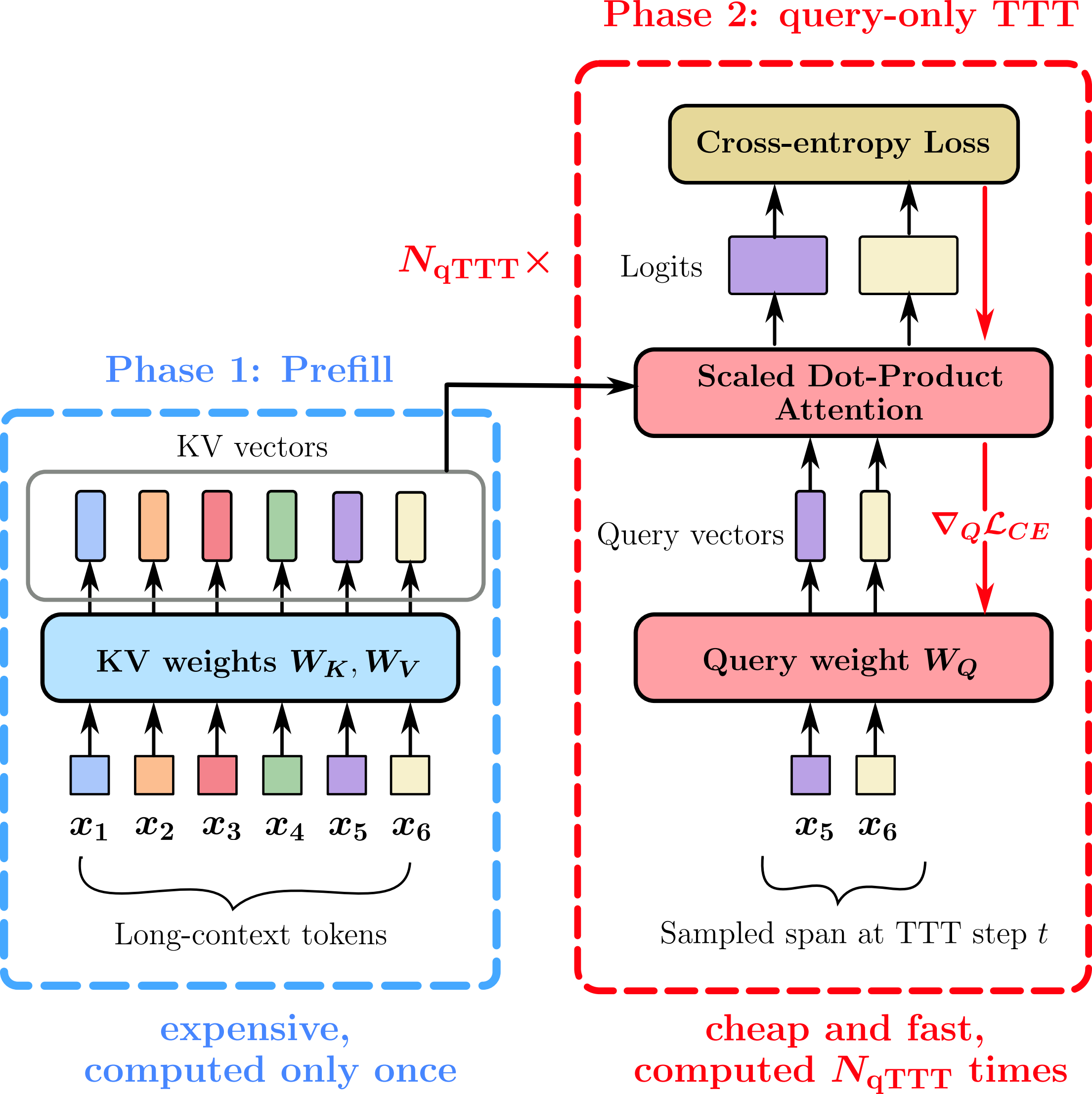
Phase 2: query-only TTT
Cross-entropy Loss
$N_{\text{qTTT}}\times$
Logits
Phase 1: Prefill
Scaled Dot-Product Attention
KV vectors
Query vectors
$\nabla_Q \mathcal{L}_{CE}$
KV weights $\boldsymbol{W}_K, \boldsymbol{W}_V$
Query weight $\boldsymbol{W}_Q$
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$ $\boldsymbol{x}_4$ $\boldsymbol{x}_5$ $\boldsymbol{x}_6$
$\boldsymbol{x}_5$ $\boldsymbol{x}_6$
Long-context tokens
Sampled span at TTT step t
expensive, computed only once
cheap and fast, computed N_{qTTT} times