RL-Trained LLM Agents Can Collapse Into Fixed Templates That Entropy Can't Detect

Kabui, Charles

2026-04-20

Read at ToKnow.ai

Researchers at Northwestern University found a hidden failure mode when training multi-turn LLM agents with reinforcement learning (RL, where models improve by trial and error against a reward signal) that every existing diagnostic misses. RAGEN-2 shows that even when entropy (the standard measure of output diversity) looks healthy, models can fall into "template collapse": generating responses that appear varied for any single input but are

actually input-agnostic. The model produces nearly identical reasoning regardless of what it's asked. To detect this, the team decomposes reasoning quality into two axes: within-input diversity (conditional entropy) and cross-input distinguishability (mutual information). Across planning, math, web navigation, and code execution tasks, mutual information correlated with final task performance far more strongly than entropy alone. The project includes Li Fei-Fei and Yejin Choi among its authors.

The fix is straightforward. RAGEN-2 proposes SNR-Aware Filtering: rank training prompts by reward variance and drop the low-signal ones. When the model scores roughly the same no matter what strategy it tries, those prompts produce weak gradients and let regularization erase cross-input differences. Filtering them out improved performance consistently across all four tested domains. Any team running GRPO or PPO-style RL can add this with minimal code changes.

This redefines how RL training should be monitored. Entropy, the standard diagnostic, is blind to an entire class of failure. Teams that evaluate only within-input diversity may be shipping agents that look capable in demos but produce input-agnostic reasoning in practice. For another angle on RL training stability, see our earlier post on self-distilled reasoning training.

Sources:

- RAGEN-2: Reasoning Collapse in Agentic RL (arXiv)
- RAGEN GitHub Repository
- RAGEN-2 Project Page
- HuggingFace Daily Papers: RAGEN-2
