

Unified Collaboration $p_f = f(p_l, p_s)$

Speculative Decoding $p_f = p_l$

Contrastive Decoding $p_f = \log p_l - \log p_s$

Proxy Tuning $p_f = p_l + \Delta p_s$...

Q1: What is the essential difference between large and small models from the perspective of decoding?

Q2: Why do small models underperform compared to large models?