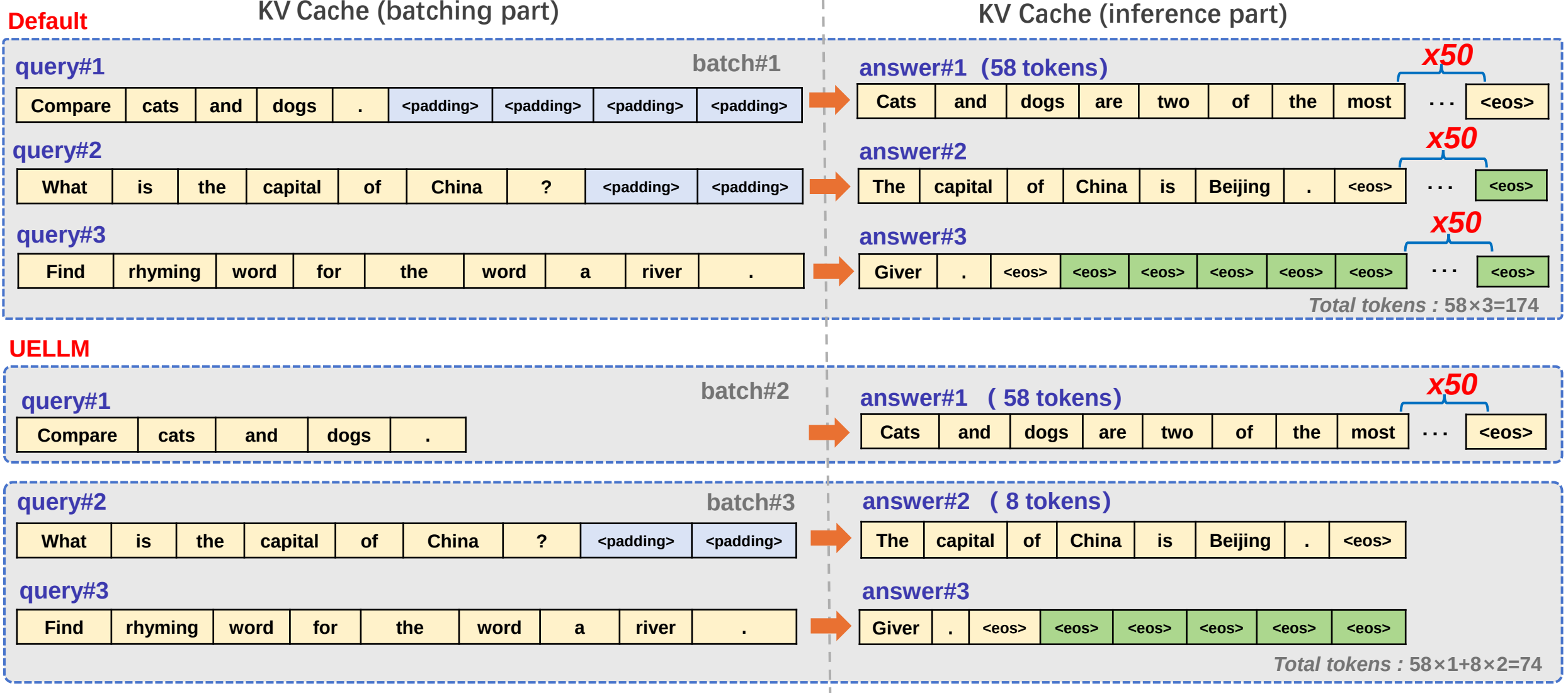
Default
KV Cache (batching part)
KV Cache (inference part)
query#1
batch#1
Compare cats and dogs . <padding> <padding> <padding> <padding>
answer#1 (58 tokens)
Cats and dogs are two of the most ... <eos>
x50
query#2
What is the capital of China ? <padding> <padding>
answer#2
The capital of China is Beijing . <eos> ... <eos>
x50
query#3
Find rhyming word for the word a river .
answer#3
Giver . <eos> <eos> <eos> <eos> <eos> <eos> ... <eos>
x50
Total tokens : 58×3=174
UELLM
query#1
batch#2
Compare cats and dogs .
answer#1 (58 tokens)
Cats and dogs are two of the most ... <eos>
x50
query#2
batch#3
What is the capital of China ? <padding> <padding>
answer#2 (8 tokens)
The capital of China is Beijing . <eos>
query#3
Find rhyming word for the word a river .
answer#3
Giver . <eos> <eos> <eos> <eos> <eos> <eos>
Total tokens : 58×1+8×2=74