PROPOSAL

Speculative MCTS: Lightweight Draft-Guided Tree Search for Efficient LLM Reasoning

1. PROBLEM STATEMENT

Current Challenges

Monte Carlo Tree Search (MCTS) has emerged as a powerful inference-time technique for improving reasoning in Large Language Models. However, existing MCTS implementations face some critical limitations:

- Computational Inefficiency: Standard MCTS requires repeated inference through large models, consuming significant computational resources and inference time.
- Accessibility Gap: MCTS methods are largely limited to research institutions with extensive computational resources.
- Quality-Speed Trade-off: Practitioners must choose between using small, fast models with lower quality or large, expensive models with better quality.
- Inference Latency: Real-time deployment of MCTS-based reasoning systems remains impractical due to high computational overhead.

Research Gap

While speculative decoding has proven effective for accelerating standard autoregressive generation, and MCTS has shown promise for complex reasoning tasks, no prior work systematically combines these two paradigms for tree-based search. This represents a significant opportunity for innovation in efficient reasoning at scale.

2. PROPOSED SOLUTION

Speculative MCTS

We propose Speculative MCTS, a novel framework that combines speculative decoding principles with Monte Carlo Tree Search to achieve:

- 2.5-3x faster inference compared to vanilla MCTS
- Maintained reasoning quality (minimal to no accuracy degradation)
- Universal applicability across different model families and sizes

Key Insight

Use a lightweight draft model(~ **1-3 B params** , **Qwen-2.5-1.5B-Instruct)** for rapid tree exploration, then selectively verify promising reasoning paths using a larger target model(**Llama-3.1-8B-Instruct**) only when needed. This creates an efficient two-tier verification system that maintains quality while dramatically reducing computation.