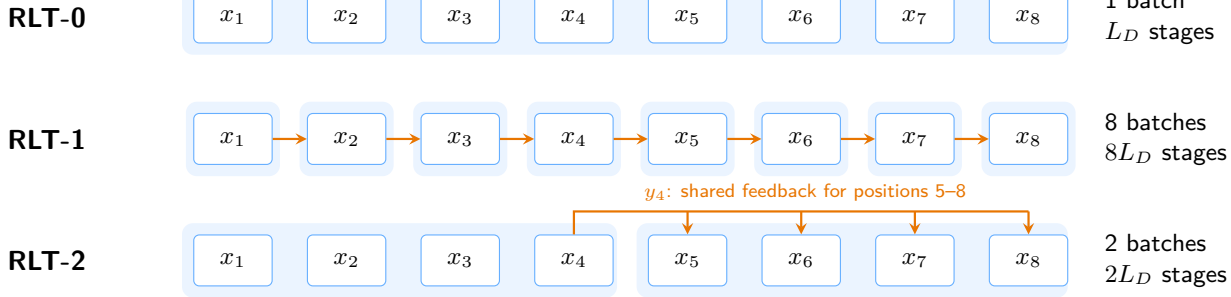


Known-token decoder batches



Within a batch: positions in parallel at each layer. Between batches: preserve causal SWA caches.