Learning to Understand and Correct: Evolving Curricular Training for Long-Context Dialogue State Tracking

Anonymous ACL submission

Abstract

Real-world dialogue systems frequently encounter long conversations involving complex goal updates, yet existing Dialogue State Tracking (DST) models struggle in these settings. In this paper, we identify and systematically analyze three fundamental challenges behind this performance degradation: (1) Context Noise, where irrelevant turns in long histories dilute key signals; (2) State Revision, where users modify previously defined slots; and (3) Error Propagation, a risk inherent to recursive tracking methods. To address these challenges, we advocate a paradigm shift from the prevailing Direct Modeling approach to Recursive Modeling, which effectively insulates the model from context noise by conditioning on the previous dialogue state. Building on this paradigm, we propose LongDST, a novel framework equipped with sophisticated self-correction capabilities. LongDST incorporates two synergistic training strategies: Progressive State Exposure, which mitigates error propagation by narrowing the training-inference gap, and Traceability-Aware Curriculum Learning, which enhances state revision accuracy by scheduling training samples based on revision difficulty. Extensive experiments on benchmark datasets demonstrate that LongDSTsuccessfully overcomes the bottlenecks of long-context DST, consistently outperforming strong baselines across multiple backbones. The source code[1] is provided for reproducibility.

1 Introduction

Task-oriented dialogue systems serve as powerful tools for assisting users with diverse tasks, such as booking trains or making hotel reservations (Huang et al., 2020). Dialogue State Tracking (DST), a crucial component of these systems, is responsible for identifying domain-slot pairs (e.g., <Hotel-Name>) and extracting the corresponding values (e.g., "Eraina") from the conversation (Peng et al., 2020; Lin et al., 2020; Zhang et al., 2025).

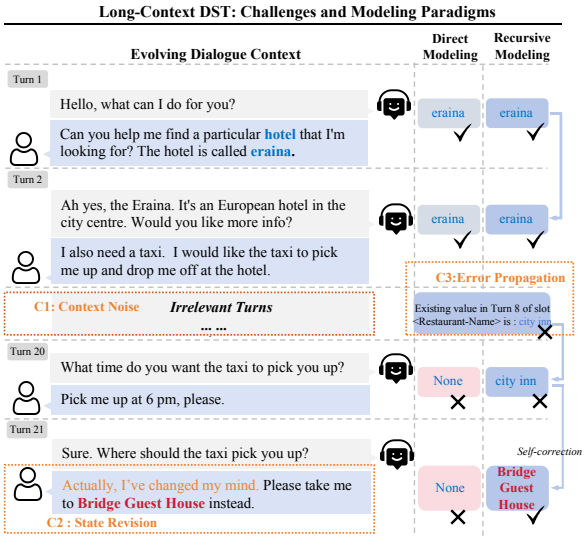

Figure 1: Challenges in Long-context DST for the <Hotel_Name> slot. The figure illustrates how *Context Noise (C1)* hinders Direct Modeling, while Recursive Modeling faces the challenge of *Error Propagation (C3)* from incorrect prior states (e.g., "city inn"). Both paradigms must also address *State Revision (C2).*

While prior DST research has primarily focused on cross-domain generalization (Wu et al., 2023; Finch et al., 2024; Tavares et al., 2023; Dong et al., 2024), real-world deployment increasingly reveals a more fundamental challenge: *Long-context Dialogue State Tracking*. As user requirements grow in complexity, models must maintain accurate slot values across evolving dialogue histories (see Figure 1). Our empirical analysis confirms this difficulty, revealing a significant degradation in Joint Goal Accuracy (JGA) as dialogue length increases (see Section 2). We attribute this decline to two inherent challenges: **(C1) Context Noise:** Long conversations often contain numerous irrelevant turns that dilute key signals, causing crucial information to be overlooked or misinterpreted (Gao

[1] https://anonymous.4open.science/r/longdst-182B/

et al., 2024; Li et al., 2025). **(C2) State Revision:** Users frequently update or modify their intents in later turns, requiring the model to dynamically track and overwrite previous values.

Paradigm Shift. To combat the context noise in long histories, we advocate transitioning from *Direct Modeling*—adopted by most existing methods (Lee et al., 2019; Heck et al., 2020; Hosseini-Asl et al., 2020; Peng et al., 2020; Feng et al., 2024; Zhang et al., 2022, 2023a), where the state relies solely on the entire dialogue history—to *Recursive Modeling* (illustrated in Figure 1). In this paradigm, the model inputs both the current dialogue history and the explicit state from turn $t-1$, which enables the model to bypass irrelevant context and focus purely on key information relevant to state updates.

However, recursive modeling is a double-edged sword: it introduces the risk of **(C3) Error Propagation**, where an incorrect prediction in an earlier turn corrupts subsequent states. Consequently, a robust model must possess sophisticated *Self-Correction* capabilities—not only to handle user-driven State Revisions (C2) but also to rectify its own past mispredictions.

To cultivate these capabilities, we propose a novel framework, **LongDST**. The framework is built on a recursive modeling paradigm that insulates the model from *context noise*, while incorporating two synergistic training strategies to address error propagation and state revision.

Specifically, we introduce *Progressive State Exposure* (PSE) to mitigate error propagation. PSE narrows the training–inference gap by gradually replacing ground-truth previous states with the model's own predictions, thereby encouraging self-correction. In addition, we propose a *Traceability-Aware Curriculum Learning* (TACL) to facilitate state revision. TACL schedules training samples according to state revision difficulty, quantified by *Revision Richness* (the extent of state updates) and *Revision Traceability* (the ease of locating revision cues). By progressing from simple extraction to complex revision scenarios, this curriculum enables more effective and stable training.

To summarize, our main contributions are:

- We identify and systematically analyze three fundamental challenges in long-context DST, namely *Context Noise*, *State Revision*, and *Error Propagation*, and empirically demonstrate their impact on tracking performance.
- We propose LongDST, a novel recursive modeling framework that mitigates context noise and equips the model with self-correction capabilities through two synergistic training strategies.
- We introduce *Progressive State Exposure* and *Traceability-Aware Curriculum Learning*, and demonstrate their effectiveness across multiple backbones and benchmark datasets for long-context DST.

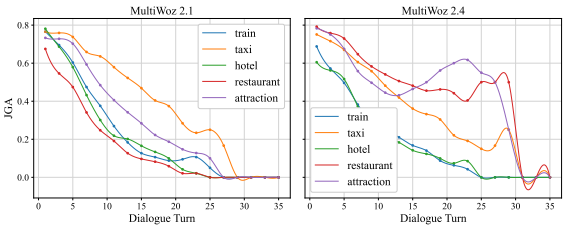

Figure 2: Long-context DST Challenge: JGA drops sharply as dialogue turns increase.

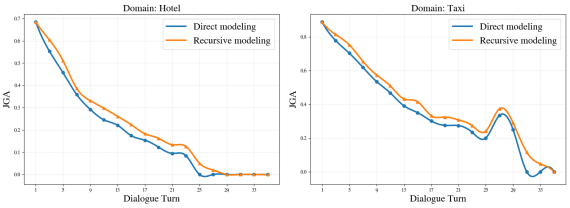

Figure 3: Performance comparison across dialogue turns, demonstrating that Recursive modeling consistently outperforms Direct modeling.

2 Analysis of Challenges in Long-Context Dialogue State Tracking

2.1 Challenge 1: Context Noise and the Need for Recursive Modeling

As dialogues extend over dozens of turns, the accumulation of irrelevant information—referred to as *Context Noise*—poses a severe challenge to understanding. As shown in Figure 2, the JGA of a T5-based DST model drops sharply as dialogue length increases, regardless of the domain, indicating a failure to distinguish critical constraints from this noisy background. To address this, we adopt **Recursive Modeling**, which explicitly conditions the input on the previous turn's state (S_{t-1}). Figure 3 demonstrates that providing an accurate S_{t-1} eliminates this performance degradation. This confirms that S_{t-1} serves as a compact context summary, effectively filtering noise and enabling the model to focus purely on information updates.

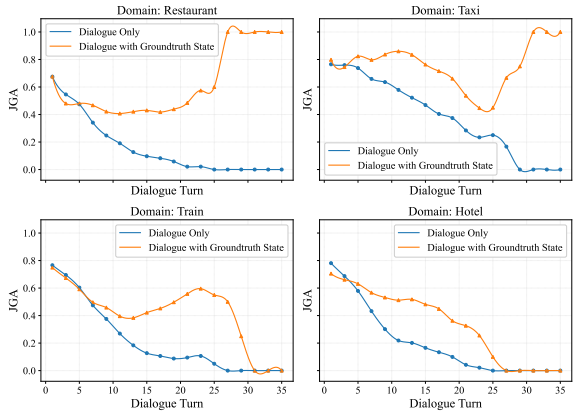

Figure 4: Incorporating the previous turn's ground-truth state into the model significantly alleviates JGA degradation over dialogue turns. This indicates that without explicit correction signals, models struggle to self-correct errors accumulated from earlier dialogue states.

2.2 Challenge 2: Error Propagation

While effective at mitigating context noise, recursive modeling is inherently susceptible to **Error Propagation**. Because each prediction conditions on previous outputs, errors made in early turns can accumulate and progressively degrade future performance. To quantify this effect, we conduct a controlled comparison on MultiWOZ 2.1 using a T5-based DST model under two inference settings (Figure 4): (1) *Ground-Truth State Inference*, where the gold state from the previous turn is provided, simulating an error-free condition; and (2) *Dialogue-Only Inference*, which reflects the realistic setting where the model relies on its own predicted state. The results show a clear divergence between the two settings. While performance under Ground-Truth State Inference remains stable, Dialogue-Only Inference exhibits a pronounced decline in JGA as dialogue length increases. The growing gap between the two curves directly isolates the effect of error propagation, indicating that without explicit correction mechanisms, the model struggles to recover from its own past mistakes.

This analysis underscores that although recursive modeling is essential for long-context DST, its effectiveness critically depends on addressing the error propagation bottleneck.

3 Proposed Method: LongDST

Problem Formulation A task-oriented dialogue consists of a sequence of utterances alternating between the assistant and user, denoted as $\{A_1, U_1, ..., A_t, U_t\}$, where A and U represent assistant and user utterances, respectively. A predefined slot set[2] $\mathcal{S} = \{S^j\}_{j=1}^{J}$ is provided, where J is the total number of slots across all domains. Given the dialogue context $C_t = \{(A_1, U_1), (A_2, U_2), ..., (A_t, U_t)\}$ at turn t, the goal of Dialogue State Tracking (DST) is to predict the dialogue state consisting of slot-value pairs (S_t^j, V_t^j) for all slots j. Formally, the model f_d aims to learn a mapping $f_d : \mathcal{C}_t \oplus S_t'^j \rightarrow V_t^j$, where $S_t'^j$ denotes enhanced slot descriptions and $\oplus$ represents concatenation.

Crucially, our training follows a *zero-shot cross-domain* setup: the model is trained on dialogues from a subset of domains $\mathcal{D}_{train}$, and evaluated on unseen domains $\mathcal{D}_{test}$, where $\mathcal{D}_{train} \cap \mathcal{D}_{test} = \emptyset$. This setting reflects practical scenarios where new domains arise without annotated training data. As different domains exhibit distinct slot structures and dialogue characteristics, the model must generalize across domains to accurately predict dialogue states in unseen domains.

Overview To address the three core challenges of long-context DST, namely Context Noise (C1), State Revision (C2), and Error Propagation (C3), we propose **LongDST**, a framework designed for robust long-context tracking. First, to combat Context Noise (C1), we adopt a *Recursive Modeling Training* paradigm, utilizing the previous state as a recursive memory to filter irrelevant history. Second, to mitigate the Error Propagation (C3) inherent in recursive models, we introduce *Progressive State Exposure*, which dynamically bridges the gap between training and inference by mixing ground-truth and predicted states. Finally, to master the complex State Revision (C2) required in dynamic dialogues, we design a *Traceability-Aware Curriculum*. As illustrated in Figure 5, this strategy organizes training samples from *Easy* to *Hard* based on the intensity of state revisions and the difficulty of tracing them, progressively enhancing the model's ability to handle evolving user intents.

Recursive Modeling Training To tackle Context Noise (C1), we transition from direct modeling to a recursive paradigm. As shown in the "Block Illustration" of Figure 5, we explicitly condition the current prediction on the previous turn's state.

[2]To specify the domain to which a slot belongs, a slot is defined as the concatenation of the domain and slot name, e.g., "<Restaurant_Area>".

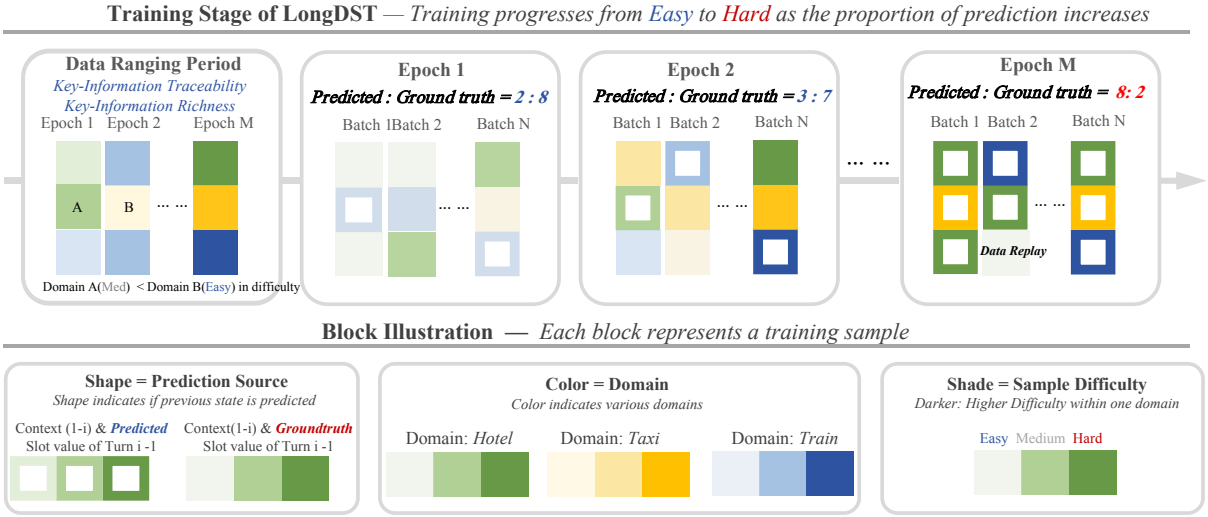

Figure 5: **Overview training process of LongDST.** Top: PSR enhances the model's ability for long-context understanding and self-correction by gradually replacing gold dialogue states with model-predicted ones over training epochs. All samples use gold states during the initial warm-up epoch; in subsequent epochs, the replacement ratio increases to introduce more training difficulty. Samples are also ordered at the data ranging period based on its revision traceability difficulty. Bottom: Each block represents a training sample. The shape indicates the source of the previous dialogue state (prediction or ground-truth), the color denotes the training domain, and the shade intensity reflects revision traceability difficulty.

Given context C_t up to turn t:

$$C_t = \{(A_1, U_1), (A_2, U_2), \ldots, (A_t, U_t)\}, \quad (1)$$

we define the enhanced input at turn t as:

$$X_t = C_t \oplus \hat{Y}_{t-1}, \quad (2)$$

where $\hat{Y}_{t-1} = \{(S^j_{t-1}, \hat{V}^j_{t-1})\}_{j=1}^{J}$ represents the predicted dialogue state from the previous turn, and $\oplus$ denotes sequence-level concatenation.

The DST model f_d is then trained to predict the current slot values:

$$\hat{V}^j_t = f_d(X_t, S^j_t), \quad (3)$$

where S^j_t is the target slot at turn t, and $\hat{V}^j_t$ is the predicted value.

Progressive State Exposure. Recursive modeling introduces Error Propagation (C3). To mitigate this, we employ *Progressive State Exposure (PSE)*, which gradually replaces ground-truth histories with model predictions. Concretely, PSE adopts a two-stage training strategy. In the initial warm-up phase, covering the first α fraction of training steps, we exclusively use ground-truth states to construct the historical input, enabling the model to learn basic instruction-following and state prediction without interference from noisy histories. Afterward, we progressively reduce the reliance on ground-truth states by linearly decaying the *state replacement rate*. Let n denote the current training step and N the total number of steps. The replacement rate γ_n decreases linearly from $\gamma_{\max}$ (typically 1.0) to $\gamma_{\min}$:

$$\gamma_n = \begin{cases} 1.0 & \text{if } n < \alpha N \\ \gamma_{\max} - \left(\frac{n-\alpha N}{(1-\alpha)N}\right) \cdot (\gamma_{\max} - \gamma_{\min}) & \text{otherwise.} \end{cases} \quad (4)$$

At each training step n, we define a state replacement ratio γ_n as described above, which determines the proportion of training samples that will use the gold dialogue state from the previous turn. Let $\mathcal{B}$ denote the current batch and $\mathcal{R}_n \subseteq \mathcal{B}$ denote the subset of samples using the gold state. Specifically, $\mathcal{R}_n$ is formed by randomly sampling each example in $\mathcal{B}$ with probability γ_n.

The model input for turn t is constructed as:

$$X^{(i)}_t = C^{(i)}_t \oplus \begin{cases} Y^{*(i)}_{t-1}, & \text{if } i \in \mathcal{R}_n \\ \hat{Y}^{(i)}_{t-1}, & \text{otherwise,} \end{cases} \quad (5)$$

where i indexes the training samples in batch $\mathcal{B}$, $C^{(i)}_t$ denotes the dialogue context, $Y^{*(i)}_{t-1}$ is the ground-truth state from turn $t-1$, and $\hat{Y}^{(i)}_{t-1}$ is the model's own prediction.

Traceability-Aware Curriculum Learning. To handle State Revision (C2), we organize training by difficulty. As shown in Figure 5, training pro-

gresses from *Easy* (high traceability) to *Hard* (complex revisions), quantified by two metrics:

1. Revision Richness. This measures the intensity of state update for slot s in dialogue d. We extract the sequence of state-changing events:

$$(t_1^{(s)}, v_1^{(s)}),\ (t_2^{(s)}, v_2^{(s)}),\ \ldots,\ (t_{K_s}^{(s)}, v_{K_s}^{(s)}),$$

where each event corresponds to an assignment or reset of the slot value. The most direct measure of state-evolution richness is the number of changes:

$$C_s(d) = K_s. \tag{6}$$

We normalize this count into $[0, 1]$:

$$\text{Rev\_Rich}(d, s) = \frac{C_s(d)}{\max_{(d',s')} C_{s'}(d')}. \tag{7}$$

2. Revision Traceability. While richness measures the amount of key information, it does not indicate how difficult it is for the model to retain and utilize such information. Traceability focuses on how well the model preserves slot-related information after each state-changing event.

Let $t_{\text{pred}}^{(s)}$ denote the prediction turn (typically the final turn), and let h_t be the model's hidden representation at turn t. For every state-changing event at turn $t_k^{(s)}$, we compute the representational drift between that event and the prediction point:

$$\Delta_k(d, s) = 1 - \cos\left(h_{t_k^{(s)}},\ h_{t_{\text{pred}}^{(s)}}\right). \tag{8}$$

The overall traceability difficulty is defined as

$$\text{Traceability}(d, s) = \frac{1}{K_s} \sum_{k=1}^{K_s} \Delta_k(d, s). \tag{9}$$

This score lies in $[0, 2]$, and larger values indicate greater difficulty. Combining these, the final difficulty is:

$$\text{Diff}(d, s) = \tfrac{1}{2}\left(\text{Rev\_Rich}(d, s) + \text{Traceability}(d, s)\right). \tag{10}$$

Samples with higher Diff scores (darker blocks in Figure 5) are introduced later in training, enabling the model to systematically master complex state revisions.

To prevent catastrophic forgetting—where the model loses proficiency on simpler patterns while adapting to harder ones—we further incorporate a *Data Replay* strategy. Specifically, in later training epochs, we periodically mix in a subset of lower-difficulty samples with a ratio of 0.2. This ensures the model maintains robustness across all difficulty levels throughout the curriculum.

4 Experiments

4.1 Experiment Setup

Dataset Our experiments use the MultiWOZ 2.1 (Eric et al., 2019) and MultiWOZ 2.4 (Ye et al., 2021) datasets, widely utilized in previous cross-domain research. MultiWOZ 2.4 builds on MultiWOZ 2.1 with improved DST evaluation and reannotated validation and test sets. We use MultiWOZ 2.4 as a reliable dataset for testing and MultiWOZ 2.1 to assess model robustness and its ability to handle annotation noise. Following prior work (Wu et al., 2019; Lin et al., 2021a,b), we select five representative domains to evaluate our approach. The dataset statistics are summarized in Appendix A.

Evaluation Metrics We evaluate DST performance using Joint Goal Accuracy (JGA) and Average Goal Accuracy (AGA), consistent with previous works (Feng et al., 2023). JGA calculates the proportion of dialogue turns where the entire state is accurately predicted, while AGA measures the mean accuracy across active slots per turn. A slot is active if its value is mentioned in the current turn and is not inherited from previous turns.

Baselines We compare our approach against existing cross-domain zero-shot approaches. These methods include TRADE (Wu et al., 2019), TransferQA (Lin et al., 2021a), T5DST (Lin et al., 2021b), D3ST (Zhao et al., 2022), CAPID (Dong et al., 2024) and LDST (Feng et al., 2023). As for MultiWOZ 2.4, we select recent baseline models including IC-DST (Hu et al., 2022), ParsingDST (Wu et al., 2023), RefPyDST (King and Flanigan, 2023), D0T (Finch et al., 2024), LDST (Feng et al., 2023) and CAPID (Dong et al., 2024). Baseline details are shown in Appendix C.

4.2 Training Details

We conduct experiments in a cross-domain setting, where models are trained on four source domains and evaluated on an unseen target domain. To reflect practical deployment scenarios, we emphasize inference efficiency in our DST setup. **As DST typically serves as a sub-module in task-oriented dialogue systems and is often deployed in latency-sensitive environments, lightweight models are commonly preferred to balance accuracy and efficiency.** Accordingly, we consider two backbone models with different scales and architectures: T5-small (60M) (Raffel et al., 2020) and LLaMA2-Chat-7B (Touvron et al., 2023). To

Method	Backbone	Params	MultiWoz 2.4											
			Hotel		Restaurant		Taxi		Attraction		Train		Average	
			JGA	AGA	JGA	AGA	JGA	AGA	JGA	AGA	JGA	AGA	JGA	AGA
Small Models (≤ 1B params)														
CAPID	T5–small	60M	38.73	77.05	29.35	67.92	73.35	88.25	47.87	**74.32**	47.88	74.43	47.43	76.39
LongDST	T5–small	60M	**65.12**	**83.55**	**61.28**	**77.03**	**91.25**	**94.15**	**60.10**	73.65	**67.45**	**78.32**	**69.04**	**81.34**
Large Models (> 1B params)														
IC–DST	Codex	12B-185B	46.69	–	57.28	–	71.35	–	59.97	–	49.37	–	56.93	–
ParsingDST	Gpt–3.5–turbo	20B	46.76	–	67.67	–	80.58	–	65.63	–	62.59	–	64.65	–
RefPyDST	LLaMa2–13B	13B	51.20	–	65.60	–	67.10	–	70.90	–	69.20	–	64.70	–
D0T	T5–11B	11B	32.00	–	72.30	–	50.60	–	68.10	–	55.80	–	55.70	–
D0T	Llama2–13B	13B	56.40	–	**78.80**	–	54.70	–	**76.80**	–	**76.10**	–	68.60	–
LongDST	LLaMa3–8B	8B	**64.25**	**82.15**	74.58	**86.40**	**92.10**	**95.12**	76.45	**84.35**	75.90	**85.28**	**76.66**	**86.66**

Table 1: Overall performance on the MultiWOZ 2.4 dataset across various backbones, evaluated using JGA and AGA. Models are categorized by parameter size: Small (≤1B) and Large (> 1B).

Method	Backbone	Params	MultiWoz 2.1											
			Hotel		Restaurant		Taxi		Attraction		Train		Average	
			JGA	AGA	JGA	AGA	JGA	AGA	JGA	AGA	JGA	AGA	JGA	AGA
Small Models (≤1B params)														
T5DST	T5–small	60M	21.21	–	21.65	–	64.62	–	33.09	–	35.43	–	35.2	–
CAPID	T5–small	60M	31.10	72.56	31.64	69.06	65.41	83.75	40.88	68.99	34.26	65.93	40.66	72.06
Prompter	PPTOD–s	∼60M	19.20	–	26.00	–	66.30	–	35.80	–	39.00	–	37.26	–
TRADE	ELMo	94M	13.70	65.32	11.52	53.43	60.58	73.92	19.87	55.53	22.37	49.31	25.61	59.50
MA–DST	ELMo	94M	16.28	–	13.56	–	59.27	–	22.46	–	22.76	–	26.87	–
TransferQA	T5–large	770M	22.72	77.84	26.28	**81.73**	61.87	86.48	31.25	60.62	36.72	**87.21**	35.77	78.78
D3ST	T5–base	220M	21.80	–	38.20	–	78.40	–	56.40	–	37.70	–	46.5	–
LongDST	T5–small	60M	**60.12**	**83.40**	**61.15**	77.52	**90.45**	**93.82**	**59.10**	**73.05**	**67.58**	79.55	**67.68**	**81.47**
Large Models (>1B params)														
FNCTOD	LLaMa2–13B	13B	46.83	–	60.27	–	67.48	–	62.24	–	60.90	–	59.54	–
LDST	LLaMa2–7B	7B	63.32	–	**73.72**	–	**91.47**	–	**75.61**	–	**75.03**	–	**75.83**	–
NL-DST	GPT-3 Small	125M	**64.80**	**89.30**	69.50	**92.10**	75.20	**94.30**	59.10	**86.70**	55.40	**85.20**	64.80	**89.52**
LongDST	LLaMa2–7B	7B	63.55	75.87	59.04	76.62	89.14	92.39	59.57	75.23	68.08	78.93	67.87	79.81

Table 2: Overall performance on the MultiWOZ 2.1 dataset across various backbones, evaluated using JGA and AGA. Models are categorized by parameter size: Small (≤1B) and Large (>1B).

further simulate limited supervision, we randomly sample 5% of the training data for T5-small and 1% for LLaMA2, which reduces computational cost while encouraging cross-domain generalization. More details are provided in Appendix D.

4.3 Main Results

Our LongDSTMethod Demonstrates Consistent Superiority and Generalization Across Backbones. LongDST consistently outperforms strong baselines across two challenging datasets, MultiWOZ 2.1 and MultiWOZ 2.4, and a wide range of backbone models, demonstrating strong generalization across diverse settings (Tables 2 and 1). Results marked with "–" are not reported in the original papers, and all baseline numbers are taken from their respective publications. Notably, LongDST exhibits strong adaptability in low-resource settings. With the LLaMA2-7B backbone trained on only 1% of the data, it achieves a competitive Joint Goal Accuracy of 67.87. Moreover, LongDST also performs well on smaller architectures such as T5-small, yielding substantial improvements over existing methods across multiple domains. For instance, on MultiWOZ 2.1, LongDST with T5-small trained on only 10% of the data attains an average JGA of 67.65. Notably, our LLaMA model achieves comparable performance to state-of-the-art baselines (e.g. NL-DST) trained on the full dataset using only 1% of the training samples.

LongDST Effectively Addresses Both Long-Context Understanding Bottleneck and Error Propagation Challenges. Figure 6 shows that LongDST (green line) consistently outperforms the baseline (blue line) across most domains on MultiWOZ 2.1, with the performance gap becoming more pronounced in later dialogue turns. This trend indicates LongDST's stronger ability to model

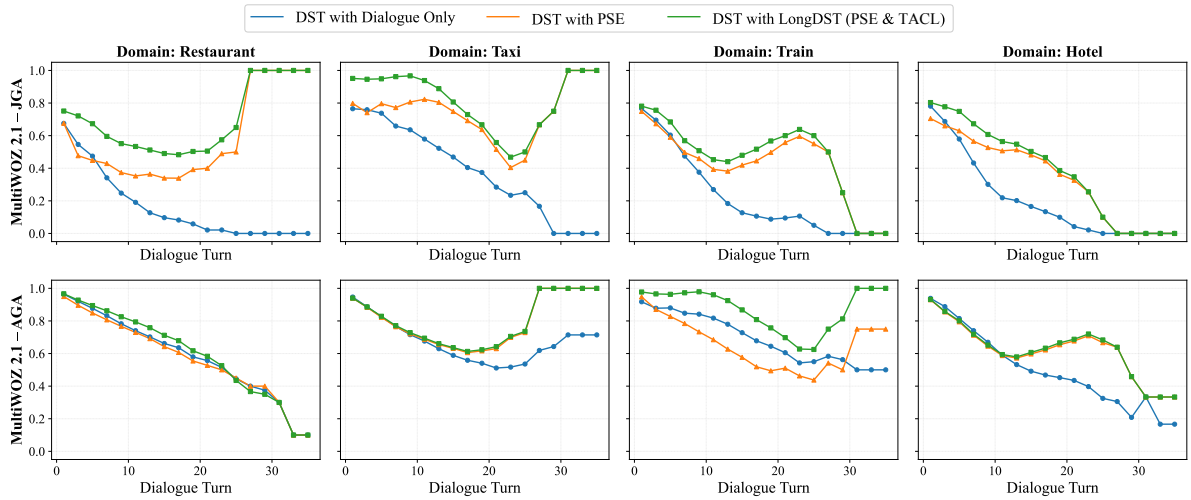

Figure 6: Turn-level comparison of JGA and AGA on MultiWOZ 2.1.

Dataset	Training Paradigm	hotel		restaurant		taxi		train		attraction		Average	
		JGA	AGA	JGA	AGA	JGA	AGA	JGA	AGA	JGA	AGA	JGA	AGA
MultiWoz 2.1	Dial. Only	42.85	80.86	34.54	75.66	63.54	83.00	44.64	71.25	53.46	72.71	47.81	76.70
	w. PSE	59.99	83.31	61.00	77.40	90.50	93.00	58.07	72.69	67.67	79.47	67.45	81.17
	w. PSE & TACL	**60.00**	**83.31**	**61.09**	**77.41**	**90.50**	**93.90**	**59.01**	**72.99**	**67.67**	**79.46**	**67.65**	**81.42**
MultiWoz 2.4	Dial. Only	36.52	78.57	38.63	71.23	57.03	80.02	59.19	73.44	63.39	77.59	50.95	76.17
	w. PSE	63.94	83.35	61.37	76.94	91.31	94.21	60.01	73.59	67.27	78.10	68.78	81.24
	w. PSE & TACL	**65.03**	**83.46**	**61.37**	**76.94**	**91.32**	**94.23**	**60.01**	**73.59**	**67.34**	**78.27**	**69.01**	**81.30**

Table 3: Ablation study showing effectiveness of Progressive State Exposure..

long-context dependencies and alleviate error accumulation over time. Similar observations on MultiWOZ 2.4 are provided in Figure 7 in the appendix.

4.4 Ablation Study

Effectiveness of Progressive State Exposure. We conduct an ablation study to validate the effectiveness of PSE. As shown in Table 3, integrating PSE into the training pipeline yields substantial improvements over the dialogue-only baseline on both MultiWOZ 2.1 and 2.4 datasets evaluated on T5 model. These consistent gains across all domains demonstrate PSE's strong ability to reduce error accumulation during long-turn dialogue modeling. In addition, as illustrated in Figure 6 and Figure 7, PSE significantly improves performance on domains with longer dialogues (e.g., taxi, train and restaurant). This suggests that PSE enhances the model's robustness in handling extended context by progressively introducing its own predictions during training, encouraging better state tracking under imperfect histories.

Effectiveness of Traceability-Aware Curriculum Learning. While PSE substantially improves long-context understanding and reduces error propagation, we observe that it introduces certain instability during early-stage training. This is primarily due to the warm-up phase governed by a hyperparameter α, which determines the number of initial steps where ground-truth states are used to help the model learn instruction-following behavior. If α is not set properly, the model may struggle to converge effectively.

To investigate this, we compare performance under two warm-up configurations: training for 3 epochs (90 warm-up steps) and 5 epochs (150 warm-up steps). As shown in Table 4, PSE without curriculum learning (TACL) shows inconsistent

Epoch = 3; Starting at Step 90					
	hotel	restaurant	taxi	train	attraction
Dial. Only	18.86	40.30	53.30	38.76	54.86
w. PSE	31.00	38.95	64.45	39.20	62.30
w. PSE & TACL	**64.32**	**61.08**	**91.10**	**59.98**	**67.05**
Epoch = 5; Starting at Step 150					
	hotel	restaurant	taxi	train	attraction
Dial. Only	42.85	34.54	63.54	44.64	53.46
w. PSE	59.99	61.00	90.50	58.07	67.67
w. PSE & TACL	**60.00**	**61.09**	**90.50**	**59.01**	**67.67**

Table 4: Ablation study demonstrating the effectiveness of Traceability-Aware Curriculum Learning.

results under the 3-epoch setting—sometimes even underperforming the dialogue-only baseline (e.g., on the restaurant domain). This suggests that with insufficient warm-up, the model lacks the initial stability needed to effectively learn from recursive supervision signals.

However, when TACL is added, the model achieves consistently strong performance across all domains, even under the shorter 3-epoch setting. This indicates that our proposed TACL not only accelerates convergence but also stabilizes training by organizing samples from easy to hard. Consequently, TACL reduces sensitivity to warm-up hyperparameters and helps mitigate the instability introduced by PSE, significantly lowering the cost of hyperparameter tuning.

5 Related Work

5.1 Dialogue State Tracking

DST is a fundamental component of TODS (Lee et al., 2019; Heck et al., 2020; Hosseini-Asl et al., 2020; Peng et al., 2020; Feng et al., 2024; Zhang et al., 2022, 2023a). It extracts and maintains a structured representation of user goals throughout a conversation (Lu et al., 2021), including constraints, requests, and preferences. Accurate DST supports downstream API calls and personalized recommendations (Liu et al., 2022, 2024a), making it critical for real-world dialogue systems.

Zero-shot Cross-Domain Dialogue State Tracking aims to generalize to unseen domains without labeled training data, a setting motivated by the high cost of dialogue annotation. The core challenge lies in handling novel slot types, values, and dialogue patterns. Early work explored transfer via auxiliary corpora (Su et al., 2022; Lin et al., 2021a) or reformulated DST as question answering (Lin et al., 2021b) or summarization (Shin et al., 2022) tasks. More recently, large language models (LLMs) have been applied to zero-shot DST due to their strong generalization ability. Prompt-based and instruction-tuned methods (Feng et al., 2023, 2024; Li et al., 2024; Heck et al., 2023; Zhao et al., 2024; Liu et al., 2024b) extract dialogue states directly from raw histories. However, empirical studies show that LLMs still struggle in cross-domain DST (Hu et al., 2022; Bang et al., 2023; Hudeček and Dušek, 2023; Zhang et al., 2023b; Chung et al., 2023), especially in complex dialogues, and their reliance on handcrafted prompts limits robustness and interpretability.

Long-Context Dialogue State Tracking. Real-world conversations often involve extensive contexts with topic shifts and revisions, exacerbating challenges like context noise and error propagation. However, prior research has largely overlooked these scenarios, particularly in zero-shot settings, often relying on performance-degrading context truncation. While recent works (Quan and Xiong, 2020; Hudeček and Dušek, 2023) have begun to explore this area, effective solutions remain limited. In this work, we explicitly target long-context DST to address key obstacles such as error propagation and insufficient long-context understanding.

5.2 Curriculum Learning

Curriculum Learning (CL) is a training strategy inspired by the human learning process, where models are trained on samples organized from easy to difficult (Bengio et al., 2009). The central idea is that presenting simpler examples early in training can help models converge faster and generalize better. CL has been explored in various NLP tasks such as machine translation (Zhang et al., 2018), text summarization (Xu et al., 2020), and question answering (Sachan and Xing, 2016). In the dialogue domain, curriculum-based methods have been adopted to improve dialogue generation (Zhao et al., 2020; Yang and et al., 2021) and dialogue state tracking (Feng et al., 2024), where sample difficulty can be measured by dialogue length, number of slots, or prediction confidence. Recent work has also proposed adaptive curricula that dynamically adjust the training distribution based on model performance (Jiang et al., 2015; Xu et al., 2020), or incorporate external signals such as reinforcement learning rewards (Matiisen et al., 2019).

6 Conclusion

In this work, we identify three key challenges that hinder dialogue state tracking in long and complex dialogues. To address these challenges, we propose LongDST, a novel framework that leverages Progressive State Exposure and traceability-aware curriculum learning. Together, these components enable more reliable state transition modeling and effective self-correction. Extensive experiments across multiple benchmarks and backbone architectures demonstrate that LongDST consistently improves DST performance, particularly in long-context and cross-domain settings.

Limitations

We discuss two limitations of our work that also suggest promising directions for future research.

First, the proposed framework introduces additional training complexity compared to standard DST pipelines, including curriculum scheduling and progressive state exposure, which increases the number of hyperparameters and requires careful tuning. Although we adopt fixed settings across datasets and backbones in our experiments, further simplification or automation of the scheduling strategy remains an important direction for future work. Second, LongDST relies on the quality of intermediate state predictions in recursive modeling. In extremely noisy dialogue settings, severely incorrect early predictions may still hinder effective recovery, despite the proposed self-correction mechanisms. Exploring uncertainty-aware modeling or confidence-based revision strategies could further enhance robustness in such cases.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 41–48.

Rafael Carranza and Mateo Alejandro Rojas. 2025. Interpretable and robust dialogue state tracking via natural language summarization with llms. *arXiv preprint arXiv:2503.08857*.

Willy Chung, Samuel Cahyawijaya, Bryan Wilie, Holy Lovenia, and Pascale Fung. 2023. Instructtods: Large language models for end-to-end task-oriented dialogue systems. *arXiv preprint arXiv:2310.08885*.

Xiaoyu Dong, Yujie Feng, Zexin Lu, Guangyuan Shi, and Xiao-Ming Wu. 2024. Zero-shot cross-domain dialogue state tracking via context-aware auto-prompting and instruction-following contrastive decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8527–8540.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

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)*, pages 1266–1279.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023. Towards llm-driven dialogue state tracking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 739–755.

James D Finch, Boxin Zhao, and Jinho D Choi. 2024. Leveraging diverse data generation for adaptable zero-shot dialogue state tracking. *arXiv preprint arXiv:2405.12468*.

Muhan Gao, TaiMing Lu, Kuai Yu, Adam Byerly, and Daniel Khashabi. 2024. Insights into llm long-context failures: when transformers know but don't tell. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7611–7625.

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*.

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*.

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.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627–2643.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Vojtěch Hudeček and Ondřej Dušek. 2023. Are llms all you need for task-oriented dialogue? *arXiv preprint arXiv:2304.06556*.

Lu Jiang, Qingming Meng, Zhengyuan Yu, and Zhenzhong Lan. 2015. Self-paced curriculum learning. In *AAAI*.

Brendan King and Jeffrey Flanigan. 2023. Diverse retrieval-augmented in-context learning for dialogue state tracking. *arXiv preprint arXiv:2307.01453*.

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

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2025. Long-context llms struggle with long in-context learning. *TMLR*.

Zekun Li, Zhiyu Zoey Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Xin Luna Dong, Adithya Sagar, Xifeng Yan, and Paul A Crook. 2024. Large language models as zero-shot dialogue state tracker through function calling. *arXiv preprint arXiv:2402.10466*.

Zhaojiang Lin, Bing Liu, Andrea Madotto, Seungwhan Moon, Zhenpeng Zhou, Paul A Crook, Zhiguang Wang, Zhou Yu, Eunjoon Cho, Rajen Subba, et al. 2021a. Zero-shot dialogue state tracking via cross-task transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7890–7900.

Zhaojiang Lin, Bing Liu, Seungwhan Moon, Paul A Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Andrea Madotto, Eunjoon Cho, and Rajen Subba. 2021b. Leveraging slot descriptions for zero-shot cross-domain dialogue statetracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5640–5648.

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*.

Qijiong Liu, Xiaoyu Dong, Jiaren Xiao, Nuo Chen, Hengchang Hu, Jieming Zhu, Chenxu Zhu, Tetsuya Sakai, and Xiao-Ming Wu. 2024a. Vector quantization for recommender systems: A review and outlook. *arXiv preprint arXiv:2405.03110*.

Qijiong Liu, Jieming Zhu, Quanyu Dai, and Xiao-Ming Wu. 2022. Boosting deep ctr prediction with a plug-and-play pre-trainer for news recommendation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2823–2833.

Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. 2024b. Multimodal pre-training, adaptation, and generation for recommendation: A survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6566–6576.

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

Tambet Matiisen, Avital Oliver, Jonathan Cohen, and John Schulman. 2019. Teacher-student curriculum learning. In *IEEE Transactions on Neural Networks and Learning Systems*.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pre-trained auto-regressive model. *arXiv preprint arXiv:2005.05298*, 3.

Jun Quan and Deyi Xiong. 2020. Modeling long context for task-oriented dialogue state generation. *arXiv preprint arXiv:2004.14080*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Mrinmaya Sachan and Eric Xing. 2016. Easy questions first? a case study on curriculum learning for question answering. In *ACL*, pages 453–463.

Jamin Shin, Hangyeol Yu, Hyeongdon Moon, Andrea Madotto, and Juneyoung Park. 2022. Dialogue summaries as dialogue states (ds2), template-guided summarization for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3824–3846.

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

Diogo Tavares, David Semedo, Alexander Rudnicky, and Joao Magalhaes. 2023. Learning to ask questions for zero-shot dialogue state tracking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2118–2122.

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*.

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*.

Yuxiang Wu, Guanting Dong, and Weiran Xu. 2023. Semantic parsing by large language models for intricate updating strategies of zero-shot dialogue state tracking. *arXiv preprint arXiv:2310.10520*.

Hu Xu, Bing Li, Yanshuai Lu, and et al. 2020. Curriculum learning for natural language understanding. In *ACL*.

Diyi Yang and et al. 2021. Learning to summarize dialogues with dialogue discourse structure via curriculum learning. In *ACL*.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *arXiv preprint arXiv:2104.00773*.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2018. An empirical study of curriculum learning for neural machine translation. In *ACL*, pages 609–617.

Haode Zhang, Haowen Liang, Liming Zhan, Xiao-Ming Wu, and Albert Lam. 2023a. Revisit few-shot intent classification with plms: Direct fine-tuning vs. continual pre-training. *arXiv preprint arXiv:2306.05278*.

Haode Zhang, Haowen Liang, Yuwei Zhang, Liming Zhan, Xiao-Ming Wu, Xiaolei Lu, and Albert Lam. 2022. Fine-tuning pre-trained language models for few-shot intent detection: Supervised pre-training and isotropization. *arXiv preprint arXiv:2205.07208*.

Shuyu Zhang, Yifan Wei, Xinru Wang, Yanmin Zhu, Yangfan He, Yixuan Weng, and Bin Li. 2025. Hicolora: Addressing context-prompt misalignment via hierarchical collaborative lora for zero-shot dst. *arXiv preprint arXiv:2509.19742*.

Xiaoying Zhang, Baolin Peng, Kun Li, Jingyan Zhou, and Helen Meng. 2023b. Sgp-tod: Building task bots effortlessly via schema-guided llm prompting. *arXiv preprint arXiv:2305.09067*.

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*.

Tiancheng Zhao, Yizhe Xie, Y-Lan Li, and Zhouhan Li. 2020. Knowledge-grounded dialogue generation with curriculum learning. In *AAAI*.

Xiangyu Zhao, Bo Liu, Qijiong Liu, Guangyuan Shi, and Xiao-Ming Wu. 2024. Easygen: Easing multimodal generation with bidiffuser and llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1351–1370.

A Dataset Statistic

We present the key statistics of the MultiWOZ 2.1 and 2.4 datasets in Table 5. Both datasets span 8 domains and contain over 10K dialogues with similar distributions in average dialogue length and token count per turn.

B Description of Prompt Templates

We provide two prompts in our work: the first prompt is used for LongDST. It guides the model to track and update the dialogue state for the current turn based on the full dialogue history and the previous slot state, as is shown in Figure 8. The second prompt (shown in Figure 9) is designed for the *Progressive State Replacement* training process, where the model is asked to generate the predicted dialogue state for the $(i-1)^{\text{th}}$ turn, enabling it to learn self-correction through its own previous predictions.

C Details of Various Baselines

To demonstrate our method's superiority in DST, we compare it with various baselines under the zero-shot cross-domain setting.

- TRADE (Wu et al., 2019). It directly generates slot values using a shared-parameter architecture across domains, and later incorporates continual learning techniques (EWC/GEM) for domain adaptation.
- TransferQA (Lin et al., 2021a). It adopts a T5-based encoder-decoder Transformer with masked language modeling pre-training to solve DST task.
- T5DST (Lin et al., 2021b). It uses a generative question-answering model that incorporates slot-type-informed descriptions for zero-shot cross-domain dialogue state tracking.
- D3ST (Zhao et al., 2022). It leverages schema descriptions as prefix prompts in a generative T5 framework, enabling zero-shot cross-domain dialogue state tracking through unified sequence generation of slots.
- ParsingDST (Wu et al., 2023). It employs a text-to-JSON semantic parsing framework, which converts dialogue context and system utterances into structured JSON representations through modular updating strategies, enabling controllable and interpretable zero-shot DST.
- RefPyDST (King and Flanigan, 2023). The baseline introduces PMI scoring within a retrieval-augmented in-context learning framework, which re-weights language model completions based on their a priori likelihood conditioned on task-specific examples.

	MultiWOZ 2.4			MultiWOZ 2.1		
	TRAIN	VAL	TEST	TRAIN	VAL	TEST
No. of domains	8	8	8	8	8	8
No. of dialogues	8 438	1 000	1 000	8 438	1 000	1 000
Total no. of turns	113 556	14 748	14 744	113 556	14 748	14 744
Avg. turns / dialogue	13.46	14.75	14.74	13.46	14.75	14.74
Avg. tokens / turn	15.34	15.69	15.69	15.34	15.69	15.69

Table 5: Statistics of the datasets used in our experiments.

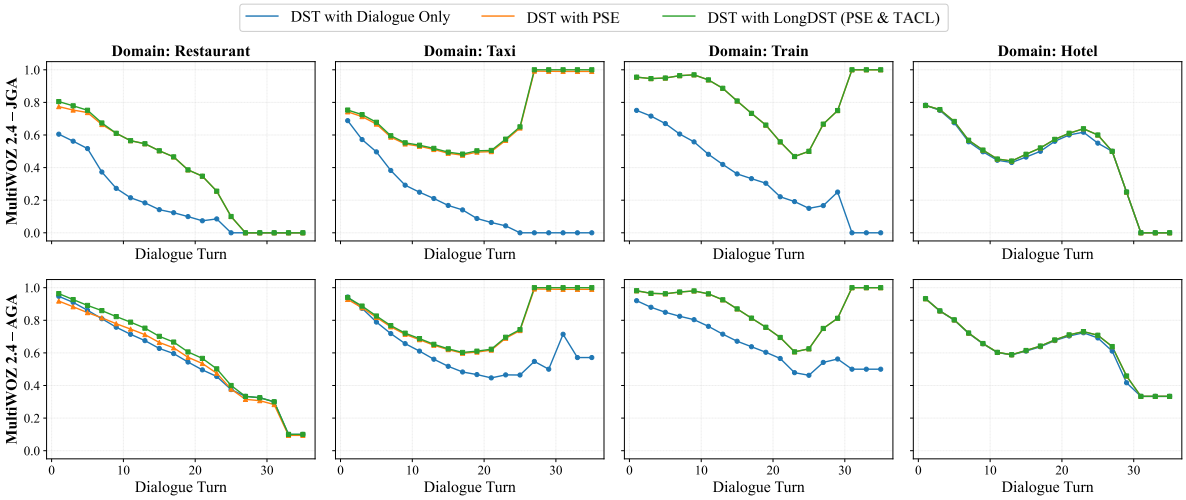

Figure 7: Turn-level comparison of JGA and AGA on MultiWOZ 2.4.

- D0T(Finch et al., 2024). This baseline method employs a QA-to-slot translation framework where GPT-generated question-answer pairs are systematically converted into structured dialogue state slots through value-type-aware prompts.

- LDST (Feng et al., 2023). This baseline employs LLM-driven dialogue state tracking (LDST) framework that utilizes fine-tuned LLaMA models with domain-slot descriptions and value lists in fixed prompt templates.

- NL-DST (Carranza and Rojas, 2025). It generates natural language state descriptions instead of structured slot-value pairs, improving robustness and interpretability.

- CAPID (Dong et al., 2024). This baseline introduces a zero-shot DST framework combining context-aware auto-prompting and instruction-following contrastive decoding.

D Implementation Details

For the T5-Small model, we train for 5 epochs with a learning rate of 3×10^{-4}, using a batch size of 16, weight decay of 0.01, and warmup steps of 50. Evaluation and checkpointing are done every 500 steps. For the LLaMA-2-7B model, we apply LoRA-based fine-tuning with parameters ($r = 8, \alpha = 16$, dropout $= 0.05$) targeting attention projections (`q_proj`, `v_proj`). We use a learning rate of 1×10^{-3}, batch size of 16 (with micro-batch size 4), and train for 3 epochs. The sequence cutoff length is set to 1536 tokens. For PSE, we set $\alpha = 0.25$ and use a scheduled teacher forcing ratio starting from 1.0 and linearly decaying to 0.5 over training steps. All experiments are performed on a single NVIDIA A100 GPU.

Track the state of the slot <Restaurant_Name> in the dialogue and Return the final slot state after this turn.

Input: You will receive the full dialogue up to the current turn and the previous slot state. Task: Update the slot state based on the current turn only. If the current turn changes a slot value, update it. If there is no relevant change, keep the previous state as is.

DIALOGUE: SYSTEM:Good morning, what can I help you? USER: Can you help me find a particular restaurant that I'm looking for? The restaurant is called eraina. SYSTEM:Ah yes, the Eraina. It's an European restaurant in the city centre. Would you like more info? USER:I also need a taxi.

Previous State of <Restaurant_Name> is: NONE

So the value of the slot <Restaurant_Name> is:

Figure 8: Prompt Template for LongDST.

Track the state of the slot <Restaurant_Name> in the dialogue and Return the final slot state after this turn.

DIALOGUE: SYSTEM:Good morning, what can I help you? USER: Can you help me find a particular restaurant that I'm looking for? The restaurant is called eraina.

So the value of the slot <Restaurant_Name> is:

Figure 9: Prompt Template for Progressive State Replacement. The Dialogue actually contains turn 1 to turn $(i-1)$.