

Continual Learning for RLVR

2026-03-05

See slide decks of some new ideas in this space:

- [overview of ideas](#)
- [continual rehearsal](#)

Important disctions:

- cross-domain: i.e., train on math, then coding, then science, ...
- within-domain, distribution-shift continual learning (i.e., new samples presented, from same domain/distribution), not cross-domain... easier and doable!

Methods comparison

Method	Training	Merging	Forgetting fix
Sequential RLVR (baseline)	Sequential	N/A	None (the problem)
GLM-5 cross-stage distillation	Sequential	Sequential (recovery stage)	Retroactive repair
Multi-objective RLVR	Simultaneous	N/A (one model)	Avoided by design
Feedback prediction rehearsal	Sequential	N/A (one model)	Concurrent maintenance loss
Teacher-based regularization	Sequential	N/A (one model)	Concurrent regularization
Population RLVR	Parallel	One-shot merge	Avoided by design
Mixture of Distillations	Parallel	Routed distillation	Avoided by design

The parallel methods (population, MoD) never forget because specialists never train on each other's tasks. The simultaneous method (multi-objective) never forgets because all tasks are optimized together. The sequential methods (rehearsal, teacher reg) accept sequential training but add a maintenance signal to slow forgetting. GLM-5 accepts forgetting and repairs it after the fact.