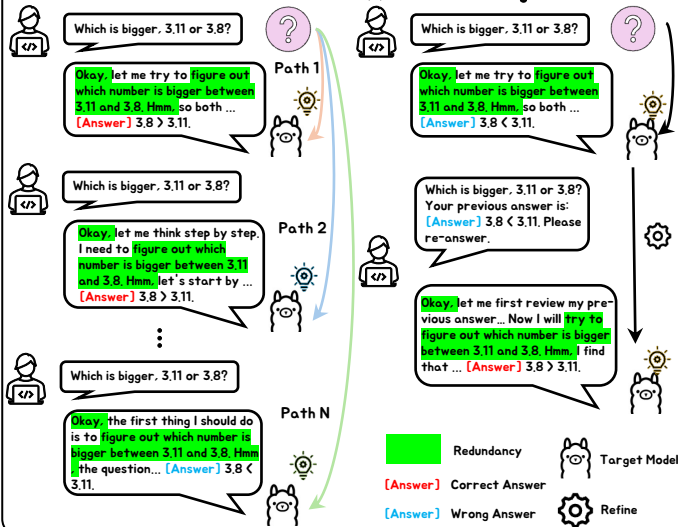


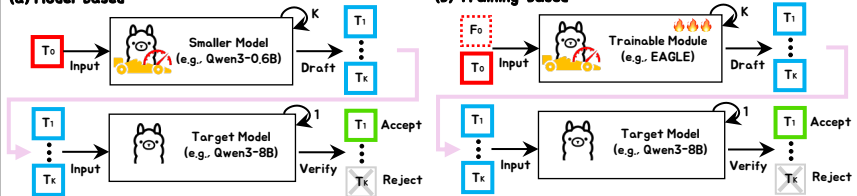
#1 Redundancy in Test-time Scaling

(a) Best-of-N



#2 Speculative Decoding

(a) Model-based



(c) N-gram-based

