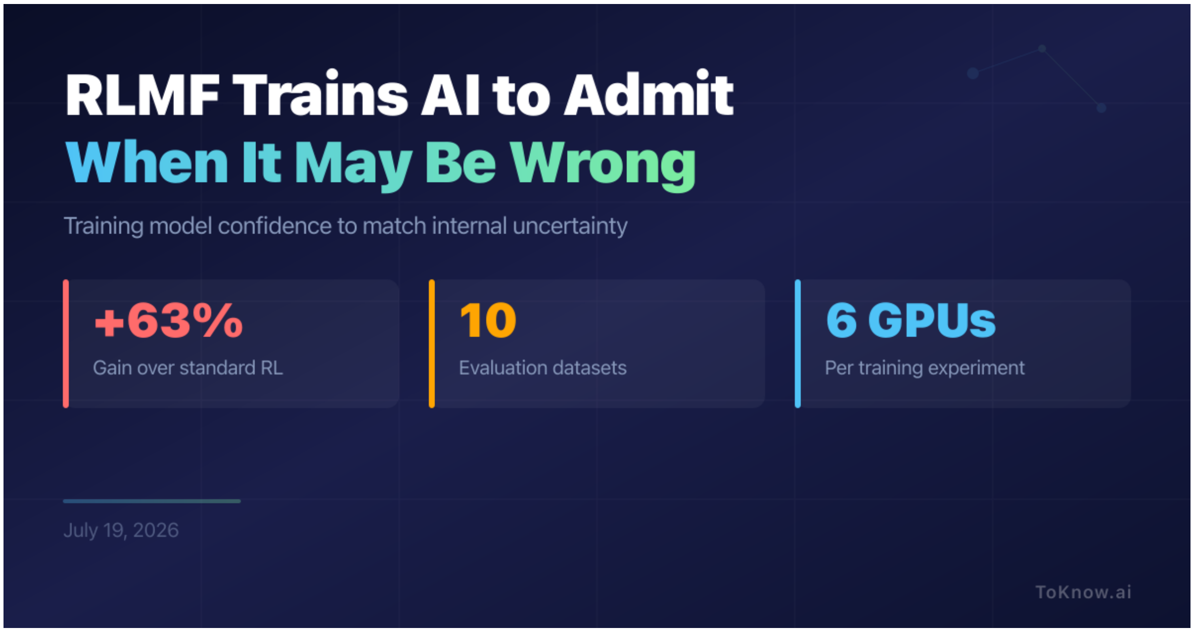
RLMF Trains AI to Admit When It May Be Wrong

Kabui, Charles

2026-07-19

Read at ToKnow.ai



Yale and Google researchers introduced Reinforcement Learning with Metacognitive Feedback, or RLMF, which trains a model to match its stated confidence to how consistently it can answer. Standard reinforcement learning rewards answer quality. RLMF also rewards accurate self-assessment, then rewrites numerical confidence as natural wording such as "I am not certain." The team first used supervised fine-tuning to teach a confidence-labelled output format, then applied RLMF to Qwen3 models from 1.7B to 8B parameters and Llama 3.1 8B using 2,000 selected PopQA examples. Tests used 1,000 examples from each of 10 datasets. The paper's corrected Mean Faithfulness Gap score rose from 0.54 to 0.83 for Qwen3 8B and from 0.60 to 0.84 for Llama. Answer accuracy rose from 0.55 to 0.57 and from 0.31 to 0.41, respectively. The reported peak gain over standard reinforcement learning is a **63% relative improvement** on this paper-specific metric for Qwen3 8B.

This could help people decide which AI answers need checking, but it is not a truth detector. The training recipe sampled 32 answers per prompt. It requires at least three GPUs; the reference setup uses **six**, including one for a judge model, one for inference, and four for training. Its stronger data-selection variant sampled 20 answers per item. A human study found 95% to 98% preference for RLMF's uncertainty wording, but it covered only 120 examples rated by three graduate researchers. The MIT-licensed code is public, so independent tests can now check whether the gains survive different judges and domains.

Earlier methods used careful prompting or fine-tuning on uncertainty phrases. RLMF instead puts self-assessment inside the training reward. That is more expensive, but it targets the gap between what a model knows and how certain it sounds.

Sources:

- RLMF paper
- RLMF code and experiments
- MetaFaith paper
- Faithful Uncertainty Tuning paper
