Self-Distillation Can Hurt LLM Reasoning by Silencing Useful Doubt

Kabui, Charles

2026-04-02

Read at ToKnow.ai

Researchers at Microsoft Research, KAIST, and Seoul National University found that self-distillation, a popular technique where a model trains on its own successful outputs, can degrade math reasoning by up to 40%. The culprit is suppression of what the authors call "epistemic verbalization": tokens like "wait," "hmm," and "maybe" that signal the model is uncertain about a step. When the teacher copy of the model sees the correct solution, it generates confident, concise traces with almost no uncertainty markers, dropping from an average of 182 epistemic tokens per response to just 9. The student learns to imitate that confident style, but at inference time it doesn't have the answer key. Across three model families, Qwen3-8B, DeepSeek-R1-Distill-Qwen-7B, and OLMo3-7B-Instruct, this leads to significant accuracy drops on out-of-distribution math problems, with AIME24 scores falling roughly 40% on DeepSeek and 15% on AMC23.

The finding matters because self-distillation is widely used to make reasoning models cheaper to run by shortening their outputs. If the shorter outputs come at the cost of silencing the model's self-correction mechanism, teams deploying distilled reasoning models may be getting faster answers that are quietly less reliable on novel problems. The code and training logs are fully open, so practitioners can check whether their own pipelines exhibit the same pattern.

This connects to a broader emerging principle: teaching models the right process, including moments of productive uncertainty, beats teaching them polished answers. A recent Google study on Bayesian teaching found the same thing from a different angle: LLMs trained on a Bayesian model's early wrong guesses generalized better than those trained on correct answers alone.

Sources:

- Why Does Self-Distillation (Sometimes) Degrade the Reasoning Capability of LLMs? (arXiv)
- Project Blog Post
- GitHub: self-distillation-analysis
- HuggingFace Paper Page
