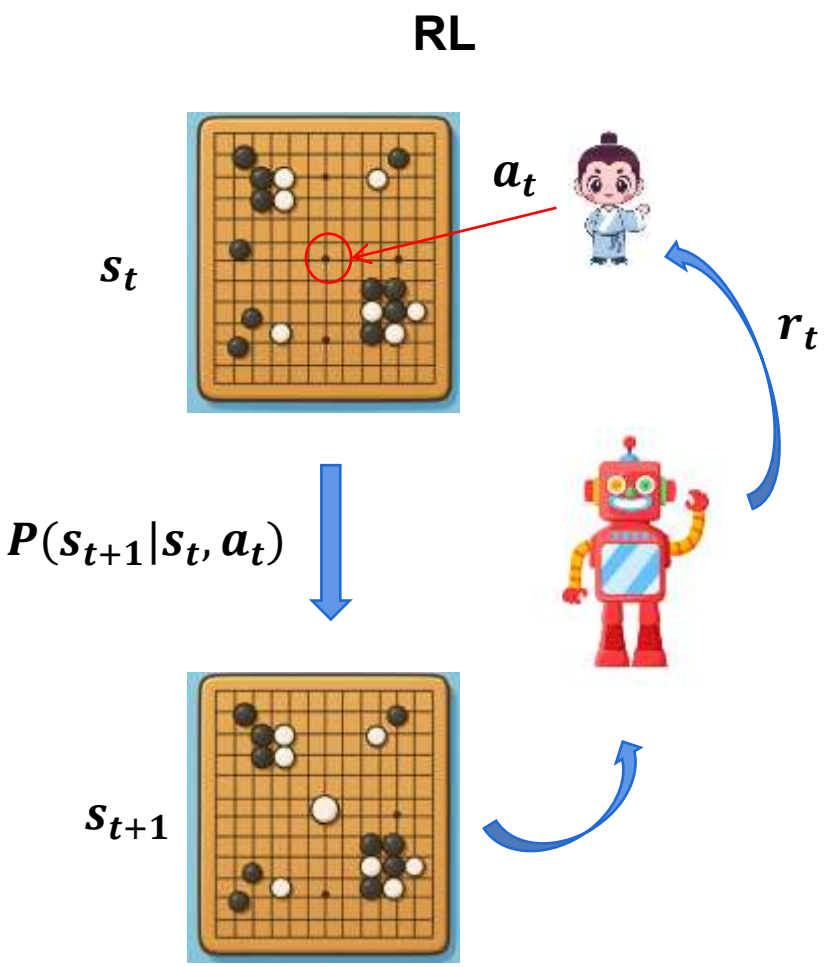
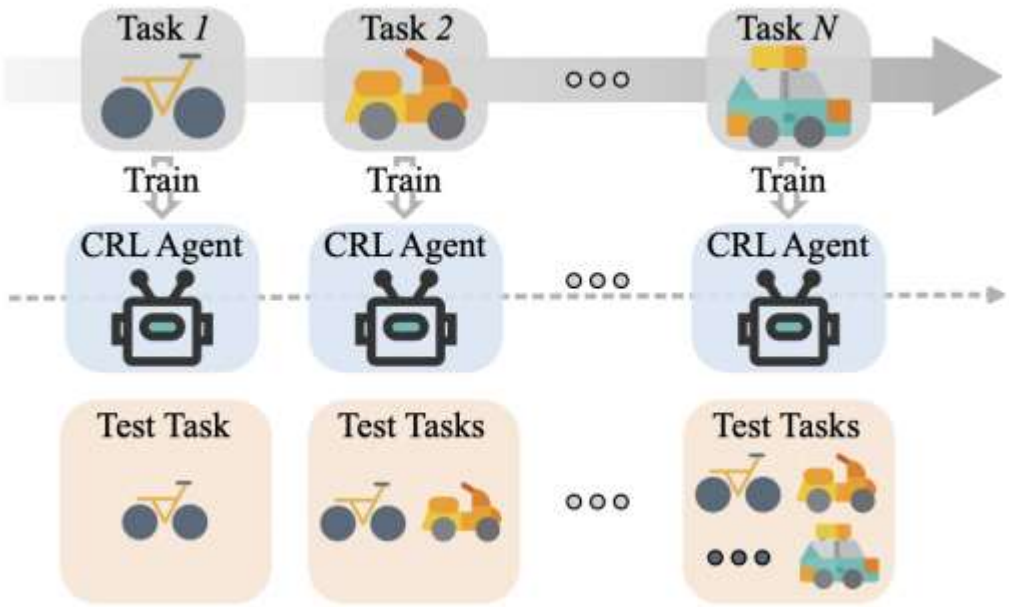
Continual Reinforcement Learning

Background



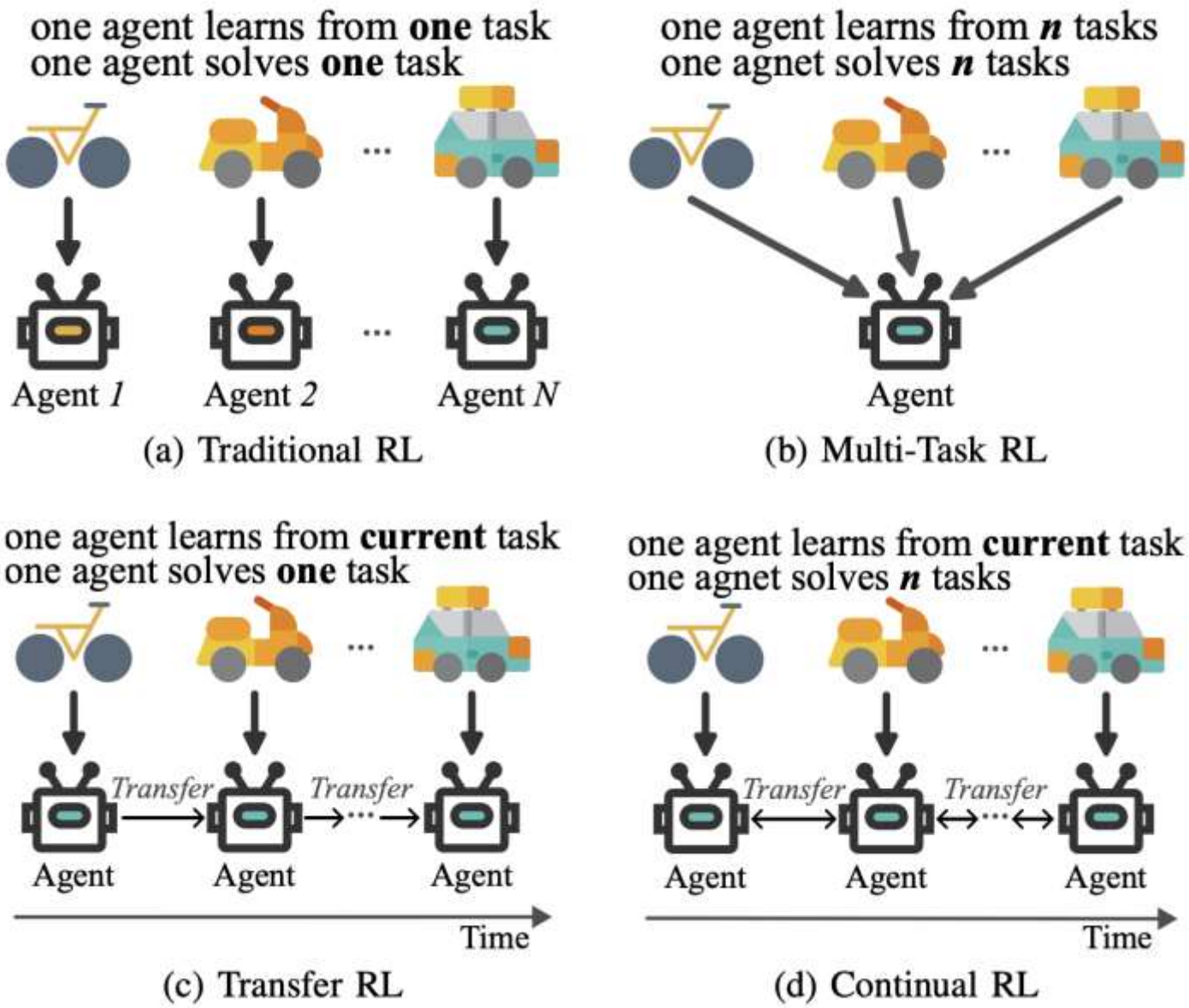
For a fixed MDP: $M = (S, A, R, P, \gamma)$



For a sequence of MDPs: $M_i = (S, A, R_i, P_i, \gamma)$

share the state and action space

Background



Challenges:

1. *Plasticity:* learn new tasks after being trained on previous tasks

2. *Stability:* maintain performance on previously learned tasks while simultaneously learning new tasks.

3. *Scalability:* learn many tasks using limited resources

PRINCIPLED FAST AND META KNOWLEDGE LEARNERS FOR CONTINUAL REINFORCEMENT LEARNING

ICLR 2026 under review

Definitions :

Definition 1. ***(MDP Distance)*** *For two finite MDPs:* $MDP_1 = (\mathcal{S}, \mathcal{A}, R_1, P_1, \gamma)$ *and* $MDP_2 = (\mathcal{S}, \mathcal{A}, R_2, P_2, \gamma)$*, we denote their optimal Q functions as* Q_1^* *and* Q_2^* *and the optimal policies as* π_1^* *and* π_2^**. The Q-value-based and policy-based MDP distances are defined as* $d_Q(Q_1^*, Q_2^*)$ *and* $d_\pi(\pi_1^*, \pi_2^*)$ *under certain divergences or distances* d_Q *and* d_π*, e.g., the* ℓ_2 *loss or the KL divergence.*

Definition 2. ***(Catastrophic Forgetting across Two Environments)*** *Denote* Q_{k-1}, Q_k *and* π_{k-1}, π_k *as Q functions and policies after training RL algorithms across the* $(k-1)$*-th and* k*-th environments sequentially. The catastrophic forgetting, denoted by CF, is defined as*

$$CF(Q_{k-1}, Q_k) = \sum_{s,a} \mu_{k-1}^{Q_{k-1}}(s) \pi^{Q_{k-1}}(a|s) d_Q \left(Q_{k-1}(s,a), Q_k(s,a)\right), \tag{1}$$

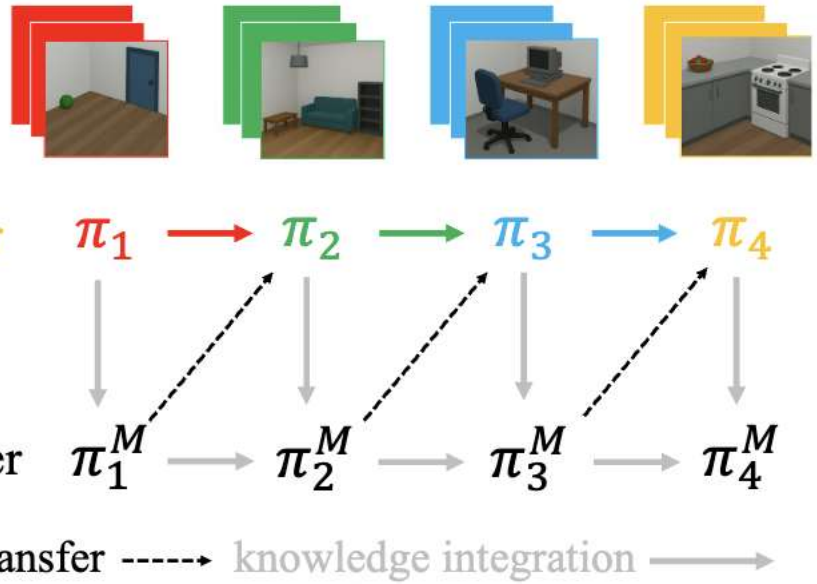
$$CF(\pi_{k-1}, \pi_k) = \sum_{s} \mu_{k-1}^{\pi_{k-1}}(s) d_\pi \left(\pi_k(\cdot|s), \pi_{k-1}(\cdot|s)\right). \tag{2}$$

Method

Q-Value-based Catastrophic Forgetting

$$Q_k^M = \arg\min_{\widetilde{Q}_k^M} \sum_{i=1}^{k} \sum_{s,a} \mu_i^{Q_i}(s)\pi^{Q_i}(a|s)\left(Q_i(s,a) - \widetilde{Q}_k^M(s,a)\right)^2$$

$$Q_k^M = \arg\min_{\widetilde{Q}_k^M} \sum_{i=1}^{k-1} \mathbb{E}_{w_i^Q}\left[\left(Q_{k-1}^M - \widetilde{Q}_k^M\right)^2\right] + \mathbb{E}_{w_k^Q}\left[\left(Q_k - \widetilde{Q}_k^M\right)^2\right]$$



Policy-based Catastrophic Forgetting

$$\pi_k^M = \arg\min_{\widetilde{\pi}_k^M} \sum_{i=1}^{k} \sum_{s} \mu_i^{\pi_i}(s) d_\pi\left(\pi_i(\cdot|s), \widetilde{\pi}_k^M(\cdot|s)\right)$$

$$\widetilde{\pi}_k^M(a|s) = \exp\left(\widetilde{Q}_k^M(a|s)/\tau\right) / \sum_{a'} \exp\left(\widetilde{Q}_k^M(a'|s)/\tau\right)$$

$$Q_k^M = \arg\min_{\widetilde{Q}_k^M} \sum_{i=1}^{k-1} \mathbb{E}_{w_i^Q}\left[\log \frac{\pi_{k-1}^M}{\widetilde{\pi}_k^M}\right] + \mathbb{E}_{w_k^Q}\left[\log \frac{\pi^{Q_k}}{\widetilde{\pi}_k^M}\right] = \arg\max_{\widetilde{Q}_k^M} \sum_{i=1}^{k} \mathbb{E}_{w_i^Q}\left[\log \widetilde{\pi}_k^M\right]$$

Method

Adaptive Meta Warm-Up

$$V_k^f = \mathbb{E}_{\pi_{k-1}}[R],\ V_k^M = \mathbb{E}_{\pi_{k-1}^M}[R],\ \text{and}\ V_k^r = \mathbb{E}_{\pi^0}[R]$$

$$H_0 : V_k^M \leq \max\left\{V_k^f, V_k^r\right\} \quad \text{vs.} \quad H_1 : V_k^M > \max\left\{V_k^f, V_k^r\right\}. \tag{6}$$

Algorithm 1 Value-based FAME Update in the k-th Environment

Initialize: Fast Buffer $\mathcal{F}$, Meta Buffer $\mathcal{M}$, Q_{k-1}^M, Q_{k-1}, Q^0, Warm-Up Step L, Estimation Step N.
Knowledge Transfer: Adaptive Meta Warm-Up
Initialize Q_k in $\{Q_{k-1}, Q_k^M, Q^0\}$ via Eq. 6 within L steps
for $t = L$ to T **do**
  Observe S_t, take action A_t, receive R_t, observe S_{t+1}
  Store (S_t, A_t, R_t, S_{t+1}) in $\mathcal{F}$
  Update Q_k
  if $t > T - N$ **then**
    Store (S_t, A_t) in $\mathcal{M}$ # To Estimate w_k^Q
  end if
end for
Reset $\mathcal{F}$
Knowledge Integration: Minimize Catastrophic Forgetting
Update Q_k^M via Eq. 5 using state-action pairs in $\mathcal{M}$

Experiments

Evaluation Metrics

Average Performance $P_K(t) = \frac{1}{K}\sum_{i=1}^{K} p_i(t)$

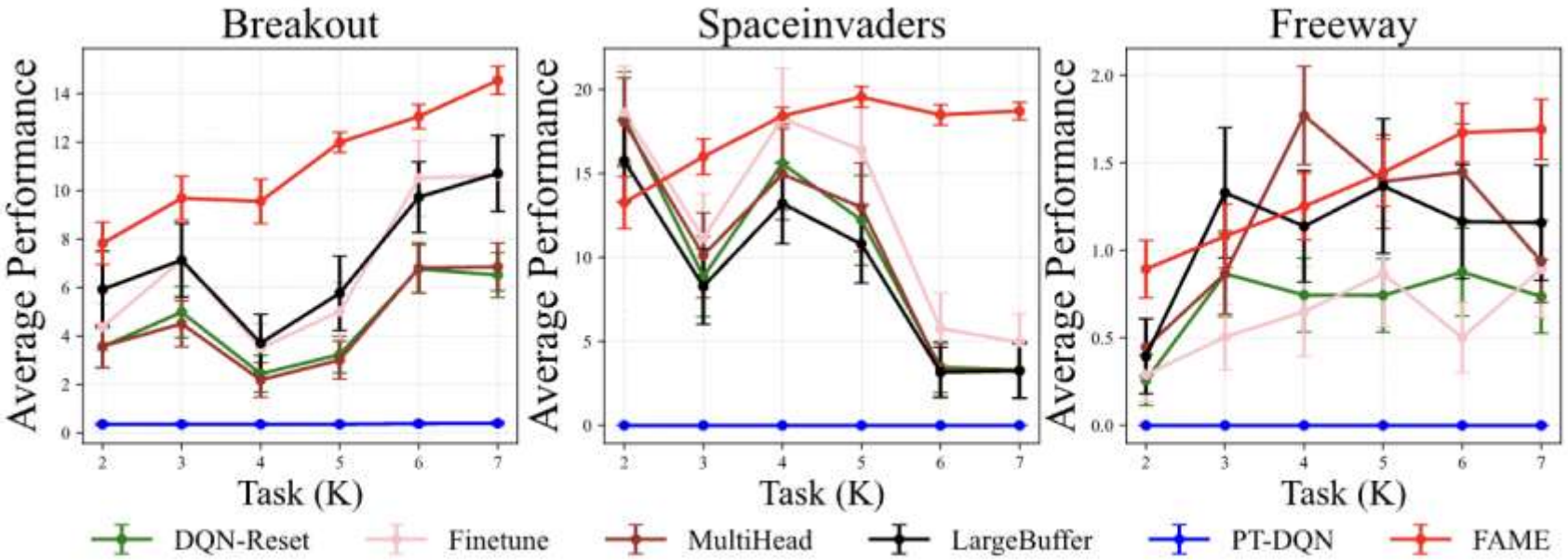
Forward Transfer (FT) $\mathrm{FT} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{FTr}_i$

$$\mathrm{FTr}_i = \frac{\mathrm{AUC}_i - \mathrm{AUC}_i^b}{1 - \mathrm{AUC}_i^b}, \quad \mathrm{AUC}_i = \frac{1}{\Delta}\int_{(i-1)\cdot\Delta}^{i\cdot\Delta} p_i(t)\mathrm{d}t, \quad \mathrm{AUC}_i^b = \frac{1}{\Delta}\int_{(i-1)\Delta}^{i\Delta} p_i^b(t)\mathrm{d}t.$$

Forgetting $F = \frac{1}{K}\sum_{i=1}^{K} F_i \text{ with } F_i = p_i(i \cdot T) - p_i(K \cdot T)$

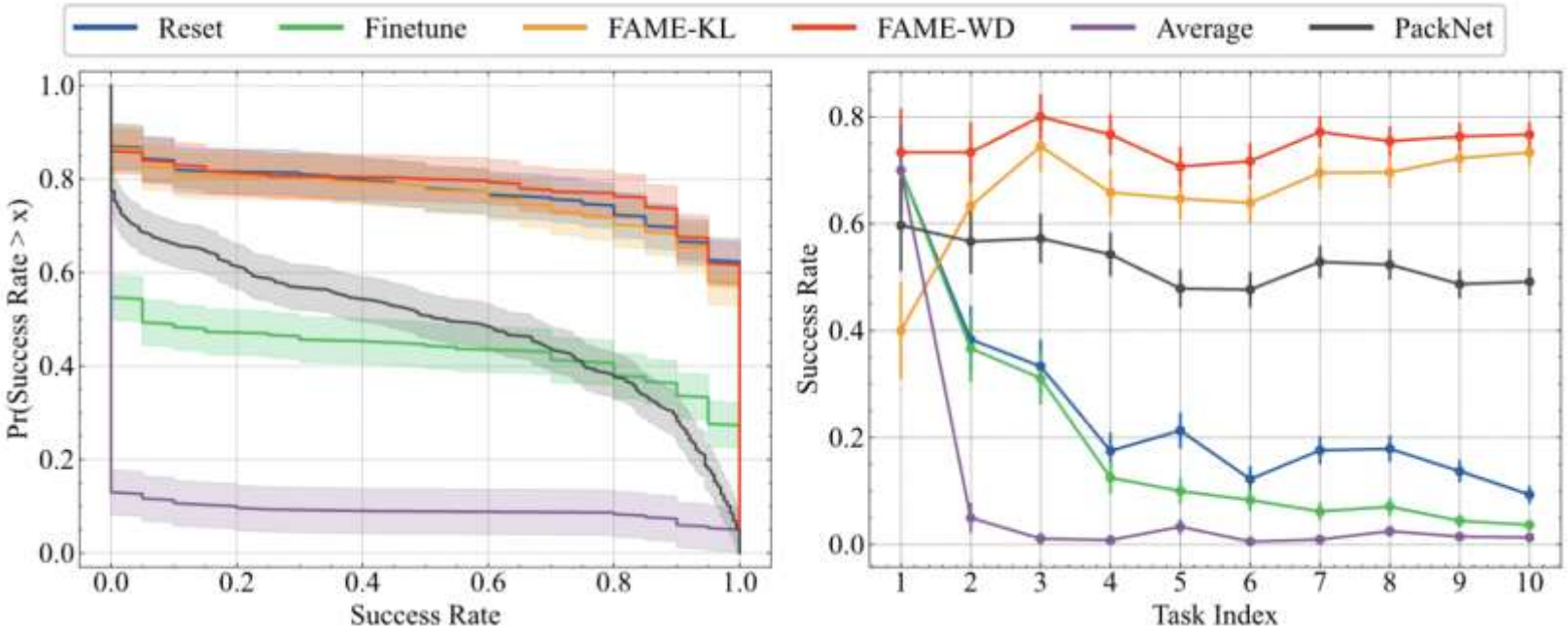
Experiments

Method	Ave. Perf ↑ Breakout	Spaceinvader	Freeway	FT ↑	Forgetting ↓
Reset	6.51 ± 1.67	3.29 ± 3.09	0.74 ± 0.38	0.00 ± 0.00	1.31 ± 0.23
Finetune	10.62 ± 2.75	4.95 ± 2.92	0.89 ± 0.49	0.13 ± 0.03	1.26 ± 0.32
MultiHead	6.85 ± 1.76	3.26 ± 2.99	0.94 ± 0.42	-0.01 ± 0.00	1.25 ± 0.22
LargeBuffer	10.71 ± 2.84	3.24 ± 2.91	1.16 ± 0.59	**0.16 ± 0.02**	1.65 ± 0.33
PT-DQN	0.39 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.07 ± 0.02	1.64 ± 0.02
FAME	**14.54 ± 0.58**	**18.72 ± 0.52**	**1.69 ± 0.17**	**0.16 ± 0.03**	**0.72 ± 0.13**



Experiments

Methods	Avg. Perf ↑	FT ↑	Forgetting ↓
Reset	0.093 ± 0.017	0.000 ± 0.000	0.710 ± 0.030
Finetune	0.037 ± 0.011	-0.265 ± 0.028	0.427 ± 0.033
Average	0.013 ± 0.007	-0.530 ± 0.024	0.070 ± 0.022
PackNet	0.491 ± 0.025	-0.194 ± 0.018	**0.000 ± 0.000**
FAME-KL	0.733 ± 0.026	**0.022 ± 0.015**	0.073 ± 0.019
FAME-WD	**0.767 ± 0.024**	-0.003 ± 0.014	0.023 ± 0.015



CRL for LLM

The state and action spaces of LLMs are inherently invariant.

$$R\colon\ human\ feedback \leftrightarrow objective\ facts$$

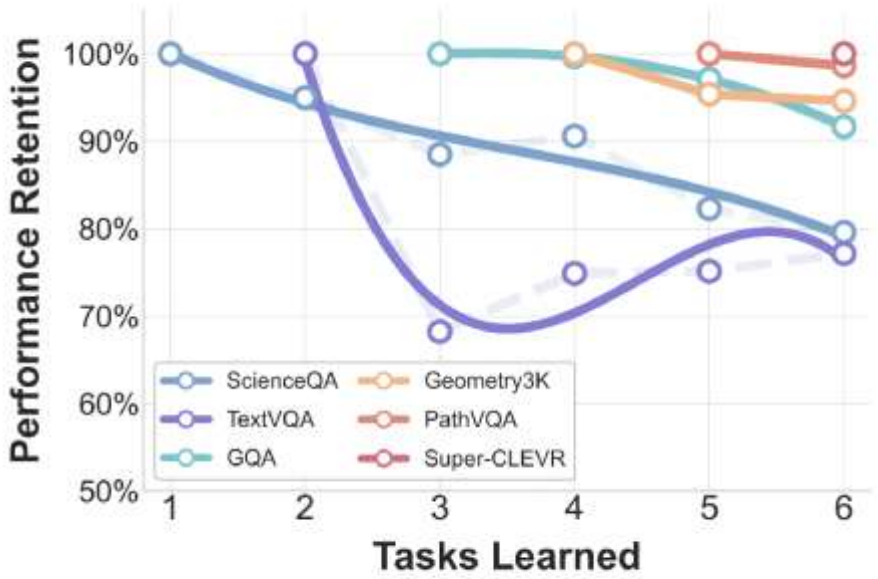
CPPO: Continual Learning for Reinforcement Learning with Human Feedback (ICLR 2024)

$$\begin{aligned}\mathbf{J}^{'}(\theta) &= L_i^{I_{D_1}\cdot CLIP+I_{D_2}\cdot KR+VF}(\theta) \\ &= \mathbb{E}_i[I_{D_1}(x)\cdot L_i^{CLIP}(\theta) - I_{D_2}(x)\cdot L_i^{KR}(\theta) - c\cdot L_i^{VF}(\theta)]\end{aligned}$$

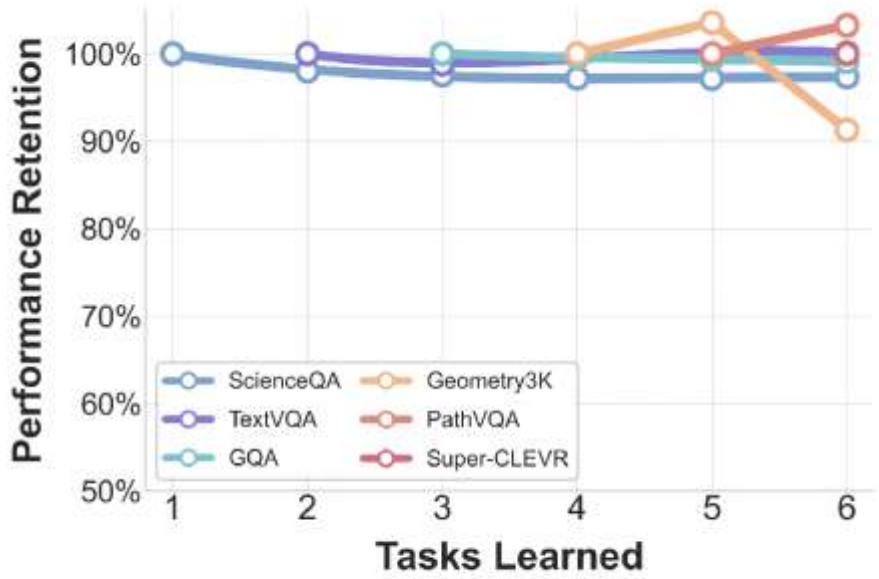
$$L_i^{KR}(\theta) = (\log P_{\pi_\theta}(x_i) - \log P_{\pi_{t-1}}(x_i))^2$$

CRL for LLM

Reinforcement Fine-Tuning Naturally Mitigates Forgetting in Continual Post-Training



(a) Supervised Fine-tuning



(b) Reinforcement Fine-tuning