Spring25 CS598YP

22.2: Prompt Cache

Yongjoo Park

University of Illinois Urbana-Champaign

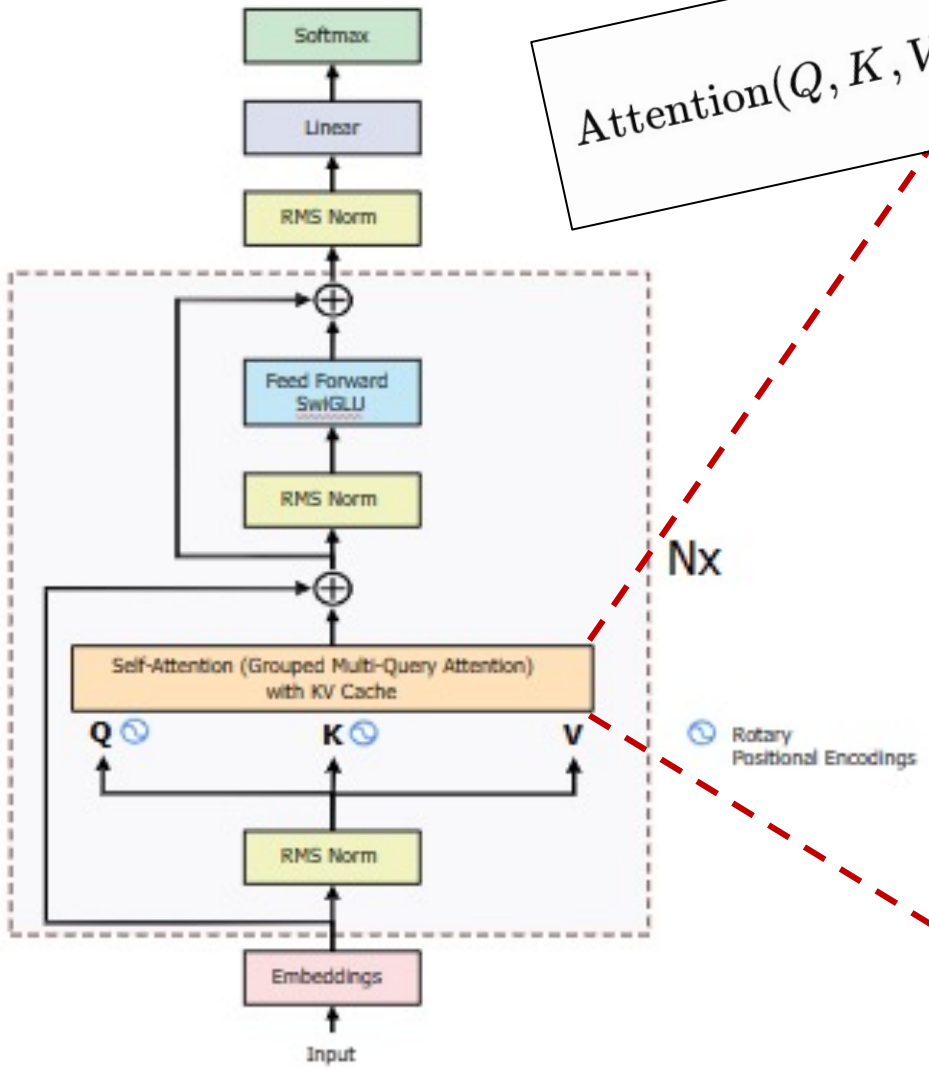
Outline

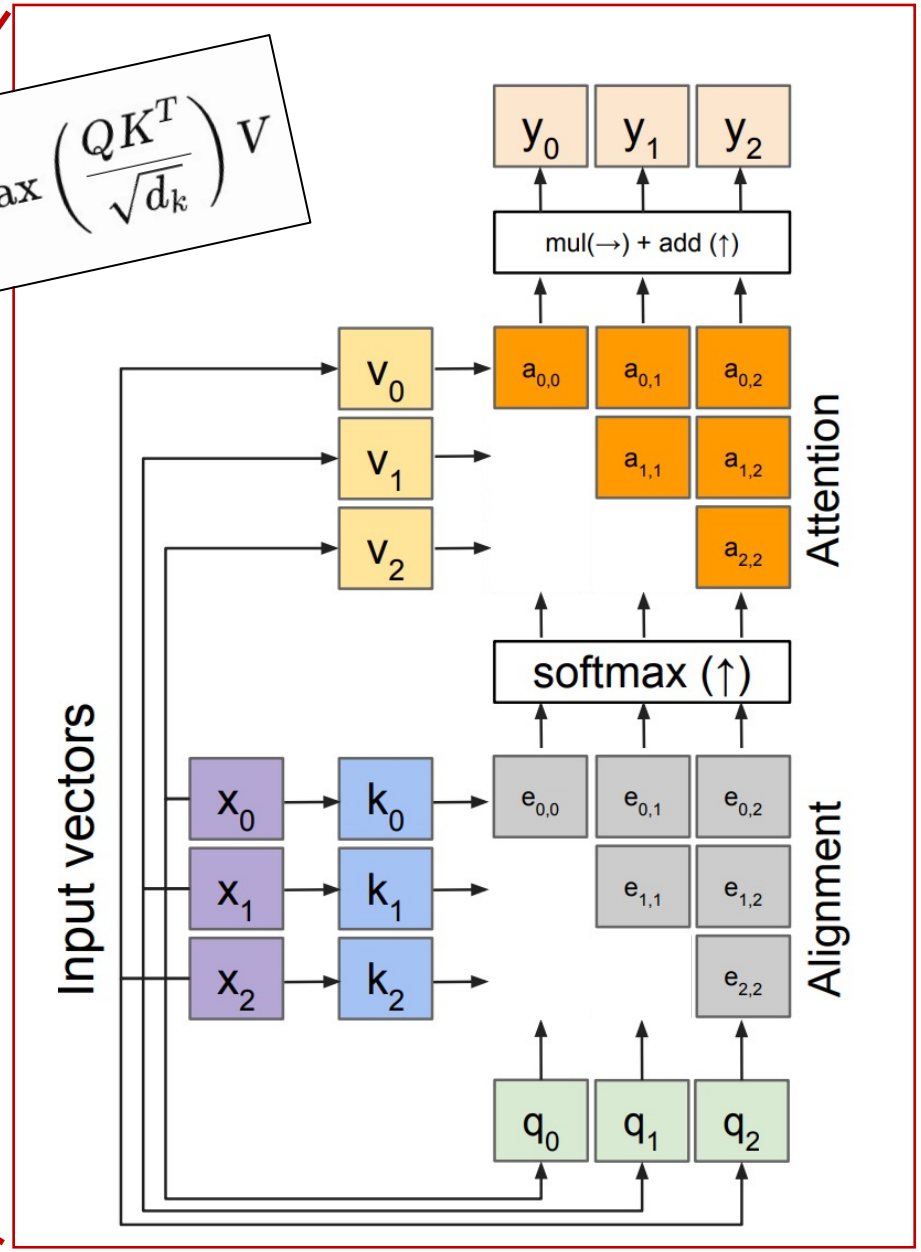
- *Recap:* ***KV Cache***
- *Typical RAG pipeline*
- *Positional Encoding*
- ***Prefix*** *caching*
- ***Prompt*** *Cache*

Attention and KV Cache

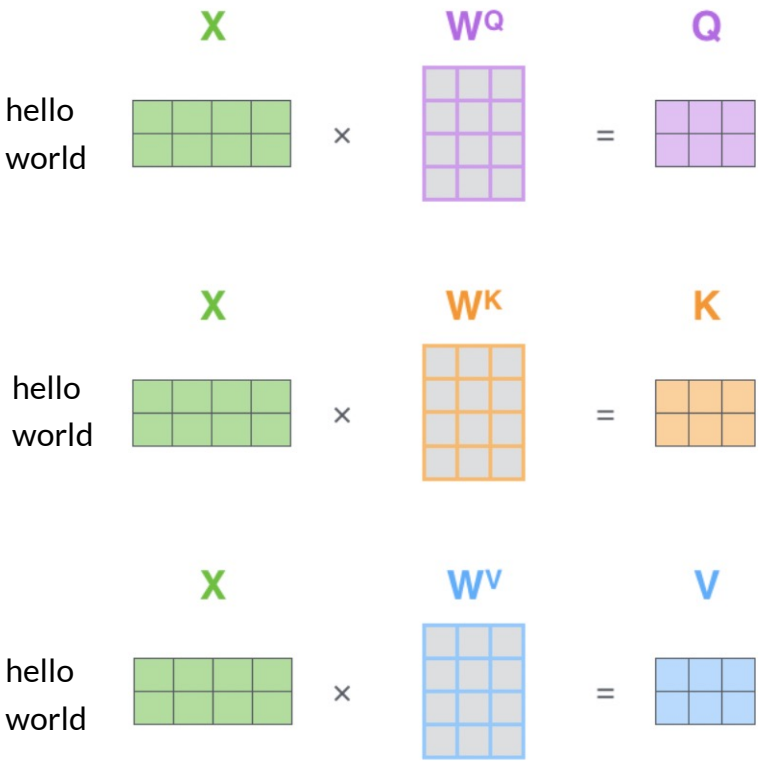
Attention zoom-ed in

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

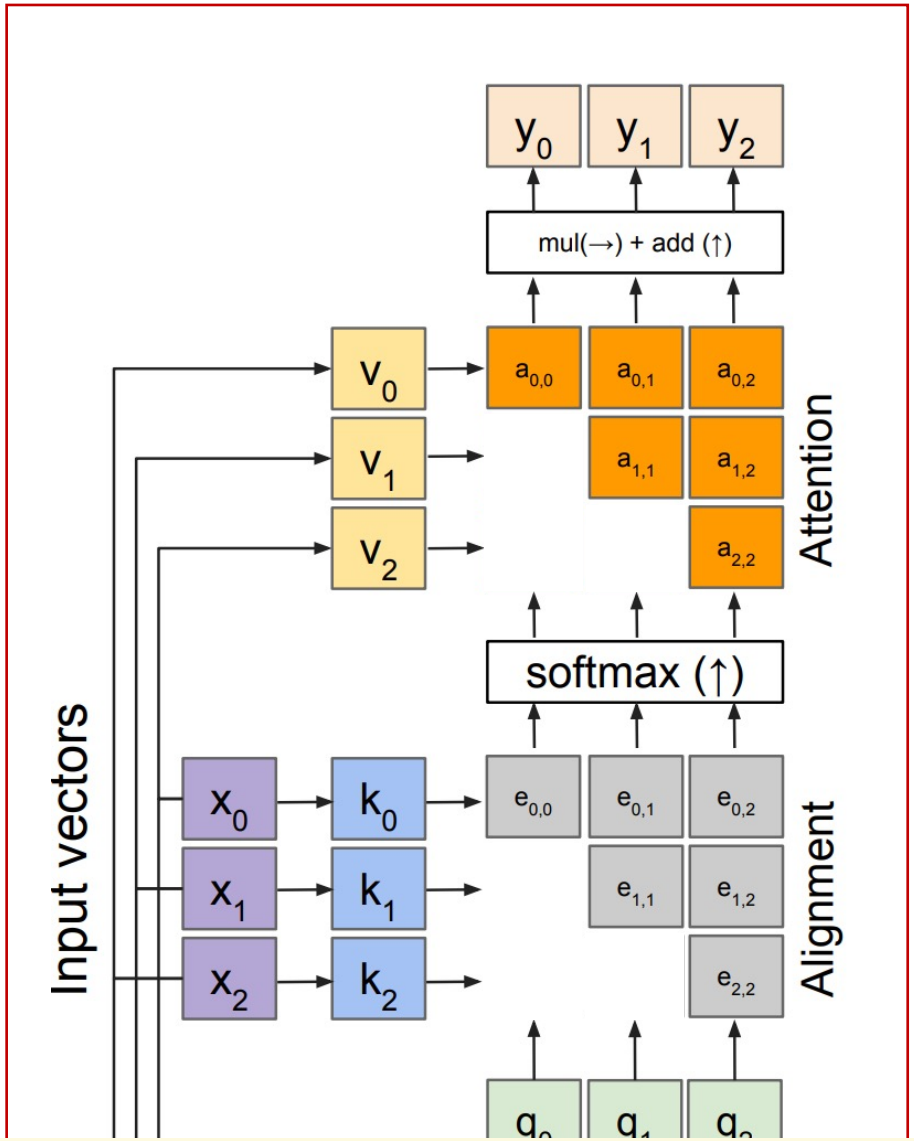




Getting K, Q, V is expensive



Dimension: 4,096 for Llama3-8B



We can re-use K and V for previous tokens -> ***KV Cache***

RAG <-> Positional Encoding

Retrieval-Augmented Generation workflow

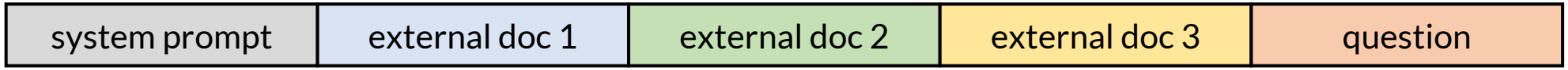
Given as detailed response as possible. Be respectful.

"Two roads diverged in a yellow wood, And sorry I could not travel both..."

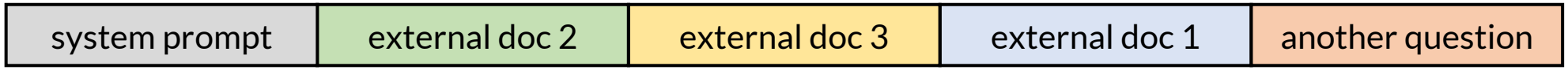
"You may trod me in the very dirt But still, like dust, I'll rise."

"Look on my Works, ye Mighty, and despair! Nothing beside remains..."

Find the poem about choices in life



Can we re-use KV cache for this request?

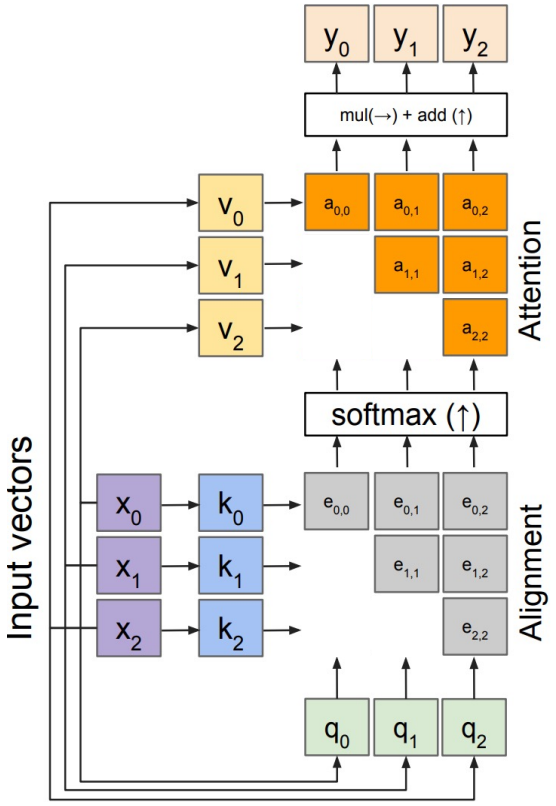


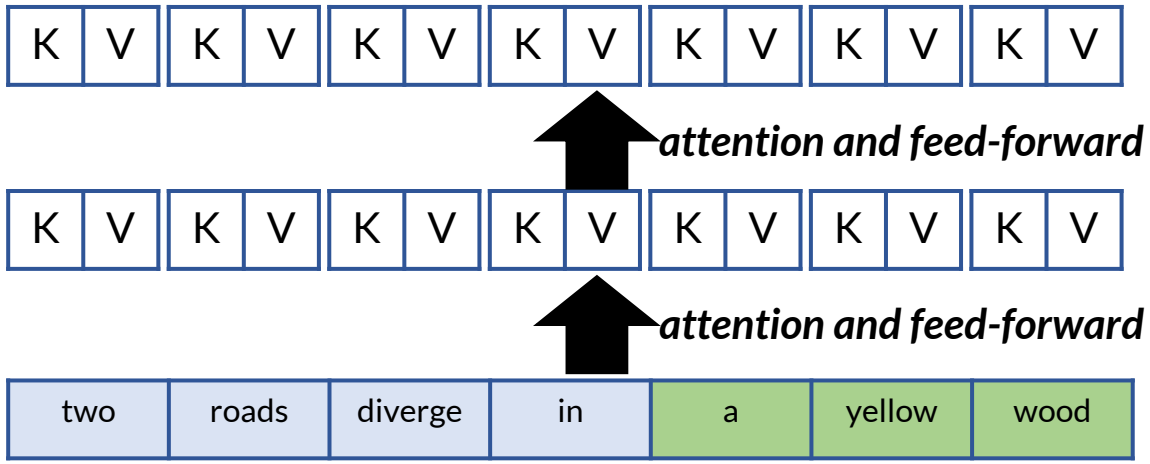
Reusing KV cache across requests is hard

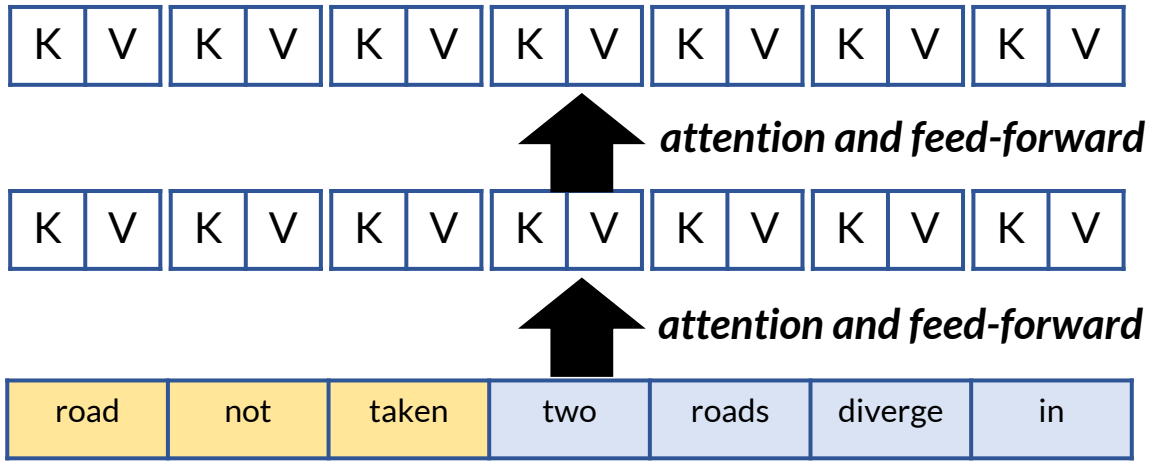
Two requests have overlapping tokens: “two roads diverge in”

However, the generated KV cache values are different

Why? two reasons

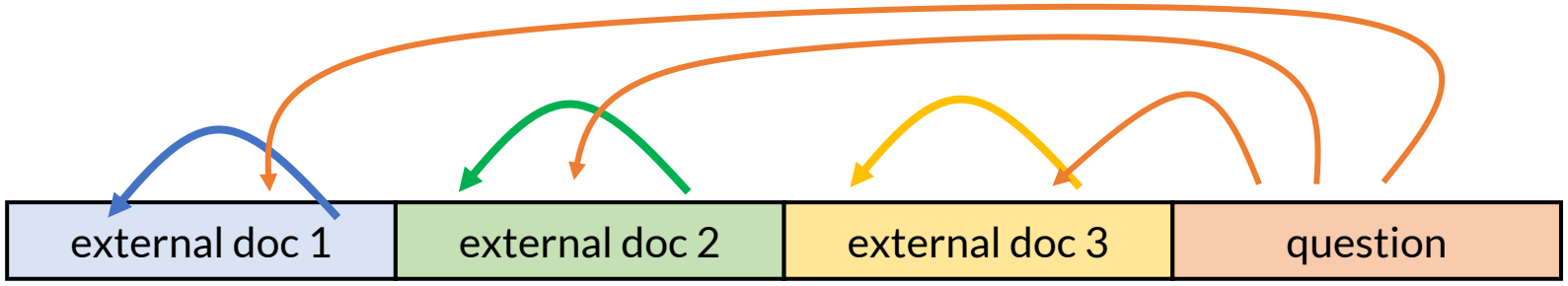


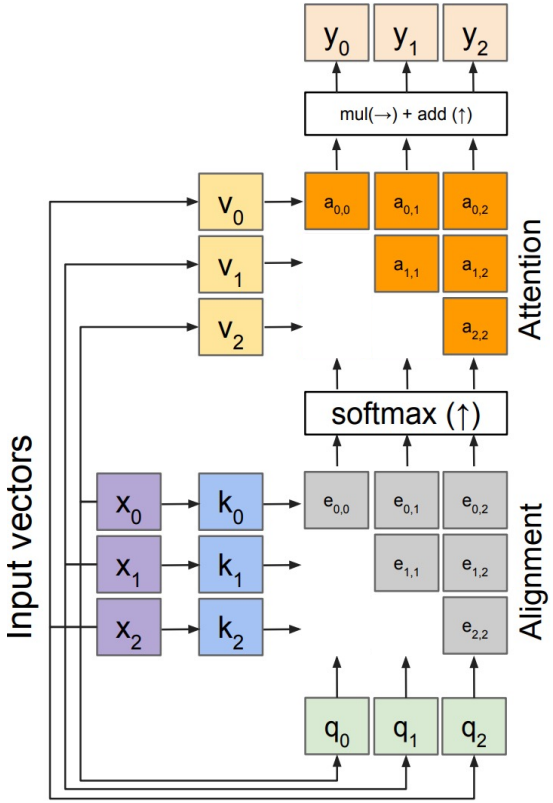


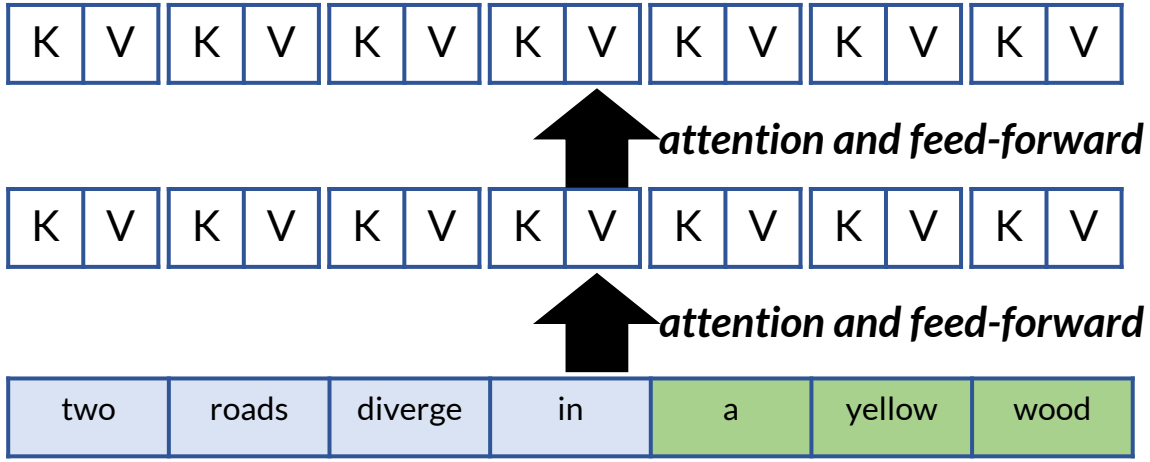


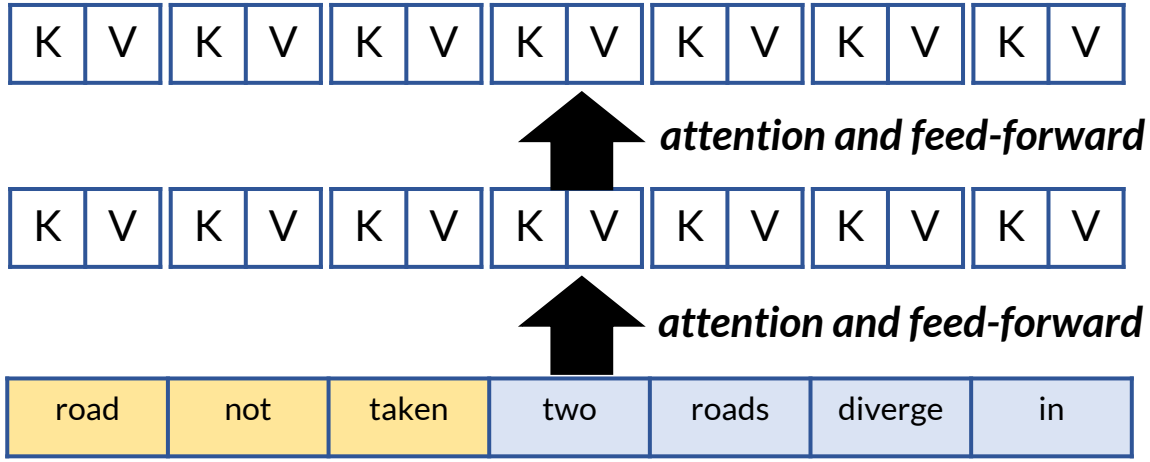
KV ***was*** different at different positions

Suppose RAG setting







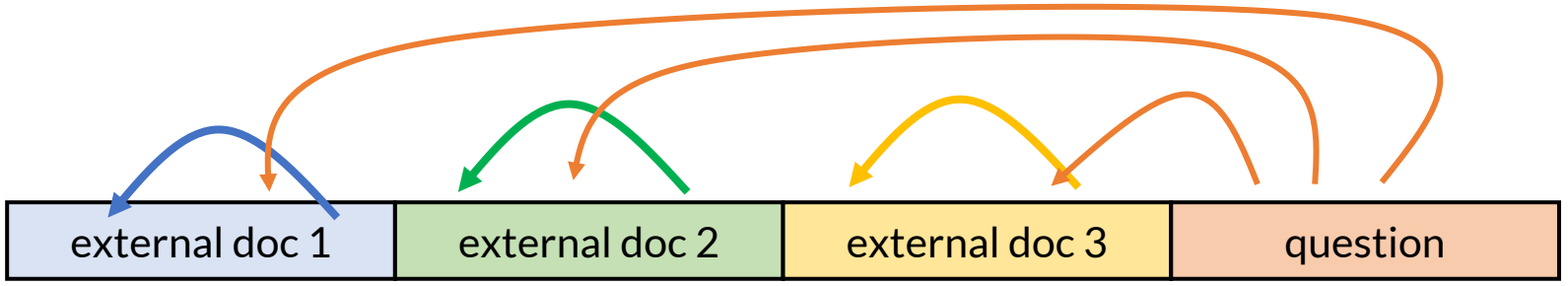


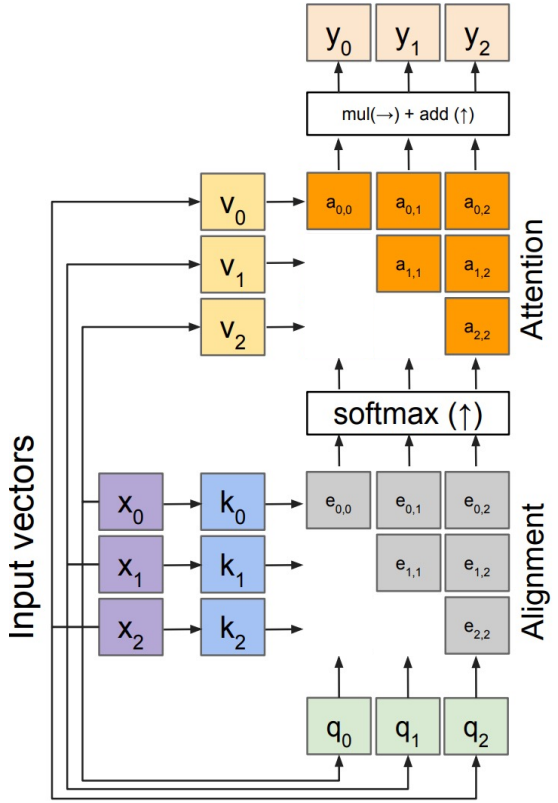
KV ***was*** different at different positions

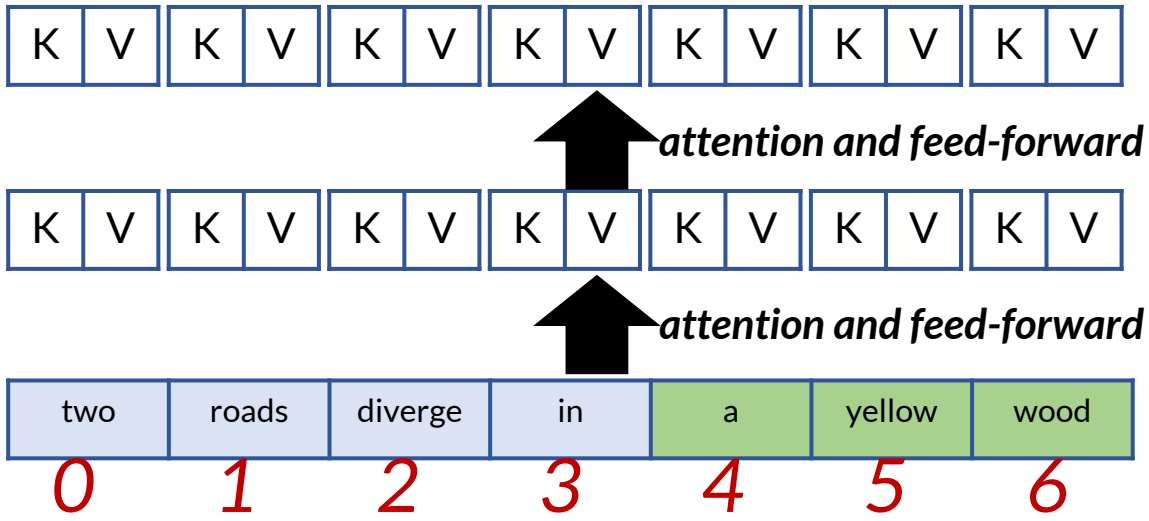
due to absolute positional encoding

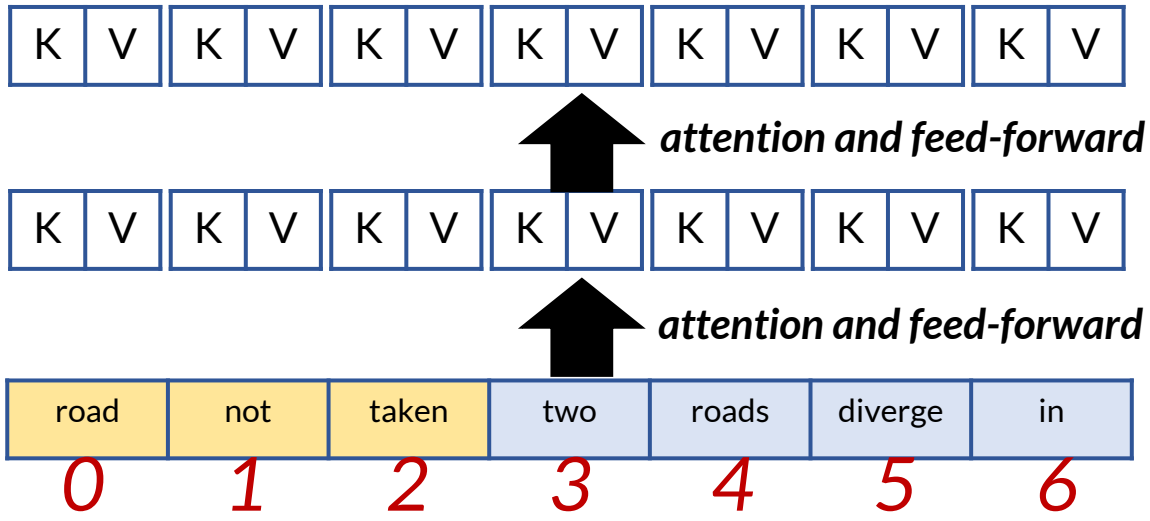
now no longer true

Suppose RAG setting



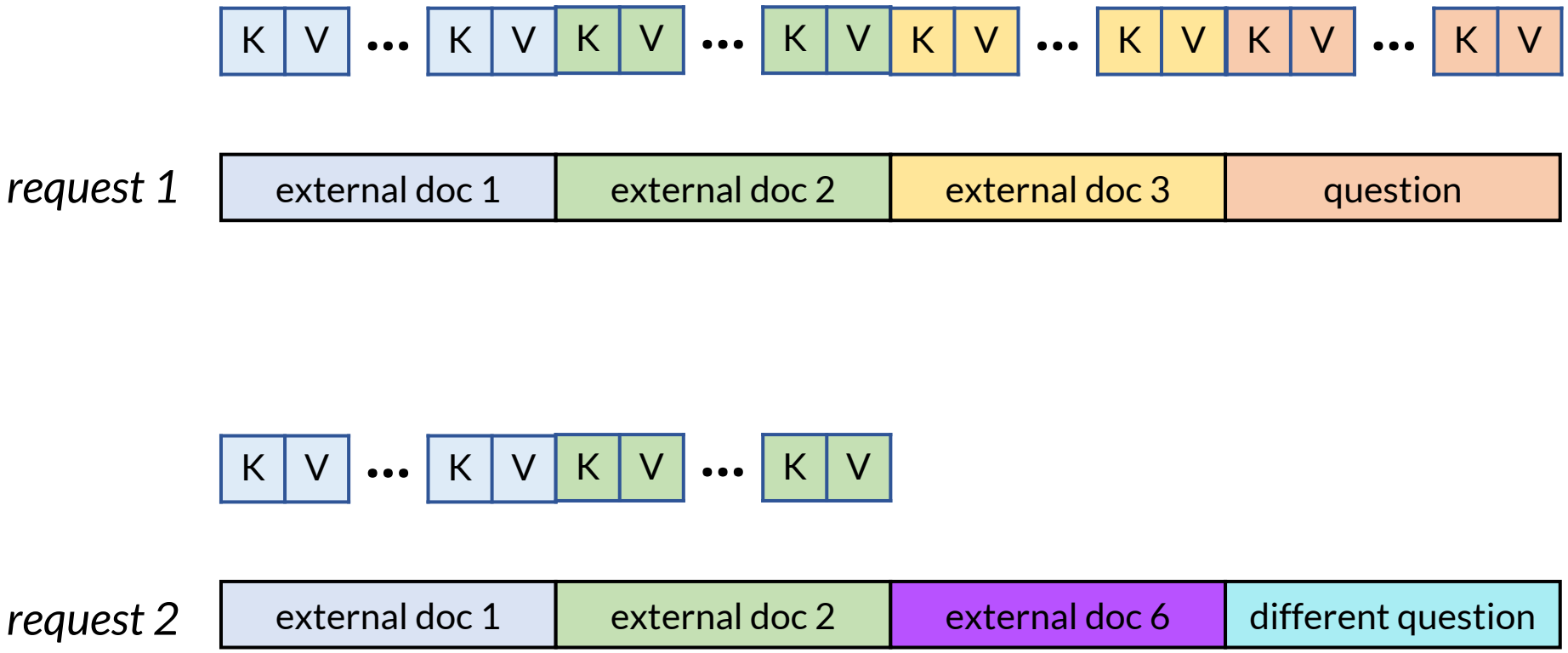






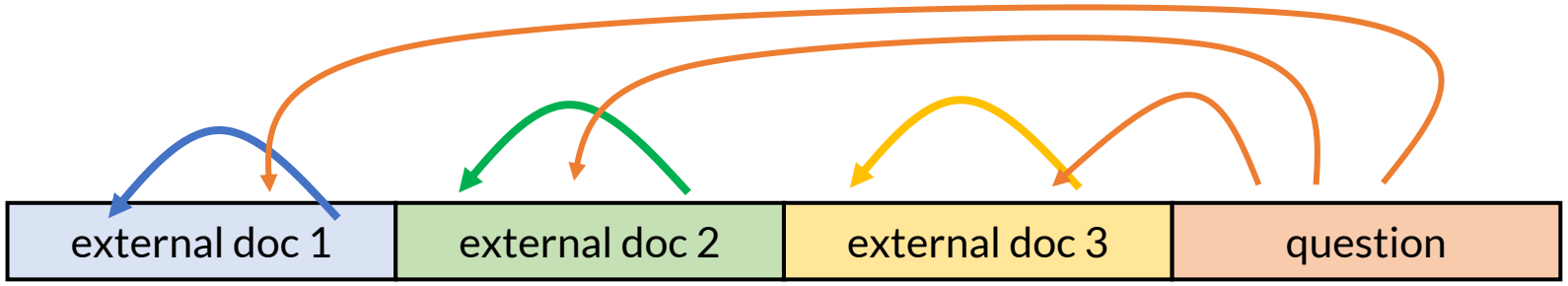
Caching

vLLM: Prefix caching



Prompt Cache allows reusing *red blocks (does it?)*

We are using this RAG setting *(Is this OK? We will study more)*



Schema

```
<schema name="cities">
  <module name="city-info">...
  </module>
  <module name="trip-plan"> ...
    <param name="duration" len=2/>
  </module>
  <module name="tokyo"> ...
  </module>
  <module name="miami"> ...
  </module> ...
</schema>
```

Prompt attention states

Prompt Cache

<city-info/> <trip-plan/> <tokyo/> <miami/> <paris/>

LLM

Prompt

```
<prompt schema="cities">
  <trip-plan duration= "3 days" /> <miami/> Highlight the surf spots.
</prompt>
```

empty space

Summary

- **KV cache** cannot be re-used across requests (in most cases)
- RAG will involve external data
- Positional encoding makes KV cache reuse harder
- **Prefix caching** is basic optimization technique
- **Prompt Cache** allows fine-grained KV cache reuse

Questions?