Reasoning Fine-Tuning Can Generalize Across Domains, but Safety Pays the Price

Kabui, Charles

2026-04-22

Read at ToKnow.ai

Researchers at AI45Research tested a common belief in LLM training: that supervised fine-tuning (SFT, training on curated examples) just memorizes while reinforcement learning (RL) generalizes. Using Qwen3 and InternLM2.5 models trained on math reasoning with long chain-of-thought traces (step-by-step reasoning paths), they found SFT can generalize across domains, but only when three conditions align. First, cross-domain performance follows a

dip-and-recovery pattern: it drops before climbing back, so early checkpoints systematically underestimate actual generalization. Second, data quality matters sharply. Verified long chain-of-thought traces yield consistent cross-domain gains, while low-quality solutions hurt performance everywhere. Third, model capability is decisive: stronger models (Qwen3-14B) internalize transferable reasoning patterns like backtracking from even a toy arithmetic game, while smaller models (1.7B) just copy surface-level verbosity without learning the underlying procedure.

Teams doing reasoning SFT should evaluate at later checkpoints, not early ones, because the dip in cross-domain performance is temporary. The safety finding is more urgent: fine-tuning on math reasoning data improves coding ability but degrades the model's ability to refuse harmful prompts. Reasoning post-training and safety alignment can't be treated as independent steps. The authors released 45 model checkpoints and 6 datasets, including 44K queries with 32 responses each and full token-level log probabilities, giving other teams the raw material to reproduce every finding.

This reframes the SFT vs. RL debate from "which method generalizes?" to "under what conditions and at what cost?" The answer depends on training duration, data quality, and model scale, not on whether you used SFT or RL. For post-training pipelines, that's a more useful question.

Read more: FIPO: Qwen's Token-Level Credit Fix That Breaks the 4K Reasoning Ceiling

Sources:

- Rethinking Generalization in Reasoning SFT (arXiv)
- GitHub Repository and Training Code
- 45 Models and 6 Datasets (Hugging Face Collection)
- 44K Response Dataset with Log Probabilities
