Continual Dialogue State Tracking via Reason-of-Select Distillation

Yujie Feng[1], Bo Liu[1], Xiaoyu Dong[1], Zexin Lu[1]
Li-Ming Zhan[1], Xiao-Ming Wu[1]*, Albert Y.S. Lam[2]

[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong S.A.R.
[2]Fano Labs, Hong Kong S.A.R.
yujie.feng@connect.polyu.hk, xiao-ming.wu@polyu.edu.hk

Abstract

An ideal dialogue system requires continuous skill acquisition and adaptation to new tasks while retaining prior knowledge. Dialogue State Tracking (DST), vital in these systems, often involves learning new services and confronting catastrophic forgetting, along with a critical capability loss termed the "Value Selection Quandary." To address these challenges, we introduce the Reason-of-Select (RoS) distillation method by enhancing smaller models with a novel 'meta-reasoning' capability. Meta-reasoning employs an enhanced multi-domain perspective, combining fragments of meta-knowledge from domain-specific dialogues during continual learning. This transcends traditional single-perspective reasoning. The domain bootstrapping process enhances the model's ability to dissect intricate dialogues from multiple possible values. Its domain-agnostic property aligns data distribution across different domains, effectively mitigating forgetting. Additionally, two novel improvements, "multi-value resolution" strategy and Semantic Contrastive Reasoning Selection method, significantly enhance RoS by generating DST-specific selection chains and mitigating hallucinations in teachers' reasoning, ensuring effective and reliable knowledge transfer. Extensive experiments validate the exceptional performance and robust generalization capabilities of our method. The source code[1] is provided for reproducibility.

1 Introduction

Practical dialogue systems require continual adaptation to new services while maintaining previous task capabilities. However, previous research in dialogue systems has focused on domain-specific offline systems, lacking adaptation abilities (Ni et al., 2023). With recent advances in large language models (LLMs), LLM-based systems have
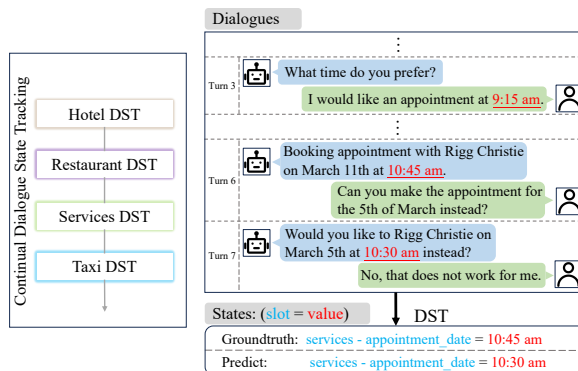

Figure 1: *Left:* Depiction of the Continual DST learning process. *Right:* An actual instance of the "Value Selection Quandary" phenomenon, demonstrating a dialogue with three mentioned date values, where the model incorrectly chooses the most recent time at turn 7 rather than the correct value at turn 6.

shown significant superiority over previous methods (Hu et al., 2023). However, the bulky sizes of LLMs make retraining models from scratch prohibitively time-consuming and challenging (Liu et al., 2023). Thus, efficient continual learning (CL) is vital for dialogue systems to obtain new skills while retaining knowledge of previous tasks. Dialog State Tracking (DST), central to task-oriented dialogue systems, dynamically updates triplets (domain, slot, value) to manage user intents (Feng et al., 2023a). The growing necessity to expand DST models for new services has spurred interest in recent Continual DST task (Cho et al., 2023).

As a specific task of CL, Continual DST grapples with catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999), where sequential learning of new tasks hinders retention of prior ones. This problem is more acute in DST due to significant data distribution shifts across domains in the CL process, exemplified by transitions from Hotel to Restaurant domains (see Figure 1). Recent advances in Continual DST, including memory replay and regularization (Zhu et al., 2022), and refor-

---

*Corresponding author.
[1]https://github.com/WoodScene/RoS

mulation as a question-answering task (Cho et al., 2023), strive to address forgetting. However, these methods face challenges like reliance on past data and high computational demands during testing, which hinder real-time applications.

Recent LLMs show impressive DST performance (Heck et al., 2023; Feng et al., 2023a), but practical deployment faces hurdles like offline computational load and online data privacy concerns. Moreover, enabling LLMs for CL demands significant resources, leading to the exploration of smaller models. After a comprehensive analysis of the current smaller DST models, we have identified a critical loss of capability brought by domain shifts in CL that models fail to output correct value when facing a set of similar candidates (named "Value Selection Quandary" as elaborated in Section 2). For example, in Figure 1, when tracking the <services-appointment_date> slot, the model incorrectly chooses the most recent time mentioned, revealing its inability to grasp contextual subtleties and favoring direct value extraction over logical reasoning. This phenomenon is because, as the CL progresses, the model needs to understand the relevant knowledge of newly emerging domains and chooses to forget domain-irrelevant or weakly relevant knowledge, such as value selection.

This paper addresses these challenges by boosting smaller models' reasoning ability, termed as **meta-reasoning**. Meta-reasoning, guided by the insight that domain-specific dialogues represent only a fragment of the underlying meta-knowledge, can be viewed as a form of multi-view augmentation from different domains in the CL process. This domain bootstrapping strategy facilitates the transfer and broadening of meta-knowledge, transcending traditional single-perspective reasoning. Meta-reasoning enriches smaller models' reasoning abilities to dissect and interpret intricate dialogue scenarios, addressing the "Value Selection Quandary" and broadening their cognitive horizons. Moreover, this domain-agnostic feature aligns data distribution shifts across different domains, effectively reducing catastrophic forgetting.

Inspired by the powerful reasoning abilities of LLMs like LLaMA-2-70B (Touvron et al., 2023) and ChatGPT [2], we introduce Reason-of-Select (RoS) distillation approach, designed to graft these advanced reasoning capabilities onto smaller models. We present a "multi-value resolution" strategy

---

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

tailored to DST, prompting teacher LLMs to generate a "selection chain" that discerns and elaborates the best value choice from various options. Then, we distill this rationale into a smaller student model to boost its meta-reasoning capabilities, making it easier to deploy to terminal systems.

Moreover, to enhance faithful reasoning and reduce hallucinations in the teacher model, we integrate a schema-guided prompt, strengthening source-target correlation and semantic coherence. Our innovative Semantic Contrastive Reasoning Selection method uses semantic similarity to select the most accurate teacher-generated rationale, suitable even in black-box scenarios. Extensive experiments with various teacher and student model sizes demonstrate our method's exceptional Continual DST performance and robust cross-dataset generalization capabilities.

To summarize, our main contributions are:

- We propose a Reason-of-Select (RoS) distillation framework for Continual DST. Through enhancing domain-agnostic meta-reasoning capabilities, ambiguous value selection and catastrophic forgetting caused by data distribution shifts across domains can be effectively mitigated.

- To alleviate the hallucination outputted by a teacher model in our annotation-free reasoning framework, we present a Semantic Contrastive Reasoning Selection strategy to obtain trustworthy rationale from multiple candidates, ensuring faithful reasoning transfer.

- Comprehensive experiments with two teacher models and four student models of varying sizes demonstrate superior performance and robust generalization of our proposed method.

2 Motivation for Boosting Meta-reasoning Capability in Continual DST

Through an initial experiment with the same setups from prior work (Zhu et al., 2022) (details in Section 4), we employed two distinct backbone models, T5-small and LLaMA-7B, and observed a performance decline in Continual DST with turn ids increasing (Figure 2 (a)). This excites two fundamental questions for obtaining better continual DST performance: (i) What crucial knowledge do current models lack for continual DST? (ii) How can this missing knowledge be effectively integrated?
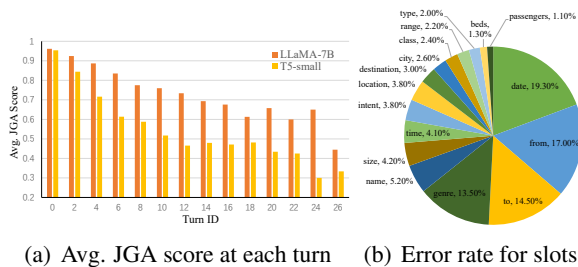
(a) Avg. JGA score at each turn   (b) Error rate for slots

Figure 2: Performance analysis of LLaMA-7B and T5-small in Continual DST task.

2.1 Pinpointing the Value Selection Quandary

Upon analyzing the 874 incorrectly predicted slot samples in long dialogues (turn id > 10), we initially found that in 94.5% of cases, the model usually returns values semantically relevant to the requested slot from the dialogue. For example, a location-related slot often results in selecting any mentioned location. Upon closer examination, we found that the model failed to select the correct value due to a significant capability loss, termed the "Value Selection Quandary". This issue occurs when the model struggles to reason out the correct value among multiple ambiguous or semantically similar values in a dialogue.

Examining 826 errors linked to this quandary, we observed a pattern: in 45% of cases, models favored the most recently mentioned value (Figure 1). In 30% of instances, they stuck with a previously chosen value, disregarding modifications in the dialogue. The remaining 25% involved random selections from available values. Moreover, Figure 2(b) classifies error rates across 49 slots, highlighting that slots regarding 'date', 'from', and 'to' are most prone to errors. This is logical, as time or location slots in dialogues often present multiple values, leading to selection challenges.

These findings highlight a critical gap in models' ability to navigate the "Value Selection Quandary", presenting a significant avenue for enhancement.

2.2 Bridging the Gap of Reasoning Ability

Inspired by the robust reasoning capabilities exhibited by models exceeding 70B parameters (Li et al., 2023; Wang et al., 2023), we evaluated their efficacy in DST task, specifically their ability to generate coherent rationales. As demonstrated in Table 1, ChatGPT's rationalizations for the <appointment_date> slot confirm their proficiency in discerning and justifying value selections.

Reasoning generated by ChatGPT	In the given dialogue, the user initially requested an appointment at 9:15 am. But the user then changed the time to 10:45 am. The system then suggested an alternative time of 10:30 am on the 5th of March, which the user rejected. Therefore, considering the user's preference and their rejection of the alternative time, '10:45 am' is the most appropriate value for the requested slot '<services-appointment_date>'.
Hallucinatory Reasoning generated by LLaMA-2 70B	The slot <services-appointment_date> has multiple possible values, but the most appropriate value is 'later today' based on the dialogue context and user intent. The user initially requests an appointment for a Psychiatrist in Napa, but then expresses interest in seeing other options. The system suggests a therapist in Santa Rosa, and the user confirms their preference for an appointment on the 8th. However, they then request to change the appointment date to today, which the system confirms. Therefore, the most appropriate value for the requested slot is 'later today' to align with the user's revised preference.

Table 1: Examples of the reasoning processes generated by different teacher models, where blue font highlights possible values in the dialogue, and red font marks hallucinatory elements in the teachers' reasoning.

Our Reason-of-Select Distillation method, leveraging LLMs' selection logic, boosts smaller models' reasoning and decision-making abilities. This domain-agnostic approach effectively reduces forgetting in continual learning by teaching models to reason across diverse tasks and domains instead of merely imparting specific information or tasks. In addition, we revealed the *hallucination* in teacher models' rationales, such as references to non-existent dialogue elements (Table 1), which motivates a novel Semantic Contrastive Selection method (Section 3.3) ensuring the accuracy and relevance of the logic transferred to student models.

3 Reason-of-Select Distillation

Problem Formulation In continual DST, we train a model $f : \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{Y}$ across a series of dialogue domains $\mathcal{T}_1, ..., \mathcal{T}_K$. This model predicts the target y based on input x and task $\mathcal{T}_k \in \mathcal{T}$. Within a specific task $\mathcal{T}_k$, a dialogue with T turns of interactions between the system and the user can be represented as $\mathcal{X}_T = \{(A_1, U_1), (A_2, U_2) ..., (A_T, U_T)\}$, where A represents the system response and U represents the user input. A predefined slot set[3] $\mathcal{S} = \{S_1, ..., S_J\}$ is provided, where J is the total number of slots for task $\mathcal{T}_k$. The objective of DST is to predict the dialogue state $\mathcal{B}_t$ based on the dialogue context $\mathcal{X}_t$. The dialogue state, $\mathcal{B}_t$, is represented as a set of (slot, value) pairs, denoted as $\mathcal{B}t = \{(S_1, V_1^t), ..., (S_J, V_J^t)\}$, where V_J^t is the value associated with slot S_J at turn t. Essentially, the DST problem is defined as training a dialogue state tracker model $f : \mathcal{X}_t \oplus S_j \rightarrow V_j^t$, where $\oplus$ denotes simple text concatenation.

---

[3]To specify the domain to which a slot belongs, a slot is defined as the concatenation of the specific domain and the slot name, e.g., "<restaurant-area>".
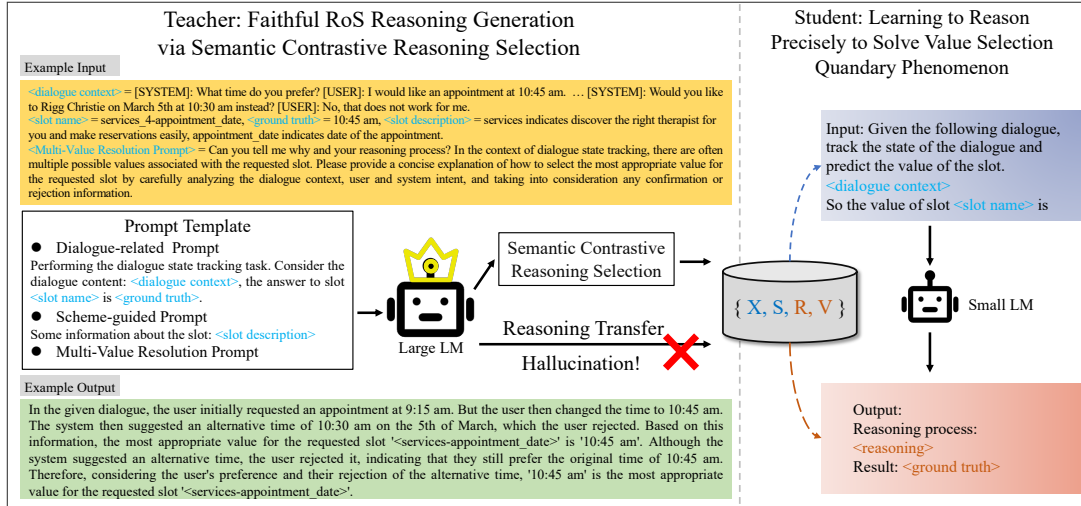
Figure 3: **Overview of the Reason-of-Select (RoS) Distillation method.** (a) Teacher: A large LM prompted to generate a faithful rationale given a dialogue context and the value for the request slot in the training set via the "multi-value resolution" strategy and Semantic Contrastive Reasoning Selection method. (b) Student: A small LM is fine-tuned to generate an accurate rationale and the corresponding value.

3.1 Overview

Our Reason-of-Select Distillation framework capitalizes on LLMs to create a faithful selection reasoning process through a 'multi-value resolution' strategy and Semantic Contrastive Reasoning Selection. This enriched knowledge is then imparted to smaller student models for training, as illustrated in Figure 3.

3.2 Teacher's Reasoning Generation

The process begins with a dialogue-centric prompt that includes dialogue content $\mathcal{X}$, the target slot S_j, and its value V_j to derive a reasoning process $\mathcal{R}$. While traditional prompts like "Tell me why $<S_j>$ is V_j" or the famous "Let's think step by step" prompt (Kojima et al., 2022) generate reasonings that tend to merely highlight the location of the correct answer in the dialogue.

Our innovative "multi-value resolution" prompt $\mathcal{P}_R$ fosters a more elaborate reasoning process, enhancing the model's capability from mere location identification to an in-depth selection reasoning process, as illustrated in Figure 3. This methodological advancement from basic identification to complex process reasoning represents a substantial leap in the model's cognitive capacities.

This stage's input-output representation is $f_{teacher} : \mathcal{X}_t \oplus S_j \oplus V_j^t \oplus \mathcal{P}_R \to \mathcal{R}_j^t$. However, neural LMs often exhibit hallucinations — generating text with tenuous ties to the input (Ji et al., 2023; Maynez et al., 2020), which our unsupervised an-

notation approach can exacerbate. To counter this, we introduce the innovative Semantic Contrastive Reasoning Selection method, pivotal for accurately aligning reasoning with corresponding answers.

3.3 Ensuring Faithful Teaching with Semantic Contrastive Reasoning Selection

Our novel strategy employs a newly schema-guided prompt (yellow background in Figure 3) to direct teacher models toward generating on-topic rationales, significantly improving the source-target correlation. Our Semantic Contrastive Reasoning Selection technique revolutionizes reasoning generation, creating a spectrum of candidates and choosing the most semantically aligned. Inspired by contrastive decoding (Li et al., 2022) in text generation task, our method adapts this concept to suit blackbox LLMs like ChatGPT. We introduce strategic input perturbations to simulate and correct reasoning errors, refining the teacher model's output.

The procedure commences with the LLM generating G diverse reasonings ($\mathcal{R}_1, \ldots, \mathcal{R}_G$). We then introduce two forms of perturbation: value-level and slot-level, as illustrated in Figure 4. While value-level perturbation subtly modifies the ground truth value, slot-level perturbation involves completely replacing the slot-value pair. This latter proves particularly effective in inducing a range of reasoning responses, from logical to nonsensical, thereby enriching our negative sample pool with N perturbated reasonings ($\mathcal{P}\mathcal{R}_1, \ldots, \mathcal{P}\mathcal{R}_N$). Positive samples are represented by the dialogue-centric
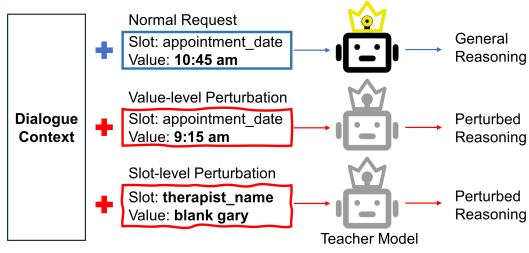
Figure 4: Demonstration of value-level and slot-level perturbations to elicit diverse negative reasonings.

Figure 5: Illustration of the Semantic Contrastive Reasoning Selection method.

prompt, denoted as $\mathcal{DC}$. Utilizing a fixed, pretrained contextual encoder, Sentence-transformers (Reimers and Gurevych, 2019), we then transform these $\mathcal{R}$, $\mathcal{PR}$ and $\mathcal{DC}$ into semantic representations within space $\mathcal{E}$, as illustrated in Figure 5.

The crux of our method lies in selecting the most appropriate reasoning from the generated G candidates, a decision based on their semantic proximity to the positive samples and divergence from the negative samples. This selection process is formulated as an elegant optimization problem:

$$\min Distance(\mathcal{R}_i, \mathcal{DC})$$
$$\text{s.t. } \max \sum_{n=1}^{N} Distance(\mathcal{R}_i, \mathcal{PR}_n) \quad (1)$$

To quantify the relationship between each reasoning and the sample sets, we introduce a sophisticated scoring mechanism:

$$Score(\mathcal{R}_i) = \frac{\exp\left(d\left(\mathcal{R}_i, \mathcal{DC}\right)/\tau\right)}{\sum_{n=1}^{N} \exp\left(d\left(\mathcal{R}_i, \mathcal{PR}_n\right)/\tau\right)} \quad (2)$$

where d is a distance function and τ is a temperature scalar. By computing a score for each reasoning, we discern the optimal choice with the lowest score for training the student model based on its alignment with actual dialog context and its divergence from perturbated content.

3.4 Training Student Models via Reasoning-Enhanced Data

Armed with the annotated dataset $\{\mathcal{X}, S, \mathcal{R}, V\}$, we proceed to train a smaller student model within the self-rationalization framework, emphasizing both predictive and explanatory skills. This approach departs from prior post-rationalization models, where rationales are formulated post-prediction or those employing a multi-task format, treating rationale creation as an auxiliary task.
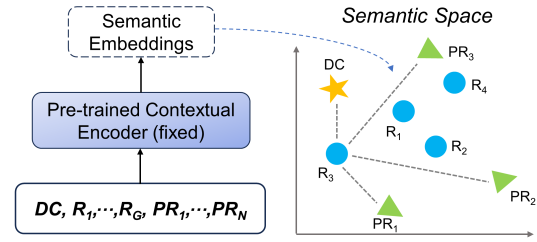
The student model is conditioned to generate a sequence that merges rationale tokens with the corresponding answer tokens in response to a given dialogue context and slot request, as demonstrated in the right part in Figure 3. This task is executed by fine-tuning a pre-trained language model, with "silver" data derived from the teacher model. We employ a standard language modeling loss termed factual reasoning loss:

$$\mathcal{L}_{factual} = -\sum_{j}^{J} \log P\left(\mathcal{R}_j, V_j \mid \mathcal{X}, S_j\right) \quad (3)$$

4 Experiments

4.1 Experimental Setup

Dataset Our experiment employs the Schema-Guided Dialog dataset (SGD) (Rastogi et al., 2020), encompassing 44 services across 19 domains with slot descriptions. Adhering to the Continual DST setup by Zhu et al. (2022), we focus on single-service dialogs, randomly selecting 15 tasks from 44. Each service varies in training sample size and slots. To ensure robustness, we experiment with five task orders via random permutations, aligning with prior studies. Appendix B details the data statistics, task selection, and orderings.

Evaluation Protocol We assess the DST performance using the widely adopted Joint Goal Accuracy (JGA) metric (Wu et al., 2019), which demands accurate predictions for all slot values. We denote $a_{j,i}$ as the JGA on the test set of task $\mathcal{T}_i$ right after training on task $\mathcal{T}_j$. CL performance is assessed using three metrics from Zhu et al. (2022): *(i)* $\mathbf{Avg.JGA} = \frac{1}{K} \sum_{i=1}^{K} a_{K,i}$, representing the average JGA across all tasks after training on the final task $\mathcal{T}_K$. *(ii)* Forward Transfer ($\mathbf{FWT}$) $= \frac{1}{K-1} \sum_{i=2}^{K} a_{i-1,i}$, evaluating generalization by

measuring zero-shot performance, and *(iii)* Backward Transfer $(\mathbf{BWT}) = \frac{1}{K-1}\sum_{i=1}^{K-1} a_{K,i} - a_{i,i}$, quantifying resistance to forgetting by assessing the influence of new learning on previous tasks.

Baselines We evaluate our model against existing Continual DST baselines: ***Fine-tuning:*** Continuously fine-tune the model on new task data. ***Replay:*** Stores $|M|$ instances per task $\mathcal{T}_i$ in memory M_i for joint training with new tasks. ***EWC (Kirkpatrick et al., 2017):*** Maintain a memory but leverage it to compute the Fisher information matrix for regularization. ***AdapterCL (Houlsby et al., 2019):*** Freeze the pre-trained model and independently train a residual Adapter for each task. ***Continual Prompt Tuning (CPT) (Zhu et al., 2022):*** Freeze the backbone model and continually train soft prompts with knowledge transfer in both forward and backward directions. ***DST-EGQA (Cho et al., 2023):*** Reformulate DST as a Question-Answering task using retrieval-augmented in-context learning.

We also include our method with memory replay and multi-task learning as performance caps.

Training Details ChatGPT (using the gpt-3.5-turbo API) and LLaMA-2-70B serve as teacher models for fine-tuning smaller student models including T5-small, T5-base, FlanT5-XL, and LLaMA-7B, utilizing generated rationales. The teacher model's temperature is set to 0.7, generating five candidate reasonings $(G = 5)$ alongside three value-level and three slot-level perturbed negative reasonings $(N = 6)$. In Eq.2, the parameter τ is set to 0.8, employing the Euclidean distance metric. The memory size per task $|M|$ is maintained at 50, in line with previous studies (Zhu et al., 2022). Detailed training specifications are provided in Appendix C.

4.2 Main Results

Table 2 presents the results from various methodologies. Key findings include:

Injecting Reasoning Knowledge Effectively Enhances CL Performance Our RoS method significantly surpasses standard fine-tuning on T5-small, raising the Avg.JGA from 44.1% to 59.0%, and showing gains in both FWT and BWT. This highlights the value of integrating reasoning into Continual DST. Compared with the erstwhile top-performing DST-EGQA, we achieve a new SOTA performance in all metrics, notably increasing Avg.
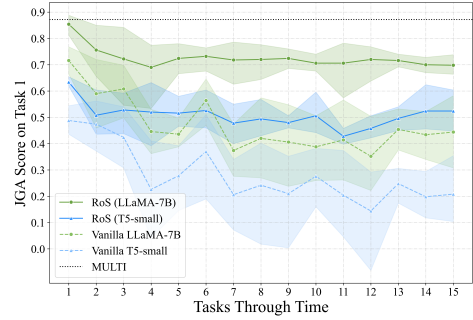

Figure 6: Task 1 performance trajectory during continual DST learning process.

JGA from 55.5% to 59.0% without extra memory, demonstrating our approach's effectiveness in mitigating historical task forgetting. Furthermore, when memory is available, RoS jumps from 59.0% to 72.1%, even outstripping the CPT's multi-task performance. Among various student models, RoS with LLaMA-7B, fine-tuned with ChatGPT's reasoning, stands out, substantially lifting the Avg. JGA from 59.0% to 68.7%, with a remarkable FWT improvement from 25.5% to 51.9%.

Figure 6 shows the models' ability to mitigate catastrophic forgetting by assessing their performance on the initial task post successive task learnings. Models with RoS reasoning skills exhibit a slower forgetting rate, with a 15% average drop in performance in both T5-small and LLaMA-7B. In contrast, vanilla backbone models show a steeper decline, averaging a 28% performance dip, highlighting the importance of domain-agnostic reasoning skills in sustaining historical task performance.

RoS Distillation Exhibits Strong Generalization Ability Our method significantly improves the FWT metric, indicative of robust generalization and zero-shot learning capabilities. It bridges the distribution gap across domains by creating a systematic reasoning chain, evident in handling semantically similar but differently described slots, like <services-appointment_date> and <hotels-check_in_date>. Unlike traditional models, ours utilizes reasoning to recognize similarities, enhancing adaptability to unseen slots.

To further evaluate this generalization, we introduce a new 16th task using MultiWOZ 2.4 data (Ye et al., 2022). The zero-shot performance on this task follows the sequential training of the first 15 tasks from the SGD, as shown in Table 3.

Compared to similar parameter-sized backbones, RoS (T5-small) outperforms the T5-base CPT

Method	Teacher	Student	Avg. JGA	FWT	BWT	+Memory	+Params	+Reg.
Fine-tune	-		$44.1_{0.9}$	$8.3_{1.0}$	$-36.6_{3.9}$	-	-	-
EWC	-		$47.9_{1.1}$	$8.4_{0.9}$	$-38.1_{4.1}$	✓	✓	✓
Replay	-		$58.6_{3.5}$	$10.9_{0.5}$	$-3.2_{2.3}$	✓	-	-
AdapterCL	-	*T5-small*	$49.8_{1.7}$	-	-	-	✓	-
CPT	-		$61.2_{2.5}$	$13.7_{0.8}$	$\mathbf{0.5}_{0.4}$	✓	✓	✓
DST-EGQA	-		$55.5_{3.5}$	$23.6_{2.1}$	$-19.1_{4.2}$	-	-	-
+ Dialogue Memory	-		$68.9_{0.3}$	$22.5_{1.8}$	$-5.9_{1.9}$	✓	-	-
RoS (ours)	*LLaMA-2-70B*	*T5-small*	$59.0_{3.9}$	$25.5_{2.0}$	$-17.9_{3.7}$	-	-	-
+ Dialogue Memory			$\mathbf{72.1}_{0.8}$	$26.7_{2.0}$	$-2.6_{1.5}$	✓	-	-
RoS (ours)	*ChatGPT*	*LLaMA-7B*	$68.7_{4.1}$	$\mathbf{51.9}_{1.7}$	$-8.5_{3.8}$	-	-	-
+ Dialogue Memory			$\mathbf{74.2}_{3.7}$	$\mathbf{52.7}_{1.5}$	$\mathbf{-2.4}_{2.7}$	✓	-	-
CPT Multi-task	-	*T5-base*	$64.0_{1.9}$	-	-	-	✓	✓
DST-EGQA Multi-task	-	*T5-small*	$74.2_{1.8}$	-	-	-	-	-
RoS Multi-task	*LLaMA-2-70B*	*T5-small*	$76.3_{0.5}$	-	-	-	-	-
	ChatGPT	*LLaMA-7B*	$78.9_{0.3}$	-	-	-	-	-

Table 2: CL metric results and reliance on other continual learning techniques. Means and standard variances are reported. We compare models sequentially trained on 15 tasks from the SGD dataset and aggregate results across five domain permutations. The last four rows provide the multi-tasking results, which serve as an upper bound. All rows that use memory are with $M = 50$.

Method	Backbone	Attraction	Restaurant	Train	Average
CPT	*T5-base*	10.05	19.37	3.34	10.92
Fine-tune	*T5-small*	8.40	12.87	2.94	8.07
	LLaMA-7B	40.47	52.73	18.75	37.32
RoS	*T5-small*	**15.74**	**24.07**	**4.05**	**14.62**
	LLaMA-7B	**42.84**	**59.94**	**25.57**	**42.78**

Table 3: Zero-Shot performance on the 16th cross-dataset task using the MultiWOZ 2.4 dataset.

Teacher Model	Initial BERTScore (%)	BERTScore After Contrastive Selection (%)
ChatGPT	19.3	14.2
LLaMA-2-70B	21.7	15.9

Table 4: Hallucination rate of the reasoning generated by teacher models after our contrastive selection method.

baseline, increasing average JGA by 3.7%, from 10.92% to 14.62% across domains. This is particularly notable in the Attraction domain, with JGA rising from 10.05% to 15.74%. Compared to various vanilla backbones, the improvement due to domain-agnostic reasoning averages a 6% increase, further confirming RoS's robust generalization.

To assess the hallucination rate of the teacher model, we utilized SelfCheckGPT (Manakul et al., 2023) and employed BERTScore for evaluation. BERTScore provides scores ranging from [0.0, 1.0], where higher values indicate a greater likelihood of non-factual content, implying an increased probability of hallucination. Our evaluation included the average BERTScore for the initial generation of five candidate reasonings by the teacher models, the BERTScore after applying the Semantic Contrastive Reasoning Selection method, and the hallucination rate of the distilled student. The summarized results are presented in the Table 4.

Table 4 illustrates that for both teacher models, ChatGPT and LLaMA-2-70B, the average

BERTScore for the initial five reasoning candidates were 19.3% and 21.7%, respectively. After applying the Semantic Contrastive Reasoning Selection method, the BERTScore decreased to 14.2% and 15.9%, affirming the effectiveness of our proposed approach in enhancing the factual of reasoning. Regarding the student models, there is no significant difference in BERTScore when distilled with different teacher models. Additionally, we observed that the quality of reasoning generated by the student models improves as the model size increases.

4.3 Ablation Study

The Role of Model Size Variation in Teacher and Student Models   Our study investigates the effect of teacher and student model sizes on performance. Figure 7 shows a clear trend: larger student models, like LLaMA-7B, outperform smaller ones, such as T5-small. We first evaluate LLaMA-7B's continual learning in DST, finding it outperforms most T5-small or T5-base-based methods. However, applying our RoS method with reasoning is the most notable enhancement, especially in smaller models. For example, integrating reasoning into T5-small
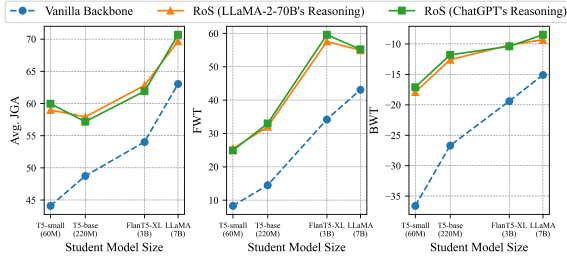
Figure 7: Comparative task performance across varying sizes of teacher and student models.

Models	Avg. JGA	FWT	BWT
LLaMA-7B (backbone)	63.04	43.10	-15.10
+ Max Score	65.72	49.05	-13.34
+ Random Selection	66.98	51.14	-11.73
+ Min Score (Ours)	**70.17**	**52.06**	**-8.90**

Table 5: Ablation study on the effectiveness of strategies to alleviate hallucination in teacher model's reasoning.

yields an impressive 15% average increase in all evaluation metrics. Contrary to our initial assumption that students guided by ChatGPT might surpass those instructed by LLaMA-2-70B, the results indicate equal reasoning quality from both teacher models, suggesting their similar effectiveness in enhancing student model learning.

Effectiveness of Semantic Contrastive Reasoning Selection Method   To assess our method's ability to reduce reasoning hallucinations, we compared three reasoning selection strategies: random selection from G candidates and selecting the highest and lowest scoring (ours) reasonings, as per Equation 2. The results are detailed in Table 5.

Initial results show that all methods significantly outperform the vanilla LLaMA, with even random selection improving Avg. JGA by 4.94%. This highlights the inherent reasoning quality in current teacher models. Yet, refining the selection to eliminate hallucinatory reasonings and aggregate higher-quality reasonings further enhanced performance. Compared to random selection, our method shows additional gains of 3.2%, 0.9%, and 2.8% in the evaluated metrics. In contrast, selecting the highest-scoring result, i.e., favoring less effective reasoning, resulted in reduced performance. These findings robustly affirm the effectiveness of our semantic similarity-based selection approach.

4.4   Case Study

Table 6 presents a key test case and various methods' predictions to showcase student models' ad-

Table 6: Exemplary dialogue from the SGD test set with two possible values, with predictions by the vanilla LLaMA-7B and our meta-reasoning augmented method applied to two distinct backbones.

vanced reasoning capabilities. The dialogue involves two values for the <pickup_time> slot: "17:15" and a later update to "half past 1 in the afternoon". The vanilla LLaMA model incorrectly selects "17:15", while our model effectively identifies all possible values and accurately determines the correct one. Interestingly, our RoS T5-small model's output of "1:30 pm" mirrors the ground truth but is marked incorrect due to current evaluation criteria. This highlights a challenge in matching model outputs with expected formats, suggesting a direction for future refinement.

5   Related Work

5.1   Knowledge Distillation from LLMs

Knowledge distillation in NLP, crucial for transferring insights from larger teacher models to smaller student models, has evolved recently, especially in extracting reasoning from LLMs. Recent studies (Hsieh et al., 2023; Wang et al., 2022; Yang et al., 2023; Xu et al., 2023b) highlight LLMs' role in bolstering smaller models. Li et al. (2023) furthered this with a symbolic Chain of Thought (CoT) distillation, enhancing smaller models through CoT prompting. Despite these advancements, the faithfulness of generated rationales, critical for student models' behavior, often must be addressed. Wang et al. (2023) proposed a contrastive decoding method to reduce hallucinations, but its reliance on teacher model logits limits its applicability for black-box LLMs. Current reasoning distillation methods are unsuitable for DST task and lead to unproductive reasoning processes.

To overcome these issues, our method focuses on the Value Selection Quandary in DST, introducing

a 'multi-value resolution' prompt and a Semantic Contrastive Reasoning Selection method, enhancing knowledge transfer accuracy and relevance.

5.2 Continual Dialogue State Tracking

Continual Learning in task-oriented dialogue systems, focusing on mitigating catastrophic forgetting, has employed various methods such as architecture-based (Shen et al., 2019; Geng et al., 2021; Xu et al., 2023a), rehearsal-based (Rebuffi et al., 2017; Hou et al., 2019; Lu et al., 2021b), and regularization-based (Li and Hoiem, 2017; Feng et al., 2024; Lu et al., 2021a). In DST, contributions like Madotto et al. (2020) and Liu et al. (2021) have utilized these CL strategies, with Zhu et al. (2022) introducing Continual Prompt Tuning (CPT) to fine-tune domain-specific soft prompts. Furthermore, DST-EGQA (Cho et al., 2023) employs a question-answering framework based on example-guided learning. However, its reliance on fixed QA templates may limit adaptability across diverse domains. Our method, in contrast, offers enhanced time efficiency and flexibility during testing, eliminating the need for sample retrieval, thus presenting a more efficient solution.

6 Conclusion

This study introduces the Reason-of-Select (RoS) distillation method to address the Continual DST task's catastrophic forgetting and "Value Selection Quandary" challenges. Enhancing smaller models with a novel domain-agnostic 'meta-reasoning' capability effectively broadens their reasoning horizon and aligns data distribution shifts across different domains. Two innovative methods, the "multi-value resolution" strategy and Semantic Contrastive Reasoning Selection, further strengthen RoS by ensuring reliable knowledge transfer. Extensive experiments demonstrate our method's superior performance and robust generalization.

Limitations

In contrast to the conventional knowledge distillation process, our approach necessitates additional computational resources during the preparation of training data and the training phase of the student model. Firstly, our Contrastive Reasoning Selection method requires the generation of multiple candidate reasonings, along with several value-level and slot-level perturbation reasonings. This process is more time-consuming than standard rational

distillation, as it involves extensive data generation by the teacher model. Secondly, training student models with rationales introduces a minor increase in computational overhead during the training period. However, it's important to note that these expenses are one-time costs. When it comes to testing, our model, already adept at reasoning, does not incur extra time for predictions, and the time involved in generating reasoning in the test phase is almost inconsequential.

Acknowledgments

References

Hyundong Cho, Andrea Madotto, Zhaojiang Lin, Khyathi Raghavi Chandu, Satwik Kottur, Jing Xu, Jonathan May, and Chinnadhurai Sankar. 2023. Continual dialogue state tracking via example-guided question answering. *arXiv preprint arXiv:2305.13721*.

Yujie Feng, Xu Chu, Yongxin Xu, Guangyuan Shi, Bo Liu, and Xiao-Ming Wu. 2024. Tasl: Continual dialog state tracking via task skill localization and consolidation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023a. Towards llm-driven dialogue state tracking. *arXiv preprint arXiv:2310.14970*.

Yujie Feng, Jiangtao Wang, Yasha Wang, and Xu Chu. 2022. Spatial-attention and demographic-augmented generative adversarial imputation network for population health data reconstruction. *IEEE Transactions on Big Data*.

Yujie Feng, Jiangtao Wang, Yasha Wang, and Xu Chu. 2023b. Towards sustainable compressive population health: a gan-based year-by-year imputation method. *ACM Transactions on Computing for Healthcare*, 4(1):1–18.

Yujie Feng, Jiangtao Wang, Yasha Wang, and Sumi Helal. 2021. Completing missing prevalence rates for multiple chronic diseases by jointly leveraging both intra-and inter-disease population health data correlations. In *Proceedings of the Web Conference 2021*, pages 183–193.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Binzong Geng, Fajie Yuan, Qiancheng Xu, Ying Shen, Ruifeng Xu, and Min Yang. 2021. Continual learning for task-oriented dialogue system with iterative network pruning, expanding and masking. *arXiv preprint arXiv:2107.08173*.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, Carel van Niekerk, and Milica Gašić. 2023. Chatgpt for zero-shot dialogue state tracking: A solution or an opportunity? *arXiv preprint arXiv:2306.01386*.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839.

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.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Zhiyuan Hu, Yue Feng, Yang Deng, Zekun Li, See-Kiong Ng, Anh Tuan Luu, and Bryan Hooi. 2023. Enhancing large language model induced task-oriented dialogue systems through look-forward motivated goals. *arXiv preprint arXiv:2309.08949*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. *arXiv preprint arXiv:2306.14050*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

Bo Liu, Liming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. 2023. How good are large language models at out-of-distribution detection? *arXiv preprint arXiv:2308.10261*.

Qingbin Liu, Pengfei Cao, Cao Liu, Jiansong Chen, Xunliang Cai, Fan Yang, Shizhu He, Kang Liu, and Jun Zhao. 2021. Domain-lifelong learning for dialogue state tracking via knowledge preservation networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2301–2311.

Zexin Lu, Keyang Ding, Yuji Zhang, Jing Li, Baolin Peng, and Lemao Liu. 2021a. Engage the public: Poll question generation for social media posts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 29–40.

Zexin Lu, Jing Li, Yingyi Zhang, and Haisong Zhang. 2021b. Getting your conversation on track: Estimation of residual life for conversations. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1036–1043. IEEE.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, and Zhiguang Wang. 2020. Continual learning in task-oriented dialogue systems. *arXiv preprint arXiv:2012.15504*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8689–8696.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Yilin Shen, Xiangyu Zeng, and Hongxia Jin. 2019. A progressive model to enable continual learning for semantic slot filling. 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 1279–1284.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. Scott: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

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.

Yongxin Xu, Xu Chu, Kai Yang, Zhiyuan Wang, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023a. Seqcare: Sequential training with external medical knowledge graph for diagnosis prediction in healthcare data. In *Proceedings of the ACM Web Conference 2023*, pages 2819–2830.

Yongxin Xu, Kai Yang, Chaohe Zhang, Peinie Zou, Zhiyuan Wang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023b. Vecocare: visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 4921–4929.

Kai Yang, Yongxin Xu, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. Kerprint: local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5357–5365.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360, Edinburgh, UK. Association for Computational Linguistics.

Qi Zhu, Bing Li, Fei Mi, Xiaoyan Zhu, and Minlie Huang. 2022. Continual prompt tuning for dialog state tracking. *arXiv preprint arXiv:2203.06654*.

A   Description of Prompt Templates

A.1   Prompt Templates for Generating Teacher Model Reasonings

Below, we present a specific example of the prompt used to elicit reasoning generation from the teacher model.

```
{
```
"instruction": Just return a concise reasoning process.
"input": Performing the dialogue state tracking task. Consider the dialogue content: "[Previous dialogue omitted ...] [USER]: I would like an appointment at 10:45 am. [SYSTEM]: When is the appointment for? [USER]: The appointment is for the 11th of March. [SYSTEM]: Booking appointment with Rigg Christie on March 11th at 10:45 am. [Remaining dialogue omitted ...]", the answer to slot <services-appointment_date> is '10:45 am'.

Some information about the slot: services indicates Discover the right therapist for you and make reservations easily, appointment_date indicates Date of the appointment.

Can you tell me why and your reasoning process? In the context of dialogue state tracking, there are often multiple possible values associated with the requested slot. Please provide a concise explanation of how to select

the most appropriate value for the requested slot by carefully analyzing the dialogue context, user and system intent, and taking into consideration any confirmation or rejection information.
}

A.2 Prompt Templates for Fine-Tuning the Student Model

Below, we provide a specific example of the prompt used for fine-tuning the student model.

```
{
"instruction": Given the following
dialogue, track the state of the dialogue
and predict the value of the slot
<alarm_1-new_alarm_name>.
"input": [USER]: I want to check the
alarms I have. [SYSTEM]: There are 2
alarms which you have set currently, with
one of the alarms being at 6:30 am, and it
is called Wake Up. [USER]: Alright, that
is good. [SYSTEM]: Are you interested to
add another alarm? [USER]: Actually I am,
I do want to add another alarm. I want
the alarm to be called Grocery run.

  [slot] alarm_1-new_alarm_name,    it
indicates Name to use for the
new alarm.   So the value of slot
<alarm_1-new_alarm_name> is
"output": grocery run
}
```

B Dataset Statistics

Here, we offer a detailed description of the dataset used in Continual DST (Feng et al., 2022). Table 7 displays the number of slots for each of the 15 services used in our experiments and the count of samples in the training, validation, and test sets. Table 8 illustrates the training sequence for these 15 tasks in the context of continual learning.

C Implementation

For reasoning generation by the teacher model (Feng et al., 2021), we utilized the following hyperparameters:

- **LLaMA-2 (70B)**: Model set as Llama-2-70B-chat-GPTQ, with temperature at 0.7, top_p at 0.9, top_k at 40, and a maximum of 512 new tokens.

- **ChatGPT**: Generation was conducted from October 21th to 30th, 2023, using the 'gpt-3.5-turbo' API. The settings were temperature = 0.7 and max_tokens = 256.

For training the student models, we applied these hyperparameters:

- **T5-small (60M)** and **T5-base (220M)**: Training was conducted with a learning rate of 3e-4, batch size of 8, maximum input length of 512 (1024 for reasoning fine-tuning), maximum target length of 128 (512 for reasoning fine-tuning), and 5 epochs.

- **T5-XL (3B)**: To optimize training time, we used LORA with a learning rate of 3e-4, batch size of 2 (8 for reasoning fine-tuning), maximum input length of 512 (1024 for reasoning fine-tuning), maximum target length of 128 (512 for reasoning fine-tuning), and 5 epochs. Lora settings included r = 8, alpha = 16, dropout = 0.05, targeting modules ['q', 'v']. For testing, we set max new tokens to 128 (512 with reasoning).

- **LLaMA (7B)**: Also utilizing LORA for efficiency, with a learning rate of 3e-4, batch size of 128, a cutoff length of 512 (1024 for reasoning fine-tuning), and 5 epochs. Lora settings were r = 8, alpha = 16, dropout = 0.05, targeting modules [[q_proj,k_proj,v_proj,o_proj]]. For testing, settings included temperature = 0.02, top_p = 0, top_k = 1, num_beams = 1, max new tokens = 128 (512 with reasoning).

Experiments are carried out using Nvidia RTX 3090 GPUs. Training durations for various models on total 15 tasks are as follows: T5-small takes about 1 hours on a single 3090 GPU, T5-base takes nearly 2 hours on a single 3090 GPU, FlanT5-XL takes about 16 hours on a single 3090 GPU, and LLaMa-7B takes 21 hours on two 3090 GPUs. Results are averaged across five different task orders and include the standard error in the tables and plots provided (Feng et al., 2023b).

Details on Perturbation Operation: We altered the ground truth value with a random value from the dataset for the same slot for value-level perturbation. However, given the robustness of the teacher model, which can often detect errors and provide correct explanations even if the misleading value does not cause significant confusion, we introduced a slot-level perturbation. This involves completely changing the slot and substituting it with a random slot-value pair from the current task. To address the "Value Selection Quandary" com-

Task ID	Service	# Slots	# Dialogs			# Samples			Avg. tokens	
			Train	*Dev*	*Test*	*Train*	*Dev*	*Test*	*Context*	*Query*
30	services_4	5	86	13	25	680	97	208	154	49
31	flights_1	10	560	80	160	4680	667	1379	168	10
32	services_3	5	131	19	38	959	143	290	143	54
33	flights_3	8	65	10	19	420	75	116	133	79
34	trains_1	7	58	9	17	415	67	117	131	76
35	homes_2	8	62	9	18	424	56	139	140	89
36	rentalcars_2	6	77	11	23	631	91	185	157	61
37	restaurants_1	9	256	37	74	2098	297	581	153	10
38	music_1	6	68	10	20	468	73	142	118	61
39	hotels_4	7	80	12	23	559	99	141	134	72
40	media_2	5	32	4	10	215	29	71	112	59
41	hotels_3	6	90	13	26	737	100	193	157	64
42	rentalcars_3	7	44	7	13	332	55	99	148	72
43	hotels_1	7	99	14	29	868	105	250	161	71
44	homes_1	7	244	35	70	1829	282	540	159	81

Table 7: Statistics of the 15 services we used in experiments.

Task order	Tasks' IDs in order														
Order1	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
Order2	39	33	36	42	40	37	38	34	32	35	41	31	30	44	43
Order3	30	41	38	31	43	39	40	33	34	44	37	36	32	35	42
Order4	43	40	44	38	30	37	31	39	32	35	41	34	33	36	42
Order5	30	33	44	31	38	32	42	40	37	43	36	39	41	35	34

Table 8: Five task orders of all our 15 tasks experiments.

mon in long dialogues and to save the time and cost of using teacher LLMs, we engage teacher models to generate reasoning for samples with turn id greater than 10, aligning with findings from Figure 2(a). For briefer dialogues, where typically only one possible value appears, our examination indicates teacher models usually return direct reasoning, like "In the given dialogue, the user explicitly mentions their destination as <1150 Webster street>. Therefore, the answer to the slot <ridesharing-destination> is 1150 Webster street.". Thus, for uniformity across all short dialogues, we've standardized the prompt to "In the given dialogue, the value of the requested slot is explicitly mentioned," enabling student models to learn from this consistent template.

D Time Complexity Comparison

Firstly, we examined the training time of the student model. When using LLaMA-7B as the back-

bone, the inclusion of additional rationales led to an approximately 1.5 times increase in training time compared to the original training. For smaller student models like T5-small, the training time remained consistent with no inclusion of reasoning. In terms of inference speed, a comparison is presented in the table 9 below: The table 9 illustrates

Inference Speed (Samples/Min)	T5-small	T5-base	FlanT5-XL	LLaMA-7B
Vanilla Method	857	222	51	41
RoS (ours)	86	24	7	5

Table 9: Inference time for different models.

a ninefold increase in time consumption due to generating rationales. For instance, during the generation of rationales, the LLaMA-7B experiences a decrease in samples per minute from 41 to 5. Despite the reduction in inference speed, this change contributes to a significant performance improvement.