Systems for Machine Learning (CMSC828G)

Inference Systems

Abhinav Bhatele

UNIVERSITY OF MARYLAND

Questions

- How many students have done from-scratch training?
- How many students have fine-tuned models?
- How many students have used pre-trained models in inference mode?

What is inference?

https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f

What is inference?

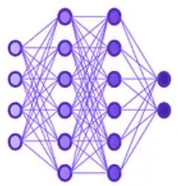
Word	Probability
speak	0.065
generate	**0.072**
politics	0.001
...	...
walk	0.003

https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f

What is inference?

A trained language model can

Input



LLM

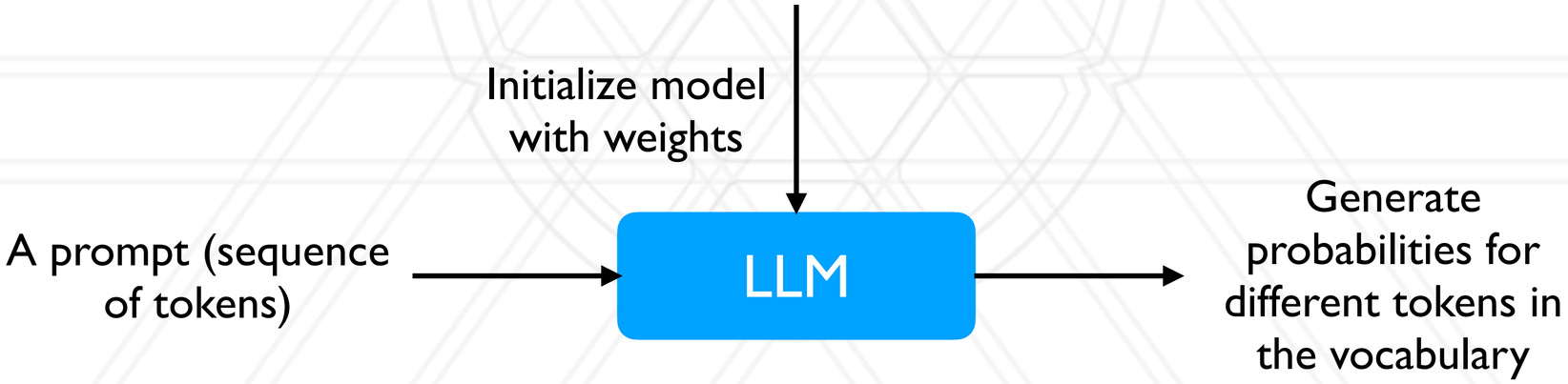
Word	Probability
speak	0.065
generate	**0.072**
politics	0.001
...	...
walk	0.003

Word	Probability
ability	0.002
text	0.084
coherent	**0.085**
...	...
ideas	0.041

https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f

How is inference done?

- Start with initializing a model on a GPU with weights from a pre-trained model
- Input: a user prompt
- Output: a generation of output tokens



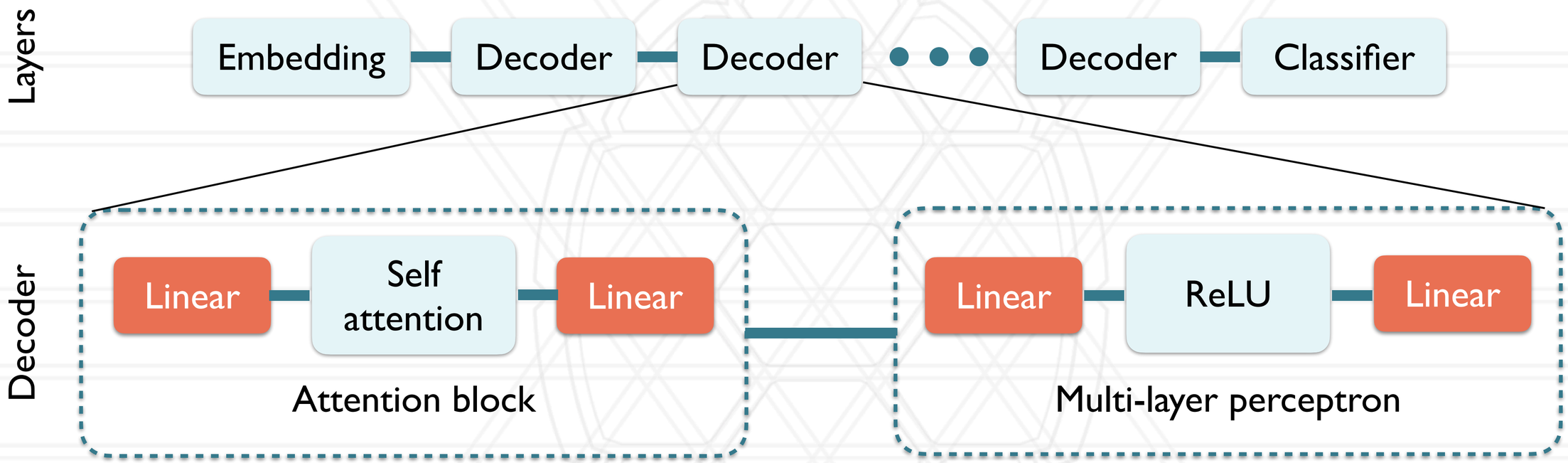
Prefill vs decode

- Prefill stage: when you process the initial set of input tokens to fill the KV cache
- Decode: when you autoregressively generate output tokens one at a time

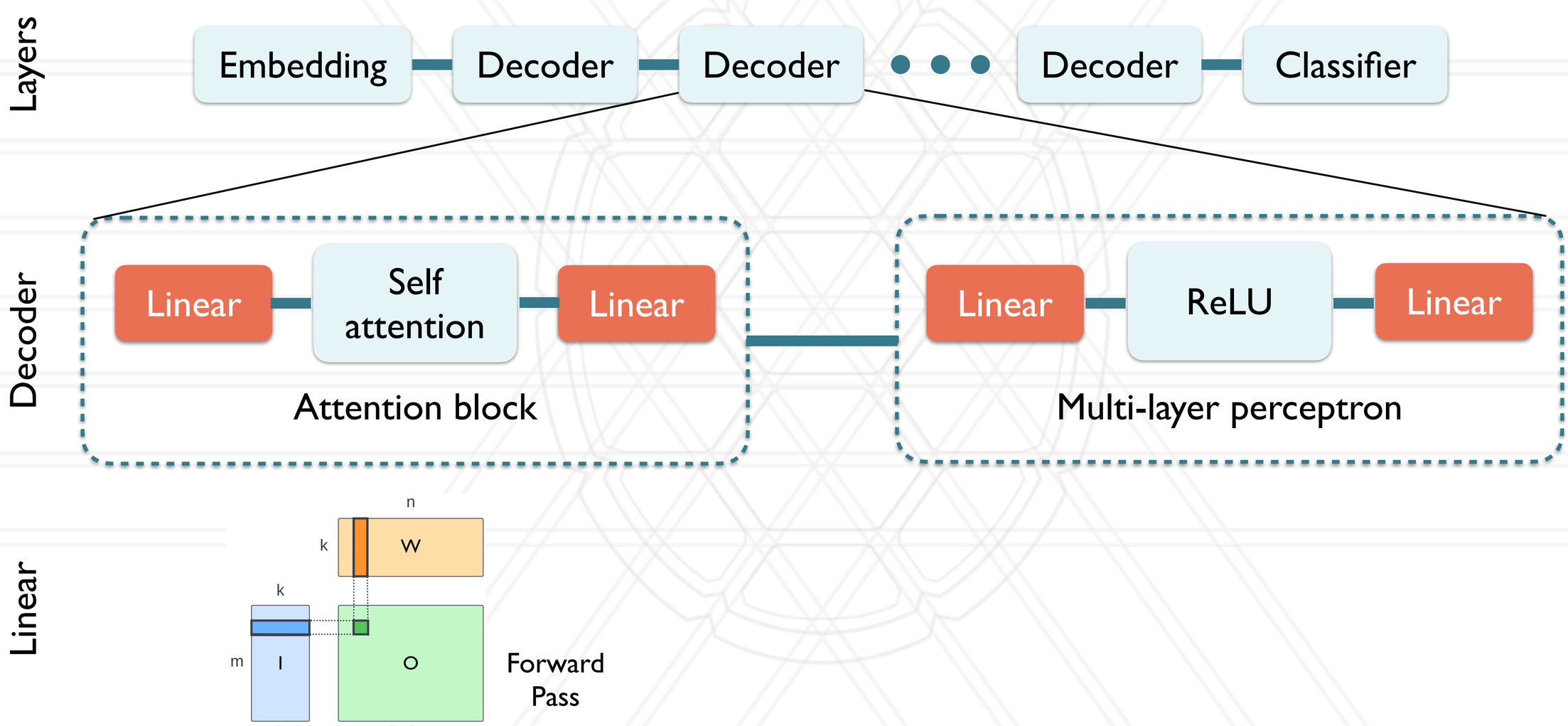
Compute work in transformer models

Layers

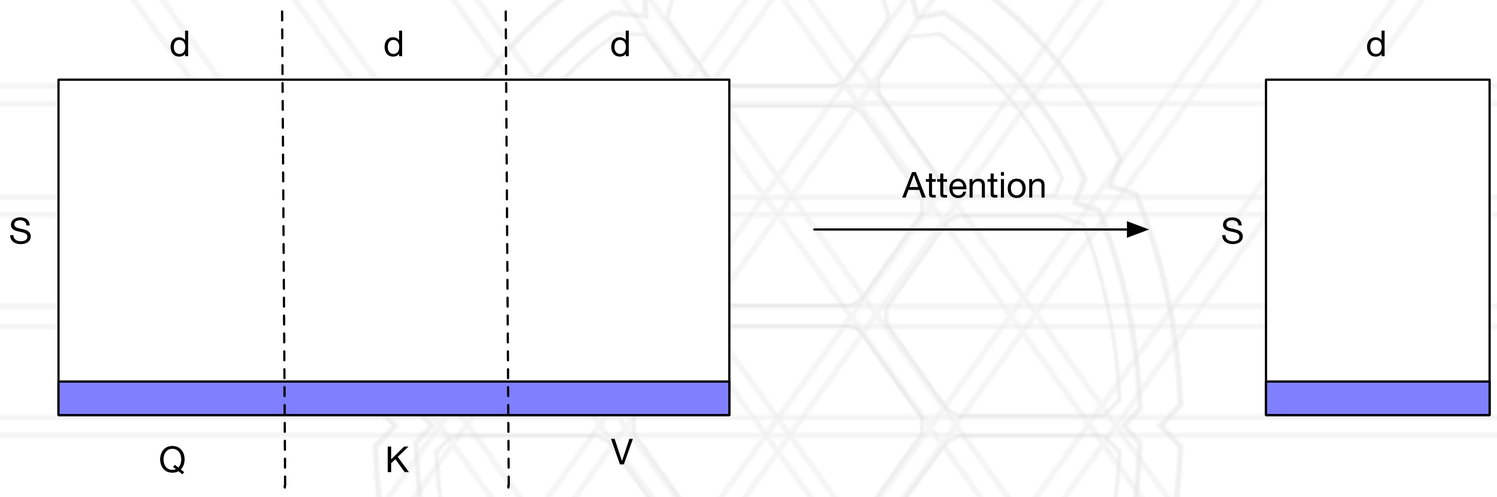
Compute work in transformer models



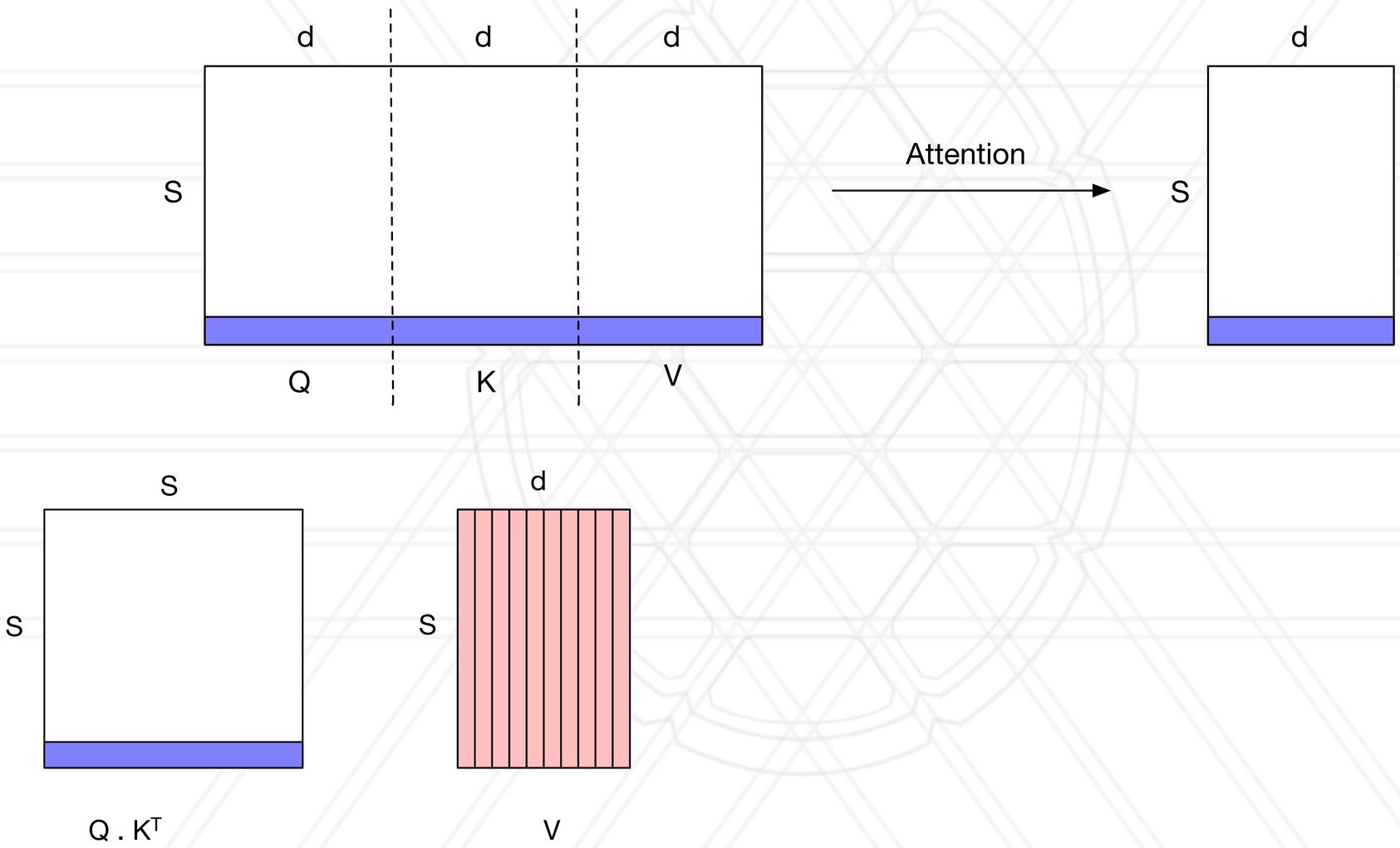
Compute work in transformer models



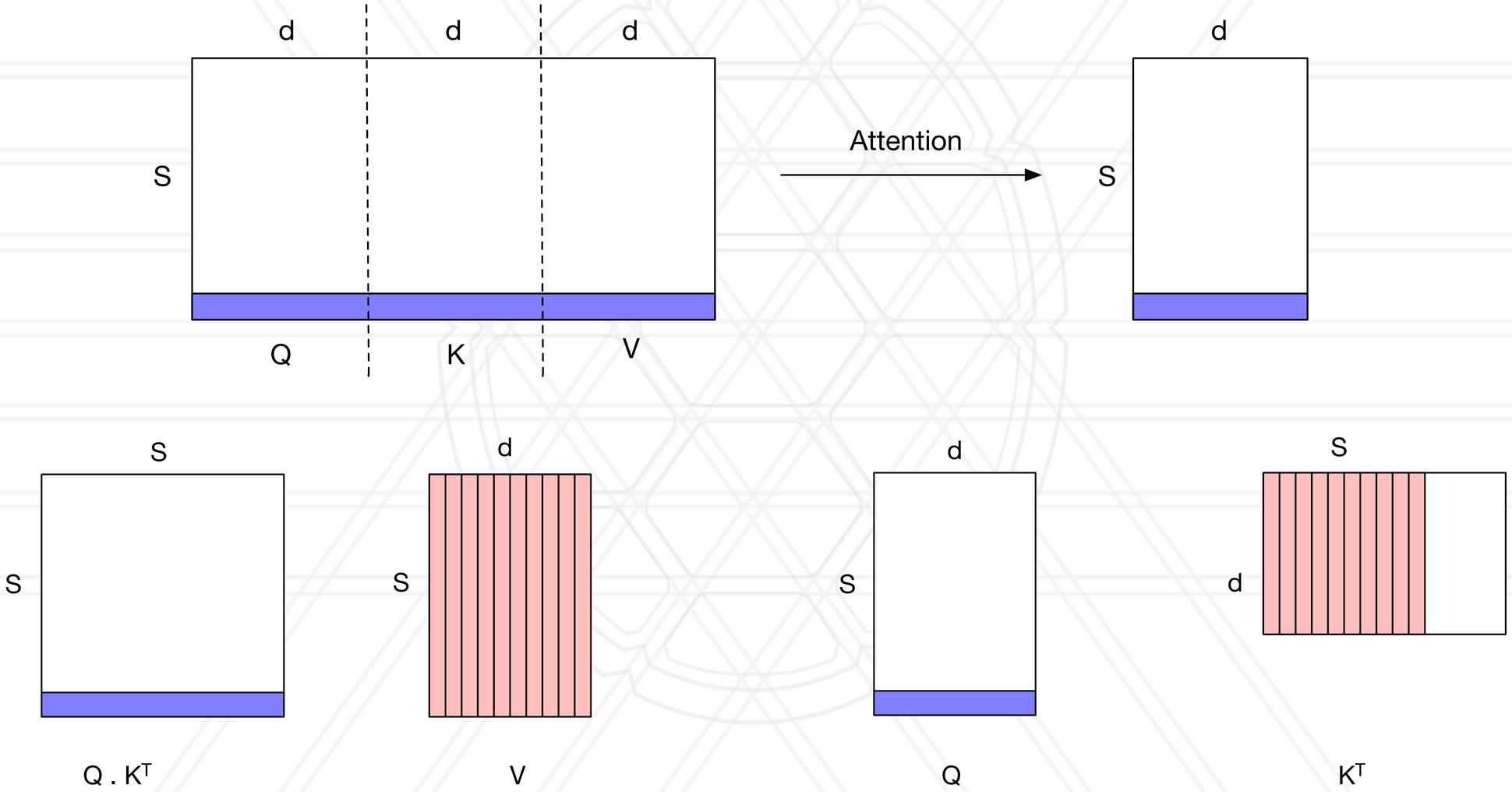
Why do we store the KV cache?



Why do we store the KV cache?



Why do we store the KV cache?



Various modes in which to run inference

- Single prompts
- Batched inference: lots of prompts put together into a batch
- Online mode: create a server
  - Single or multiple users interact with the server

Performance metrics for inference

- Latency:
  - Time to first token
  - Latency between tokens
- Throughput: tokens generated per second

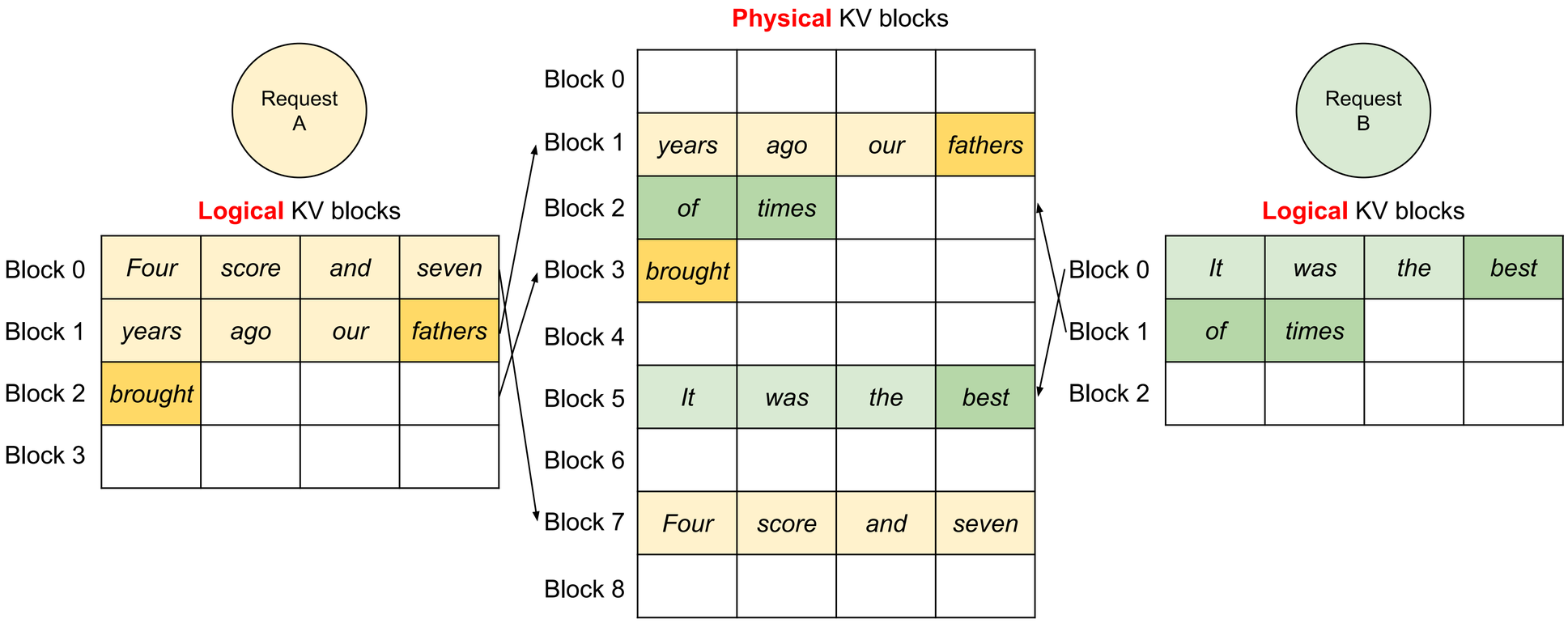
Inference frameworks

- vLLM
- SGLang
- ORCA

Distributed inference

- What do you do when the model does not fit in memory?
- Tensor parallelism
- Pipeline parallelism
- TP + PP

Paged attention in vLLM



UNIVERSITY OF
MARYLAND