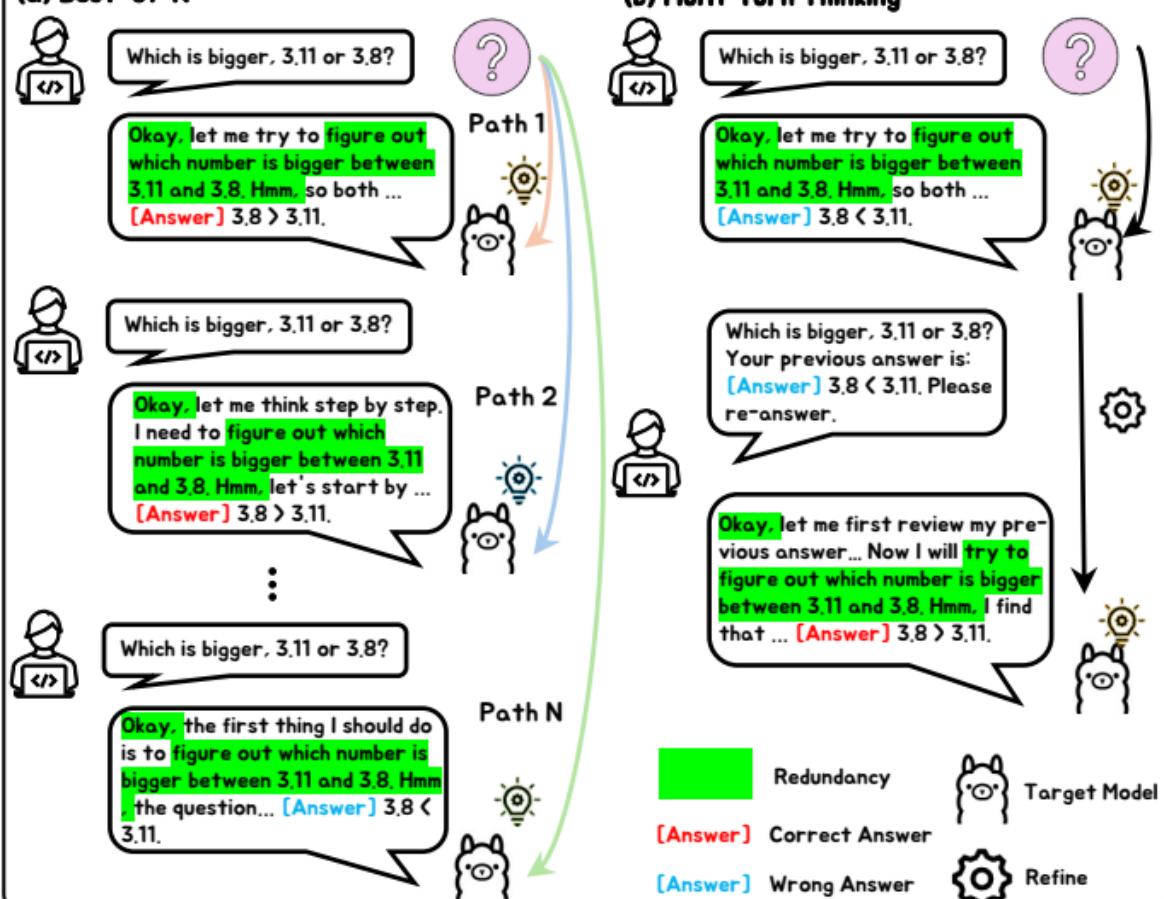
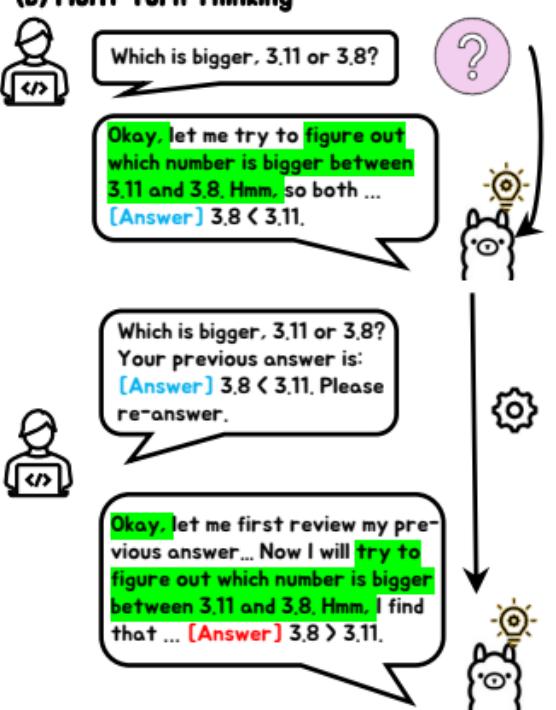


#1 Redundancy in Test-time Scaling

(a) Best-of-N

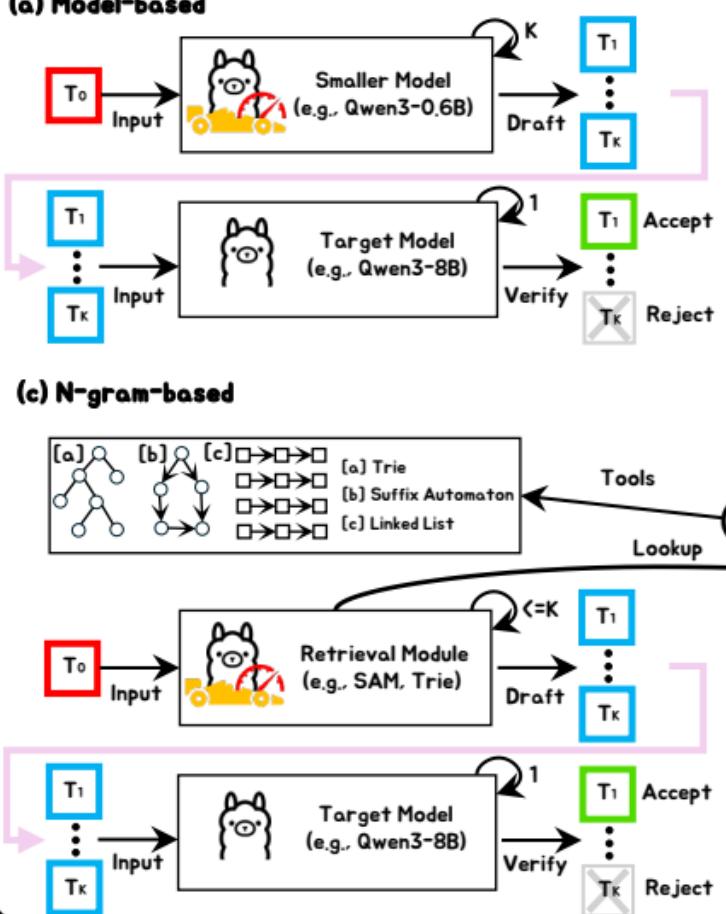


(b) Multi-turn Thinking

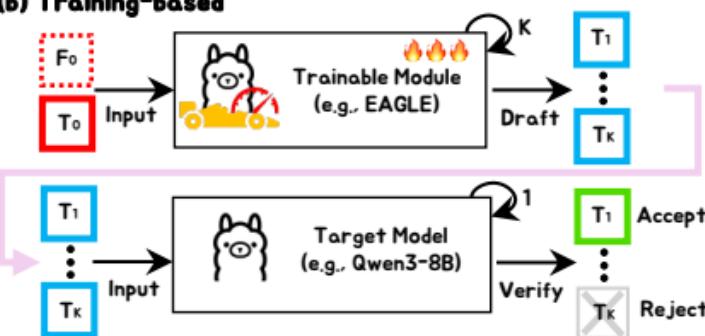


#2 Speculative Decoding

(a) Model-based



(b) Training-based



(c) N-gram-based

