

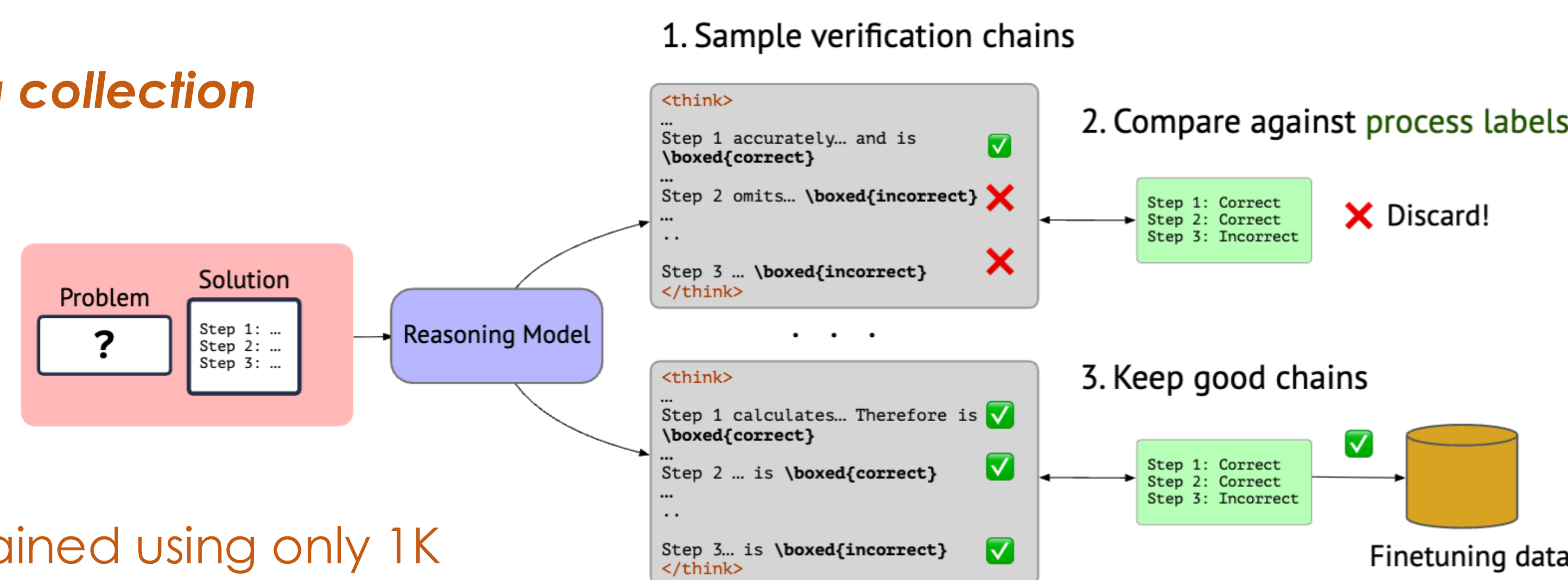
A generative PRM that outperforms baselines with ~1% process labels



Core Idea & Model Training

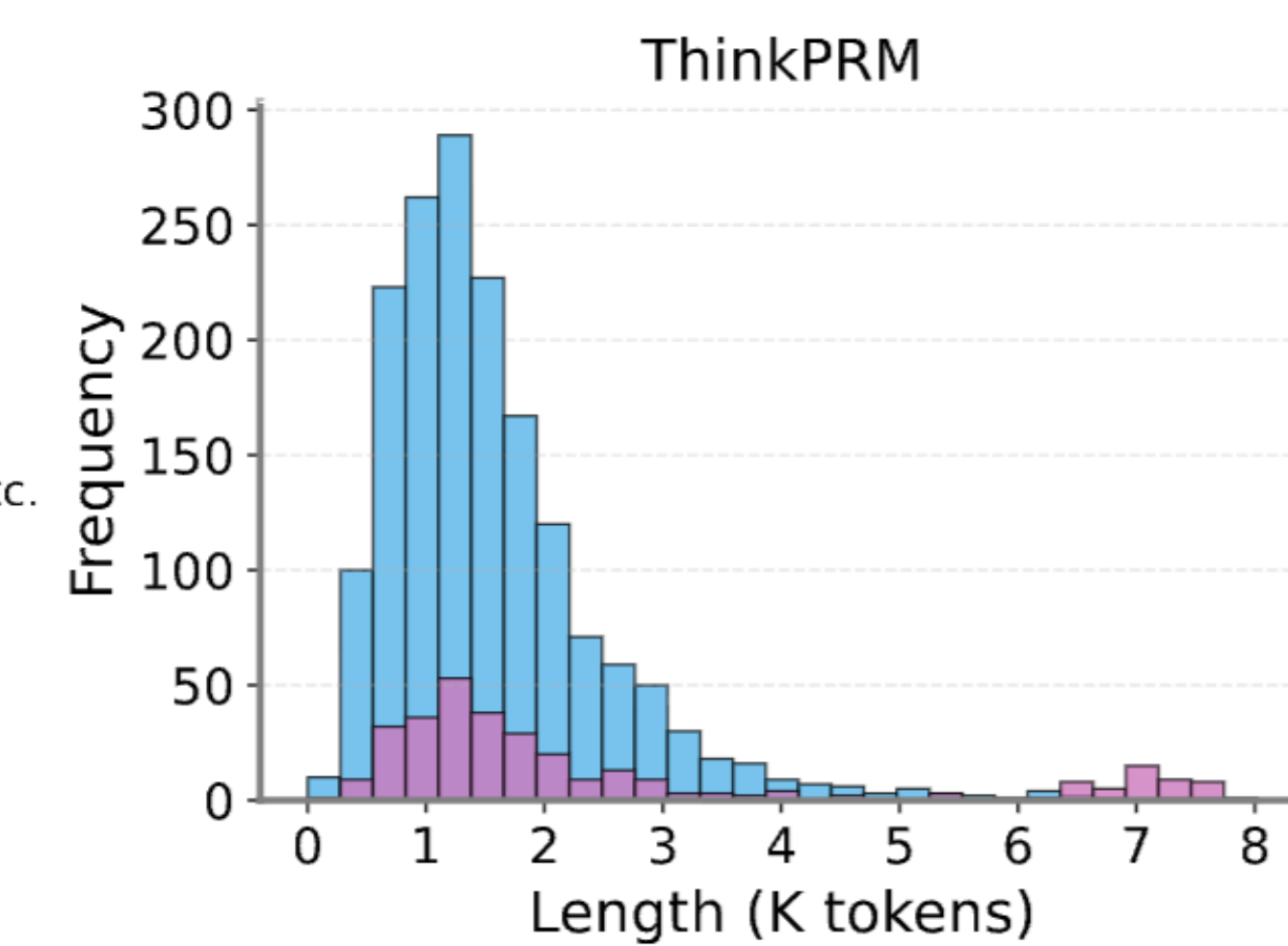
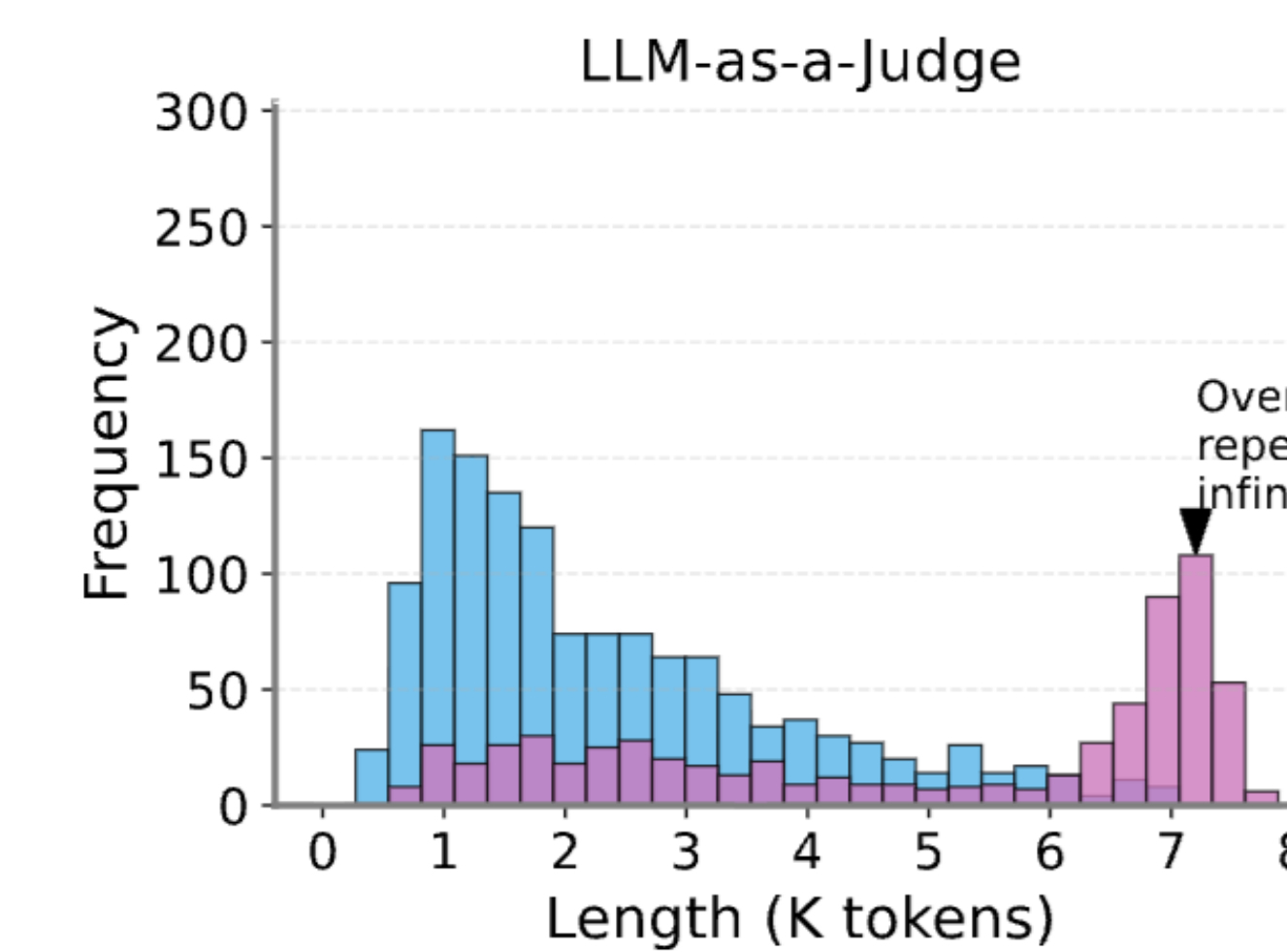
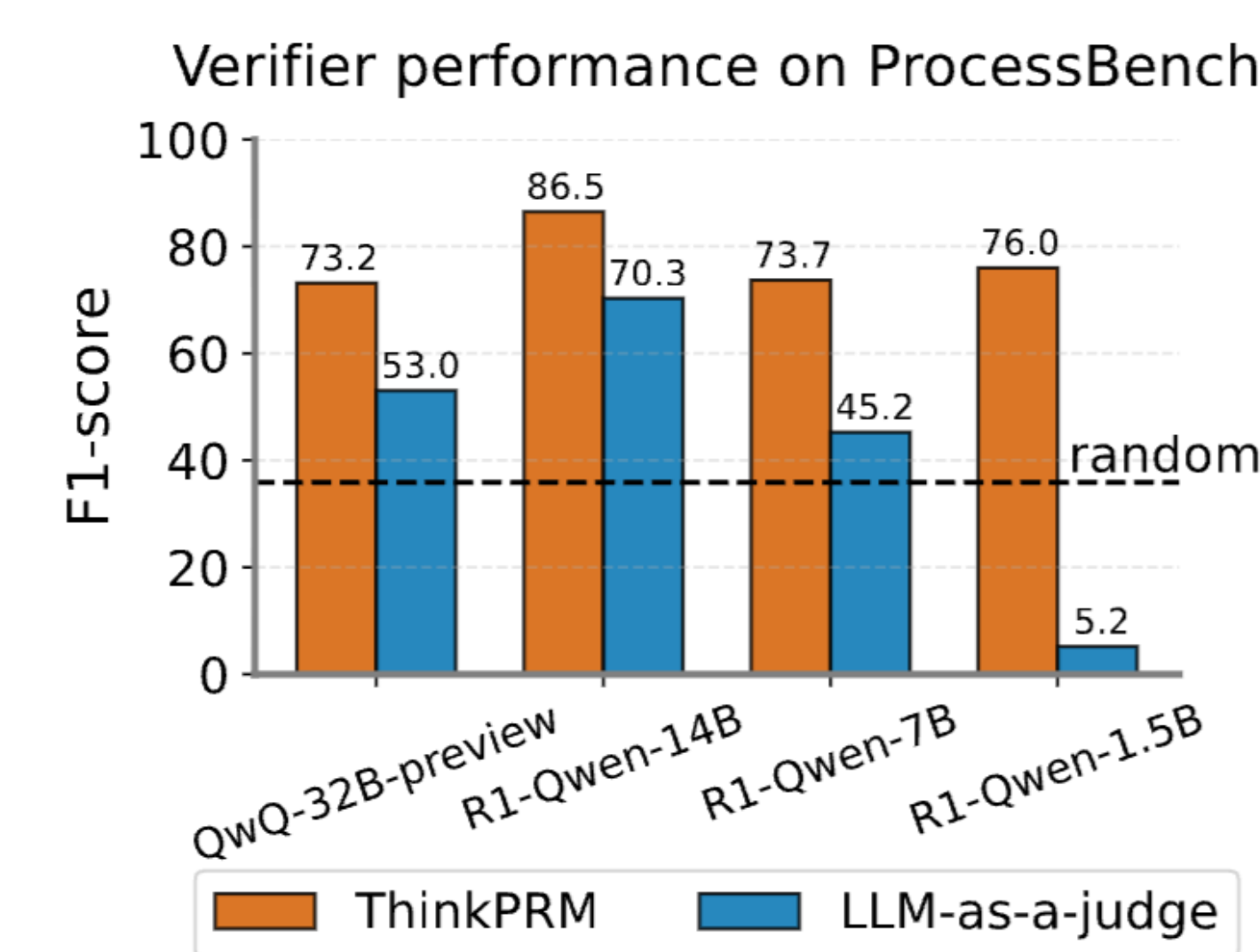
- Test-time scaling needs a strong **process reward model (PRM)**, but training standard PRMs requires **expensive step-level supervision**.
- ThinkPRM**: a generative long-CoT process verifier that thinks through verification.
- Training**: Fine-tune on **high-quality verification CoTs** filtered to match gold step labels

Synthetic data collection



ThinkPRM is trained using only 1K CoTs including ~8K process labels!

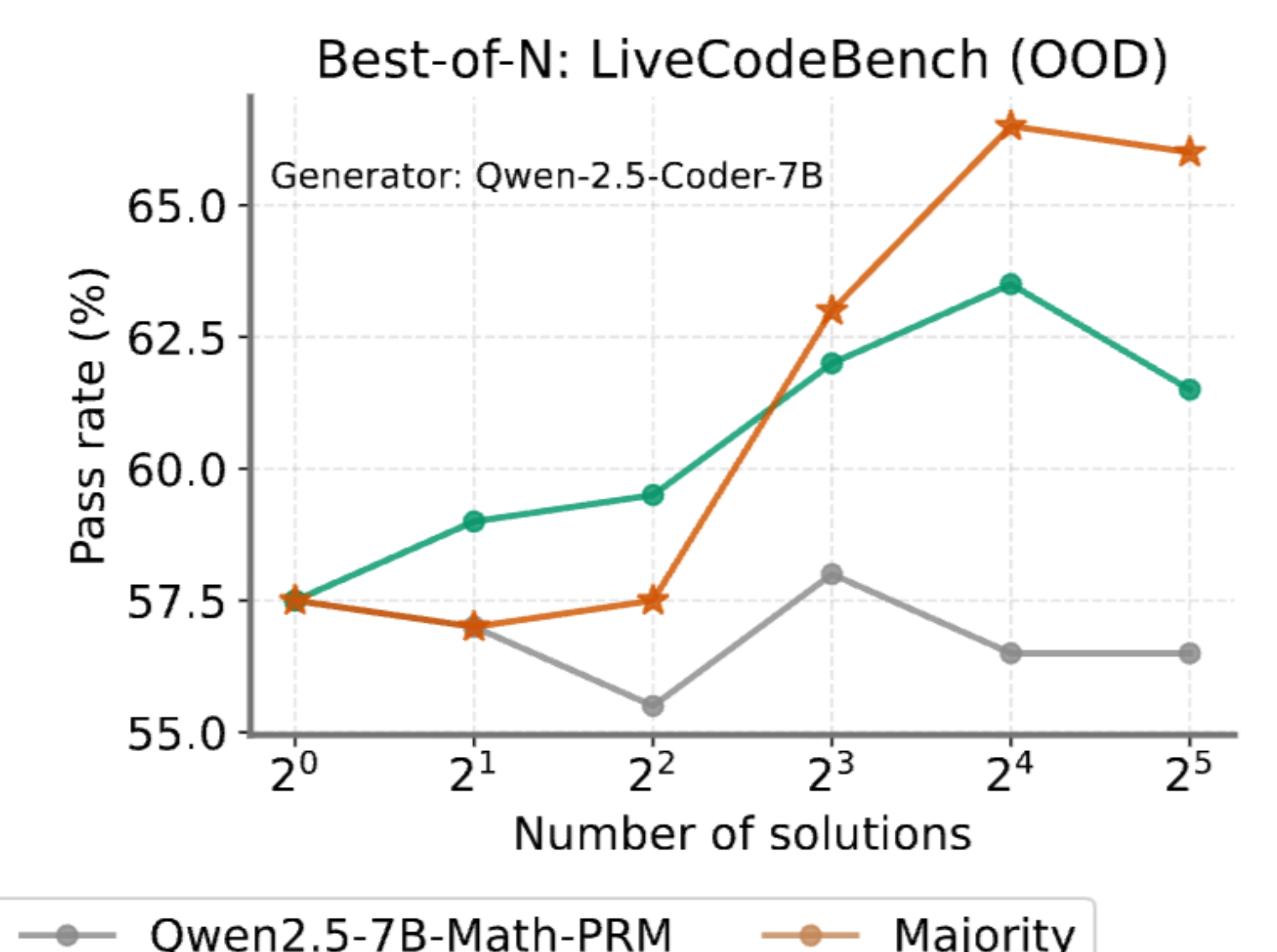
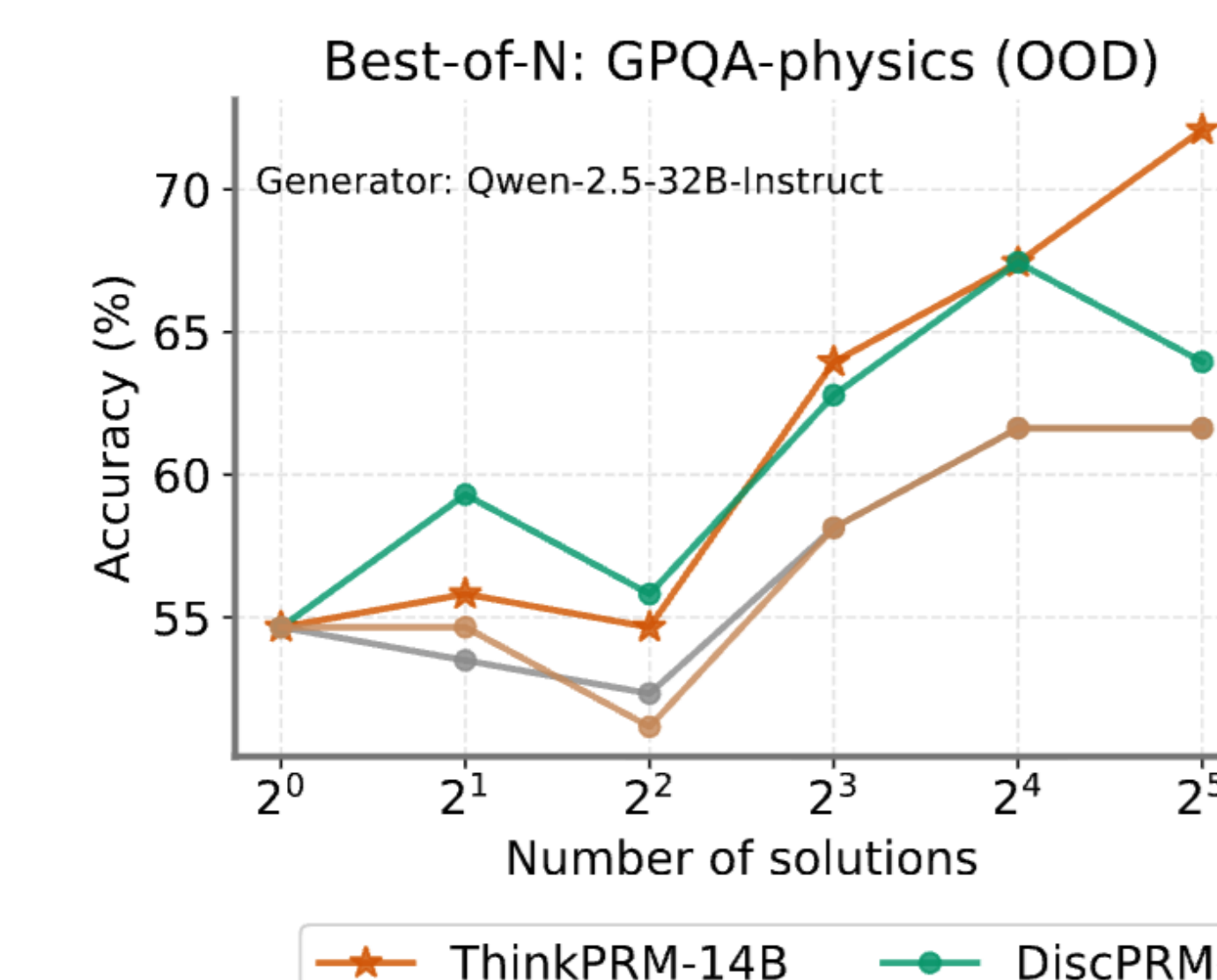
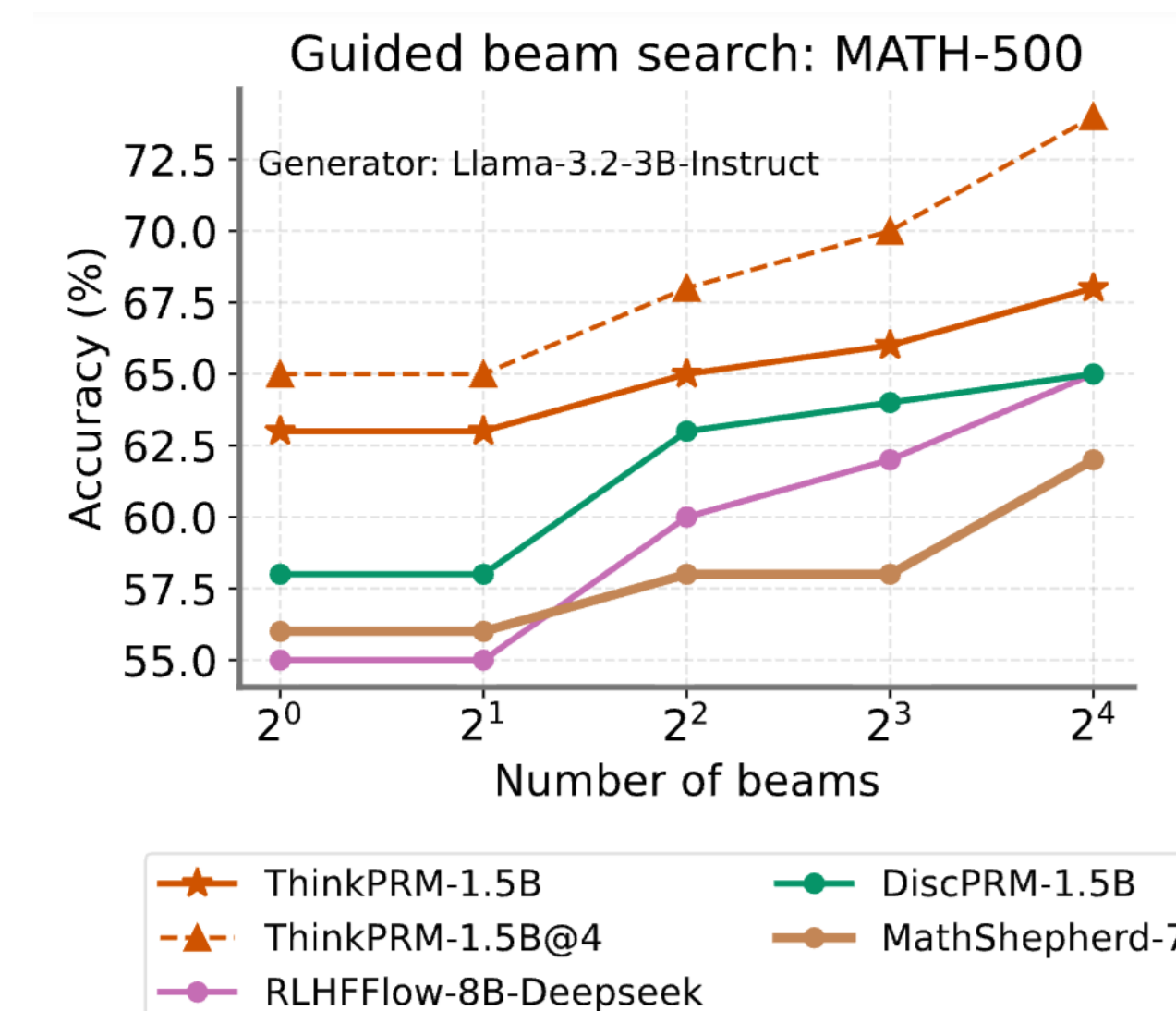
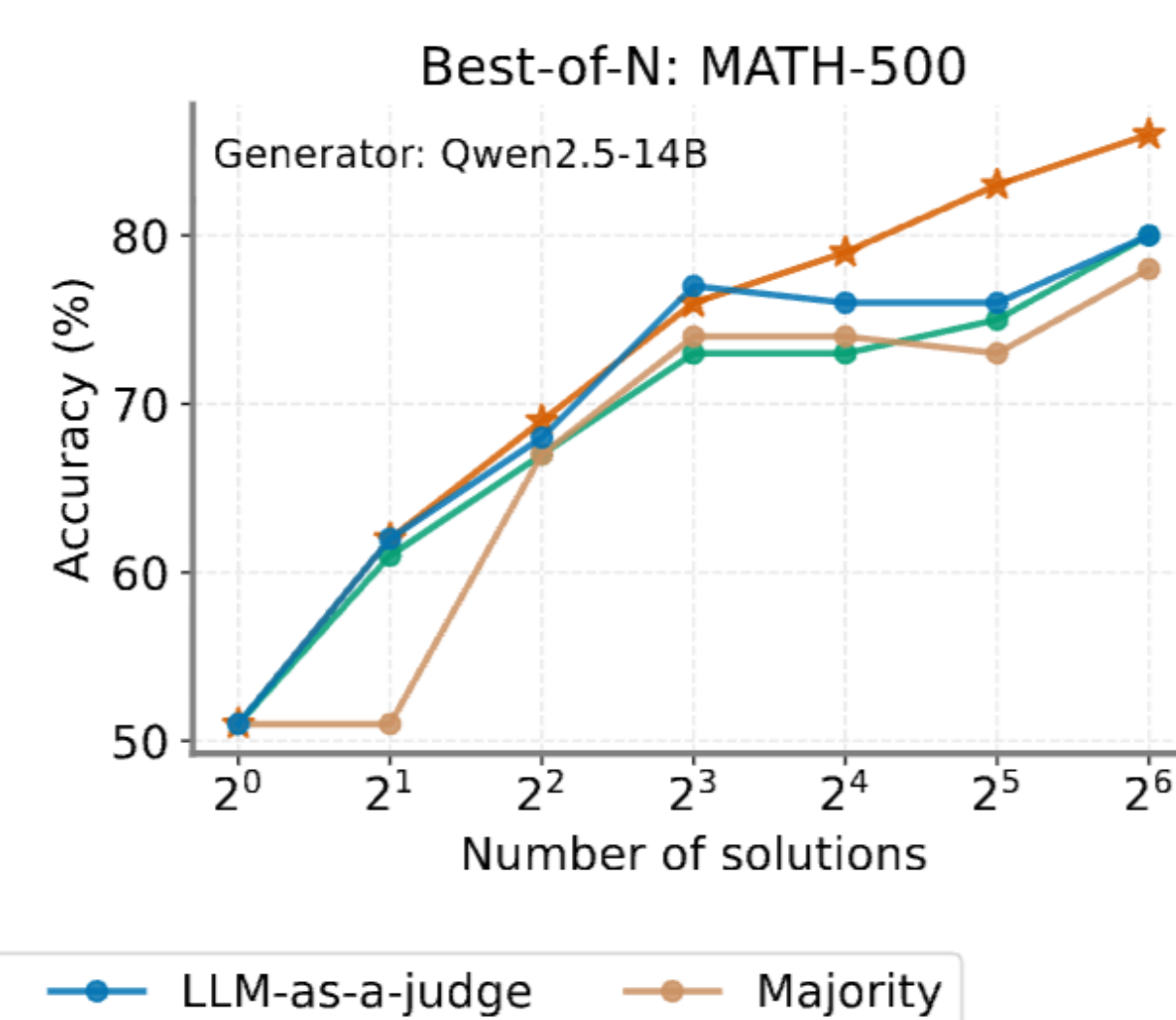
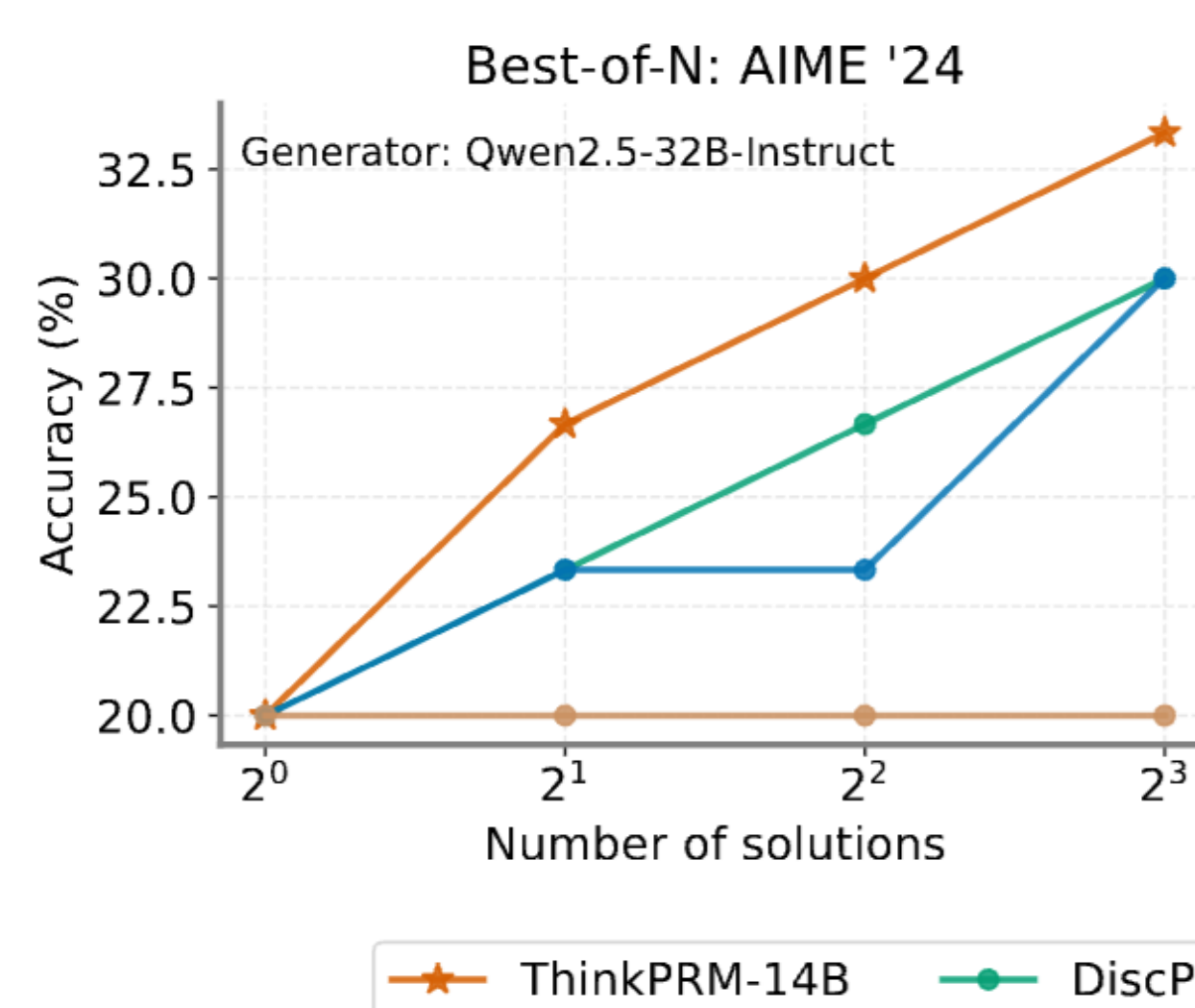
Comparison to LLM-as-a-Judge



Improves substantially on zero-shot LLM-as-a-Judge

Reduces repetition, looping, getting stuck issues that happen with LLM-as-a-Judge

Test-time scaling with ThinkPRM

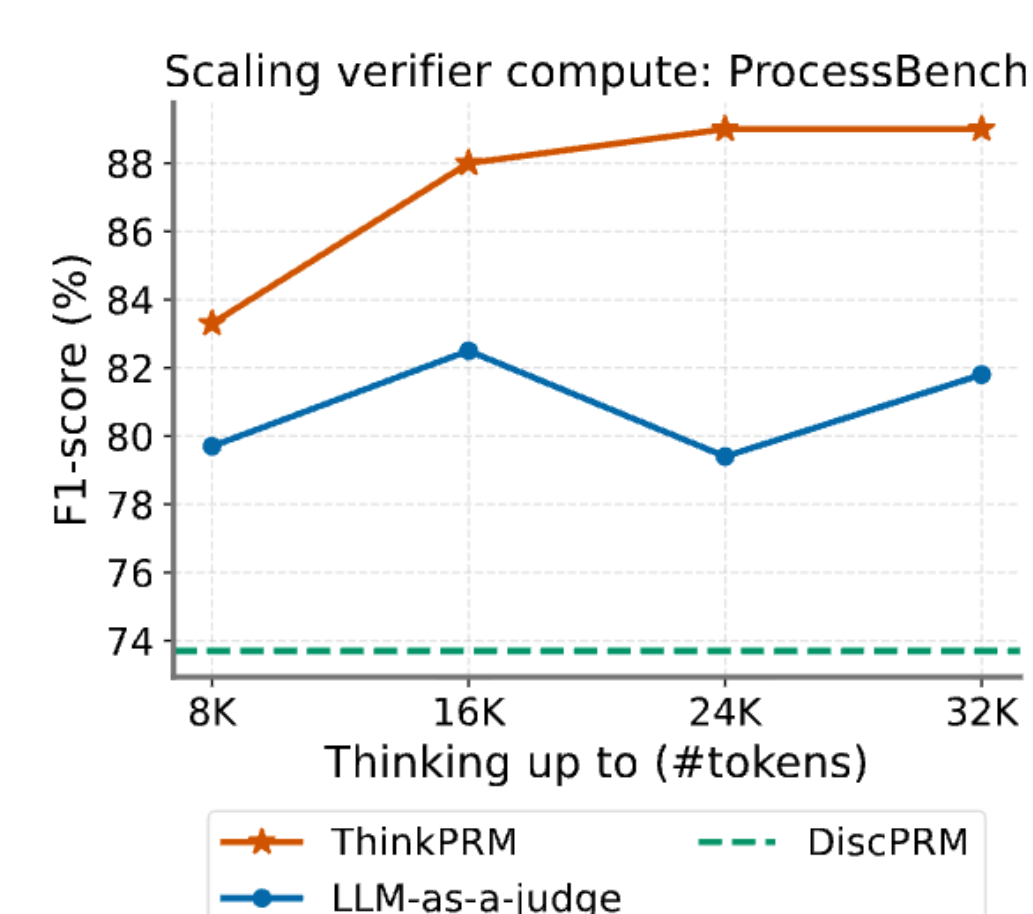


Scales compute better under best-of-N

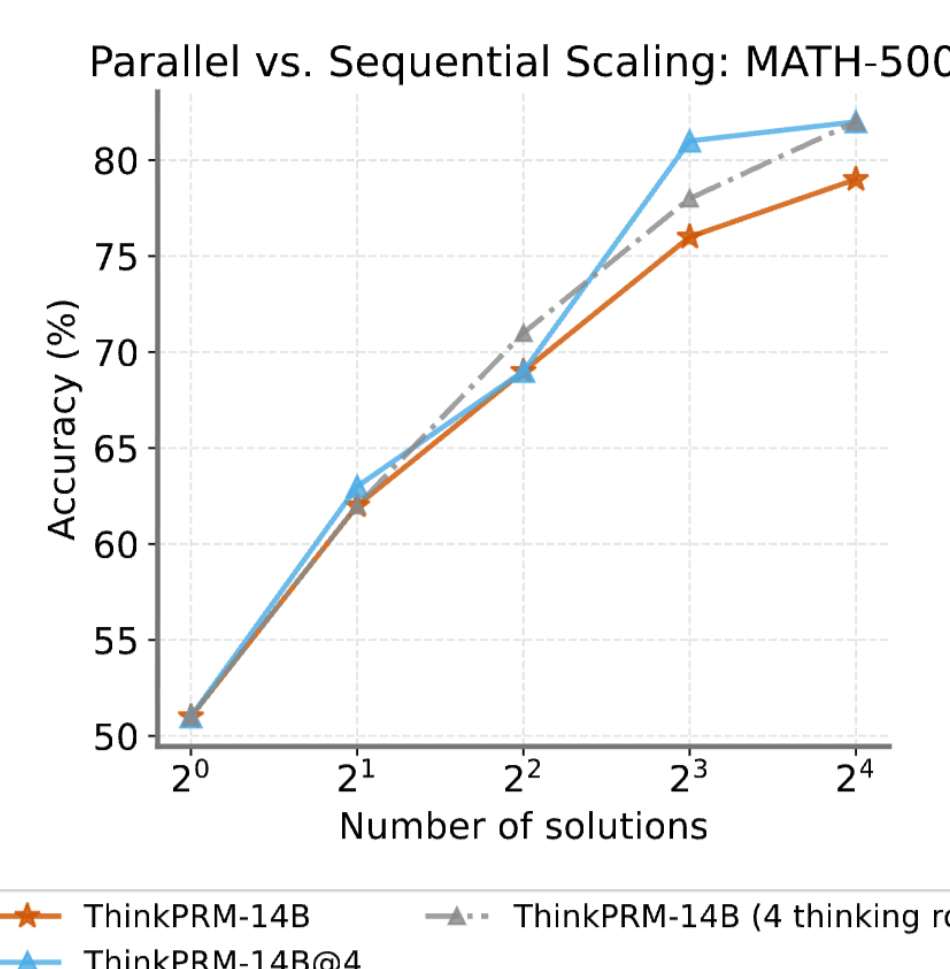
Outperforms off-the-shelf PRMs under beam search

Generalizes to OOD tasks

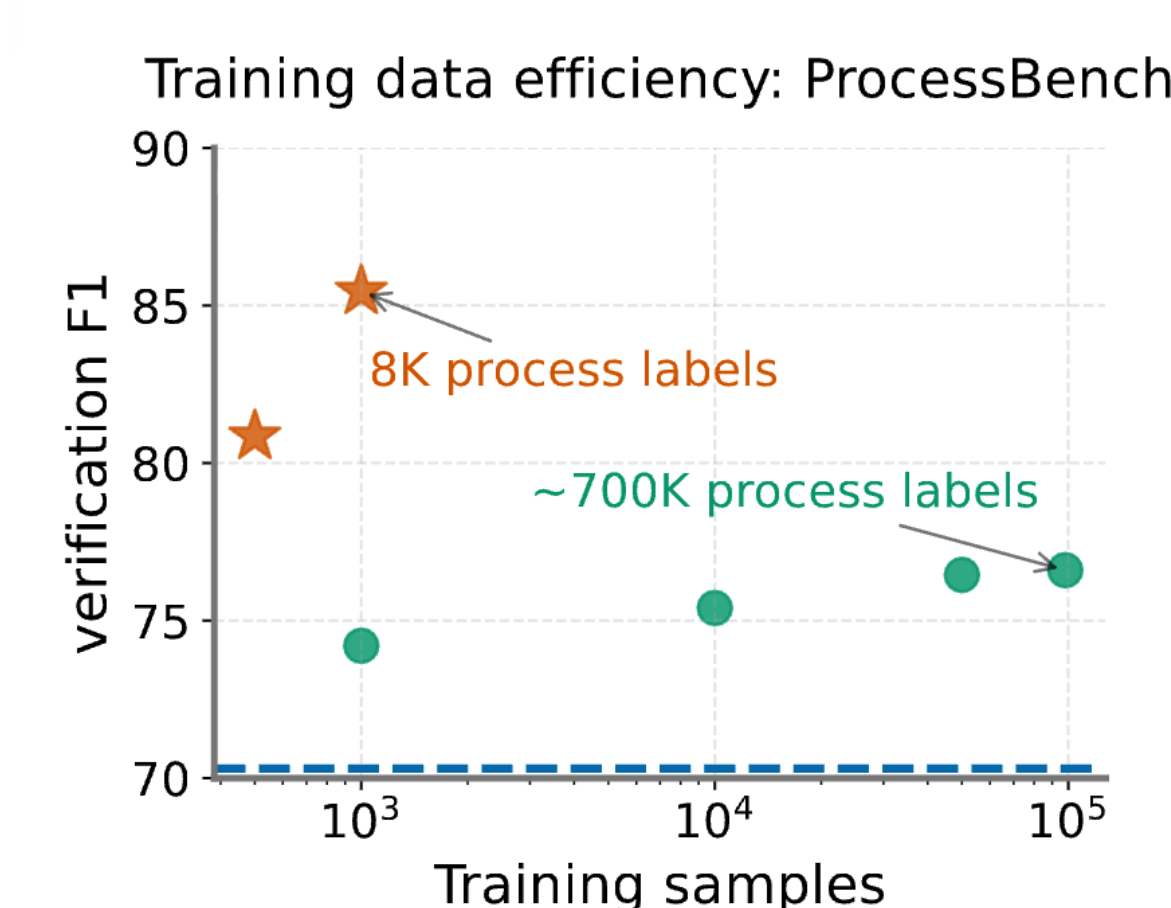
Extra Analysis



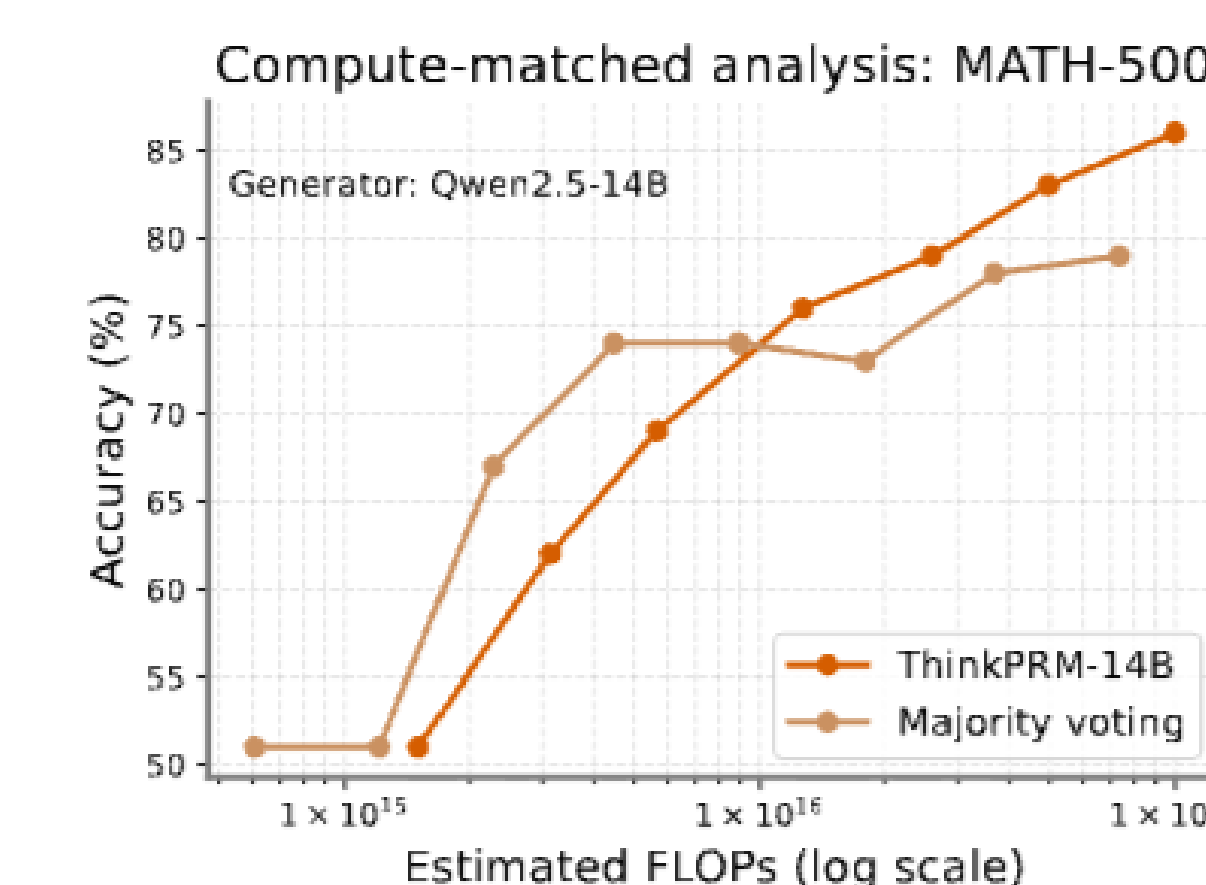
Thinking longer improves verification



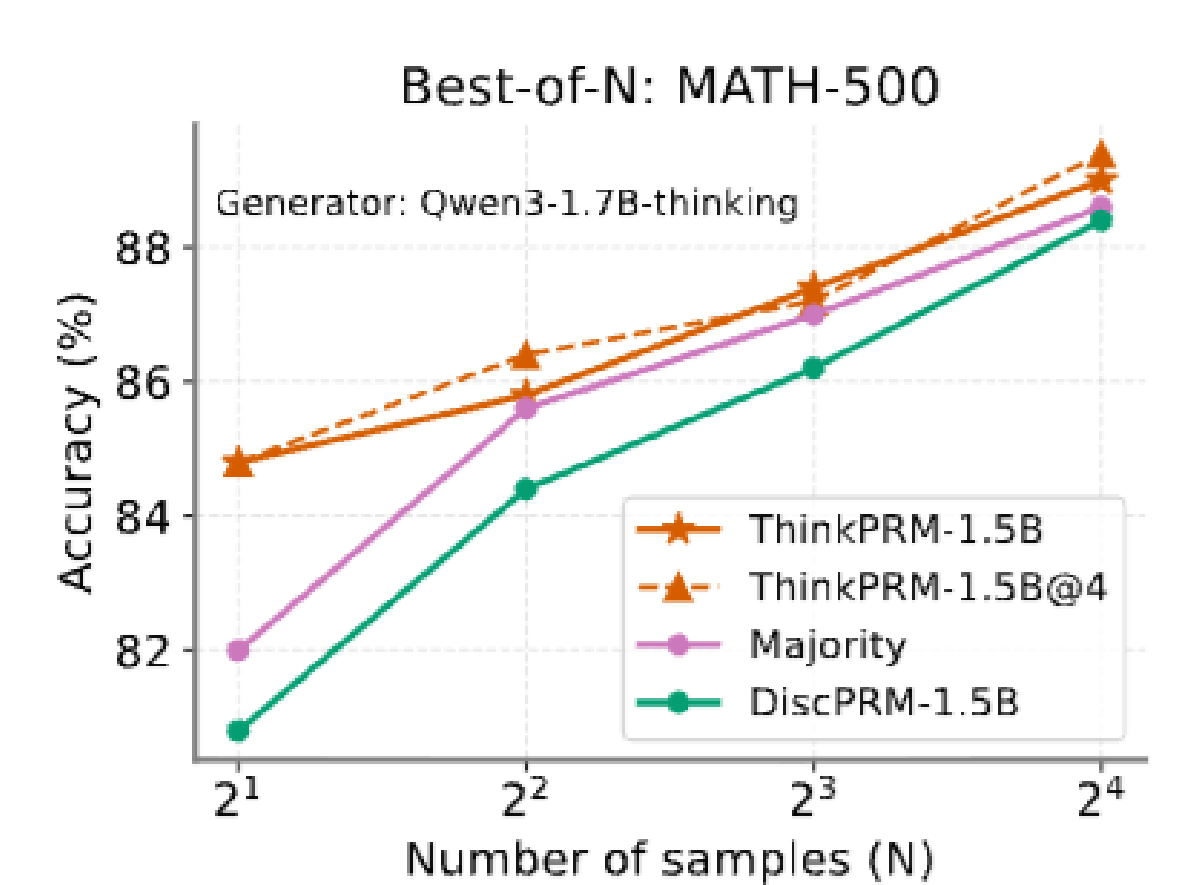
Scales compute both in parallel and sequentially



Outperforms Discriminative PRM with ~1% labels



Scales compute better than self-consistency



Can also verify long CoTs even though trained on short CoT