RLSD: Combining Verifiable Rewards With Self-Distillation for Stable Reasoning Training

Kabui, Charles

2026-04-19

Read at ToKnow.ai

Researchers from CASIA, Microsoft, and Shanghai AI Lab released RLSD (RLVR with Self-Distillation), a method that mixes two competing approaches for teaching language models to reason. The first, Reinforcement Learning with Verifiable Rewards (RLVR), only checks whether the final answer passes a verifier, then gives every token in the trajectory the same credit or blame. Reliable, but coarse. The second, on-policy self-distillation (OPSD), uses

the same model as both teacher and student, with the teacher fed privileged information like the reference answer. It produces dense per-token signals but, as the paper shows, leaks the answer into the gradient and destabilises long-term training. RLSD splits the job: the verifier sets the *direction* of every update, while the stop-gradient, clipped teacher signal only scales the *magnitude* per token. The reported outcome is a higher convergence ceiling and more stable runs than either method alone.

For teams training reasoning models, this is a drop-in fix for a real problem. Pure RLVR wastes signal because correct and incorrect chains get blanket credit, so training is slow. Pure self-distillation speeds it up but quietly overfits to the teacher's hints, and the model collapses later. RLSD plugs into existing GRPO-style pipelines without a second model and without changing the reward source.

The wider shift is that reasoning-training is moving away from "pick one signal and scale it" toward decomposing what each signal is good for. Rewards tell you *whether* you were right; teacher policies tell you *how much* each token mattered.

Read More: Qwen's FIPO attacks the same coarse-credit problem from a different angle.

Sources:

- Self-Distilled RLVR (arXiv:2604.03128)
- Hugging Face Daily Papers discussion
- DeepSeekMath: introduction of GRPO and RLVR
- Related work: HDPO (Hybrid Distillation Policy Optimization)
