Learning What Matters: Dynamic Experience Prioritization for Task-Oriented Dialogue Policy via Stage-aware Experience Management

Anonymous ACL submission

Abstract

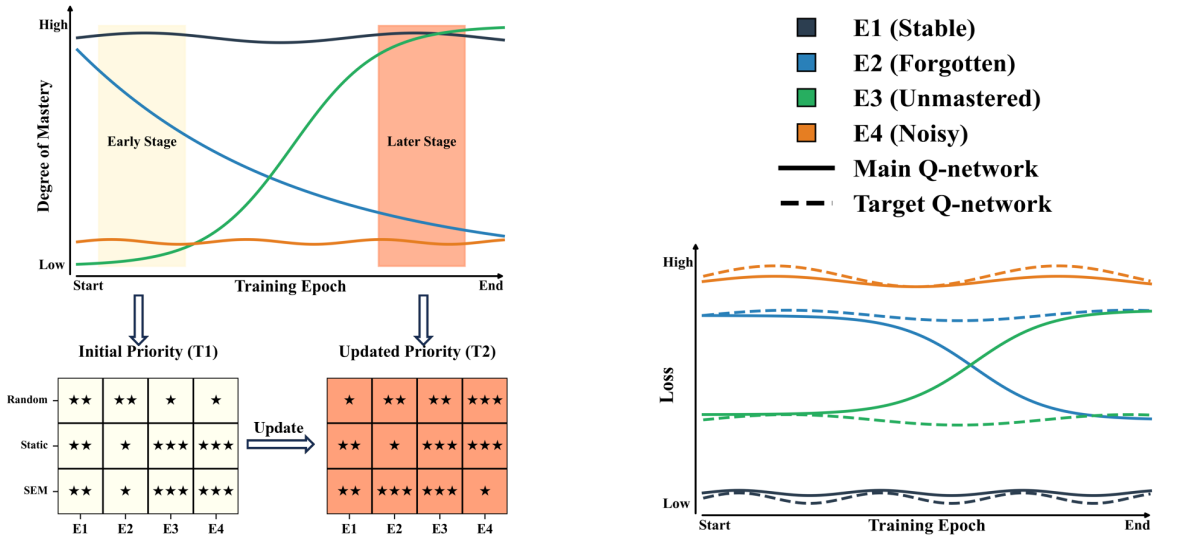
Experience replay plays a pivotal role in enhancing sample efficiency for reinforcement learning-based dialogue policy optimization. However, traditional random sampling or static heuristic strategies fail to dynamically exploit critical experiences following policy learning stages, resulting in inefficient sampling and noise propagation. To address this issue, this paper presents a dynamic Stage-aware Experience Management (SEM) framework that establishes quantitative mapping between policy learning stages and experience states to adjust replay priorities adaptively. This framework adopts a quadripartite experience state paradigm to characterize the stages of policy learning and provide a quantitative basis for experience management decisions. Moreover, a dual Q-network structure is employed to monitor loss discrepancies and trends in real-time, discriminating each experience as stable, forgotten, unmastered, or noisy. Benefiting from this dynamic stage-aware mechanism, the SEM prioritizes replaying critical experiences in forgotten and unmastered experiences to strengthen weak links while suppressing noisy samples to reduce interference. Experiments on four public dialogue datasets verify the effectiveness and generalizability of the SEM in dynamic priority management.

1 Introduction

As the decision core of task-oriented dialogue (TOD) systems, dialogue policies (DPs) aim to accurately infer user intents and efficiently accomplish domain-specific goals through multi-turn interactions (Algherairy and Ahmed, 2025). Though large language models (LLMs) have demonstrated strong power in linguistic tasks, their lack of explicit long-term value estimation compromises DP convergence stability in multi-step decision optimization (Yi et al., 2024). Even with reasoning-enhanced techniques like chain-of-thought prompting, LLMs still struggle to explore ambiguous solution spaces in complex dialogues efficiently (Yi et al., 2024), while the scarcity of domain-specific TOD datasets further exacerbates their adaptability challenges (Kamuni et al., 2024). Consequently, off-policy reinforcement learning (RL) which leverages experience replay to reuse historical interaction trajectories for policy learning has emerged as the mainstream technique for DP optimization (Algherairy and Ahmed, 2024).

The management efficiency of experience replay buffers directly impacts the training outcomes, where the key challenge resides in accurately quantifying the experience sampling priorities at every training stage. Existing methods rely on random sampling or static heuristic strategies. However, both fail to dynamically adjust the experience sampling priorities as policies evolve (Zhao et al., 2024). As shown in Fig.1(a), the random sampling method assigns experiences indiscriminately at stage T1, ignoring the empirical state discrimination. The static heuristic assigns the maximum priority to sample E4 (Noisy Experience) at stage T1 and maintains this assignment continuously at stage T2. Nevertheless, this assignment is likely no longer optimal at this stage since sample E4 is a noisy sample. That is, traditional strategies struggle to exploit critical experiences required at different training stages. This results in two critical deficiencies: (i) high-value experiences are prematurely diluted due to fixed priorities, and (ii) noisy experiences continuously interfere with the training process. The underlying cause lies in the absence of dynamic quantification linking policy learning phases to the actual contribution of experiences.

To address these issues, this paper proposes a dynamic experience prioritization framework with a Stage-aware Experience Management (SEM) mechanism to learn the latent rhythms between experience significance and policy evolution for dialogue policy optimization. This mechanism quan-



(a) The priority-mastery correlation in the SEM framework

(b) Four States differentiation under dual-network loss

Figure 1: (a) This diagram illustrates the correlation between priority and mastery within the SEM framework. It shows the transition from initial priorities (T1) to updated priorities (T2), highlighting SEM's adaptive mechanism for reassessing and adjusting priorities based on mastery levels. (b) This graph depicts the representational differentiation of four experience states under dual-network loss. It tracks the evolution of mastery (solid lines) and priority (dashed lines) over training epochs for each experience state (E1–E4), showcasing SEM's dynamic response and adjustment based on the quality of experiences.

titatively maps the evolving stages of policy learning to distinct experience states. Specifically, four quantifiable states are first defined: *Stable* (consistently mastered), *Forgotten* (previously mastered but recently degraded), *Unmastered* (not effectively learned), and *Noisy* (containing unreliable or misleading information). Then, a dual Q-network structure consisting of a main network and a target network is maintained to identify these states. The main network loss reflects the current instantaneous mastery of policy, while the target network loss represents historical learning outcomes. The dynamic trends of their discrepancies reveal the underlying state of each experience and guide the adjustment of priority at different training stages: increasing priorities for *Forgotten* and *Unmastered* states to reinforce critical knowledge gaps, while reducing priorities for *Noisy* states to suppress interference. As shown in Fig.1(a), the SEM assigns the lowest priority to *Forgotten* sample E2 and the highest to *Noisy* sample E4 at stage T1. Subsequently, it evaluates experience states continuously at each training stage and updates the priorities of E2 (*Forgotten*) and E4 (*Noisy*) to the highest and lowest at stage T2, which are optimal assignments for this stage.

Moreover, a hierarchical sum tree is employed for experience storage, enabling local updates, fast localization of high-frequency sampling regions, and low-complexity real-time priority adjustment.

- A quadripartite experience state quantification paradigm is proposed, which establishes the first dynamic mapping between policy learning stages and experience values to provide a quantitative decision-making basis for experience management.
- A stage-aware experience management framework is presented, enabling real-time experience state classification and dynamic priority adjustment with plug-and-play lightweight adaptability to off-policy RL algorithms.
- The effectiveness and generalizability of the proposed SEM are validated across four public dialogue datasets, demonstrating its capabilities in experience state identification and dynamic sampling priority management.

2 Related Work

Our work focuses on improving experience management in off-policy RL-based DP optimization. Existing relevant studies predominantly rely on two paradigms: uniform random sampling and static heuristic strategies (Zhang et al., 2024).

Uniform random sampling: The foundational Deep Q-Networks (DQN) (Mnih et al., 2015) employed experience replay with uniform sampling.

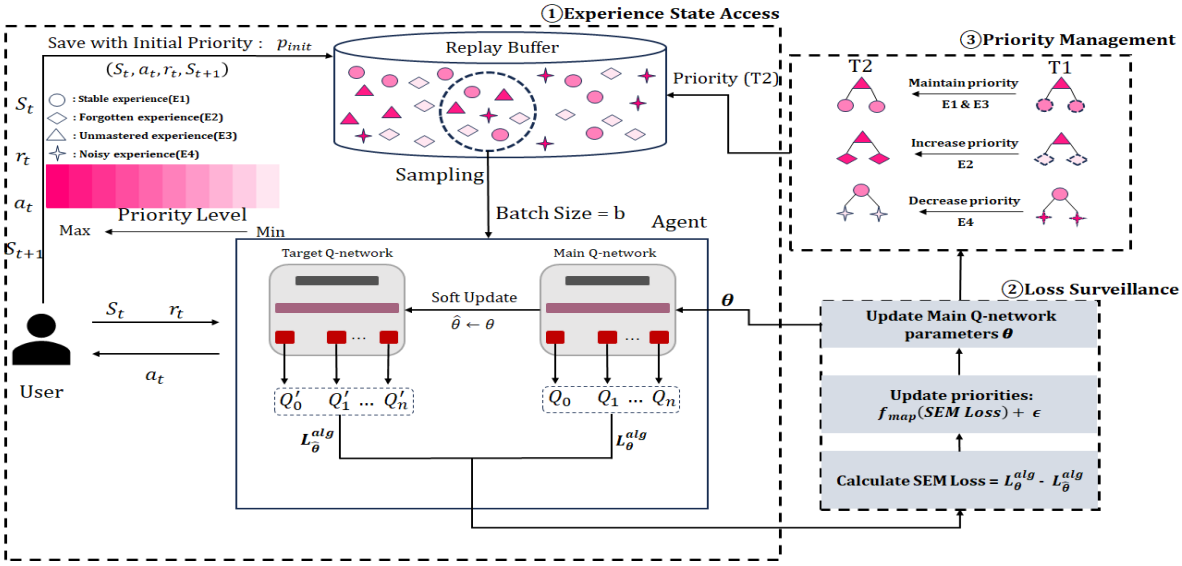


Figure 2: Dialogue policy optimization experience management process under the SEM framework.

It treats all transitions equally regardless of their intrinsic value. Though unbiased, this method cannot distinguish valuable experiences from noisy ones, resulting in inefficient learning (Liu et al., 2024). Subsequent studies confirm its insufficiency for DP optimization where strategic experience selection proves crucial (Yang et al., 2024).

Static heuristic strategies: Efforts to overcome random sampling limitations have converged on three static heuristic categories: (i) Priority Weighting: Prioritized Experience Replay (PER) (Schaul et al., 2016) utilized temporal difference (TD) errors as fixed importance metrics, while variants (Mei et al., 2023; Oh et al., 2022) incorporated auxiliary reward signals. Though improving initial sampling efficiency through priority-based selection, it fundamentally operates within a binary state paradigm – identifying only consistently mastered and unmastered experiences. (ii) Noise Filtering: Approaches such as conversational dead-end detection (Zhao et al., 2024) and adversarial filtering (Yu et al., 2024) aimed to suppress low-quality samples. However, their binary keep/discard decisions often exclude borderline experiences with partial utility, exacerbating forgetting (Gu et al., 2017); (iii) Memory Augmentation: Topological Experience Replay (TER) (Hong et al., 2022) organized experiences via graph structures, while multi-buffer strategies (Lu et al., 2023) preserved historical samples. Despite organizational benefits, these methods incurred prohibitive computational overhead and struggled with dynamic policy adaptation (Yang et al., 2022).

These methods suffered from dual priority distortion: premature high-value discard from static prioritization metrics (e.g., PER's outdated TD errors estimates (Horgan et al., 2018)) and noise amplification from rigid filtering thresholds that propagate residual artifacts through training iterations (Vezhnevets et al., 2017). Hybrid methods combining priority weighting and filtering (Buzzega et al., 2020) partially mitigated but retained heuristic-based limitations. In contrast, our SEM breaks this paradigm through dynamic priority calibration via dual Q-network discrepancy analysis.

Unlike existing memory-augmented approaches requiring complex architectures, the proposed SEM achieves: (i) **Adaptive Reassessment**: Continuous priority updates aligned with policy evolution; (ii) **Noise-Resilient Selection**: Probabilistic suppression rather than binary filtering; (iii) **Computational Efficiency**: Linear-time complexity versus TER's quadratic overhead. This systematic approach overcomes the experience prioritization distortion inherent in existing methods while maintaining plug-and-play compatibility with standard off-policy RL pipelines.

3 Methodology

As shown in Fig. 2, the SEM method consists of three steps: 1) *Experience States Access*, which captures agent-environment interactions, systematically storing transition tuples in replay buffers partitioned into four distinct state categories; 2) *Loss Surveillance*, which identifies four experience states by calculating loss discrepancies between the main and the target Q-networks outputs; 3) *Priority Management*, which employs dynamic ex-

perience priority adjustment via a sum-tree architecture, enabling hierarchical storage and weighted sampling based on categorized experience states.

3.1 Experience State Access

The SEM method establishes a quadripartite experience state quantification paradigm to dynamically align sampling priorities with policy learning demands. In this method, a dual Q-network architecture is employed, where the main network loss indicates current mastery, and the target network loss reflects historical learning. Experiences exhibit distinct temporal loss patterns across two networks, reflecting varying levels of policy mastery.

- **Stable Experiences (E1 in Fig.1(b)** exhibit *consistently low losses* in both main and target Q-networks, indicating mastered knowledge. These experiences should maintain baseline sampling priority to preserve policy stability.
- **Forgotten Experiences (E2 in Fig.1(b)** show *high main network loss* paired with *low target network loss*, revealing knowledge degradation. Their rising main loss triggers priority elevation to reinforce fading skills.
- **Unmastered Experiences (E3 in Fig.1(b)** display *synchronized high losses* in both networks, signaling unlearned patterns. These receive progressive priority boosts to accelerate initial acquisition.
- **Noisy Experiences (E4 in Fig.1(b)** manifest *persistent high losses* regardless of training progress. Their priority undergoes exponential decay to mitigate interference.

This taxonomy enables the SEM adaptive experience management. During agent-environment interaction, the DQN agent[1] generates experience tuples $\langle \boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{r}_t, \boldsymbol{s}_{t+1} \rangle$ through ϵ-greedy exploration. The dual Q-network architecture, comprising a rapidly updated main network $\boldsymbol{\theta}$ and a slowly evolving target network $\hat{\boldsymbol{\theta}}$, provides temporal loss signals for state classification:

$$\hat{\theta} \leftarrow \tau\theta + (1-\tau)\hat{\theta} \tag{1}$$

$$L_{\theta}^{alg}, L_{\hat{\theta}}^{alg} = \text{TD-loss}(Q_{\theta}, Q_{\hat{\theta}}) \tag{2}$$

where $\boldsymbol{\tau} \in (\mathbf{0}, \mathbf{1}]$ is the soft update coefficient. The losses $\boldsymbol{L}_{\boldsymbol{\theta}}^{\boldsymbol{alg}}$ and $\boldsymbol{L}_{\hat{\boldsymbol{\theta}}}^{\boldsymbol{alg}}$ are computed for the SEM loss and priority update (Eq. 3).

3.2 Loss Surveillance

The loss surveillance phase monitors the SEM loss of each sampled experience. It classifies the experience and updates the priority accordingly. This process consists of two steps:

1) SEM Loss: The SEM loss is the metric to measure the priority of experience. It is defined by the loss discrepancy between the main and the target Q-networks, formulated as:

$$\boldsymbol{SEMLoss}_i = \boldsymbol{L}_{\boldsymbol{\theta}}^{\boldsymbol{alg}}(i) - \boldsymbol{L}_{\hat{\boldsymbol{\theta}}}^{\boldsymbol{alg}}(i) \tag{3}$$

where $\boldsymbol{L}_{\boldsymbol{\theta}}^{\boldsymbol{alg}}(i)$ and $\boldsymbol{L}_{\hat{\boldsymbol{\theta}}}^{\boldsymbol{alg}}(i)$ are the network losses for the experience $\boldsymbol{i}$ in the main and the target Q-networks, respectively. The SEM loss provides an objective metric for prioritizing experiences by dynamically assessing agent mastery via loss discrepancies between two Q-network outputs[2]. This design naturally supports stage-aware prioritization during policy learning.

2) Priority Update: Based on the SEM loss, we update the experience priority employing a mapping function $\boldsymbol{f_{map}}$. The form is as follows:

$$\boldsymbol{p}_i = \boldsymbol{f_{map}}(\boldsymbol{SEMLoss}_i) + \boldsymbol{\epsilon} \tag{4}$$

where $\boldsymbol{f_{map}}$ maps the SEM loss to non-negative values and adds a small positive value ϵ to ensure all experiences have positive sampling probabilities, preventing sampling dead zones (Lee et al., 2019).

After the priority is updated, the main Q-network parameters $\boldsymbol{\theta}$ will be updated.

3.3 Priority Management

The priority management phase manages the experience priorities through a hierarchical sum tree storage structure.

The sum tree is a binary tree data structure designed to manage experience priorities. Its characteristic is that each node stores the sum of its children's priorities, which allows weighted sampling and priority updates in $\mathcal{O}(\log n)$ time. Moreover, when an experience priority changes, only the path from the corresponding leaf to the root needs to be updated, avoiding full array recalculations and improving update efficiency: **i) Leaf nodes:** store

[1]We adopt DQN as our illustrative baseline, given its seminal role in deep RL and its widespread use in task-oriented dialogue policy research. This choice ensures clear exposition and empirical validation against a well-understood off-policy RL framework. Importantly, our approach is algorithm-agnostic and can be seamlessly extended to other off-policy RL–based methods, as detailed in App. D.

[2]It is worth noting that the priority calculation in PER differs from our SEM loss. PER calculates priority based on the TD errors, defined as $p_i = |\delta_i| + \epsilon$, where δ_i is the TD errors and ϵ is a small constant added to prevent zero priority. In contrast, our method adopts a different approach.

Algorithm 1: SEM Implementation

Require:

- Off-policy RL algorithm $\mathcal{A}$ with loss function L^{alg}
- Main Q-network parameters $\boldsymbol{\theta}$
- Target Q-network parameters $\hat{\boldsymbol{\theta}}$
- Experience replay pool B
- Initial priority P_{init}
- SEM Loss Normalization Function f_{map}

1: Initialize Experience replay pool B as empty;

2: Initialize Target Q-network $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$;

3: for $t = 1$ ***to*** T **do**

a. Interact with environment:;

- Observe state $\boldsymbol{s}_t$ from the environment;
- Compute action $\boldsymbol{a}_t$ from the agent;
- Execute action $\boldsymbol{a}_t$, observe reward $\boldsymbol{r}_t$, and next state $\boldsymbol{s}_{t+1}$;
- Store transition $\langle \boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{r}_t, \boldsymbol{s}_{t+1} \rangle$ in B with P_{init};

b. for *each Iteration step from 1 to* T_{iter} **do**

- Sample minibatch of size b from B;
- Compute $L^{\text{alg}}_{\boldsymbol{\theta}}$ and update $\boldsymbol{\theta}$;
- Compute $L^{\text{alg}}_{\hat{\boldsymbol{\theta}}}$ and calculate the SEM Loss;
- Update minibatch priorities: $f_{\text{map}}(\text{SEM Loss}) + \boldsymbol{\epsilon}$;

4: Update Target Q-network following $\mathcal{A}$;

the priority $\boldsymbol{p}_i$ of each experience. **ii) Intermediate nodes:** store the sum of the priorities of all their child nodes. **iii) Root node:** stores the sum of leaf node priorities, representing the total priority of the experience pool $\sum_{i=1}^{n} \boldsymbol{p}_i$. With this structure, a weighted sampling mechanism can efficiently sample experiences from the replay buffer, balancing the latent rhythms between experience importance and policy evolution. The operation is as follows: The operation is as follows: A random value $\boldsymbol{u}$ is generated within the range $[0, \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{p}_i]$. Starting from the root node, we recursively compare $\boldsymbol{u}$ with the priorities of the left and right child nodes and eventually locate the matching leaf node (i.e., the corresponding experience tuple). It ensures that each experience is sampled with a probability $\boldsymbol{P}(\boldsymbol{i})$ proportional to its priority $\boldsymbol{p}_i$, i.e.,

$$\boldsymbol{P}(\boldsymbol{i}) = \frac{\boldsymbol{p}_i}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \boldsymbol{p}_j} \tag{5}$$

The updated priorities are stored in the leaf nodes of the Sum Tree, and the priority sums of the intermediate and root nodes are recursively updated to maintain the consistency of the entire structure.

The procedure of the SEM is described in Alg. 1.

4 Experiments

The objectives of this experiment are [3]: i) Assess the effectiveness of the SEM over baseline methods in simulated (4.3) and human evaluations (4.6); ii) Examine the performance of the SEM in addressing four experience states by the visualization of experience distributions (4.4.1) and priority trends (4.4.2); iii) Validating the generality of the SEM in off-policy RL-based DP algorithms (4.5).

4.1 Baselines

We comprehensively evaluate the effectiveness of the SEM framework against six state-of-the-art experience-management baselines: DQN with Random Experience Replay (RER) (Mnih et al., 2015), which uniformly samples experiences from the replay buffer without any prioritization; PER (Schaul et al., 2016), which prioritizes experiences based on high TD errors to enhance learning efficiency; TER (Hong et al., 2022), which organizes experiences into a state-dependency graph and performs value backups via breadth-first search from terminal nodes; DDR (Zhao et al., 2024), which identifies dialogue dead-ends, provides corrective rescue actions, and augments the buffer with penalty experiences to steer exploration; LLM_DA (Yi et al., 2024), which employs a LLM[4] to replace the TOD system's DP module, which replaces the DP module with a LLM to generate dialogue actions that are then responded by an NLG component; LLM_Word (Yi et al., 2024), which extends this by using an LLM to directly select words and produce end-to-end responses.

4.2 Experimental Settings

4.2.1 Datasets

Four public datasets widely used in TODs research, including both single-domain (Li et al., 2018) and multi-domain datasets (Budzianowski et al., 2018) are employed for evaluation. The domain and feature information of datasets are shown in App. A.

4.2.2 Implementation Details

For experimental fairness, all dialogue agents employ identical DQN with synchronized parameter initialization. Each agent undergoes from-scratch

[3] We will release the code on GitHub after the anonymity period.

[4] We adopt GPT-4.0 as LLM-based agents for its superior generative performance on dialogue tasks. Since ChatGPT-4.0 is closed-source and cannot be fine-tuned, we instead drive it via a carefully engineered prompt, detailed in App. C.

Table 1: Performance comparison of agents on four datasets with 10% noise[5], with top performers in each column bolded. All results are statistically significant (t-test, $p < 0.05$). Epochs denote early(50), mid(250), and post-convergence(500) stages. ChatGPT-4.0's numbers reflect its post-convergence evaluation (fine-tuning is not possible), while all other agents were trained from scratch. Therefore, comparisons with LLM_DA&LLM_Word focus on the converged results on 500 epoch.

Domain	Agent	Venue	Epoch=50			Epoch=250			Epoch=500		
			Success↑	**Reward↑**	**Turn↓**	**Success↑**	**Reward↑**	**Turn↓**	**Success↑**	**Reward↑**	**Turn↓**
Movie	RER	Nature 2015	0.0193	-52.49	31.62	0.4058	-4.29	27.98	0.6001	21.73	22.56
	PER	ICLR 2016	0.0163	-52.48	30.88	0.4422	0.19	27.75	0.6382	26.67	21.87
	TER	ICLR 2022	0.0152	-54.79	35.23	0.4228	-2.12	27.71	0.6698	30.90	20.96
	DDR	TACL 2024	0.0526	-49.77	32.67	0.4788	5.53	26.89	0.7173	37.63	20.11
	LLM_DA	CoRR 2024	**0.4140**	**-4.03**	**24.88**	0.4140	-4.03	24.88	0.4140	-4.03	24.88
	LLM_Word	CoRR 2024	0.2720	-28.58	31.35	0.2720	-28.58	31.35	0.2720	-28.58	31.35
	SEM	-----	0.0847	-43.03	28.39	**0.5806**	**18.57**	**24.21**	**0.7708**	**44.21**	**18.57**
Rest.	RER	Nature 2015	0.0004	-43.81	29.67	0.0549	-37.85	27.58	0.2566	-18.60	25.38
	PER	ICLR 2016	0.0002	-42.10	26.23	0.1097	-32.48	26.71	0.3784	-6.39	22.90
	TER	ICLR 2022	0.0003	-41.14	24.32	0.1784	-25.86	25.83	0.3950	-4.85	22.81
	DDR	TACL 2024	0.0020	-42.69	26.26	0.1967	-24.58	26.33	0.4601	1.80	22.26
	LLM_DA	CoRR 2024	**0.3080**	**-6.41**	**21.33**	**0.3080**	**-6.41**	**21.33**	0.3080	-6.41	21.33
	LLM_Word	CoRR 2024	0.2530	-20.31	31.63	0.2530	-20.31	31.63	0.2530	-20.31	31.63
	SEM	-----	0.0021	-42.63	27.65	0.1826	-25.60	26.07	**0.5040**	**5.78**	**21.15**
Taxi	RER	Nature 2015	0.0446	-39.61	29.25	0.1569	-29.52	29.29	0.2411	-21.71	28.82
	PER	ICLR 2016	0.0226	-41.51	29.08	0.1356	-31.18	28.77	0.2636	-19.21	27.87
	TER	ICLR 2022	0.0421	-39.68	28.93	0.2408	-21.41	28.17	0.3304	-13.25	27.98
	DDR	TACL 2024	0.0265	-42.11	29.65	0.1919	-26.30	28.85	0.3550	-10.50	27.51
	LLM_DA	CoRR 2024	**0.2940**	**-12.35**	**26.75**	0.2940	**-12.35**	**26.75**	0.2940	-12.35	26.75
	LLM_Word	CoRR 2024	0.2130	-19.68	29.31	0.2130	-19.68	29.31	0.2130	-19.68	29.31
	SEM	-----	0.0434	-39.63	29.08	**0.2949**	-16.22	27.53	**0.4551**	**-1.29**	**26.50**
MultiWOZ	RER	Nature 2015	0.0144	-45.73	31.06	0.0287	-36.76	30.47	0.0366	-30.19	30.11
	PER	ICLR 2016	0.0156	-43.69	32.14	0.0624	-35.92	31.59	0.0842	-30.13	31.05
	TER	ICLR 2022	0.0048	-46.18	33.97	0.0533	-36.71	32.18	0.0763	-29.86	31.74
	DDR	TACL 2024	0.0212	-41.07	30.93	0.0849	-33.42	29.17	0.1171	-25.64	29.73
	LLM_DA	CoRR 2024	**0.1220**	**-20.47**	**30.80**	**0.1220**	**-20.47**	30.80	0.1220	-20.47	30.80
	LLM_Word	CoRR 2024	0.1040	-22.57	32.78	0.1040	-22.57	32.78	0.1040	-22.57	32.78
	SEM	-----	0.0192	-42.38	31.07	0.0995	-30.43	**28.94**	**0.1584**	**-17.91**	**27.48**

training with uniform experience budgets (100 dialogues for warm-start initialization, 1 dialogue per training epoch) [6]. A multi-layer perceptron with two hidden layers of 80 neurons each is employed across all DQN-based algorithms. For details on the experimental parameters and the implementation of the f_{map} function, please refer to the App. B.

4.3 Main Results

Table 1 benchmarks baseline performance across four datasets. RER performs worst because its uniform sampling fails to prioritize high-value experiences, hurting DP learning efficiency. PER's static reliance on TD errors induces gradient bias and delayed priority updates, degrading long-term performance. TER is unstable in practice due to its sensitivity to shifting policies. Although achieving second-best performance, DDR remains constrained by its rule-based dead-end detection mechanism that lacks online adaptability. LM_DA achieves moderate initial success thanks to its restricted action space but cannot improve further due to hallucination issues and its closed-source nature. LLM_WORD yields the poorest LLM-based results due to combinatorial action space explosion amplified by unconstrained generative hallucination tendencies. In contrast, the SEM discriminates between different experience states and adaptively adjusts their priorities, achieving best performance.

4.4 Visualization Analysis

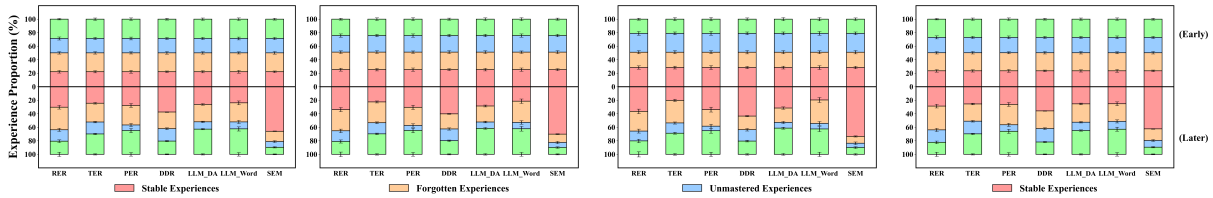
Due to space constraints, we present the experience priority trend results on the most challenging dataset MultiWOZ 2.1 in the main text. Results on other datasets are provided in the app.G.

4.4.1 Evolution of Experience State Distributions

We analyze the evolution of four experience states distribution across different baselines during early and late training phases, as illustrated in Fig-

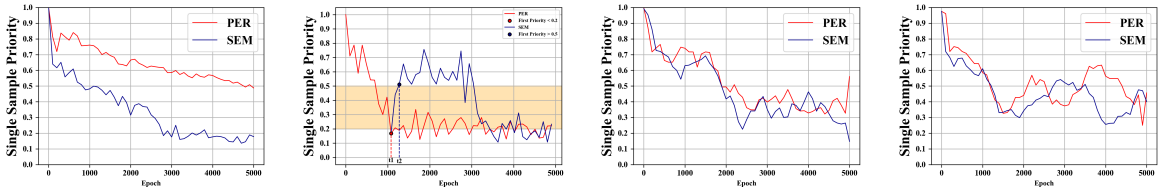
[6]Current methodological reporting gaps in dialogue policy learning research include insufficient specification of epoch-wise data scheduling parameters, which our work addresses by strictly following the standardized training protocol detailed in Zhao et al. (2024). The benchmark policies in the Covlab Platform, despite achieving 0.89 success rates, utilize hybrid training strategies involving MLP pretraining (initial success rate=0.56) followed by RL refinement. Our full retraining paradigm removes this pretraining advantage, with the SEM demonstrating statistically significant improvements over all baselines under this initialization condition (see App. E).

[5]10% noise approximates real-world dialogue system errors (8-12% range); other noise-level results appear in App. F.



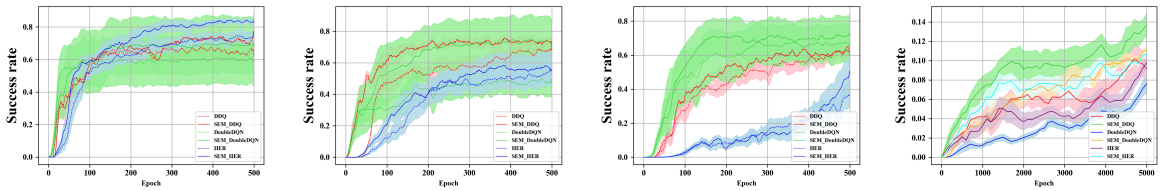
(a) Movie (b) Restaurant (c) Taxi (d) MultiWOZ 2.1

Figure 3: Evolution of experience states distribution across different baselines.



(a) Noisy (b) Forgotten (c) Stable (d) Unmastered

Figure 4: Sampling priorities for four experience states across MultiWOZ 2.1 dataset. The results on the movie, taxi, and restaurant datasets are presented in App.G



(a) Movie (b) Restaurant (c) Taxi (d) MultiWOZ 2.1

Figure 5: Incorporate different off-policy RL-based dialogue policy approaches with our SEM on four datasets. The standalone learning curves are presented in App. D.

ure 3. Both RER and TER exhibit initial uniformity across experience states, reflecting undifferentiated state recognition capabilities in early training. This homogeneity gradually gives way to increased distribution variance during later phases, with stable states ultimately dominating through spontaneous convergence mechanisms. PER demonstrates targeted prioritization by differentially processing stable and unmastered experiences. However, its inability to distinguish between noisy and forgotten states leads to concurrent amplification of both states, revealing fundamental pattern recognition limitations. While DDR effectively suppresses noise and forgotten states through dead-end detection and data augmentation, it shows limited efficacy on unmastered experiences. LLM-based agents introduce significant noise artifacts due to inherent hallucination effects, consistent with large model training literature. The proposed SEM addresses these limitations by dynamically prioritizing different experience states, reducing noisy and forgotten experiences while enhancing unmastered ones. This adaptive approach facilitates the progressive conversion of various experiences into stable states during later training phases.

4.4.2 Differentiated Efficacy Experience States

This section systematically evaluates the SEM's scheduling capability across four distinct experience states: noisy, forgotten, stable, and unmastered. We adopt PER as the baseline for two principal considerations: (i) As a representative dynamic sampling method, PER has demonstrated proven effectiveness in enhancing sample efficiency and training stability across diverse tasks; (ii) As a typical priority-based sampling method, PER shares the same priority mechanism as the SEM, making it a suitable baseline for highlighting the SEM's unique advantages. (iii) Empirical results reveal that the SEM and PER exhibit consistent priority evolution patterns for stable and unmastered experiences (see Fig. 4(c) and (d)). This equivalence establishes PER as an ideal baseline to isolate the SEM's unique advantages in noise suppression and forgotten experience reactivation.

Noisy Experience Filtering: To validate the SEM's noise robustness, we inject random noise into training experiences and dynamically track their sampling priority evolution. As shown in Fig. 4(a), the SEM's dual-error joint evaluation mechanism (via main and target Q-Networks) effectively identifies noisy experiences through persistent high error signals, subsequently suppressing their sampling priorities. In contrast, TD-errors-based PER fails to distinguish between meaningful gradients and noise fluctuations, persistently misidentifying noisy samples as high-priority candidates. This leads to repeated sampling of noisy experiences and degraded learning efficiency.

Forgotten Experience Reactivation: We construct a noise-free environment to isolate knowledge-forgetting, normalizing experience priorities from both methods. Thresholds ($A = 0.2, B = 0.5$) categorize experiences into *forgotten* (priority $< A$) and *reactivated* (priority $> B$). Fig.4(b) demonstrates PER's priorities drop below $A = 0.2$ at t_1, indicating gradual knowledge loss. Conversely, the SEM reactivates forgotten experiences via dual Q-network loss analysis, elevating their priorities above $B = 0.5$ at t_2. This benefit is especially pronounced in the complex taxi task and the MultiWOZ 2.1 multi-domain dialogue task.

Stable/Unmastered Experience Optimization: We visualize priority trends for stable and unmastered experiences during training. As depicted in Fig. 4(c) and (d), both methods appropriately reduce priorities for stable experiences while maintaining high priorities for unmastered ones. This alignment confirms the SEM retains PER's advantages in handling Stable/Unmastered experience states, focusing innovations on addressing PER's key limitations in noise and forgetting scenarios.

In summary, by establishing a quantitative mapping between dialogue experience states and network mastery phases, the SEM inherits PER's strengths in stable and unmastered experience management and significantly improves noise suppression and forgotten-experience recovery, thereby enhancing overall training efficiency and robustness.

4.5 Generality Evaluation

To verify the universality of the SEM, we conduct a generality evaluation by integrating it with representative off-policy RL dialogue policy methods, including DDQ (Peng et al., 2018), Double DQN (van Hasselt et al., 2016) and HER (Lu et al., 2019). As illustrated in Fig. 5, the SEM exhibits method-agnostic characteristics, achieving consistent performance improvements. The superior performance metrics stem from the SEM's dual advantages in experience state differentiation and dynamic priority allocation, which complement rather than conflict with existing RL methodologies.

4.6 Human Evaluation

To complement simulation limitations in assessing dialogue naturalness and coherence, we conducted human evaluations following the dataset platform's standardized protocol. Fifty-six evaluators (28 domain experts and 28 general users) interacted with trained models, scoring them on two metrics: Success Rate (SR, binary task completion) and Human Score (HS, 1-5 scale for naturalness/coherence). As shown in Tab. 2, the SEM achieves superior performance, aligning with simulation results.

Table 2: Human evaluation of different agents in different domains. For a fair comparison, all models in single-domain are trained for 500 epochs, while multi-domain (MultiWOZ 2.1) tests utilized RL-fine-tuned models initialized with MLE-pretrained models (corresponding simulation results are detailed in App. E). LLM-based agents are excluded from multi-domain comparisons due to their closed-source architectural constraints.

Model	Movie		Restaurant		Taxi		MultiWOZ 2.1	
	SR	HS	SR	HS	SR	HS	SR	HS
RER	0.4643	2.33	0.1786	2.55	0.2143	2.02	0.3263	2.47
PER	0.4107	2.65	0.3036	2.01	0.2321	2.14	0.3861	2.86
TER	0.5357	2.58	0.3214	2.46	0.2500	2.10	0.3790	2.82
DDR	0.4931	2.77	0.3323	2.56	0.2776	2.45	0.4028	3.30
LLM_DA	0.1968	1.68	0.1268	1.46	0.0937	1.73	–	–
LLM_Word	0.0890	1.20	0.1050	1.23	0.0366	1.09	–	–
SEM	**0.5714**	**3.05**	**0.3571**	**2.98**	**0.3036**	**2.55**	**0.4489**	**3.83**

5 Conclusion

This paper proposes a novel SEM approach that captures the latent rhythms between experience significance and policy evolution by quantitatively mapping dialogue experience states to policy learning phases, offering new methodological insights for experience management. By analyzing loss discrepancies between the main and the target Q-networks, the SEM dynamically evaluates experience states and adjusts sampling priorities to enhance sampling efficiency. As a model-agnostic framework, the SEM can effectively integrate off-policy RL-based dialogue policy algorithms to improve performance. Extensive evaluations across four TOD datasets under various noise configurations validate the SEM's effectiveness and generalizability. Visual analytics on the sampling priorities of experiences further confirms its capability to optimize priority assignments adaptively.

Limitations

The SEM demonstrates strong effectiveness in advancing prioritized experience replay, with two main limitations: First, due to its design for priority calculation, the SEM is primarily focused on optimizing experience replay in off-policy RL. In off-policy learning, historical experiences are the sole source of learning, and optimizing the priority of these experiences is crucial for improving learning efficiency and stability. However, in Online RL, real-time interactions and dynamic updates to the experience pool limit the applicability of the SEM. Future efforts will extend our method to accommodate scenarios involving Online RL. Second, the SEM dynamically adjusts the priorities of the four different experience categories, particularly emphasising the differentiated treatment of noisy and forgotten experiences. Although the SEM is an effective experience management mechanism that alleviates prioritization distortion in experience replay, it still faces the challenge of not eliminating the impact of negative experiences, as shown in Figure 3. Future research will enhance the SEM's ability to handle these situations more effectively.

Ethics Statement

We have carefully considered the potential ethical implications of our work and conclude that it does not pose significant ethical risks. Specifically:

No sensitive data involved: All experiments are conducted using publicly available or synthetic datasets that contain no sensitive personal information or private user interaction data.

Responsible human evaluation: Although human evaluation is included, it was conducted without involving identifiable individuals or personally sensitive content. No formal ethical approval was required, as the evaluation process posed no risk or harm to participants.

Low risk of misuse: The methods and data presented in this work do not enable or facilitate the generation of harmful, offensive, or otherwise unsafe content. We do not foresee any substantial risk of misuse.

Based on these considerations, our work adheres to standard ethical research practices and does not require further ethical review.

References

Atheer Algherairy and Moataz Ahmed. 2024. A review of dialogue systems: current trends and future directions. *Neural Computing and Applications*, 36(12):6325–6351.

Atheer Algherairy and Moataz Ahmed. 2025. Prompting large language models for user simulation in task-oriented dialogue systems. *Comput. Speech Lang.*, 89:101697.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026. Association for Computational Linguistics.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, and Simone Calderara. 2020. Rethinking experience replay: a bag of tricks for continual learning. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 2180–2187. IEEE.

Shixiang Gu, Timothy P. Lillicrap, Zoubin Ghahramani, Richard E. Turner, and Sergey Levine. 2017. Q-prop: Sample-efficient policy gradient with an off-policy critic. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Zhang-Wei Hong, Tao Chen, Yen-Chen Lin, Joni Pajarinen, and Pulkit Agrawal. 2022. Topological experience replay. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Dan Horgan, John Quan, David Budden, Gabriel Barth-Maron, Matteo Hessel, Hado van Hasselt, and David Silver. 2018. Distributed prioritized experience replay. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Navin Kamuni, Hardik Shah, Sathishkumar Chintala, Naveen Kunchakuri, and Sujatha Alla Old Dominion. 2024. Enhancing end-to-end multi-task dialogue systems: A study on intrinsic motivation reinforcement learning algorithms for improved training and adaptability. In *18th IEEE International Conference on Semantic Computing, ICSC 2024, Laguna Hills, CA, USA, February 5-7, 2024*, pages 335–340. IEEE.

Su Young Lee, Sung-Ik Choi, and Sae-Young Chung. 2019. Sample-efficient deep reinforcement learning via episodic backward update. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2110–2119.

Xiujun Li, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *CoRR*, abs/1807.11125.

Minsong Liu, Yuanheng Zhu, Yaran Chen, and Dongbin Zhao. 2024. Enhancing reinforcement learning via transformer-based state predictive representations. *IEEE Trans. Artif. Intell.*, 5(9):4364–4375.

Cong Lu, Philip J. Ball, Yee Whye Teh, and Jack Parker-Holder. 2023. Synthetic experience replay. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Keting Lu, Shiqi Zhang, and Xiaoping Chen. 2019. Goal-oriented dialogue policy learning from failures. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 2596–2603. AAAI Press.

Yongsheng Mei, Hanhan Zhou, Tian Lan, Guru Venkataramani, and Peng Wei. 2023. MAC-PO: multi-agent experience replay via collective priority optimization. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, pages 466–475. ACM.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nat.*, 518(7540):529–533.

Youngmin Oh, Jinwoo Shin, Eunho Yang, and Sung Ju Hwang. 2022. Model-augmented prioritized experience replay. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. 2018. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2182–2192. Association for Computational Linguistics.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized experience replay. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Hado van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2094–2100. AAAI Press.

Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. 2017. Feudal networks for hierarchical reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3540–3549. PMLR.

Biao Yang, Fucheng Fan, Rongrong Ni, Jie Li, Chu Kiong Loo, and Xiaofeng Liu. 2022. Continual learning-based trajectory prediction with memory augmented networks. *Knowl. Based Syst.*, 258:110022.

Zhen Yang, Ming Ding, Tinglin Huang, Yukuo Cen, Junshuai Song, Bin Xu, Yuxiao Dong, and Jie Tang. 2024. Does negative sampling matter? a review with insights into its theory and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5692–5711.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *CoRR*, abs/2402.18013.

Jiayu Yu, Jingyao Li, Shuai Lü, and Shuai Han. 2024. Mixed experience sampling for off-policy reinforcement learning. *Expert Syst. Appl.*, 251:124017.

Ye Zhang, Wang Zhao, Jingyu Wang, and Yuan Yuan. 2024. Recent progress, challenges and future prospects of applied deep reinforcement learning: A practical perspective in path planning. *Neurocomputing*, 608:128423.

Yangyang Zhao, Mehdi Dastani, Jinchuan Long, Zhenyu Wang, and Shihan Wang. 2024. Rescue conversations from dead-ends: Efficient exploration for task-oriented dialogue policy optimization. *Trans. Assoc. Comput. Linguistics*, 12:1578–1596.

Yangyang Zhao, Zhenyu Wang, Kai Yin, Rui Zhang, Zhenhua Huang, and Pei Wang. 2020. Dynamic reward-based dueling deep dyna-q: Robust policy learning in noisy environments. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9676–9684. AAAI Press.

A Dataset Details

MultiWOZ 2.1 is a large-scale, multi-domain TOD dataset that includes dialogues from multiple domains such as restaurant booking, hotel booking, taxi booking, and tourist attraction recommendations. The dataset provides detailed user intent, slot annotations, and dialogue context, making it suitable for evaluating core tasks such as dialogue management, intent recognition, and slot filling. The scale and complexity of MultiWOZ 2.1 make it an ideal choice for testing cross-domain generalization ability. The Microsoft Dialogue Challenge focuses on daily conversations and customer support, offering diverse dialogue scenarios across three domains: movie-ticket booking, restaurant booking, and taxi booking, making it suitable for multi-task learning and sentiment analysis research. By training with these datasets, this study can validate the effectiveness of the proposed SEM loss method in multi-domain, multi-task environments, particularly in handling noisy experiences, forgetting experiences, and human evaluation performance.

Table 3: Dataset statistics for various dialogue tasks.

Dataset	Domains	Scale
MultiWOZ 2.1	7	Dialogue scale: 8,438; Dialogue rounds: 115,424; Avg. rounds: 13.68; Slots: 25
Movie	1	Dialogue scale: 2,890; Intention: 11; Slots: 29
Restaurant	1	Dialogue scale: 4,103; Intention: 11; Slots: 30
Taxi	1	Dialogue scale: 3,094; Intention: 11; Slots: 29

B Experimental setup details

Table 4: Experimental Settings

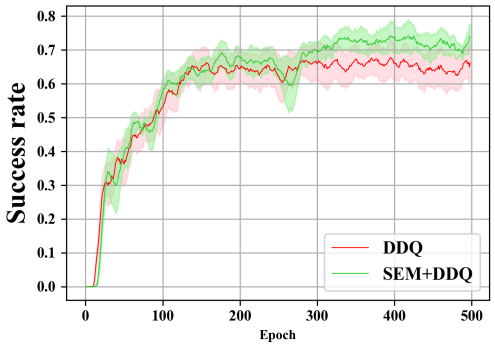
Parameter	Value / Description
Exploration rate (ϵ)	Initial: 0.1; decayed to 0.01 during training
Update coefficient (τ)	1×10^{-2} (soft update for target network)
L2 regularization coefficient (regc)	1×10^{-3}
Discount factor (γ)	0.99
Batch size	16
Learning rate	0.001
Replay buffer size	10,000
Operating system	Ubuntu 24.01
Python version	3.9
Toolkit	ConvLab-3
Simulator	Rule-based simulator from ConvLab-3
Model architecture	Modular neural models
Fmap	ReLU(x)=max(0,x)

C Prompt Design

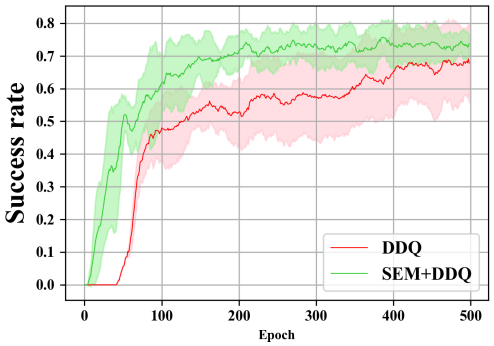
Table 5: Descriptions of Prompts used for LLM-based baselines.

Model	Prompt
LLM_DA	1. System role definition: Function as the policy component in of a task-oriented dialogue system, you need to produce system responses according to dialogue context. 2. Processing user dialogue state: Process the provided dialogue state representation to guide response selection. 3. Generate system actions: Given the user's dialogue state, generate system actions in the format: [["ActionType", "Domain", "Slot", "Value"]]. Here, 'ActionType' refers to the system's intended operation (e.g., Request, Inform, Confirm), 'Domain' indicates the relevant area (such as restaurant, taxi, or hotel), 'Slot' specifies the information field (like name, area, or type), and 'Value' holds the related content or remains empty if not provided. 4. Compliance requirements: Output must strictly conform to JSON schema without extraneous content.
LLM_Word	1. System role definition: Function as the dialog policy module and natural language generation module of task-oriented dialogue systems, you need to determine system behaviors based on real-time conversational state analysis. 2. Processing user dialogue state: Process the provided dialogue state representation to guide response selection. 3. Generate system actions: make decisions based on the current state of the dialogue and formulate natural language responses to the user. 4. Compliance requirements: Output must strictly conform to JSON schema without extraneous content.

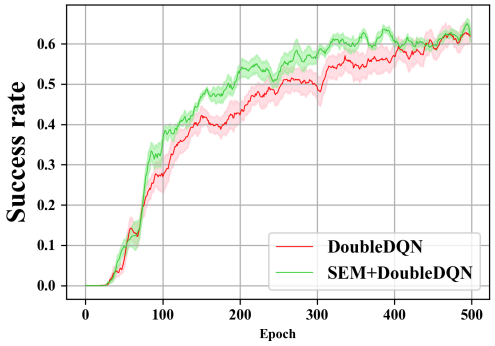
D Generality Evaluation of SEM-based Methods



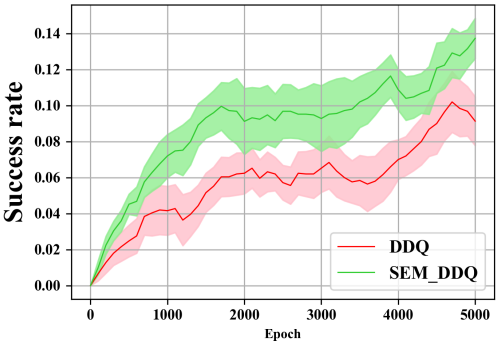
(a) Movie



(b) Restaurant

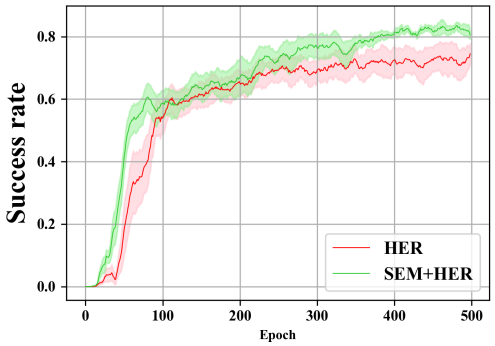


(c) Taxi

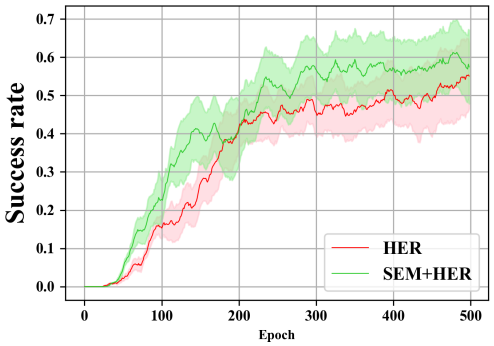


(d) MultiWOZ 2.1

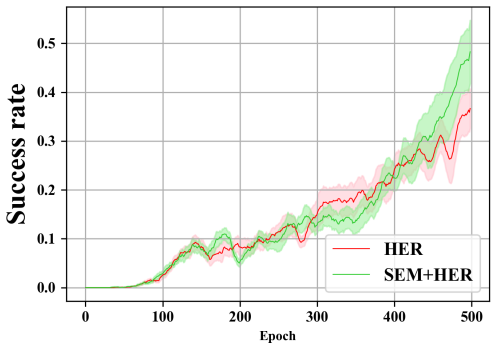
Figure 6: Incorporating DDQ with our SEM on four datasets.



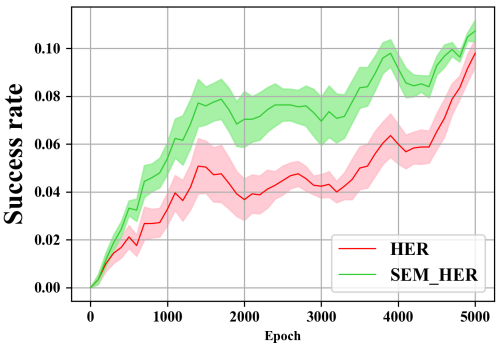
(a) Movie



(b) Restaurant

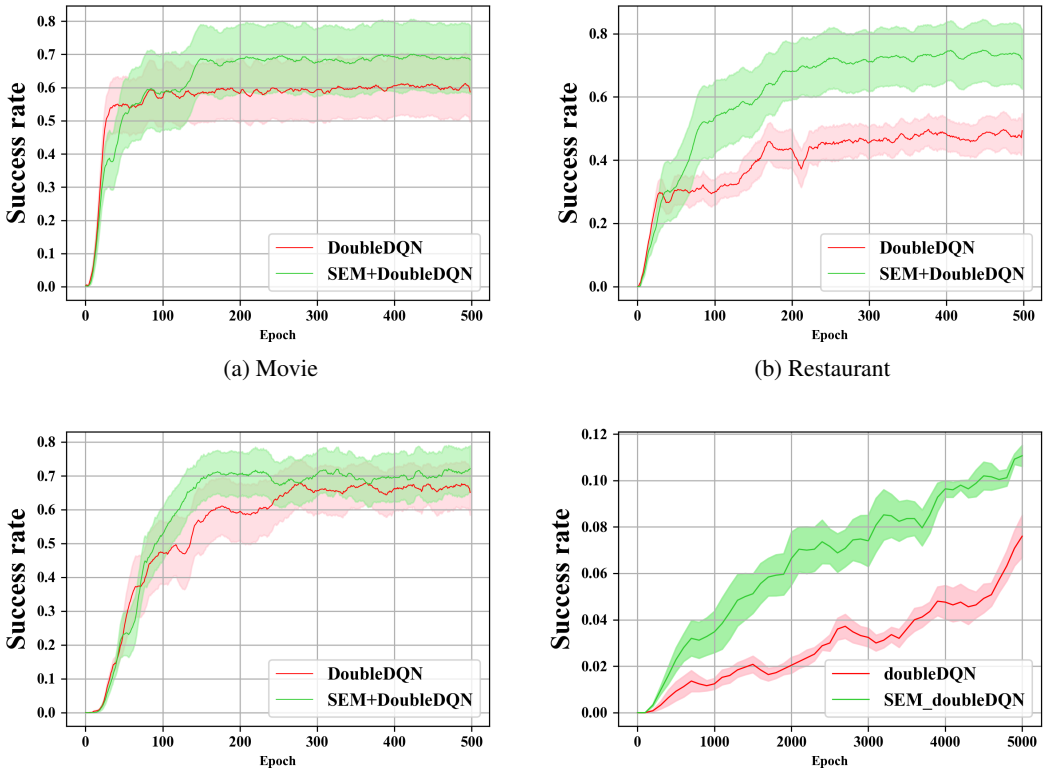


(c) Taxi



(d) MultiWOZ 2.1

Figure 7: Incorporating HER with our SEM on four datasets.



(a) Movie (b) Restaurant

(c) Taxi (d) MultiWOZ 2.1

Figure 8: Incorporating Double DQN with our SEM on four datasets.

E Model Performance Initialized with the Pretrained MLE Model

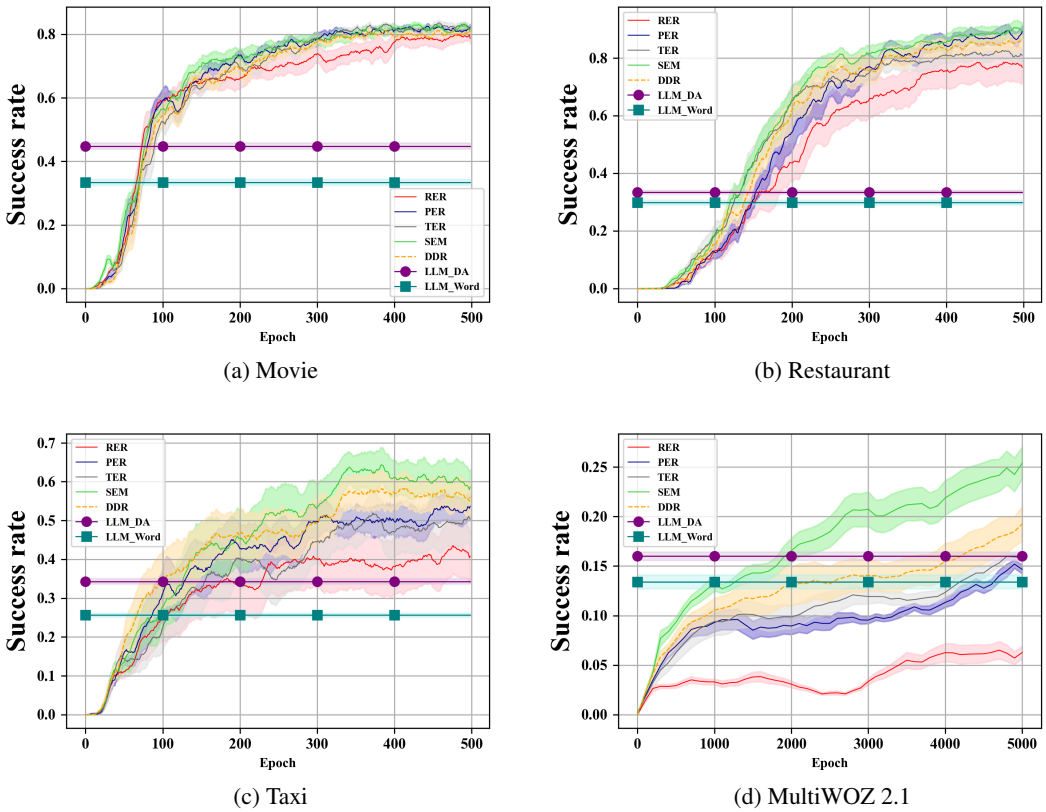
Table 6: Performance Comparison of Different Models Initialized with the Pretrained MLE Model

Model	Success ↑	Rewards ↑	Turns ↓
MLE	0.56	20.9	24.1
RER	0.67	25.8	21.5
PER	0.71	35.8	22.3
TER	0.74	32.1	26.4
DDR	0.82	37.2	25.3
LLM_DA	–	–	–
LLM_Word	–	–	–
SEM	**0.89**	**44.6**	**19.0**

The results provided by the ConvLab platform are based on MLE models that have been pretrained and subsequently fine-tuned using RL, rather than models trained from scratch. Therefore, directly comparing those results to our own models trained from scratch would be unfair. To ensure a fair comparison under the same conditions, we initialized our training with the pretrained MLE model provided by ConvLab (which achieves a success rate of 0.56). We then apply RL-based baselines for fine-tuning over 2000 epochs, all uniformly based on the PPO architecture. LLM-based baselines (e.g., GPT4.0) are not available in this comparison due to their closed-source nature, which prevents us from performing any fine-tuning or reinforcement learning-based adaptation. The results, as shown in Tab. 6, further confirm that our proposed method consistently outperforms others—whether trained from scratch or fine-tuned from pretrained models.

F Performance Comparison under Varying Levels of Noise

As shown in Fig. 10, the SEM outperforms all other methods in all scenarios, demonstrating higher success rates and better stability, especially in high noisy environments (15%&20%). It highlights its ability to identify and avoid noisy experiences effectively. Although PER performs better than RER, its performance still lags behind the SEM. The increased probability of noisy experiences in such environments leads PER to repeatedly prioritize this experience, which is confusing training. In contrast, the SEM adjusts the priority of noisy experience by comparing losses between the main and marget Q-networks, ensuring that its priority decreases as training progresses, avoiding confusion. For TER, using hash tables to construct the graph optimization experience replay process is better than RER to a certain extent. Still, its performance will be limited when the task state and action space increase sharply and are disturbed by noisy experiences (Zhao et al., 2020).

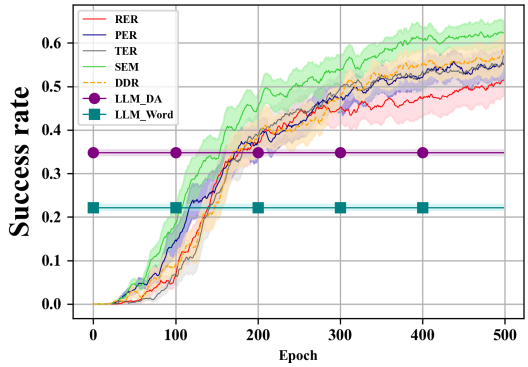


(a) Movie (b) Restaurant

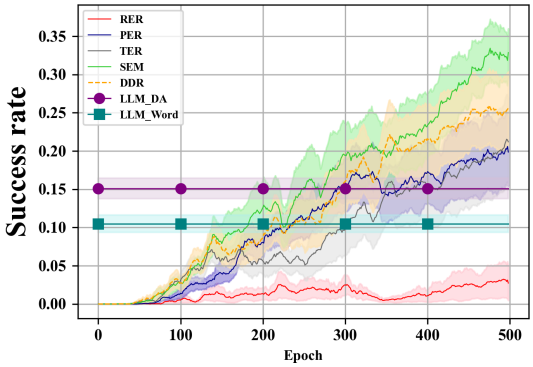
(c) Taxi (d) MultiWOZ 2.1

Figure 9: Performance comparison of agents under noisy-free environment) across four datasets.

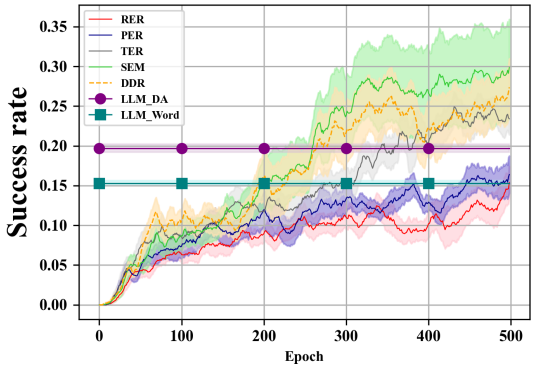
Furthermore, we conducted experiments under ideal conditions without the presence of noise, as depicted in Fig. 9. Even in such scenarios, our method continues to deliver superior performance. This outcome eliminates the influence of noisy experiences, highlighting the impact of other experience types, particularly forgotten experiences.



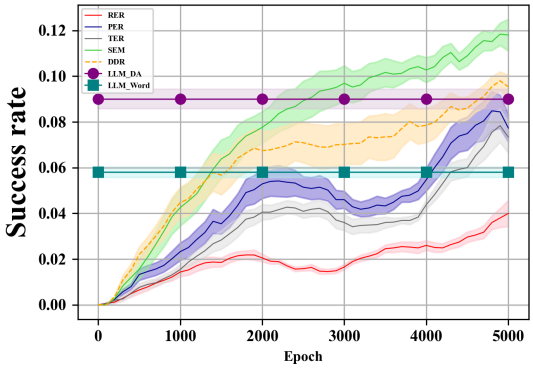
(a) Movie (15%)



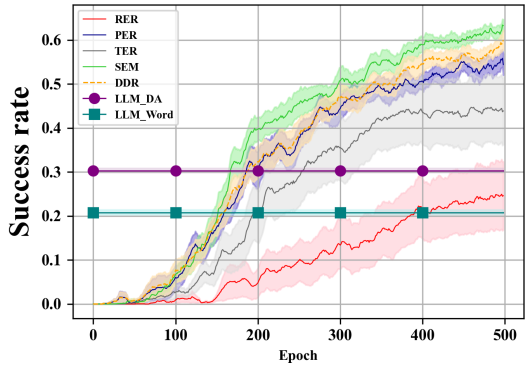
(b) Restaurant (15%)



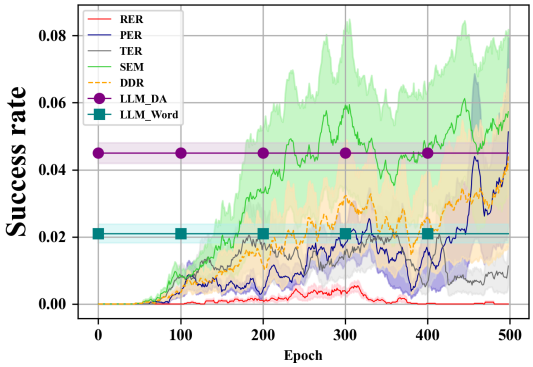
(c) Taxi (15%)



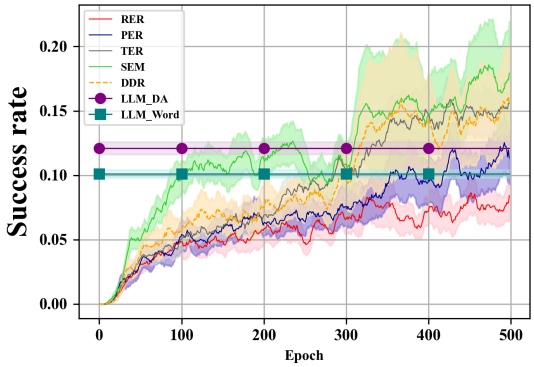
(d) MultiWOZ 2.1 (15%)



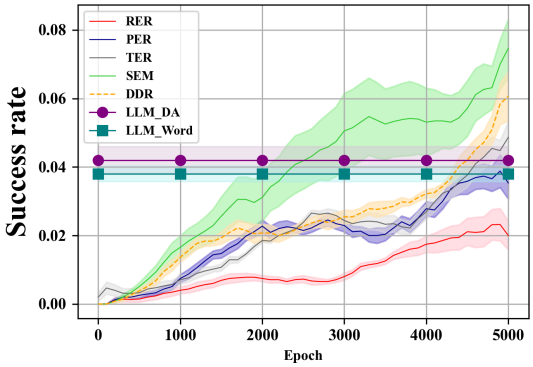
(e) Movie (20%)



(f) Restaurant (20%)



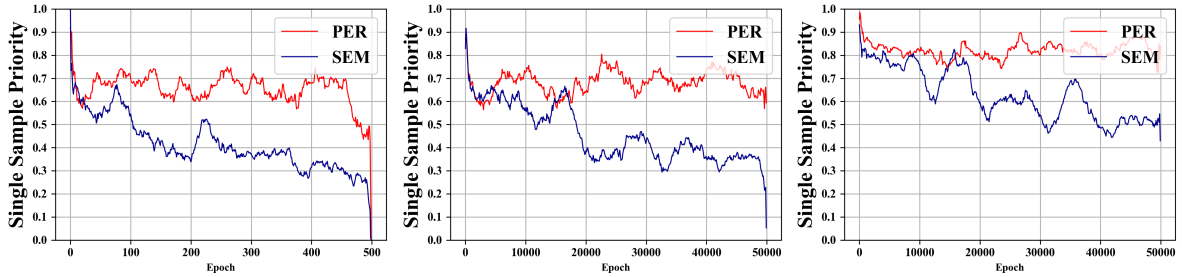
(g) Taxi (20%)



(h) MultiWOZ 2.1 (20%)

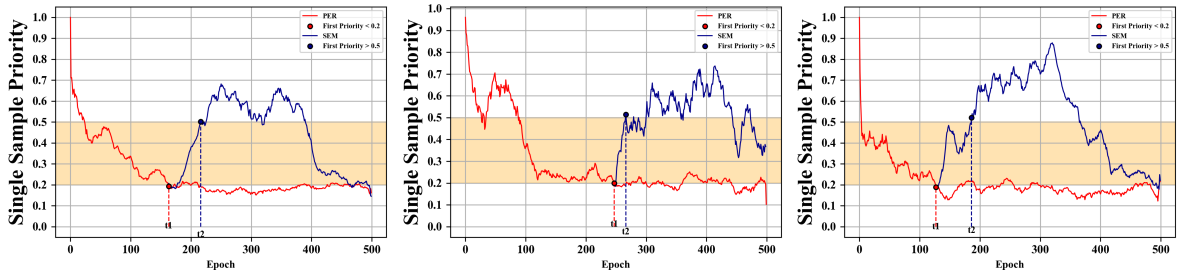
Figure 10: Performance comparison of agents under noisy environments (15% and 20% noise) across four datasets.

G Sampling priorities for four experience states across Movie, Restaurant, Taxi datasets



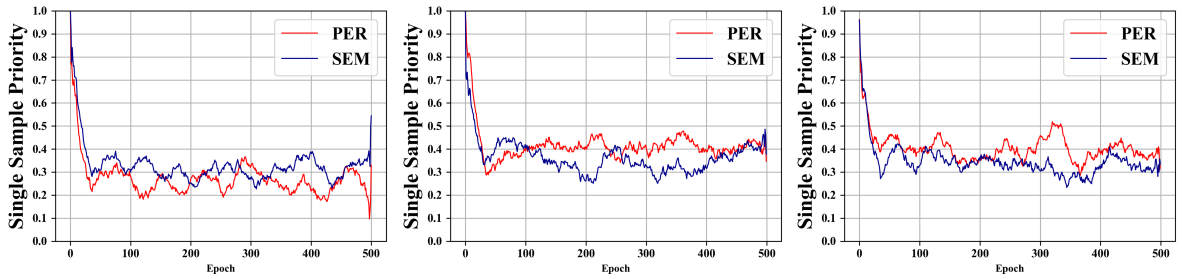
Movie Restaurant Taxi

(a) Sampling priorities for noisy experiences across three datasets.



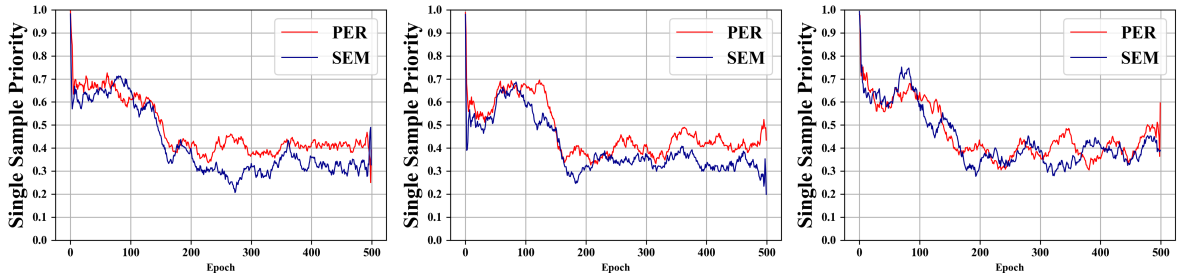
Movie Restaurant Taxi

(b) Sampling priorities for forgotten experiences across three datasets.



Movie Restaurant Taxi

(c) Sampling priorities for stable experiences across three datasets.



Movie Restaurant Taxi

(d) Sampling priorities for unmastered experiences across three datasets.

Figure 11: Comparative analysis of priority trends for four experience states in SEM and PER.