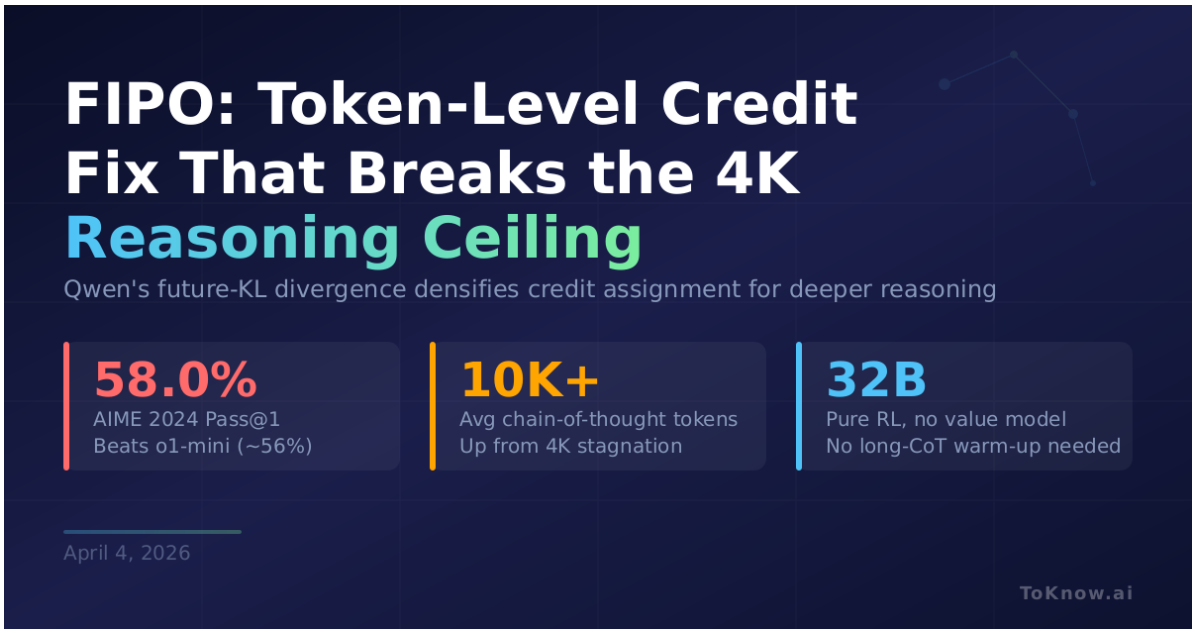
FIPO: Qwen's Token-Level Credit Fix That Breaks the 4K Reasoning Ceiling

Kabui, Charles

2026-04-04

Read at ToKnow.ai



Alibaba's Qwen team released FIPO (Future-KL Influenced Policy Optimization), a reinforcement learning algorithm that fixes a specific bottleneck in how reasoning models are trained. Standard GRPO-style training assigns the same reward signal to every token in a reasoning chain, whether it is a critical logical pivot or filler text. This coarse credit assignment causes chain-of-thought length to stagnate around 4,000 tokens, no matter how hard the problem

is. FIPO replaces that with a dense, per-token advantage: it measures how much each token shifts the model's future behavior using discounted future-KL divergence, then amplifies the reward for tokens that cause major reasoning shifts and dampens it for tokens that don't. Applied to Qwen2.5-32B-Base, a model with no prior long-reasoning training, FIPO pushes average chain-of-thought length past 10,000 tokens and reaches 58.0% Pass@1 on AIME 2024, outperforming both DeepSeek-R1-Zero-Math-32B (~47%) and o1-mini (~56%).

What makes this useful beyond benchmarks: the extra tokens aren't padding. As training progresses, the model develops self-reflection and multi-pass verification, re-deriving answers through alternative methods. Generation length now correlates strongly with accuracy, meaning longer chains genuinely improve answers. FIPO does all this without a separate value model or synthetic long-reasoning warm-up data, keeping the training pipeline simple and the overhead low. The full 32B model checkpoint, training code, and recipes are open-source.

The broader pattern here is clear: the bottleneck in RL-based reasoning isn't model size or data volume, it's credit assignment granularity. FIPO shows that a relatively simple mathematical fix, measuring each token's downstream causal influence, can unlock capabilities that complex critic models and expensive value networks were supposed to provide.

Read more: Nabla-Reasoner: Gradient Descent at Inference Time Makes LLMs Think Harder

Sources:

- FIPO Paper (arXiv)
- FIPO Blog (Qwen Pilot)
- FIPO GitHub Repository
- FIPO 32B Model (Hugging Face)
