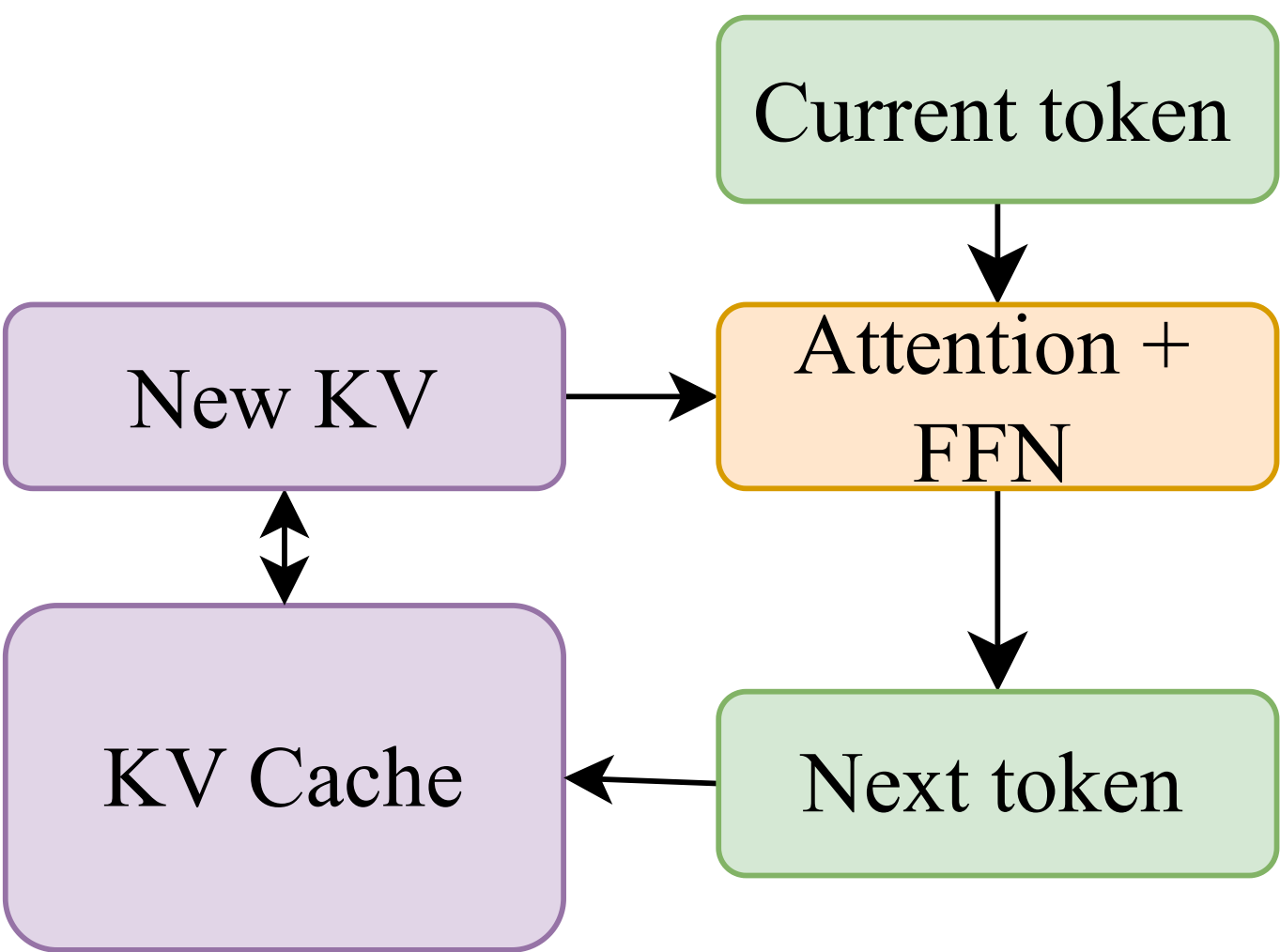
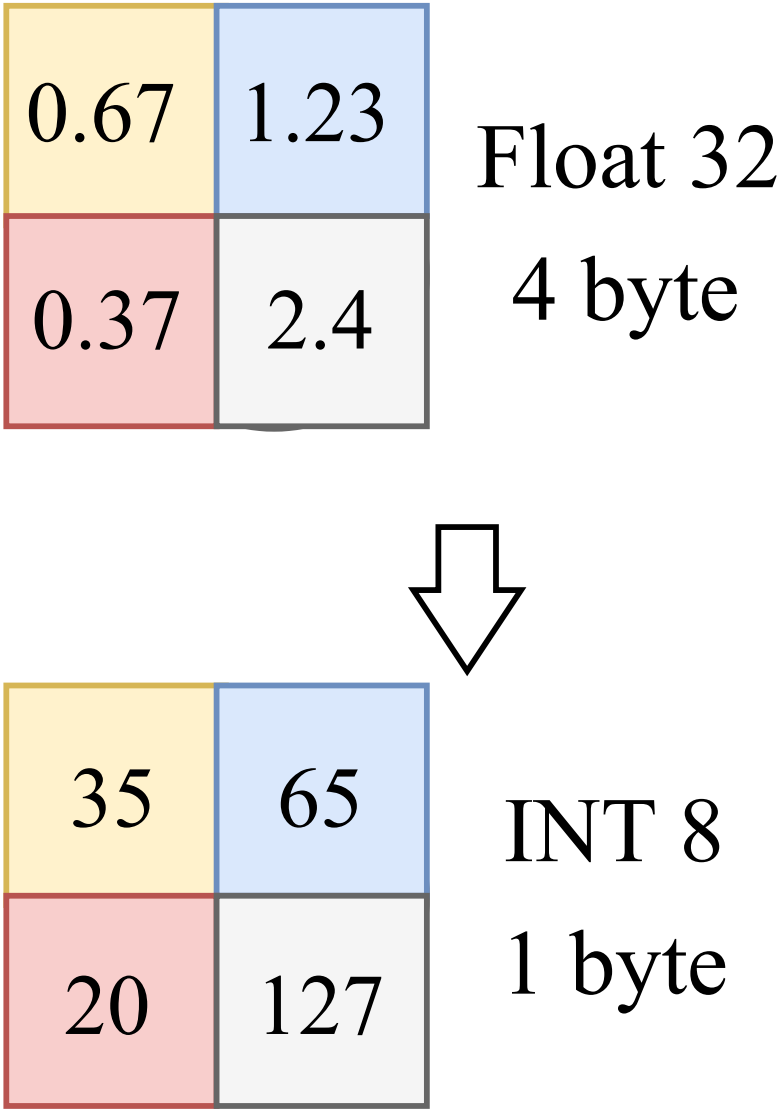


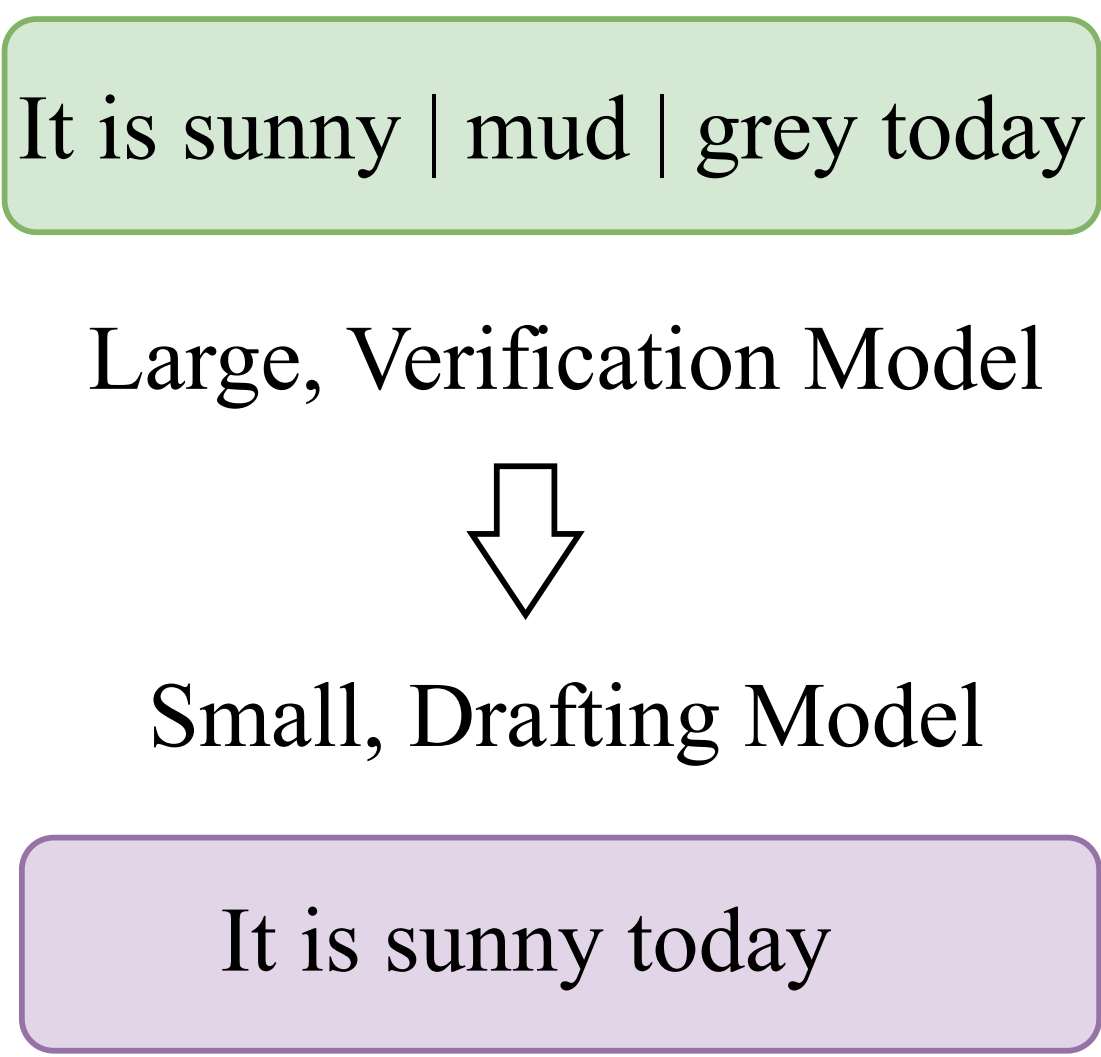
Cache Reuse



Quantization



Speculative Decoding

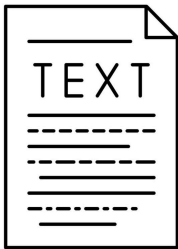
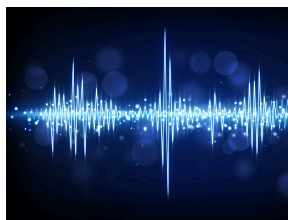


Token Compression



Encode Frozen weights with compressed tokens

Inputs



...