Many Hands Make Light Work: Task-Oriented Dialogue System with Module-Based Mixture-of-Experts

Anonymous ACL submission

Abstract

Task-oriented dialogue systems are broadly used in virtual assistants and other automated services, providing interfaces between users and machines to facilitate specific tasks. For example, in the context of hotel reservations, these systems not only recommend hotels that align with user preferences but also retain user requirements for future reference. Corresponding to a wide range of properties and applications of task-oriented dialogue systems, their outputs may also be diverse. Nowadays, task-oriented dialogue systems have benefited greatly from pre-trained language models (PLMs). While being effective and performant, scaling these models is expensive and complex. To address these challenges, we propose SMETOD to generate diverse natural language outputs, which scales the capacity of a task-oriented dialogue system while maintaining efficient inference. We extensively evaluate our model on dialogue state tracking, dialogue response generation, and intent prediction. Experimental results demonstrate that SMETOD consistently achieves state-of-the-art or comparable performance on all evaluated datasets. Furthermore, SMETOD shows an advantage in the cost of inference compared to existing approaches.

1 Introduction

Task-oriented dialogue systems play a crucial role in virtual assistants and various automated services through human-machine interactions. The fundamental objective of a task-oriented dialogue system is to aid users in completing specific services or tasks all achieved through natural language dialogues (Wen et al., 2017). Considering a broad range of applications, task-oriented dialogue systems should generate diverse types of outputs for processing information, evaluating user intentions, or retaining for future reference. In real-world scenarios, useful information processed from dialogue could be presented in various formats, including form-based (Goddeau et al., 1996; Eric and Manning, 2017b), probability-based (Thomson and Young, 2010; Mrkšić et al., 2016; Lee et al., 2019), or text-based (Hosseini-Asl et al., 2020; Wang et al., 2022). Typically, several components are responsible for managing a variety of information: natural language understanding (**NLU**) for comprehending and translating user intent into either natural language or a format suitable for machine processing, dialogue state tracking (**DST**) for discerning the user's requirements and providing a foundation for subsequent decisions, and natural language generation (**NLG**) generate a natural language response to the user based on the machine's decision of the next move.

This leads to two predominant system designs, namely pipeline-based and end-to-end, divided by whether the machine-generated response is based on dialogue utterances or processed information from other components only. Either system design presents its own set of limitations in effectively addressing diverse output objectives (Takanobu et al., 2020). Drawbacks of pipeline-based systems lie in the potential for error propagation from one module to another, and local decisions can have adverse global effects (Su et al., 2016). End-to-end dialogue systems, on the other hand, raise concerns about missing all essential information that may be required other than responses. Moreover, diagnosing and considering component-flow characteristics can be challenging in end-to-end systems (Bang et al., 2023).

Despite the limitations in dialogue-system designs, there are also significant constraints in terms of scaling dialogue models with efficiency. Recent advancements have leveraged the transfer learning capabilities of pre-trained language models (PLMs) (Devlin et al., 2018; Dong et al., 2019; Radford et al., 2019; Raffel et al., 2020b) by fine-tuning (Budzianowski and Vulić, 2019; Hosseini-Asl et al., 2020; Heck et al., 2020) or pre-training di-

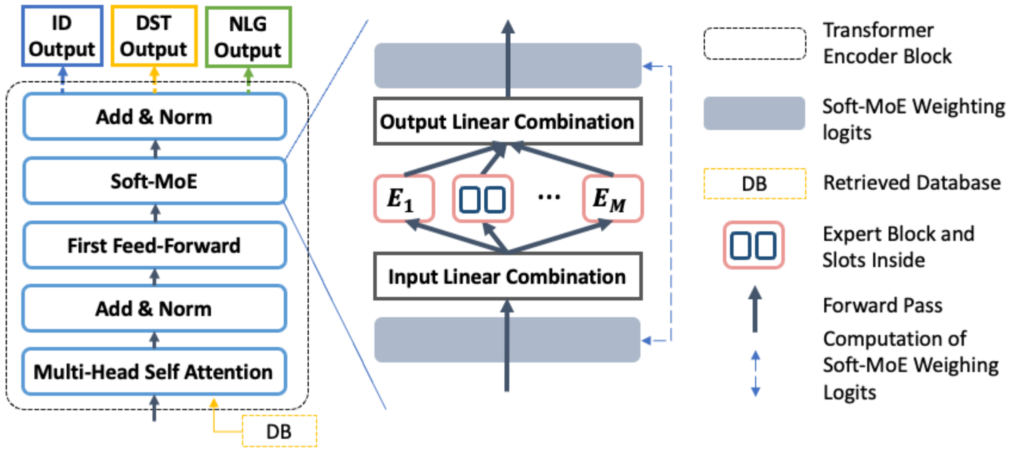

Figure 1: Architecture of the **SMETOD** as in Transformer (Vaswani et al., 2017) encoders. The result from the DB derived from the output of DST is used for NLG inference. All of the expert layers share the same architecture. The input is ensembled by experts in the Soft-MoE layer for improving model capacity without the cost of efficiency. The model is fine-tuned by maximizing the likelihood of predicting the next token for NLU, DST, and NLG outputs.

alogue models (Wu et al., 2020; Zhang et al., 2020; Peng et al., 2021; He et al., 2022b). However, their remarkable performance is at the cost of significant computational resources, especially as the sizes of PLMs continue to grow. Recently, parameter-efficient adapters raised that freeze the PLM while only allowing a small number of parameters updated for downstream models (Houlsby et al., 2019; Li and Liang, 2021; Lester et al., 2021), and have gained popularity in dialogue systems (Bang et al., 2023; Wang et al., 2023). Nevertheless, the model capacity (i.e. number of parameters) is limited by the number of downstream models, and the addition of adapters can become computationally expensive due to their sequential processing (Rücklé et al., 2020). We also argue that the issue of inference time scaling with model complexity becomes more prominent considering the time sensitivity associated with the deployment of dialogue systems.

To address these issues, we propose a **S**oft **M**ixture-of-**E**xpert **T**ask-**O**riented **D**ialogue system (**SMETOD**) which scales the model capacity for diverse outputs of dialogue systems with significantly less training and inference cost. Specifically, we leverage Soft MoE (Puigcerver et al., 2023) to improve model capacity and leverage the effectiveness and performance of considerably larger models with significantly lower computational costs. We present a task-oriented dialogue system as a multi-module end-to-end text generation to bridge the gap between traditional pipeline-based and end-to-end response generation systems, and optimize NLU, DST, and NLG, respectively, as in (Su et al., 2022; Bang et al., 2021). We formulate NLU, DST, and NLG as the text generation problems, which take dialogue history sequence as model input and generate spans as the output. In the cases of NLG, we predict the DST output to obtain the database (DB) state, which becomes incorporated into its input. With T5-small (Raffel et al., 2020a) and T5-base (Raffel et al., 2020a) as the backbone PLM, we evaluate our method on MultiWOZ (Eric et al., 2019; Zang et al., 2020) and NLU (Casanueva et al., 2020; Larson et al., 2019; Liu et al., 2019) datasets. We show that our method achieves significant improvement in multi-domain DST on MultiWOZ 2.1 and NLG on both benchmarks.

Our contribution is as follows:

- We propose SMETOD, a task-oriented dialogue system for diverse outputs, which first leverages Soft-MoE in text generation and dialogue systems to improve model capacity with efficiency.

- Experimental results demonstrate the effectiveness of our model by improving the performance of NLU and DST on all evaluation benchmarks and achieving comparable performance for NLG.

- Our study of time efficiency and the architect of Soft-MoE proves the significant improvement of efficiency as model complexity continues to grow, promoting future study on dialogue system design with efficiency.

2 Preliminaries

Soft Mixture-of-Experts. Mixture-of-Experts (MoE)-based models have shown advantages in scaling model capacity without large increases in training or inference costs. There has been work on scaling sparsely activated MoE architectures. In the context of modern deep learning architectures, it was firstly found effective by Shazeer et al. (2017) by stacking MoE between LSTM (Hochreiter and Schmidhuber, 1997) and resulted in the state-of-the-art in language modeling and machine translation. Shazeer et al. introduced MoE Transformer where MoE layers are a substitute for the FFN layers (2018).

We adopt Soft-MoE (Puigcerver et al., 2023), which scales model capacity without the loss of fine-tuning efficiency and is fully differentiable and balanced compared to conventional efficient MoEs (Lepikhin et al., 2020; Fedus et al., 2022; Du et al., 2022; Zhou et al., 2022; Puigcerver et al., 2023). Specifically, it performs a soft assignment on experts to each input token, achieving similar training costs and much lower inference costs at a larger model capacity. We use $f(\cdot;\theta)$ to denote a mapping f associated with the parameter θ from the input sample to an output space. $\sigma(\cdot)$ is the Softmax function. We denote $\{f(\cdot;\theta_i)\}_{i=1}^{m}$ as $\boldsymbol{m}$ **experts** with identical architectures; their weights $\theta_1,\ldots,\theta_m$ applied to individual tokens. Each expert has $\boldsymbol{p}$ **slots**, each of which is a weighted average of input. Slots in the same expert apply the same weights. Given input and output tokens $\boldsymbol{x} = \{x_1,\ldots,x_l\}$ and $\boldsymbol{y} = \{y_1,\ldots,y_l\}$ at the length l. Each expert will process p slots with parameters denoted as $\Psi = \{\psi^{(1)},\ldots,\psi^{(m\times p)}\}$. The input of experts, $\tilde{\boldsymbol{x}}$, is defined as the result of convex combinations of input tokens.

$$\tilde{\boldsymbol{x}}_j = (\sigma(\boldsymbol{x}\psi^{(j)}))^T\boldsymbol{x} \tag{1}$$

where j is the index of the slot in experts and $j \in [1,\ldots,m\times p]$. The corresponding expert function is applied on each slot to obtain the output slots:

$$\tilde{\boldsymbol{y}}_j = f(\tilde{\boldsymbol{x}}_j;\theta_{\lfloor j/p\rfloor}) \tag{2}$$

Given $\tilde{\boldsymbol{y}} = \{\tilde{\boldsymbol{y}}_j\}_{j=1}^{m\times p}$, the output of Soft-MoE layer, y_i, is computed as a convex combination of all $(m\times p)$ output slots over the expert dimension (i.e. the rows of $\boldsymbol{x}\Psi$):

$$y_i = \sigma(x_i\Psi)\,\tilde{\boldsymbol{y}} \tag{3}$$

End-to-end task-oriented dialogue system. End-to-end learning was found effective in training and optimizing the map directly from input to output (Wen et al., 2017; Liu and Lane, 2018; Eric and Manning, 2017a; Williams et al., 2017). Later on, a lot of endeavor was given to fine-tuning pre-trained language models and adapting their generalization capacities for an end-to-end system of task-oriented dialogues (Budzianowski and Vulić, 2019; Casanueva et al., 2020; Mehri et al., 2020; Hosseini-Asl et al., 2020). In recent years, pre-trained task-oriented dialogue models have emerged as strong contenders, surpassing traditional fine-tuning approaches and showcasing competitive generalization capabilities, particularly in multi-objective scenarios (Wu et al., 2020; Zhang et al., 2020; Peng et al., 2021; He et al., 2022b). However, it's worth noting that they require a large amount of dialogue data to train the backbone models and without an interface to optimize sub-modules.

Efficient transfer learning. To reduce the effort in tuning large PLMs and promote the scalability of model adaptation, there is a line of work that fixes the entire PLM and introduces a small number of new trainable parameters. Notable examples in this category include adapters (Houlsby et al., 2019; Pfeiffer et al., 2021; Karimi Mahabadi et al., 2021), prefix-tuning (Li and Liang, 2021) and prompt-tuning (Lester et al., 2021), *etc*. In-context learning prepends related task examples to condition on the generated dialogue states (Hu et al., 2022; Gupta et al., 2022; Venkateswaran et al., 2022). In end-to-end dialogue systems, a line of work prompts with specific text to generate desired outputs (Su et al., 2022) or injecting adapters to capture the knowledge of different functionalities (Bang et al., 2023; Mo et al., 2023). GPT-3 (Brown et al., 2020) and ChatGPT[1] are also successful and efficient open-domain dialogue systems. On the other hand, the MoE approach focuses on improving performance by efficiently scaling model sizes. Recent work on MoE develops more efficient routing implementations of Mixture-of-Experts in scaling language models (Lepikhin et al., 2020; Fedus et al., 2022; Du et al., 2022; Zhou et al., 2022; Puigcerver et al., 2023; Ma et al., 2018).

3 Method

We introduce **SMETOD**, a multi-objective dialogue system for NLU, DST, and NLG in task-

[1]https://chat.openai.com/chat

oriented dialogues, scaling model capacities while maintaining computational efficiency with Soft MoE (Puigcerver et al., 2023). The overall architecture is illustrated in Figure 1.

3.1 Problem Formulation

We define the dialogue history $h = [u_1^{sys}, u_1^{usr}, \ldots, u_t^{sys}, u_t^{usr}]$ as the concatenation of the system and user utterances in previous turns, where t is the number of current turns in the dialogue. h has all the dialogue history without the last system utterance, denoted as r. NLU outputs an I which is an intent or the API-name. The objective of DST is to output user goals, the tasks or purposes that the user wants to accomplish through the dialogue. user goals are typically represented as a set of pre-defined slot-value pairs that consist of the required information to query the dialogue system, i.e. $y_{API} = \{(s_1, v_1), \ldots, (s_n, v_n)\}$, where n is the number of slot-value pairs. Finally, NLG will generate S with the previous output: $h + y_{DB} \rightarrow r$, where y_{DB} is the items in the database retrieved by y_{API}. Given a pair of training examples (x', y'), we elaborate x' and y' corresponding to different modules of the dialogue system in the following Table.

	x'	y'
NLU	h	I
DST	h	y_{API}
NLG	$h + y_{DB}$	r

3.2 Soft Mixture-of-Expert Layer

We implement the Soft-MoE layer to replace the second Feed-Forward Layer in each Transformer (Vaswani et al., 2017) Encoder block, as illustrated in Figure 1. Mathematically, we denote the output out the first Feed-Forward layer of the k-the encoder is $g(\cdot; \phi_k)$, then $\boldsymbol{x} = g(x'; \phi_k) \in \mathbb{R}^{l \times d_{ff}}$ in Eq. 1, denoting d_{ff} as the dimension between the first and second Feed-Forward layer and d as model's hidden dimension, and l is the length of tokens. $\psi^{(j)} \in \mathbb{R}^{d_{ff}}$ is d_{ff}-dimensional vector of parameters corresponding to each slot of experts.

The mapping $f(\cdot; \theta_i)$ in Eq. 2 is simply a linear mapping corresponding to each expert, and p is the slots per expert having the same weights. Therefore, the output of the k-th encoder layer, $y'^{(k)}$, can be represented as

$$y'^{(k)} = f(g(x'; \phi_k); \Theta_k, \Psi_k) \quad (4)$$

For fine-tuning, we replicate the pre-trained weights from the second Feed-Forward layer of encoders and assign them to each expert, leveraging the contextual learning abilities inherent in pre-trained models.

3.3 Training Objectives

We optimize the generation outputs of NLU, DST, NLG, respectively, following Su et al. (2022). Given a pair of training samples as (x', y'), the loss function is defined to maximize the log-likelihood of the token to predict given the current context:

$$\mathcal{L}_{\{NLU,DST,NLG\}} = -\frac{1}{l}\sum_{q=1}^{l} \log P(y'_q | y'_{<q}; x') \quad (5)$$

4 Experiment

4.1 Data

We evaluate our models for NLU on Banking77 (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), and HWU64 (Liu et al., 2019); DST and NLG are evaluated on the task-oriented dialogue benchmarks MultiWOZ 2.1 (Eric et al., 2019) and MultiWOZ 2.2 (Zang et al., 2020). Banking77 contains 13,083 customer service queries labeled with 77 distinct intents for distinguishing between intents among queries related to similar tasks. CLINC150, consists of a comprehensive dataset comprising 23,700 examples, annotated with 150 intents across 10 distinct domains. HWU64 is collected from the home robot that has 25,716 examples for 64 intents spanning 21 domains.

MultiWOZ 2.1 (Eric et al., 2019) consists of multi-turn task-oriented dialogues across several domains, where 8,438 dialogues are for training and 1,0000 for dev and test. MultiWOZ 2.2 (Zang et al., 2020) improves MultiWOZ 2.1 by correcting annotation errors and adding dialogue act annotations. In MultiWOZ, the generation of response is not only related to the dialogue context but also grounded on the database (DB) state. The DB state is automatically retrieved from a pre-defined database using the generated dialogue states. SMETOD adopts a two-step approach during inference (Su et al., 2022; Bang et al., 2023). Firstly, it predicts the DST results to access the

Model	Banking77	HWU64	CLINC150
BERT-FIXED$^{\diamond *}$	87.19	85.77	91.79
CONVBERT-DG +Pre+Multi*	92.99	92.94	97.11
CONVBERT +Pre+Multi*	93.44	92.38	97.11
BERT-TUNED$^{\diamond *}$	93.66	92.10	96.93
CONVERT$^{\diamond *}$	93.01	91.24	97.16
USE+CONVERT$^{\diamond *}$	93.36	92.62	97.16
SPACE-2$^{\sharp *}$	94.77	**94.33***	97.80
SPACE-3*	**94.94***	94.14	97.89
TOATOD$_{small}$	92.40	90.42	98.45
TOATOD$_{base}$	92.17	90.79	98.01
SMETOD$_{small}$	92.47	90.88	98.12
SMETOD$_{base}$	**93.02**	**92.56**	**98.64**

Table 1: Accuracy (%) on three intent prediction datasets with full-data experiments. $^{\diamond}$ comes from Casanueva et al.(2020). $^{\sharp}$ are obtained from DialoGLUE leaderboard[2]. All others are reported as in the original papers. Models with * are classification-based.

DB state. Subsequently, it utilizes the retrieved DB state and the current dialogue context to generate the NLG results.

4.2 Training & Inference Details

All models are fine-tuned respectively using PPTOD (Su et al., 2022), the pre-trained dialogue models based on T5-small (60M parameters) (Raffel et al., 2020b) and T5-base (220M parameters) (Raffel et al., 2020b), as the backbone. T5-small has 6 encoders and decoders with hidden size $d = 512$ and $d_{ff} = 2048$. While T5-large has 12 encoders and decoders and $d = 768$, $d_{ff} = 3072$. For models' architecture, we replace the second Feed-Forward layer in all encoder blocks with the illustrated Soft-MoE layers, and copy pre-trained weights to each expert in the Soft-MoE layers. We augment T5 with 8 experts and 2 slots per expert for DST, and 16 experts with 2 slots per expert for NLU and NLG.

We fine-tuned all model parameters on the full-shot training setting. The linear combination weights in Soft-MoE layer are initialized by Kaiming initialization (He et al., 2015). The initial learning rate is set to 0.001 for NLU, and 0.0001 DST, NLG, respectively. We use the Adafactor (Shazeer and Stern, 2018) optimizer and the training batch size is set to 64 on Nvidia A10 GPUs. We tried a wide range of learning rates from 1e-2 to 1e-6 then set the initial training rate to 1e-4 in all training. Our code is developed based on *Soft-Mixture-of-Experts*[3] and *TOATOD*[4]. Code repository will be released to the public soon.

Because different batch sizes will result in different padded lengths, inference results are slightly changed by batch sizes due to Softmax over input tokens in the Soft-MoE layer. We make inferences on several selected batch sizes and report average scores. We found out that different batch sizes in our experiments have negligible influence on the inference results[5].

5 Results & Discussion

We show the effectiveness of our models on **NLU** (Sec. 5.1), **DST** (Sec. 5.2), and **NLG** (Sec. 5.3) in task-oriented dialogue systems compared to plenty of strong baselines. In the experiments, we fine-tune SMETOD using the small and base versions of PPTOD(Su et al., 2022), which continues pre-training T5 (Raffel et al., 2020b) on large dialogue corpora, as the start point. We observe that SMETOD is state-of-the-art on NLU and DST and comparable with existing baselines on NLG. We also study the improvement of efficiency with SMETOD (Sec. 5.4). In Sec. 5.5, we investigate model performance when the Soft-MoE layers are in different architectures.

5.1 Intent Prediction

The goal of intent prediction, known as NLU in a task-oriented dialogue system, is to identify the user's intention based on the user's utterance. We conduct experiments on three benchmarks: Banking77 (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), and HWU64 (Liu et al., 2019). We report Accuracy (%) of predicting an intention correctly for evaluation.

5.1.1 Baselines

Baselines have a wide range from BERT-based models: CONVBERT (Mehri et al., 2020), CONVERT (Casanueva et al., 2020), UniLM-based models: SPACE-2 (He et al., 2022a), SPACE-3 (He et al., 2022b), to T5-based TOATOD (Bang et al., 2023). All baseline models utilizing BERT and UniLM follow a classification-based approach, employing a classifier featuring a Softmax layer to make predictions from a predefined set of intents.

[3] https://github.com/fkodom/soft-mixture-of-experts.git

[4] https://github.com/sogang-isds/TOATOD.git

[5] We conducted a hypothesis test and found out p-value < 0.01 for scores changed by batch size. Statistics are summarized in Appendix A, Table 5.

Model	Pre-Trained Model	MultiWOZ2.1	MultiWOZ2.2
TRADE	-	45.6	45.4
TripPy	BERT-base	55.29	-
TripPy$_{+SaCLog}$	BERT-base	60.61	-
CONVBERT-DG	BERT-base	55.29	-
SimpleTOD	DistilGPT-2	55.76	-
SOLOIST	GPT-2	56.85	-
AG-DST	PLATO-2	57.26	57.26
UniLM‡	UniLM	54.25	54.25
SPACE-3	UniLM	57.50	57.50
PPTOD$_{base}$	T5-base	57.10	-
PPTOD$_{large}$	T5-large	57.45	-
D3ST$_{base}$	T5-base	54.2	56.1
D3ST$_{large}$	T5-large	54.5	54.2
D3ST$_{XXL}$	T5-XXL	57.80	58.7
T5DST$_{+desc}$	T5-base	56.66	57.6
TOATOD$_{small}$ †	T5-small	59.49	59.33
TOATOD$_{base}$ †	T5-base	59.51	**60.02**
SMETOD$_{small}$	T5-small	59.69	59.60
SMETOD$_{base}$	T5-base	**60.36**	**60.08**

Table 2: Joint Goal Accuracy (%) for DST on MultiWOZ 2.1 and 2.2. Results with ‡ are from He et al.(2022b). † represents the results of our re-implementation. All others are reported as in the original papers.

5.1.2 Evaluation Results

Table 1 shows that our approaches perform state-of-art on CLINC150, which has the most number of intent types. On the other two benchmarks, our approaches have the highest accuracy compared to other generation-based approaches. Classification-based approaches are better which may benefit from smaller numbers of intents to choose from. Compared to classification models, SMETOD copes with the classification task as a generation problem by directly generating the text label. Therefore, when adapting to a new classification task, SMETOD is more scalable to new domains and tasks and can predict intents that are not in the ontology.

5.2 Dialogue State Tracking

As a crucial component in task-oriented dialogue systems, DST determines the user goals based on the history of dialogue turns. For the evaluation of DST models, we use joint goal accuracy (JGA) which is the average accuracy of predicting all slot-values for the current turn correctly.

5.2.1 Baselines

In Table 2, we compare SMETOD with a wide range of classification-based approaches: TRADE (Wu et al., 2019), TripPy (Heck et al., 2020), TripPy + SaCLog (Dai et al., 2021), CONVBERT-DG (Mehri et al., 2020), SimpleTOD (Hosseini-Asl et al., 2020), SOLOIST (Peng et al., 2021), AG-DST (Tian et al., 2021), SPACE-3 (He et al., 2022b), and generation-based approaches: PPTOD (Su et al., 2022), D3ST (Zhao et al., 2022), T5DST (Lee et al., 2021), and TOATOD (Bang et al., 2023).

5.2.2 Evaluation Results

Compared to other approaches, SMETOD obtains state-of-the-art JGA on MultiWOZ 2.1 and 2.2 among all generation-based approaches. Our model is more flexible to generate slot-value pairs while classification-based models are limited to the pre-defined ontology. The results show that our model can benefit from not only the transfer learning capacities of per-trained models but also the improvement of model size.

5.3 End-to-End Response Generation

End-to-end dialogue response generation, aiming at evaluating the model in the most realistic, fully end-to-end setting, where the generated dialogue states are used for the database search and response generation (Hosseini-Asl et al., 2020; Su et al., 2022), is NLG in task-oriented dialogue system. Our models evaluated on MultiWOZ generates responses not only related to the dialogue history but also grounded on the database (DB) state.

Model	Backbone	MultiWOZ2.1				MultiWOZ2.2			
		Inform	*Success*	*BLEU*	*Combined*	*Inform*	*Success*	*BLEU*	*Combined*
DOTS	BERT-base	86.65	74.18	15.90	96.32	-	-	-	-
DiactTOD	S-BERT	-	-	-	-	89.5	**84.2**	17.5	**104.4**
SimpleTOD	DistilGPT-2	85.00	70.50	15.23	92.98	-	-	-	-
SOLOIST	GPT-2	-	-	-	-	82.3	72.4	13.6	90.9
UBAR$^{\triangle}$	GPT-2	95.70	81.80	16.50	105.25	83.4	70.3	17.6	94.4
MinTL$^{\triangle}$	BART$_{large}$	-	-	-	-	73.7	65.4	19.4	89.0
RewardNet$^{\triangle}$	BART$_{large}$	-	-	-	-	87.6	81.5	17.6	102.2
GALAXY	UniLM	95.30	86.20	**20.01**	110.76	85.4	75.7	**19.64**	100.2
PPTOD$_{base}$	T5-base	87.09	79.08	19.17	102.26	-	-	-	-
MTTOD$^{\natural}$	T5-base	90.99	82.08	19.68	106.22	85.9	76.5	19.0	100.2
RSTOD$^{\natural}$	T5-small	93.50	84.70	19.24	108.34	83.5	75.0	18.0	97.3
TOATOD$_{small}$	T5-small	92.10	80.40	18.29	104.54	85.80	74.00	18.00	97.90
TOATOD$_{base}$	T5-base	**97.00**	**87.40**	17.12	**109.32**	**90.00**	79.80	17.04	101.94
KRLS	T5-base	-	-	-	-	89.2	80.3	19.0	103.8
SMETOD$_{small}$	T5-small	92.50	74.00	16.89	100.14	89.6	76.2	17.1	100.1
SMETOD$_{base}$	T5-base	92.30	78.80	16.88	102.43	89.0	76.0	17.6	99.7

Table 3: Evaluation of NLG on Inform, Success, BLEU, and Combined Scores, where Combined = (Inform + Success) × 0.5 + BLEU. $^{\natural}$ means the NLG results on MultiWOZ 2.1 is from Cholakov and Kolev (2022). All other results are from MultiWOZ leaderboards[6]. $^{\triangle}$ shows models that require oracle dialogue states for prediction.

5.3.1 Metrics

For evaluation, we follow the individual and combined metrics in Hosseini-Asl et al. (2020): Inform, Success, and BLEU, and Combined score which is defined as Combined = (Inform + Success) × 0.5 + BLEU. Specifically, Inform rate measures the correctness of entities in the response. Success rate success rate assesses attribute fulfillment requested by user. BLUE score is used to measure the fluency of the generated responses.

5.3.2 Baselines

In Table 3, we compare our model with several strong baselines: DOTS (Jeon and Lee, 2021), DiactTOD (Wu et al., 2023), SimpleTOD (Hosseini-Asl et al., 2020), SOLOIST (Peng et al., 2021), UBAR (Yang et al., 2021), MinTL (Lin et al., 2020), RewardNet (Feng et al., 2023), GALAXY (He et al., 2022c), PPTOD (Su et al., 2022), RSTOD (Cholakov and Kolev, 2022), MTTOD (Lee, 2021), TOATOD (Bang et al., 2023), KRLS (Xiao Yu, 2022).

5.3.3 Evaluation Results

On both MultiWOZ 2.1 and 2.2 datasets, SMETOD performs, though not the best, comparable to T5-based models except TOATOD$_{base}$. We hypothesize that metrics hinder each other from being improved together and may require a mechanism to promote performance towards specific metrics, for example, REINFORCE (Sutton et al., 1999). Besides, we observe that only replacing the Feed-Forward layer in Transformer encoders as in Puigcerver et al. (2023) without copying weights to experts doesn't generate the best results in our dialogue system. It might be because their implementation requires a large amount of data to pretrain, which is inappropriate in the task-oriented scenario. It demonstrates that by duplicating pretrained weights and fine-tuning, SMETOD optimizes well for DST and NLG, respectively, maintaining the prior knowledge learned from the pretrained model.

Model	Small↓	Base↓
PPTOD	1×	3.163×
TOATOD	1.116×	3.519×
SMETOD	1.005×	3.095×

Table 4: Comparison of the inference time with small and base-size models of PPTOD and TOATOD for NLG on MultiWOZ 2.1. All models are experimented with 5 same and randomly sampled batch sizes. Average time is reported. ↓: Smaller is better.

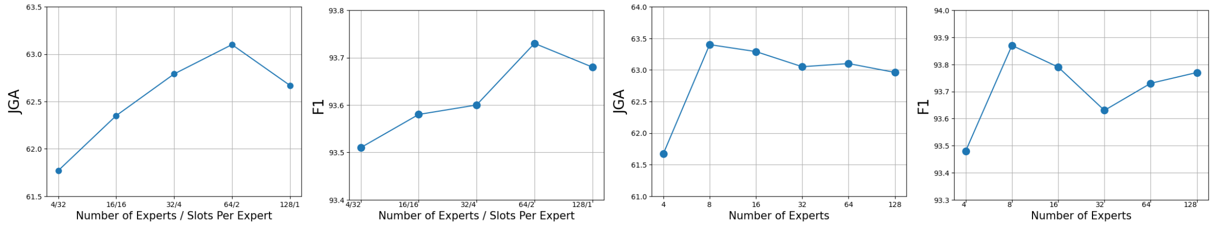

Figure 2: (**Left**) Performance of SMETOD as a function of the number of experts, for models with a fixed number of experts × slots-per-expert. (**Right**) Performance of SMETOD trained with increased experts and 2 slots per expert. JGA and F1 scores are on MultiWOZ 2.1 dev set for DST.

5.4 Time Complexity Analysis

According to Puigcerver et al. (2023), the time complexity of the Soft-MoE layer can be reduced to $O(l^2d + lk)$, given input token length l, model hidden dimension d, and the cost of applying an expert per token $O(k)$. Thus, the time complexity is constant and the same as the single-headed self-attention cost by increasing the number of experts m and scaling slots per expert $p = O(l/m)$ accordingly, which will not become a bottleneck in Transformer.

We show in Table 4 that SMETOD could make inferences without bringing about much extra time. SMETOD$_{small}$ is 3.5 times larger than PPTOD and TOATOD while achieving a similar inference speed as the former. Our SMETOD$_{base}$ has even less inference time while its model size is 4 times of PPTOD$_{base}$. It proves that we can achieve much better scaling while cost is roughly constant (Puigcerver et al., 2023), with the benefit of improved performance.

5.5 Impact of Expert Numbers

We investigate the impact of expert and slot numbers in our models on the development set of MultiWOZ 2.1 for DST as illustrated in Figure 2. First, we fix the total number of slots to 128 and vary expert numbers {4, 16, 32, 64, 128} by scaling slot numbers per expert. Results suggest the best configuration is 64 experts and 2 slots per expert. Then, we set the number of slots per expert to one and evaluate performance with regard to the number of experts. The number of experts 8 and 16 perform better than others. It should be mentioned that the model size scales with increasing expert numbers only. Meanwhile, we observe performance is not always increasing with the number of experts, indicating there is a trade-off between model size and the amount of training data.

6 Conclusion

We propose an efficient fine-tuning approach based on Soft-MoE to satisfy requirements on diverse outputs in task-oriented dialogue systems. We demonstrate that incorporating Soft-MoE to our dialogue system achieves remarkable success on MultiWOZ baselines and optimizes outputs of each submodule, showing it powerful technique for task-oriented dialogue systems with better scaling performance while maintaining time efficiency.

7 Limitations

Limitations related to adopting Soft-MoE: This work is a practice of leveraging Soft-MoE (Puigcerver et al., 2023) in downstream models with supervised, while the original practice requires unsupervised pre-training. We consider per-taining experts on larger dialogue corpus, for example, Lin et al. (2021); Hu et al. (2022) for better generality performance in the future. Furthermore, we didn't evaluate our approach to NLP datasets which have more diverse example lengths. Unlike Soft-MoE used in computer vision, the weights over tokens are inconstant due to the variety of length of input tokens, which leads to inconsistent inference with different batch sizes. Although we observe negligible influence in our experiments, variations of lengths require further study. We should also have experimented with more expert numbers and investigated the performance on NLG as well to study how performance is improved with model size. Last, scaling up model sizes requires a lot of computational memory.

Limitations related to datasets: DST and NLG evaluations are on MultiWOZ, which are English and have limited domains. More generalized and larger-scale dialogue corpus need considering, such as DialoGLUE (Mehri et al., 2020), SGD (Lee et al., 2022), or multi-lingual datasets (Ding et al., 2021). NLU evaluations are only on single-utterance benchmarks, CamRest676 (Quan and Xiong, 2019), In-Car Assistant (Eric and Manning, 2017b).

Limitations related to training time: Recently, Adapters and prompt approaches have been proposed that update fewer parameters in models compared with our fine-tuning approaches. Although we didn't observe longer training time explicitly compared to adapter-based models with similar sizes, empirical study on this issue is not covered in this work. We have shown in Sec. 5.4 that the forward pass of our approach is faster. It has been shown that original adapters should backpropagate through the entire model only except the first components (Rücklé et al., 2020). Moreover, we argue that performance and inference efficiency are more important regarding the deployment of task-oriented dialogue systems.

Limitations related to GPT3 or ChatGPT (LLM) as baselines: We did not include evaluation with the above models due to the following reasons. First, we consider the generation problem in this paper to generate diverse outputs given the same input. The quality of prompts will have a significant impact on LLM results, making it hard to make a fair comparison. Second, our training is in full-shot scenarios, while GPT3 or ChatGPT is usually considered as a zero-shot or few-shot baseline. Last, there is a high probability that LLMs have contaminated public benchmarks used in this paper.

8 Potential Risks

Using public dialogue benchmarks introduces the potential for biases stemming from the data collection method. Models trained on such datasets might encounter challenges when attempting to generalize to real-world scenarios or specific domains, as the data may not accurately represent these situations. Additionally, public dialogue datasets frequently lack essential context or metadata, rendering it difficult to comprehend the circumstances surrounding the conversations.

In our approach, we also rely on open-source code repositories. However, these repositories can present issues related to security vulnerabilities and compatibility. Furthermore, their often incomplete documentation can pose additional hurdles for further development. Given the absence of reliable support or comprehensive documentation, these factors can impede troubleshooting and hinder the overall development process.

References

Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023. Task-optimized adapters for an end-to-end task-oriented dialogue system. *arXiv preprint arXiv:2305.02468*.

Yejin Bang, Nayeon Lee, Etsuko Ishii, Andrea Madotto, and Pascale Fung. 2021. Assessing political prudence of open-domain chatbots. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 548–555, Singapore and Online. Association for Computational Linguistics.

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*, 33:1877–1901.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

Radostin Cholakov and Todor Kolev. 2022. Efficient task-oriented dialogue systems with response selection as an auxiliary task. *arXiv preprint arXiv:2208.07097*.

Yinpei Dai, Hangyu Li, Yongbin Li, Jian Sun, Fei Huang, Luo Si, and Xiaodan Zhu. 2021. Preview, attend and review: Schema-aware curriculum learning for multi-domain dialog state tracking. *arXiv preprint arXiv:2106.00291*.

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*.

Bosheng Ding, Junjie Hu, Lidong Bing, Sharifah Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. Globalwoz: Globalizing multiwoz to develop multilingual task-oriented dialogue systems. *arXiv preprint arXiv:2110.07679*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Mihail Eric and Christopher D Manning. 2017a. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. *arXiv preprint arXiv:1701.04024*.

Mihail Eric and Christopher D Manning. 2017b. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270.

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*.

D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai. 1996. A form-based dialogue manager for spoken language applications. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 2, pages 701–704 vol.2.

Raghav Gupta, Harrison Lee, Jeffrey Zhao, Abhinav Rastogi, Yuan Cao, and Yonghui Wu. 2022. Show, don't tell: Demonstrations outperform descriptions for schema-guided task-oriented dialogue. *arXiv preprint arXiv:2204.04327*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Wanwei He, Yinpei Dai, Binyuan Hui, Min Yang, Zheng Cao, Jianbo Dong, Fei Huang, Luo Si, and Yongbin Li. 2022a. SPACE-2: Tree-structured semi-supervised contrastive pre-training for task-oriented dialog understanding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 553–569, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022b. 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. 2022c. 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 Gašić. 2020. Trippy: A triple copy strategy for value independent neural dialog state tracking. *arXiv preprint arXiv:2005.02877*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. *arXiv preprint arXiv:2203.08568*.

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

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural*

Language Processing, pages 4937–4949, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*, volume 36, pages 10938–10946.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. Sumbt: Slot-utterance matching for universal and scalable belief tracking. *arXiv preprint arXiv:1907.07421*.

Yohan Lee. 2021. Improving end-to-end task-oriented dialog system with a simple auxiliary task. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1296–1303.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Zhaojiang Lin, Bing Liu, Andrea Madotto, Seungwhan Moon, Paul Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Eunjoon Cho, Rajen Subba, et al. 2021. Zero-shot dialogue state tracking via cross-task transfer. *arXiv preprint arXiv:2109.04655*.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. *arXiv preprint arXiv:2009.12005*.

Bing Liu and Ian Lane. 2018. End-to-end learning of task-oriented dialogs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 67–73.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. *arXiv preprint arXiv:1903.05566*.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*.

Yunho Mo, Joon Yoo, and Sangwoo Kang. 2023. Parameter-efficient fine-tuning method for task-oriented dialogue systems. *Mathematics*, 11(14):3048.

Nikola Mrkšić, Diarmuid O Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2016. Neural belief tracker: Data-driven dialogue state tracking. *arXiv preprint arXiv:1606.03777*.

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. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. 2023. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*.

Jun Quan and Deyi Xiong. 2019. Effective data augmentation approaches to end-to-end task-oriented dialogue. In *2019 International Conference on Asian Language Processing (IALP)*, pages 47–52. IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

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

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2020. Adapterdrop: On the efficiency of adapters in transformers. *arXiv preprint arXiv:2010.11918*.

Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee, Mingsheng Hong, Cliff

Young, et al. 2018. Mesh-tensorflow: Deep learning for supercomputers. *Advances in neural information processing systems*, 31.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

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

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.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. 2020. Is your goal-oriented dialog model performing really well? empirical analysis of system-wise evaluation. *arXiv preprint arXiv:2005.07362*.

Blaise Thomson and Steve Young. 2010. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech & Language*, 24(4):562–588.

Xin Tian, Liankai Huang, Yingzhan Lin, Siqi Bao, Huang He, Yunyi Yang, Hua Wu, Fan Wang, and Shuqi Sun. 2021. Amendable generation for dialogue state tracking. *arXiv preprint arXiv:2110.15659*.

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.

Praveen Venkateswaran, Evelyn Duesterwald, and Vatche Isahagian. 2022. District: Dialogue state tracking with retriever driven in-context tuning. *arXiv preprint arXiv:2212.02851*.

Qingyue Wang, Liang Ding, Yanan Cao, Yibing Zhan, Zheng Lin, Shi Wang, Dacheng Tao, and Li Guo. 2023. Divide, conquer, and combine: Mixture of semantic-independent experts for zero-shot dialogue state tracking. *arXiv preprint arXiv:2306.00434*.

Weizhi Wang, Zhirui Zhang, Junliang Guo, Yinpei Dai, Boxing Chen, and Weihua Luo. 2022. Task-oriented dialogue system as natural language generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2698–2703.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, 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, Valencia, Spain. Association for Computational Linguistics.

Jason D Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv preprint arXiv:1702.03274*.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1905.08743*.

Qingyang Wu, James Gung, Raphael Shu, and Yi Zhang. 2023. Diacttod: Learning generalizable latent dialogue acts for controllable task-oriented dialogue systems. *arXiv preprint arXiv:2308.00878*.

Kun Qian Zhou Yu Xiao Yu, Qingyang Wu. 2022. Krls: Improving end-to-end response generation in task oriented dialog with reinforced keywords learning. *arXiv preprint arXiv:2211.16773*.

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.

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, ACL 2020*, pages 109–117.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Jeffrey Zhao, Raghav Gupta, Yuan Cao, Dian Yu, Mingqiu Wang, Harrison Lee, Abhinav Rastogi, Izhak Shafran, and Yonghui Wu. 2022. Description-driven task-oriented dialog modeling. *arXiv preprint arXiv:2201.08904*.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V Le, and James Laudon. 2022. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*.

A Statistics of Results

Model	Dataset	Module	Metric	Mean	Std
T5$_{small}$	**MultiWOZ 2.1**	**NLU**	JGA	59.69	0.028
		NLG	Inform	92.50	0.167
			Sucess	74.00	0.335
			BLEU	16.89	0.019
			Combined	100.14	-
	MultiWOZ 2.2	**NLU**	JGA	59.60	0.026
		NLG	Inform	89.6	0.207
			Sucess	76.2	0.349
			BLEU	17.1	0.031
			Combined	100.1	-
T5$_{base}$	**MultiWOZ 2.1**	**NLU**	JGA	60.36	0.017
		NLG	Inform	92.3	0.071
			Sucess	78.8	0.217
			BLEU	16.88	0.011
			Combined	102.43	-
	MultiWOZ 2.2	**NLU**	JGA	60.08	0.026
		NLG	Inform	89.0	0.182
			Sucess	76.0	0.349
			BLEU	17.6	0.013
			Combined	99.7	-

Table 5: Mean and standard deviation of all reported scores in Table 2 and Table 3 using 5 randomly sampled batch sizes, which are the same for all models and datasets. Student paired t-test shows $p < 0.01$ for scores changed by batch size. Combined = (Inform + Success) × 0.5 + BLEU.