our progressive goal-oriented reward mechanism. As shown in Table 3, when the understanding reward R_u is removed, leaving only the generation reward, the combined score drops significantly to 105.2. This substantial decrease highlights the crucial role of immediate feedback during dialogue state tracking. When we remove the generation reward R_g, the combined score decreases by 6.1 points. This suggests that the generation reward is also important for task completion. If R_u and R_g are all removed, that is the result of only SFT, the performance degrades to a lower value. Overall, the study demonstrates that our progressive reward mechanism greatly improves the system's ability to perform understanding and generation tasks.

6 Analysis and Discussion

In this section, we first explore how well our model performs in low-resource settings (Section 6.1). Besides, we show that our approach can be integrated into recent state-of-the-art LLMs for better performance (Section 6.2). Lastly, we conduct human evaluation on the generated outputs (Section 6.3).

6.1 Low-Resource Evaluation

Due to the challenge of creating extensive, well-annotated dialogue datasets for real-world applications, we also explore the performance of our approach with limited training samples. We train models using the In-Car dataset and randomly sample 5%, 10%, 20%, 30%, 40%, and 50% of the training data. Our approach is benchmarked against two robust baselines, SPACE-3 and GALAXY. Both models are initialized with their pre-trained versions and subsequently fine-tuned using the sam-

Model	Inform	Succ.	BLEU	Comb.
Ours	96.1	92.4	17.2	111.5
$-R_u$	91.2	87.0	16.1	105.2
$-R_g$	92.1	87.5	15.6	105.4
$-R_u - R_g$	86.0	81.8	17.2	101.1

Table 3: Ablation results on MultiWOZ2.0.

pled datasets from the In-Car dataset. Our model is trained through SFT and RL stages both with the sampled data. To ensure fairness, all models are trained for 30 epochs. Figure 3 presents the experimental results. As illustrated, our approach consistently outperforms the baselines across all sample sizes on the metrics of Match, SuccF1, and BLEU. The performance advantage is especially prominent when the training data is limited. This suggests that our model demonstrates enhanced generalizability and is more apt for tackling new TOD tasks. It is noted that results between Ours-large and Ours-base are similar. This may be because, in a low-data setting, the smaller model (Ours-base) better utilizes the available data. In contrast, the larger model (Ours-large) may not have been fully trained with the limited data, leading to no significant performance improvement.

6.2 Integration with LLMs

Recently, LLMs have led to remarkable advancements in NLP, demonstrating impressive emergent abilities. However, LLMs often underperform compared to specialized models for TOD tasks (Hudeček and Dušek, 2023; Li et al., 2023). We utilize few-shot dialogue examples of the training set of MultiWOZ2.0 to prompt LLMs. As shown in the top half of Table 4, the powerful representatives of LLMs, including Codex (code-davinci-002) (Chen et al., 2021), ChatGPT (gpt-3.5-turbo) (Ouyang et al., 2022), Claude (Claude 3 sonnet) (Anthropic, 2024), and GPT-4o[1], do not perform as well as our model. Additionally, fine-tuning LLMs for TOD systems can also be resource-intensive and computationally inefficient. There has been a surge of research interest that combines LLMs with small models for specific applications. We follow the recent work DSP (Li et al., 2023), which utilizes a small tunable model for dialogue policy learning to generate dialogue acts. The dialogue acts are used as hints to prompt LLMs to generate the final response. To demonstrate the advantages of our approach, we employ our generation reward func-

---

[1] https://openai.com/index/hello-gpt-4o/

Method	MultiWOZ2.0			
	Inform	Succ.	BLEU	Comb.
Codex[†]	76.7	41.5	7.7	66.8
ChatGPT[†]	71.8	44.1	10.5	68.4
Claude	78.3	41.2	2.9	62.7
GPT-4o	77.0	53.1	5.2	70.3
DSP w/ ChatGPT[†]	95.3	82.3	10.9	99.6
Ours w/ ChatGPT	95.1	91.2	9.8	102.9
Ours_{large}	96.1	92.4	17.2	111.5

Table 4: Performance comparison on MultiWOZ2.0 based on different LLMs. [†]: The results are reported in the work of DSP(Li et al., 2023).

tion to enhance the policy learning for the small model. To be fair, we use the same data setup as the DSP method -10% of the training data for SFT and RL training. Our approach shows superior performance on the MultiWOZ2.0 dataset. The Success rate is 8.9% higher than the DSP results reported in their work. This demonstrates that our method has superior generation capabilities, leading to more effective task completion.
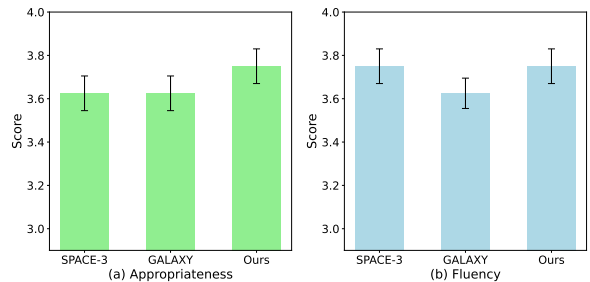

Figure 4: The human evaluation results regarding appropriateness and fluency. The numbers represent the average and the standard deviation for each method.

6.3 Human Evaluation

The automatic evaluation metric like BLEU might not be able to accurately evaluate the generation quality (Freitag et al., 2021, 2022; Wu and Aji, 2023). To thoroughly evaluate our approach, we conduct a human evaluation of the generated responses using our developed platform. We compare our model with the top two baselines, SPACE-3 and GALAXY, from Table 1 on the In-Car dataset. Following previous works (Zhang et al., 2020b; Ramachandran et al., 2022; Jang et al., 2022), we use two metrics: 1) Appropriateness: how well the response fits the dialogue context, and 2) Fluency: the clarity and coherence of the response. We randomly select 50 dialogue turns from the test set,

showing each turn and its history to 6 evaluators. Evaluators score each response on a 5-point Likert Scale (1 to 5), where 1 represents the lowest quality and 5 represents the highest quality. Importantly, the evaluators are unaware of the model identities to ensure unbiased judgments. As shown in Figure 4, our model outperforms the baselines in terms of appropriateness and matches SPACE-3 in fluency. This aligns with the results in Table 1 and further validates that our approach can not only achieve superior dialogue-task completion performance but also ensures high-quality responses.

7   Conclusion

We introduce a new approach for incorporating RL into TOD systems. Our approach focuses on improving both understanding and generation tasks by addressing challenges related to sparse and delayed rewards. We devise a progressive reward mechanism that combines understanding and generation rewards at the token level, facilitating gradual learning. Through extensive experiments on standard benchmarks using Flan-T5-Base and Flan-T5-Large backbones, we demonstrate the effectiveness of our approach and achieve state-of-the-art results on three widely used datasets.

8   Limitations

While our proposed approach using step-by-step rewards has shown promising results, it may struggle to capture all the nuances of TOD tasks fully. As a result, biases could be unintentionally introduced, causing the model to learn suboptimal strategies. In the future, developing a comprehensive reward model grounded in our reward function would be beneficial. Such a model can learn intricate patterns and enhance flexibility and adaptability.

Moreover, the reward design in our approach relies on predefined *informable* and *requestable* lists in the dialogue schema. While this is a common practice in task-oriented dialogue systems, it is limited when extending to open-domain dialogues. Open-domain dialogues typically lack fixed slots and values, which makes it challenging to apply this reward mechanism effectively. In the future, it would be valuable to have a more generalizable approach that supports both task-oriented and open-domain dialogues in conversational agents.

9   Acknowledgments

References

Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. 2023. LMRL gym: Benchmarks for multi-turn reinforcement learning with language models. *CoRR*, abs/2311.18232.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Saurabh Arora and Prashant Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. 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. Association for Computational Linguistics.

Iñigo Casanueva, Paweł Budzianowski, Pei-Hao Su, Stefan Ultes, Lina M Rojas Barahona, Bo-Hsiang Tseng, and Milica Gasic. 2018. Feudal reinforcement learning for dialogue management in large domains. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 714–719.

Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7521–7528.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen

Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Xiuyi Chen, Jiaming Xu, and Bo Xu. 2019. A working memory model for task-oriented dialog response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2687–2693.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Kumar Goyal, Peter Ku, and Dilek Hakkani-Tür. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 422–428. European Language Resources Association.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

Yihao Feng, Shentao Yang, Shujian Zhang, Jianguo Zhang, Caiming Xiong, Mingyuan Zhou, and Huan Wang. 2023. Fantastic rewards and how to tame them: A case study on reward learning for task-oriented dialogue systems. *arXiv preprint arXiv:2302.10342*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1371–1374.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–273.

Jinyu Guo, Kai Shuang, Kaihang Zhang, Yixuan Liu, Jijie Li, and Zihan Wang. 2023. Learning to imagine: distillation-based interactive context exploitation for dialogue state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12845–12853.

Dhawal Gupta, Yash Chandak, Scott Jordan, Philip S Thomas, and Bruno C da Silva. 2024. Behavior alignment via reward function optimization. *Advances in Neural Information Processing Systems*, 36.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022a. Unified dialog model pre-training for task-oriented dialog understanding and generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 187–200.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022b. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10749–10757.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting. Association for Computational Linguistics.

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*, 33:20179–20191.

Vojtěch Hudeček and Ondřej Dušek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting*

of the Special Interest Group on Discourse and Dialogue, pages 216–228.

Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. 2022. GPT-critic: Offline reinforcement learning for end-to-end task-oriented dialogue systems. In *International Conference on Learning Representations*.

Hyunmin Jeon and Gary Geunbae Lee. 2021. Domain state tracking for a simplified dialogue system. *arXiv preprint arXiv:2103.06648*.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2020. Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 567–582, Online. Association for Computational Linguistics.

Wai-Chung Kwan, Hong-Ru Wang, Hui-Min Wang, and Kam-Fai Wong. 2023. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research*, 20(3):318–334.

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023. Guiding large language models via directional stimulus prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 62630–62656. Curran Associates, Inc.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2020. Rethinking supervised learning and reinforcement learning in task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3537–3546.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

Jianfeng Liu, Feiyang Pan, and Ling Luo. 2020. Gochat: Goal-oriented chatbots with hierarchical reinforcement learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1793–1796.

Yajiao Liu, Xin Jiang, Yichun Yin, Yasheng Wang, Fei Mi, Qun Liu, Xiang Wan, and Benyou Wang. 2023. One cannot stand for everyone! leveraging multiple user simulators to train task-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–21.

Keting Lu, Shiqi Zhang, and Xiaoping Chen. 2019. Goal-oriented dialogue policy learning from failures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2596–2603.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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 of the Association for Computational Linguistics*, pages 311–318.

Jiahuan Pei, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2020. Retrospective and prospective mixture-of-generators for task-oriented dialogue response generation. In *ECAI 2020*, pages 2148–2155. IOS Press.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. SOLOIST: building task bots at scale with transfer learning and machine teaching. *Trans. Assoc. Comput. Linguistics*, 9:907–824.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Kam-Fai Wong, and Shang-Yu Su. 2018. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. *arXiv preprint arXiv:1801.06176*.

Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2231–2240, Copenhagen, Denmark. Association for Computational Linguistics.

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

Govardana Sachithanandam Ramachandran, Kazuma Hashimoto, and Caiming Xiong. 2022. [CASPI] causal-aware safe policy improvement for task-oriented dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 92–102,

Dublin, Ireland. Association for Computational Linguistics.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *The Eleventh International Conference on Learning Representations*.

Liliang Ren, Jianmo Ni, and Julian McAuley. 2019. Scalable and accurate dialogue state tracking via hierarchical sequence generation. 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)*, pages 1876–1885, Hong Kong, China. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Charlie Victor Snell, Ilya Kostrikov, Yi Su, Sherry Yang, and Sergey Levine. 2023. Offline rl for natural language generation with implicit language q learning. In *The Eleventh International Conference on Learning Representations*.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

Ryuichi Takanobu, Runze Liang, and Minlie Huang. 2020. Multi-agent task-oriented dialog policy learning with role-aware reward decomposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 625–638. Association for Computational Linguistics.

Ryuichi Takanobu, Hanlin Zhu, and Minlie Huang. 2019. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog. *arXiv preprint arXiv:1908.10719*.

Da Tang, Xiujun Li, Jianfeng Gao, Chong Wang, Lihong Li, and Tony Jebara. 2018. Subgoal discovery for hierarchical dialogue policy learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2298–2309.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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*, pages 438–449.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *CoRR*, abs/2307.03025.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George F. Foster, and Gholamreza Haffari. 2024a. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2024b. LaMini-LM: A diverse herd of distilled models from large-scale instructions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 944–964, St. Julian's, Malta. Association for Computational Linguistics.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024c. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *arXiv preprint arXiv:2405.11804*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14230–14238.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Xiao Yu, Qingyang Wu, Kun Qian, and Zhou Yu. 2023. Krls: Improving end-to-end response generation in task oriented dialog with reinforced keywords learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yichi Zhang, Zhijian Ou, Min Hu, and Junlan Feng. 2020a. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9207–9219, Online. Association for Computational Linguistics.

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

Zhirui Zhang, Xiujun Li, Jianfeng Gao, and Enhong Chen. 2019. Budgeted policy learning for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3742–3751.

Yangyang Zhao, Mehdi Dastani, and Shihan Wang. 2024. Bootstrapped policy learning: Goal shaping for efficient task-oriented dialogue policy learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2615–2617.

Li Zhou and Kevin Small. 2019. Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering. *arXiv preprint arXiv:1911.06192*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.

A Implementation Details

SFT Details. In the supervised fine-tuning stage, we use a train batch size of 8 and an evaluation batch size of 16. We set the learning rate to 2×10^{-5} and train the model for a total of 30 epochs. For generation settings of inference, the maximum length is set to 256 tokens for the Multi-WOZ dataset and 168 tokens for the In-Car dataset respectively. The shorter maximum length for the In-Car dataset is due to the relatively shorter utterances compared to those in the MultiWOZ datasets. It is important to note that the In-Car dataset does not include dialogue act annotations. Hence, we predict only the belief state and system response, formatting the input as $I_t = [\text{prefix} : u_{t-1} : bs_{t-1} : sr_{t-1} : u_t]$.

RL Details. The policy network is trained for 20k episodes, with 5 epochs per batch for enforcement learning. The batch size is set to 8, and the learning rate is 2×10^{-6}. We employ sampling with a top-k value of 50 during training. Following (Ziegler et al., 2019), the KL coefficient β in Equation 5 is dynamically adapt during training:

$$e_t = \text{clip}\left(\frac{D_{KL}(\pi \| \pi_o) - \text{KL}_t}{\text{KL}_t}, -0.2, 0.2 \right), \quad (6)$$

$$\beta_{t+1} = \beta_t(1 + K_\beta e_t), \quad (7)$$

where KL_t is the KL divergence between initial model π_o and current policy π. β are initially set to 0.01. K_β is the update rate which we set to 0.2 in our experiments.

Model and Implementation Details. We use Flan-T5 base ($\sim$250M parameters) and Flan-T5 large ($\sim$780M parameters) models as the backbone, which are the extensions of the T5 model designed to enhance performance on a wide range of natural language processing tasks. Our experiments are all run on a server equipped with 8 NVIDIA A800.

B Reward Curve

Figure 5 shows how rewards increase incrementally during token-by-token generation when completing slot-values or values. The plateau phases represent the process of generating a complete slot value or value. We present the reward patterns for three tasks: DSP, DPL, and RG. The curves demonstrate that our approach provides gradually increasing dense rewards for end-to-end models, effectively supporting understanding and generation tasks.
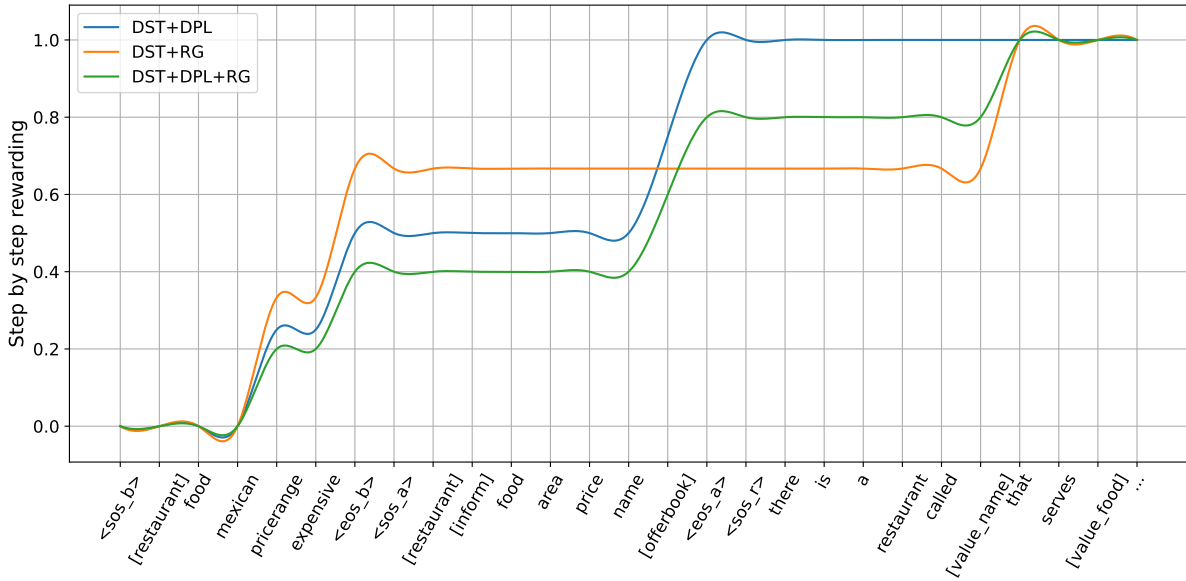
Figure 5: Reward accumulation for different tasks: DST+DPL, DST+RG, and DST+DPL+RG during token generation. Our reward function progressively provides important feedback for understanding and generation tasks.

C Case Study

To evaluate the effectiveness of our dialogue system, we develop a user interface using the Streamlit[2] as shown in Figure 6. The interface allows users to select a dialogue goal and interact with the system according to that goal. Users assess the system's responses utilizing the evaluation methodology detailed in Section 6.3.

We provide an example comparison from our model and GALAXY in Figure 7. It illustrates a scenario in which our model generates more accurate and comprehensive results compared to GALAXY.

D Error Examples

We present a representative error example in our predicted results in Figure 8. We observe that the response of our model includes all the necessary value information for the task, but it lacks conversational fluency. This indicates that our designed reward function prioritizes task completion efficiency over dialogue naturalness. Future work could explore integrating metrics like BLEU into the reward function to enhance both task completion and conversational fluency.
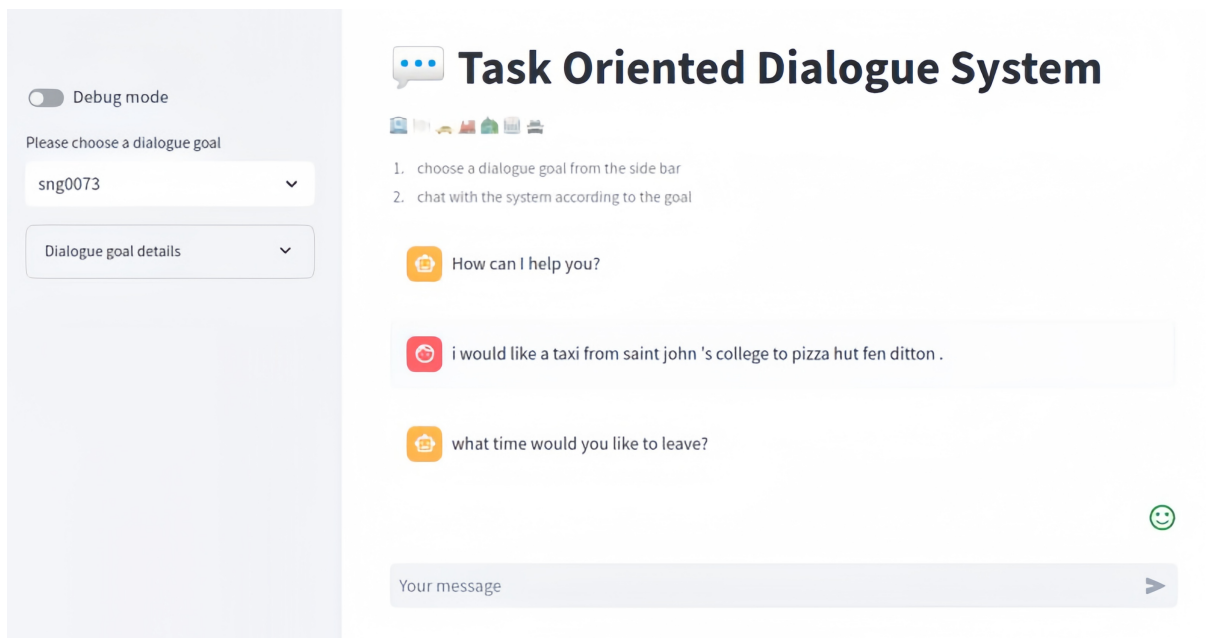
---

[2]https://streamlit.io/

Figure 6: Our user interface allows users to evaluate the dialogue system. They can select dialogue goals, interact with the system, and provide feedback on the responses.

pmul0187

i need a resturant that serves british food .

there are [value_choice] [value_food] restaurant -s . what price range -s do you want ?

i am open to any price range , but am looking for somewhere located in the centre area of town .

i would recommend [value_name] , they are [value_price] , would you like more info on them ?

yes can i please get more info .

BS OURS: [restaurant] food british area centre name cotto
BS GALAXY: [restaurant] food british area centre name cotto

DA OURS: [restaurant] [inform] address postcode phone[general] [reqmore]
DA GALAXY: [restaurant] [inform] phone

SR OURS: sure, their address is [value_address], their postcode is [value_postcode], and their phone number is [value_phone]. is there anything else i can help you with?

SR GALAXY: their phone number is [value_phone]

may i also get the address for the restaurant ?

absolutely , there address is [value_address] [value_postcode] . is there anything else i may help you with today ?

Figure 7: An example showing our model's effectiveness in predicting comprehensive restaurant information. Green, purple, and red text represent predicted results of our model and GALAXY for BS, DA, and SR respectively.

mul0317

i am looking for an expensive restaurant in the south part of town .

[value_name] is [value_price] and in the [value_area] .

actually , i am looking for an expensive restaurant in the south part of town
. [value_name] is [value_price] and in the [value_area] .

[value_name] is an [value_food] food restaurant in the [value_price] price rang
and the [value_area] .

what is its postcode and address ?

BS PRED: [restaurant] pricerange expensive area south food italian
BS GOLD: [restaurant] pricerange expensive area south food italian

DA PRED: [restaurant] [inform] postcode address
DA GOLD: [general] [greet] [restaurant] [inform] address postcode [general] [reqmore]

SR PRED: the postcode is [value_postcode] and the address is [value_address].
SR GOLD: sure . their address is [value_address] and their postcode is [value_postcode]
. would you like their telephone number as well ?

Figure 8: Error Example. Black text is the input context. Green, purple, and red text represent predicted (PRED) and ground truth (GOLD) for BS, DA, and SR. Red highlights indicate incorrect or missing key tokens.