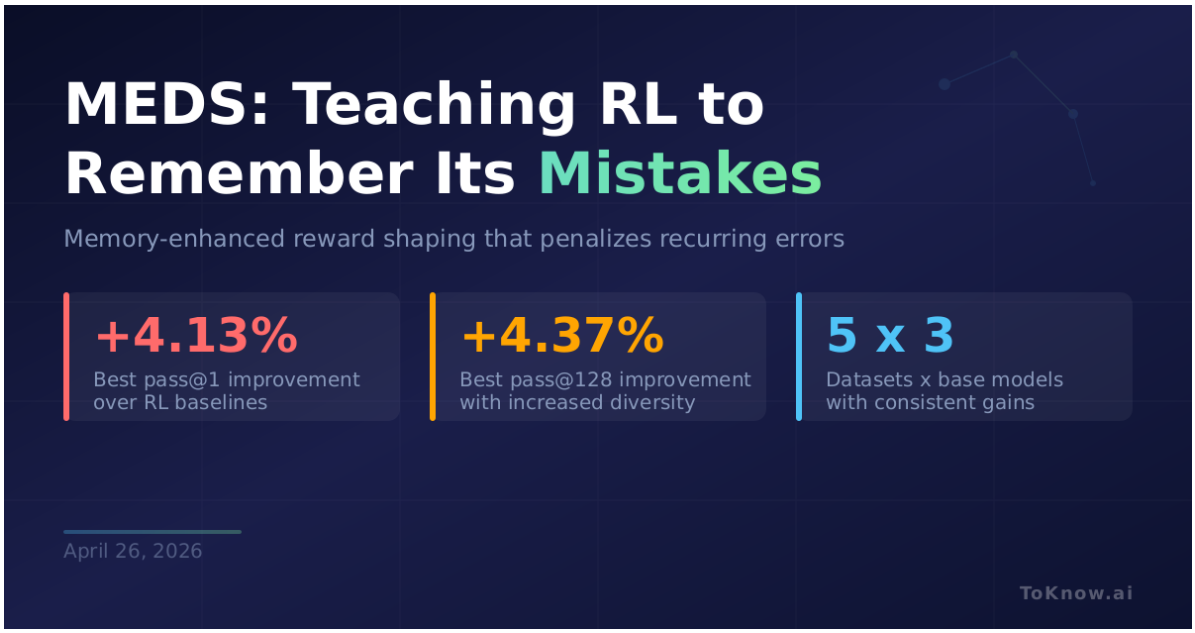
MEDS: Teaching RL to Remember Its Mistakes Instead of Repeating Them

Kabui, Charles

2026-04-26

Read at ToKnow.ai



Researchers at Fudan University released MEDS (Memory-Enhanced Dynamic Reward Shaping), a framework that gives RL training a memory of past mistakes. Standard reward functions in RL for LLMs are stateless: they score each rollout in isolation, so the model keeps repeating the same errors across training iterations without realizing it has seen them before.

MEDS fixes this by storing layer-wise logits from the model's forward pass as lightweight "reasoning fingerprints," then using HDBSCAN density-based clustering to group recurring error patterns. Rollouts that fall into more common error clusters get penalized more heavily, pushing the model toward new reasoning paths instead of revisiting known dead ends. Across five math reasoning benchmarks and three base models (Qwen2.5 and Qwen3 variants), MEDS improves pass@1 by up to +4.13% and pass@128 by +4.37% over baselines, while measurably increasing the diversity of sampled solutions.

MEDS adds no new models or human annotations. The reasoning fingerprints come from logits the forward pass already computes, and the clustering step is lightweight. For teams running GRPO-style RL training, this plugs directly into a veRL-based pipeline, making the reward function track which failures keep recurring and penalize them proportionally. The diversity gain matters as much as accuracy: diverse sampling is how pass@k finds better solutions at inference time.

This fits a broader pattern. Between FIPO's per-token credit assignment and RAGEN-2's mutual information diagnostics, the field is converging on a theme: stateless reward functions waste compute by letting models rediscover the same failures. Making rewards remember, even crudely, produces outsized gains.

Sources:

- MEDS: Memory-Enhanced Dynamic Reward Shaping (arXiv:2604.11297)
- MEDS GitHub Repository
- veRL: Reinforcement Learning Framework for LLMs
- HDBSCAN: Hierarchical Density-Based Clustering
