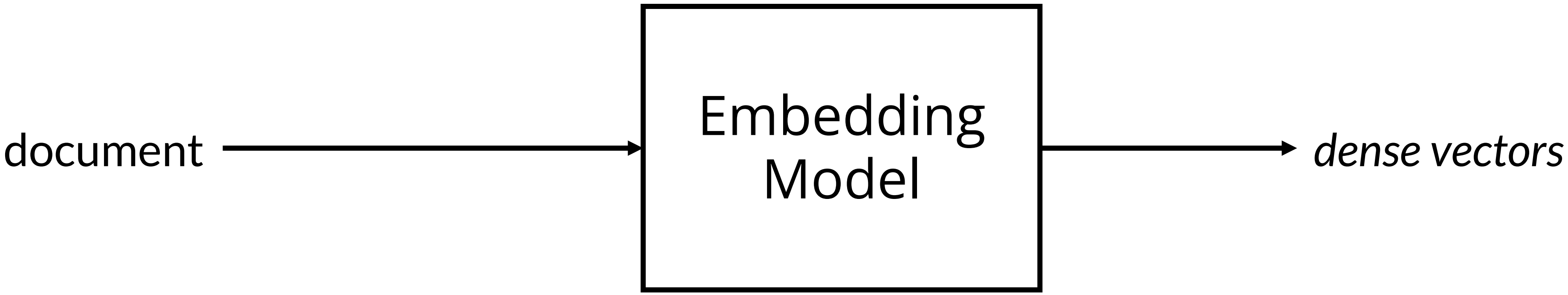
Spring25 CS598YP

18.2: LLM Embedding

Yongjoo Park

University of Illinois Urbana-Champaign

High-level objective: document -> vector



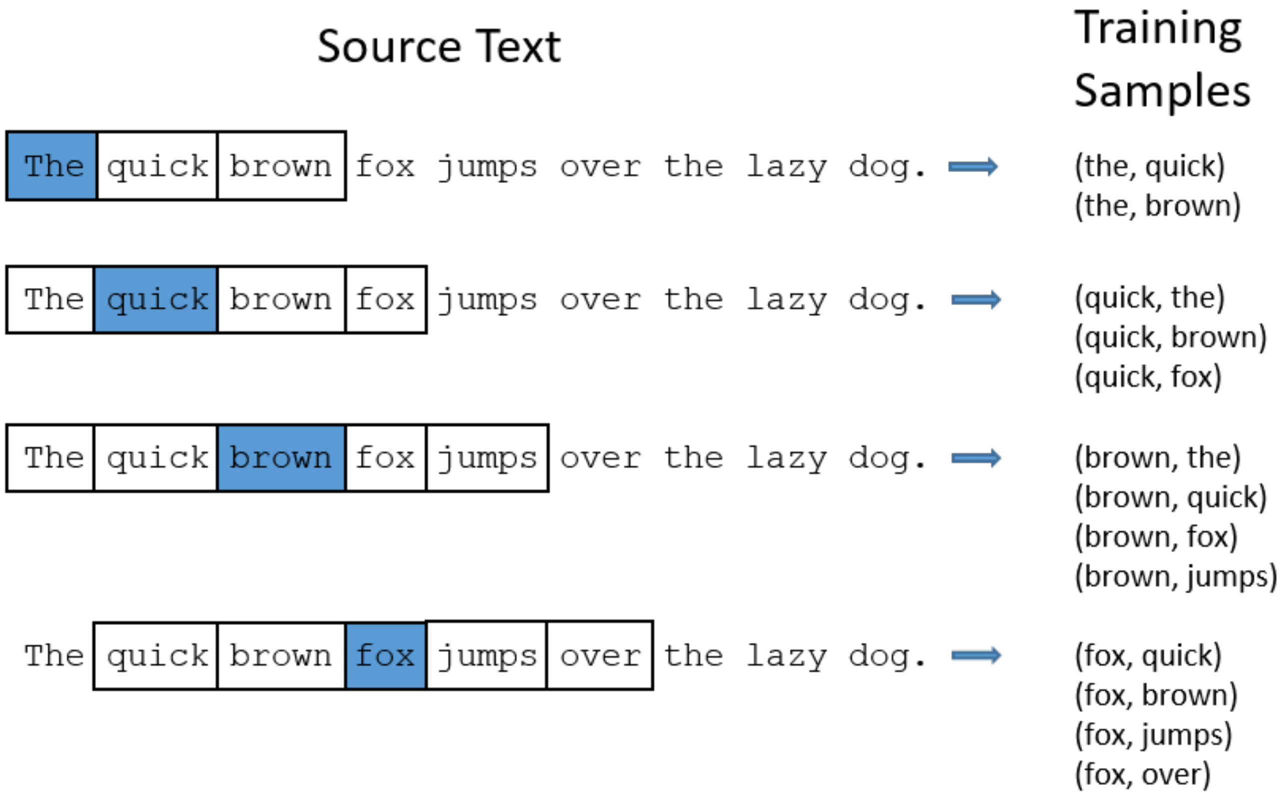
Outline

- Transformer architecture
- Pooling methods: mean, EOS, trainable layer

Transformer

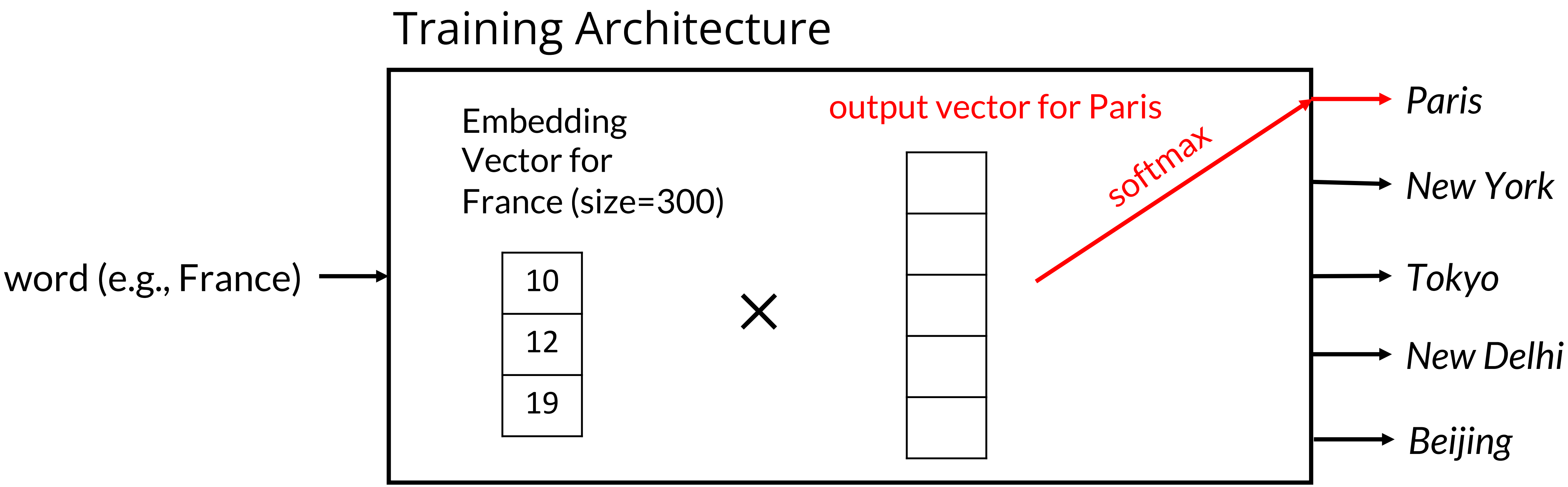
Word2Vec skip-gram task

Task: Given **blue**, predict other words in the window



Word2vec uses C=10 past and future words

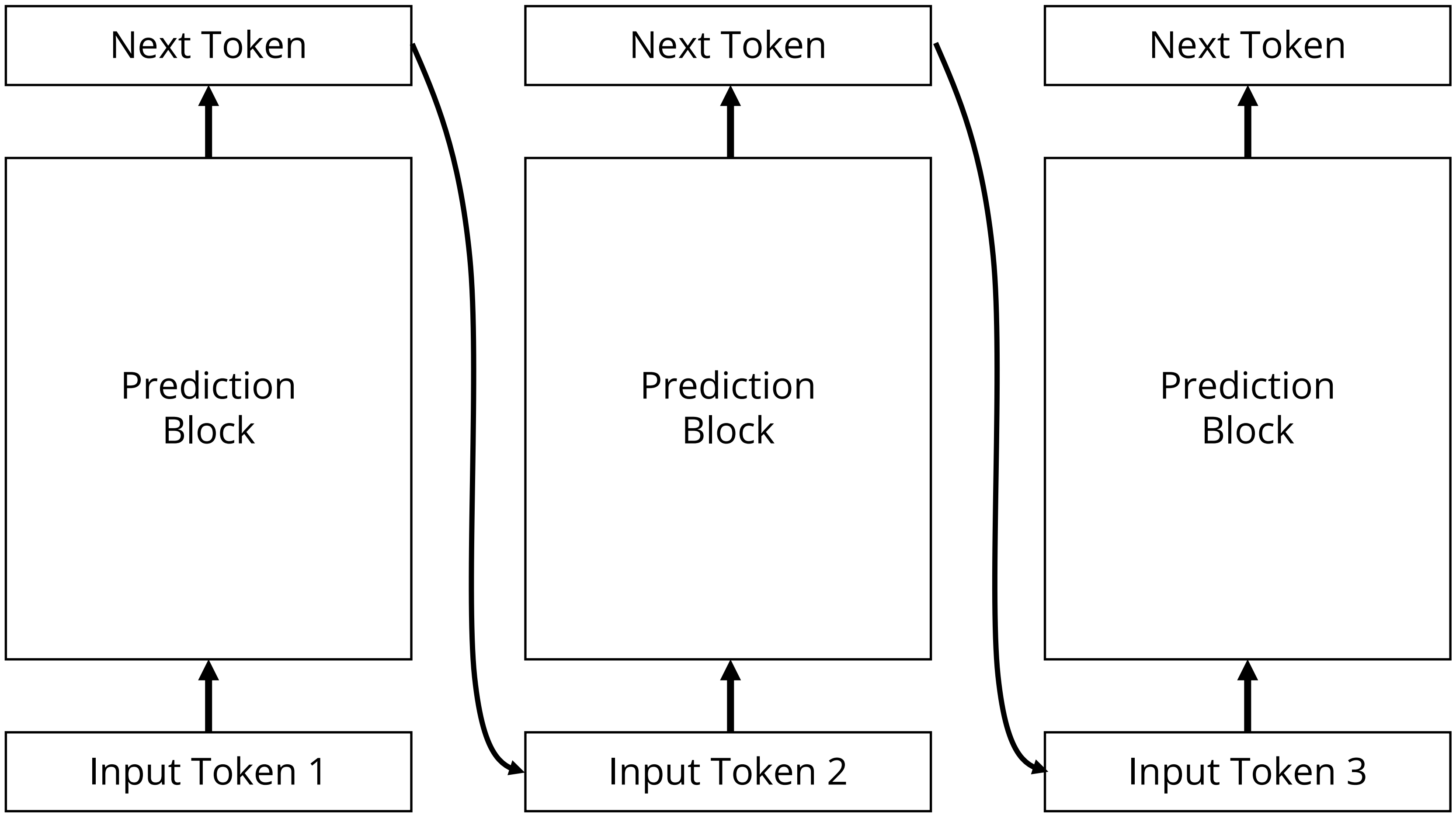
Word2Vec: **Mat-mul** for predicting another word



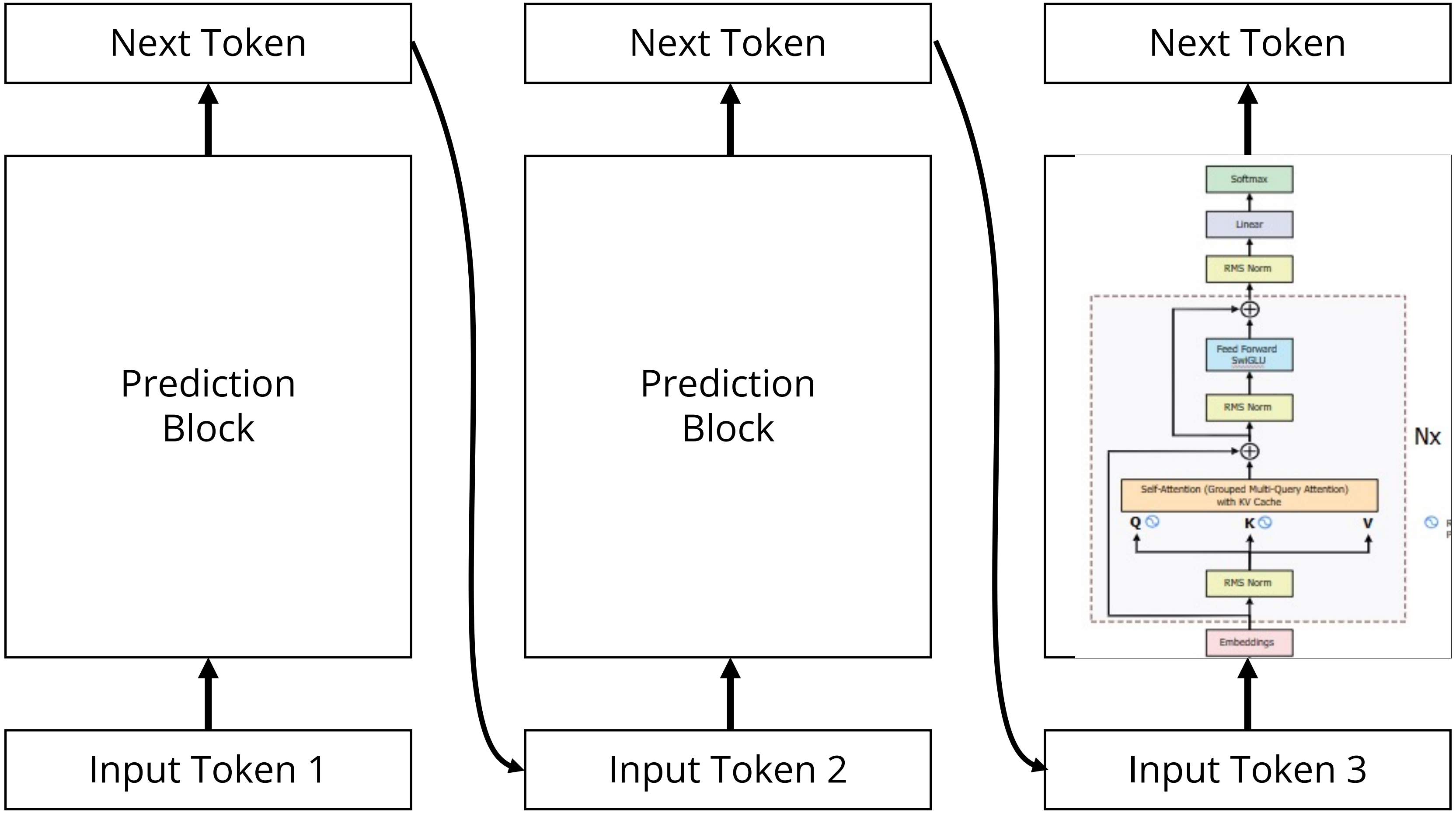
$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = \begin{bmatrix} 10 & 12 & 19 \end{bmatrix}$$

There are as many embedding vectors as the vocabulary size (e.g., 10K)

Decoding-only task

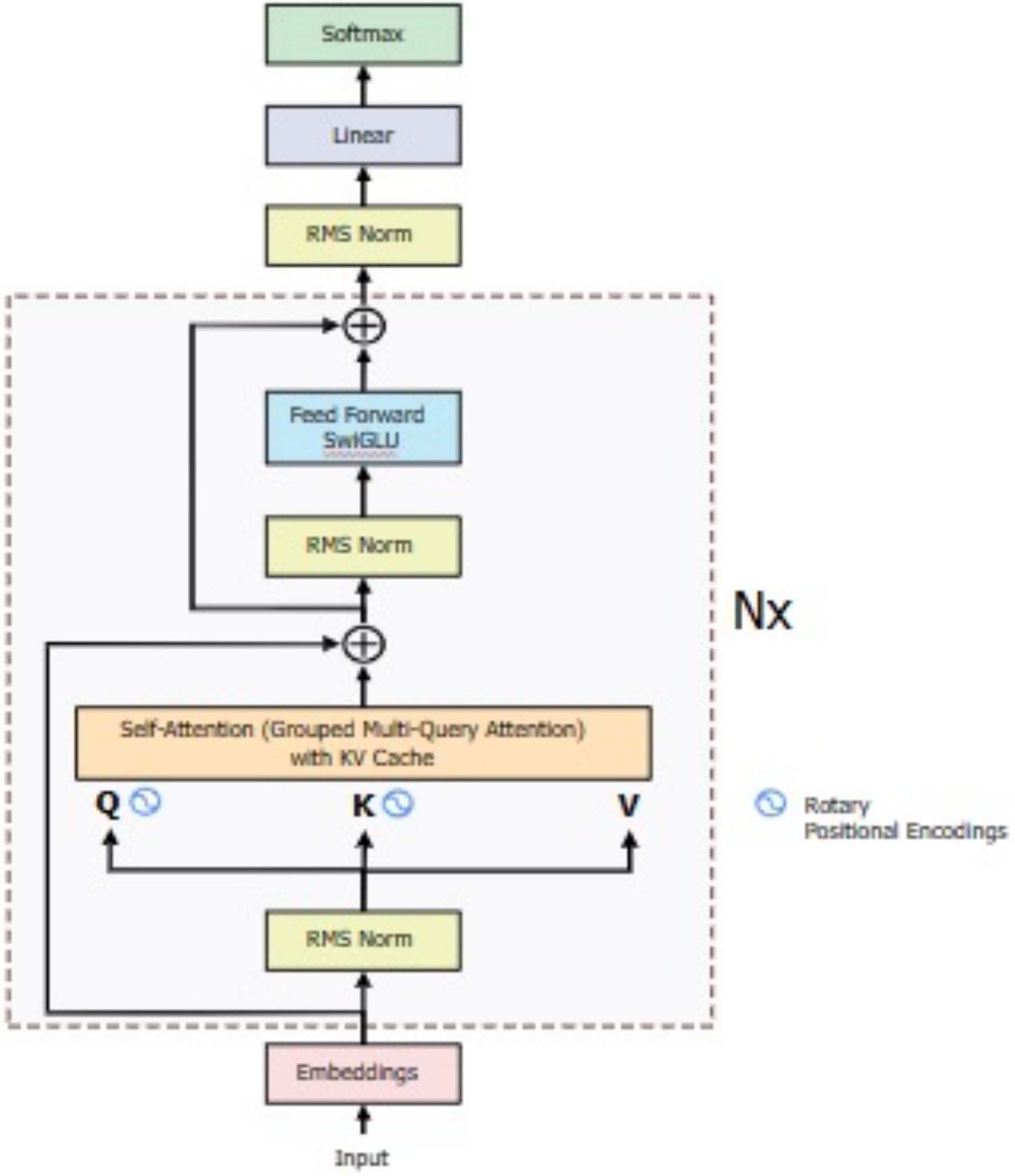


Llama architecture

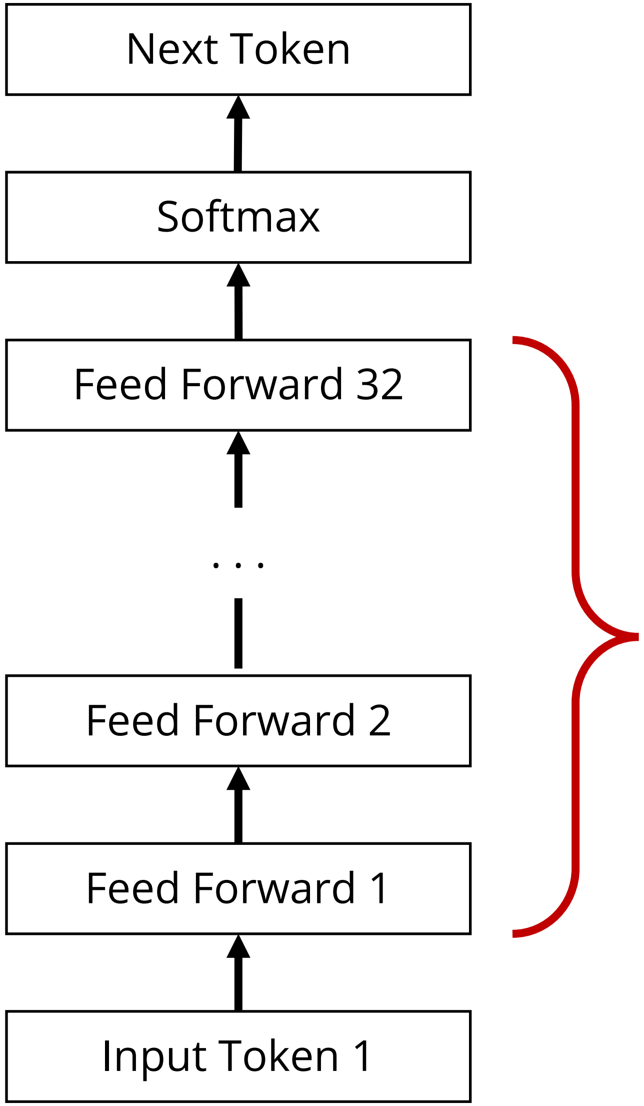


Llama architecture

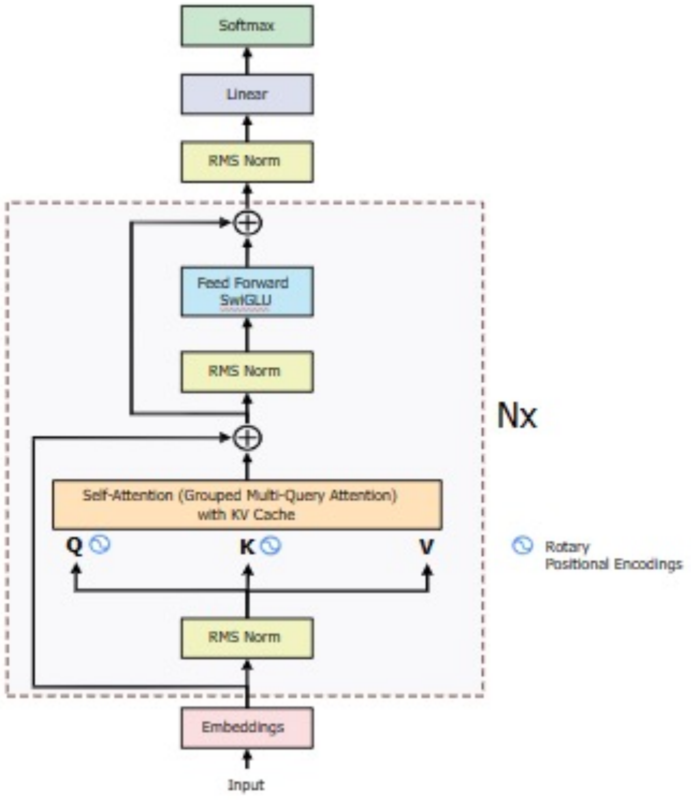
- Llama 3 8B has 32 layers
  (i.e., transformer blocks)
- Property 1: Deep
- Property 2: Attention



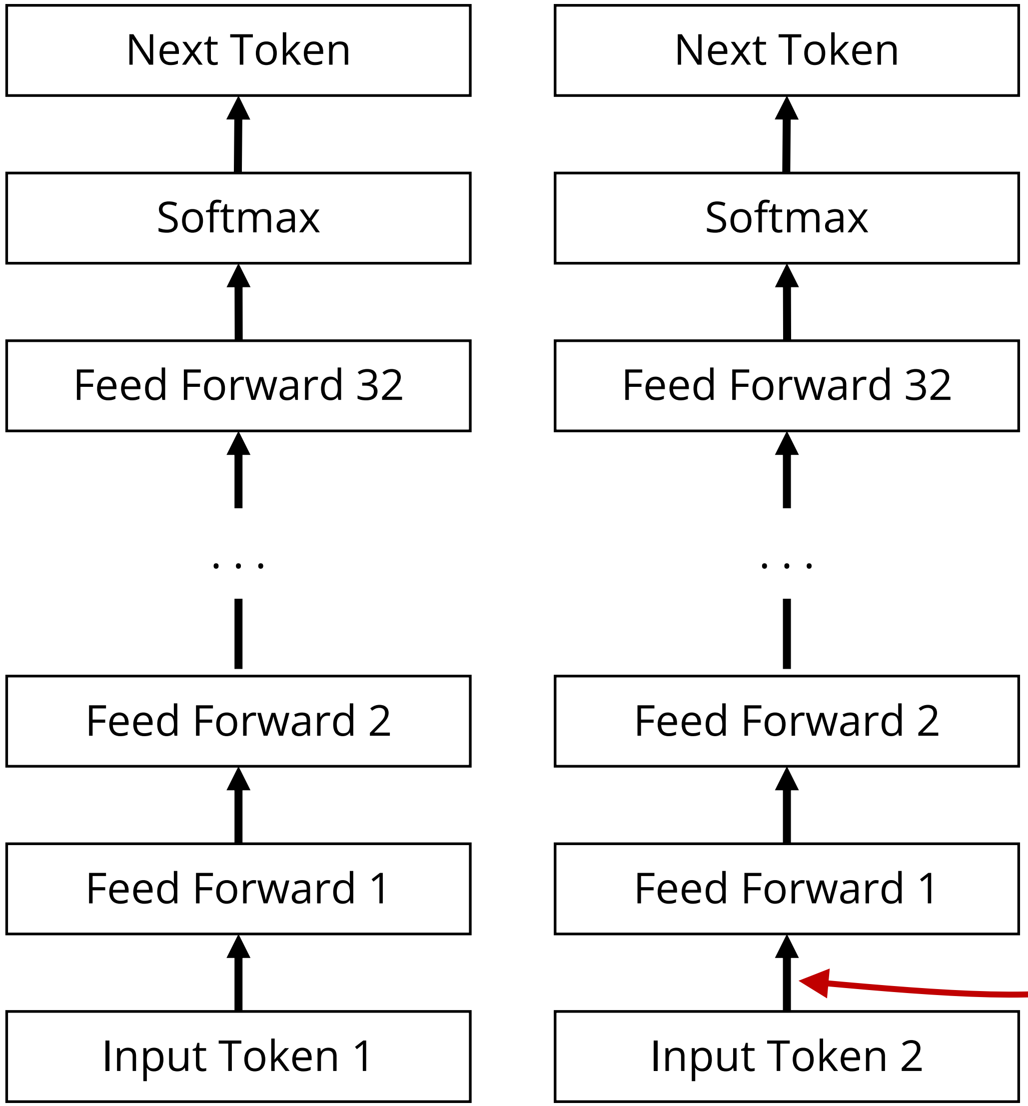
Property 1: Deep neural network



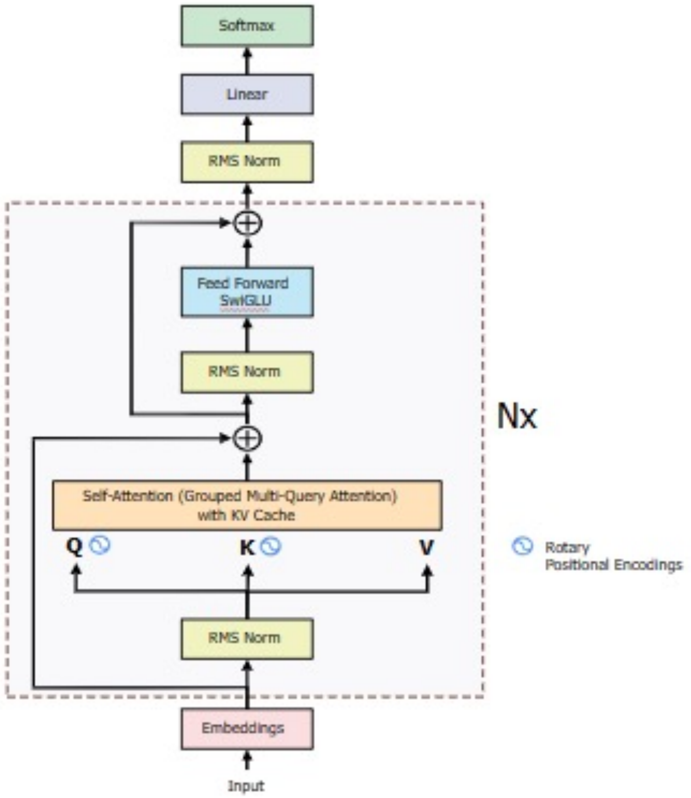
Captures semantics



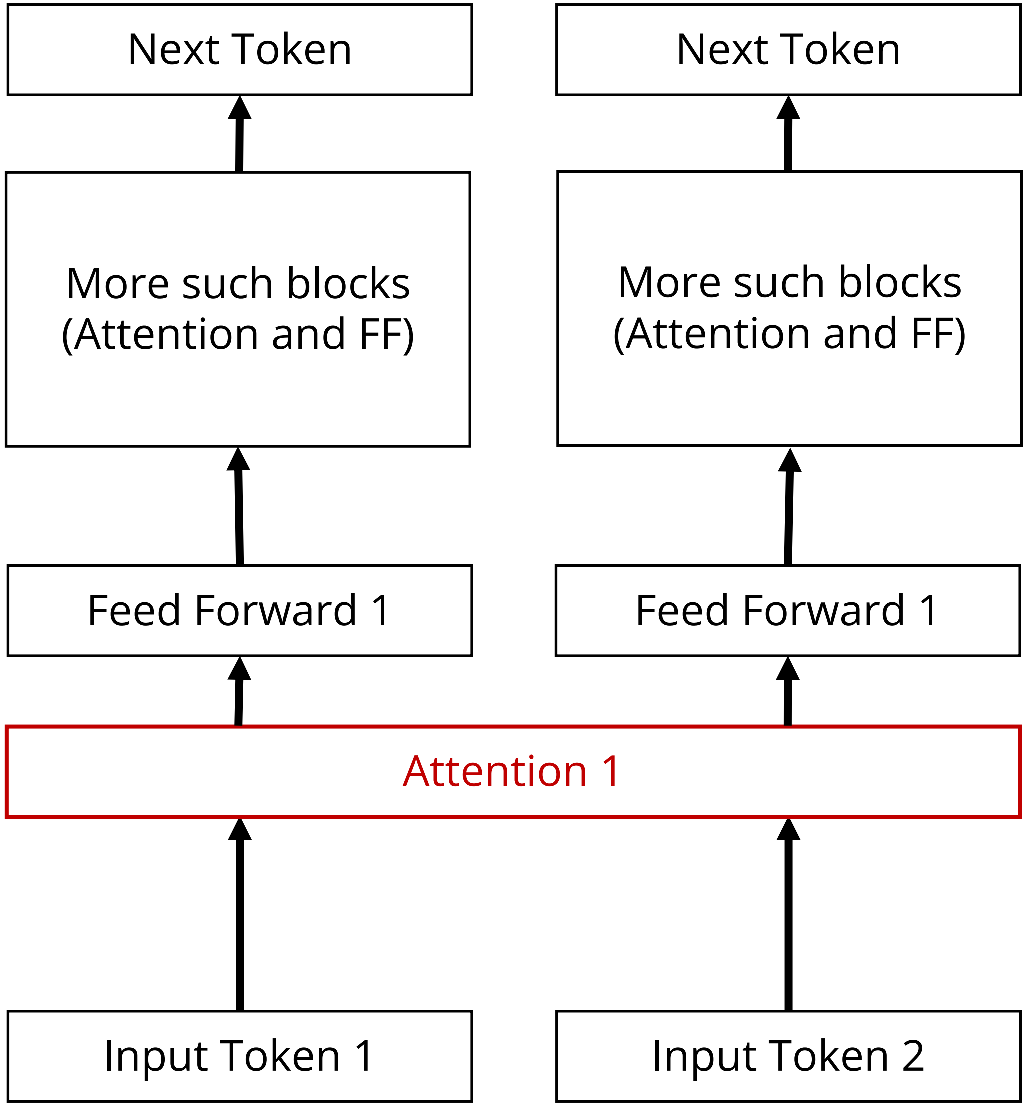
Property 2: Captures dependency



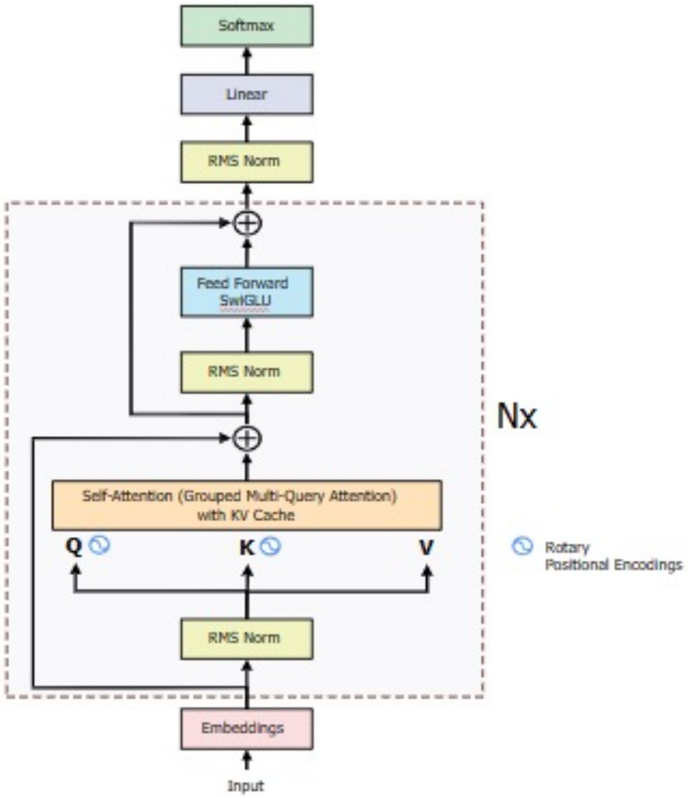
Something before each Feed Forward



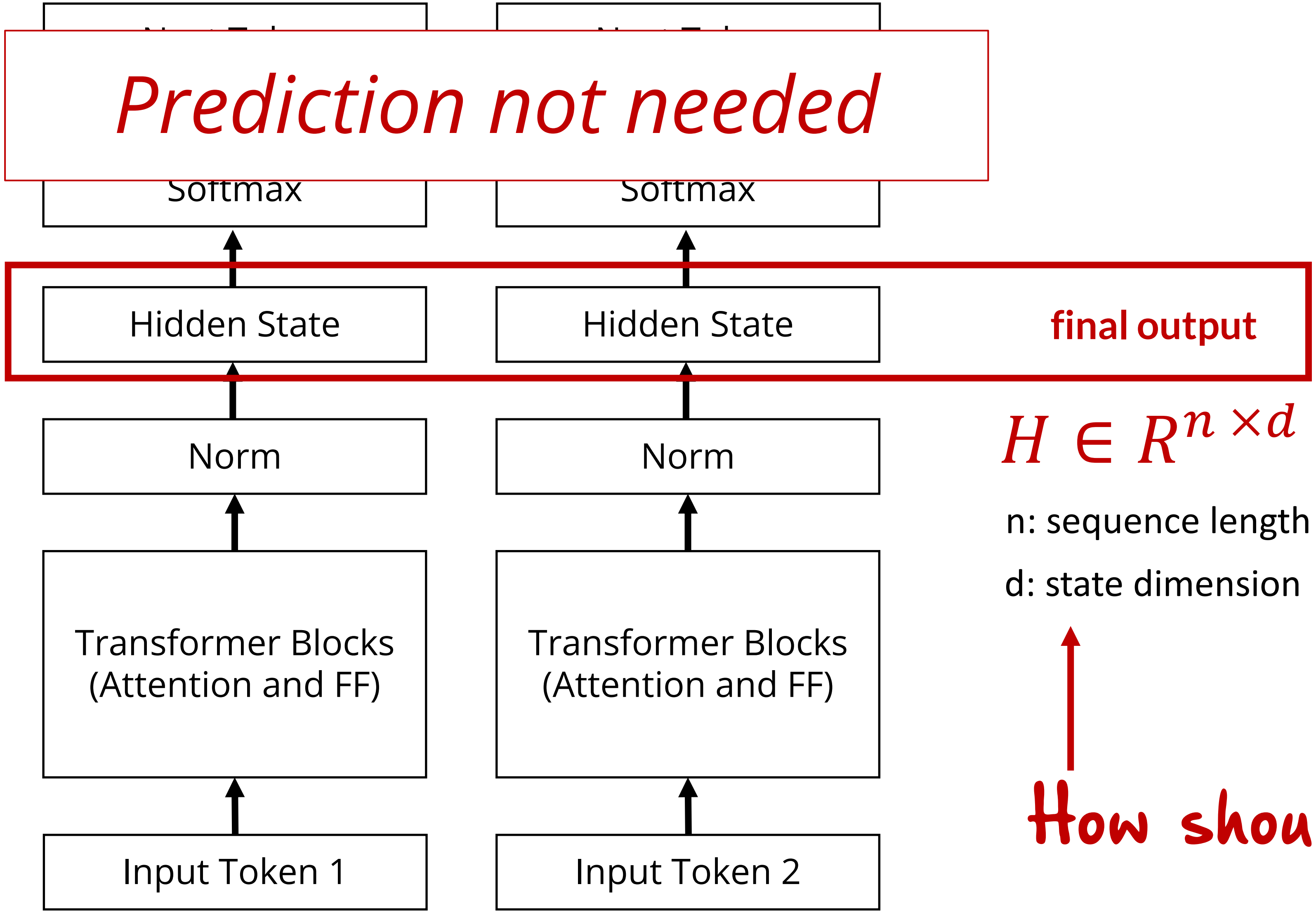
Property 2: Captures dependency



Something before each Feed Forward



Result: Final output captures ***cumulative*** meaning

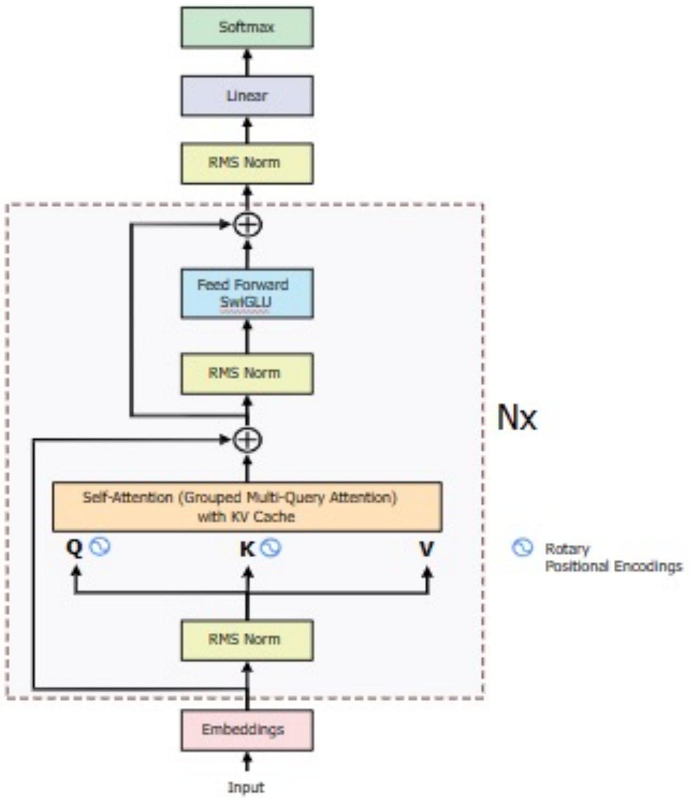


$H \in R^{n \times d}$

n: sequence length

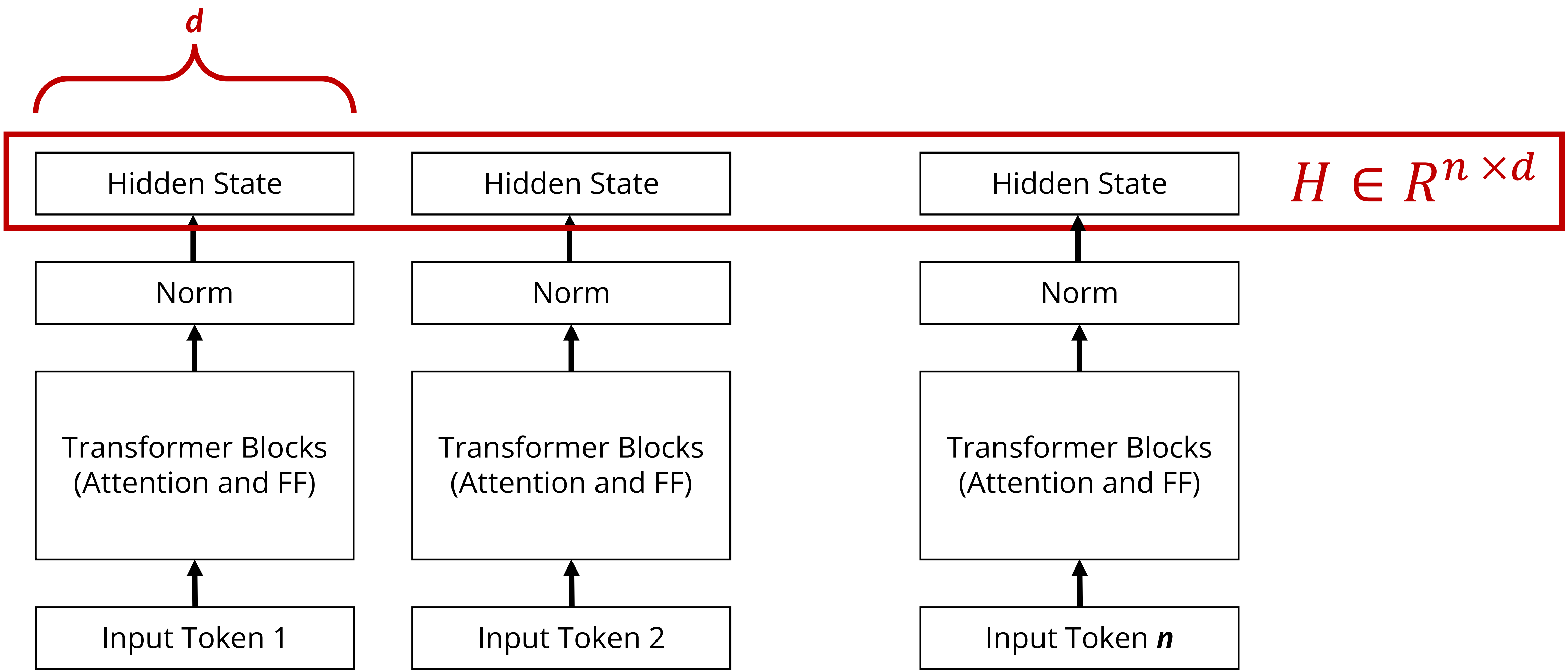
d: state dimension

How should we use it?

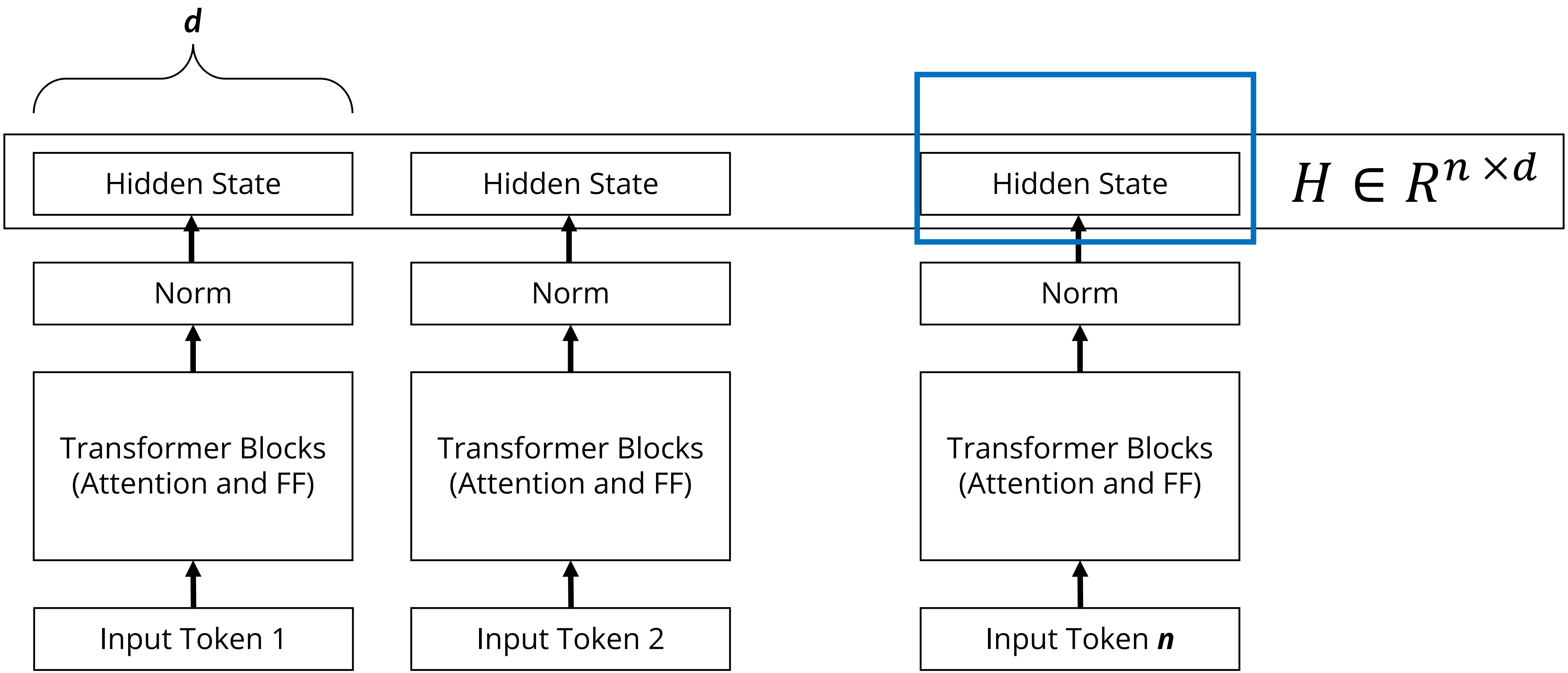


Pooling Methods

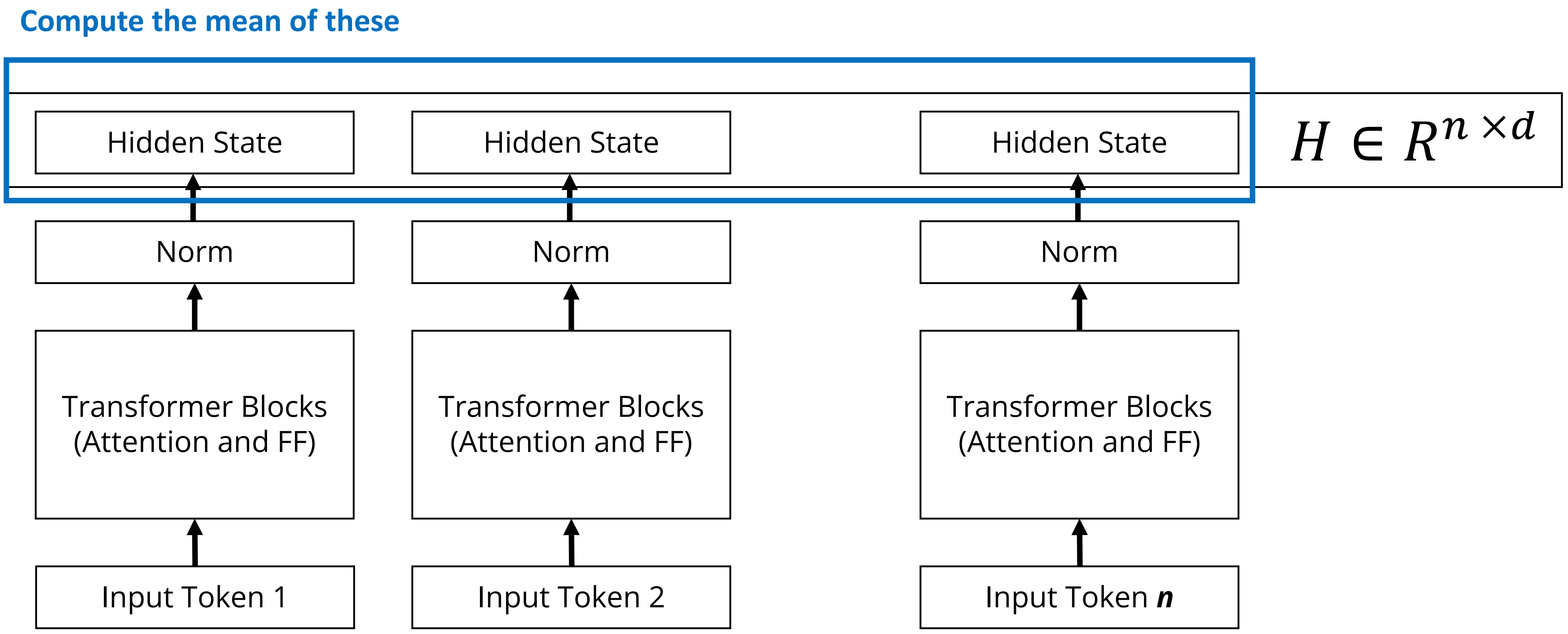
Output hidden states: we will obtain ***size-d*** vector



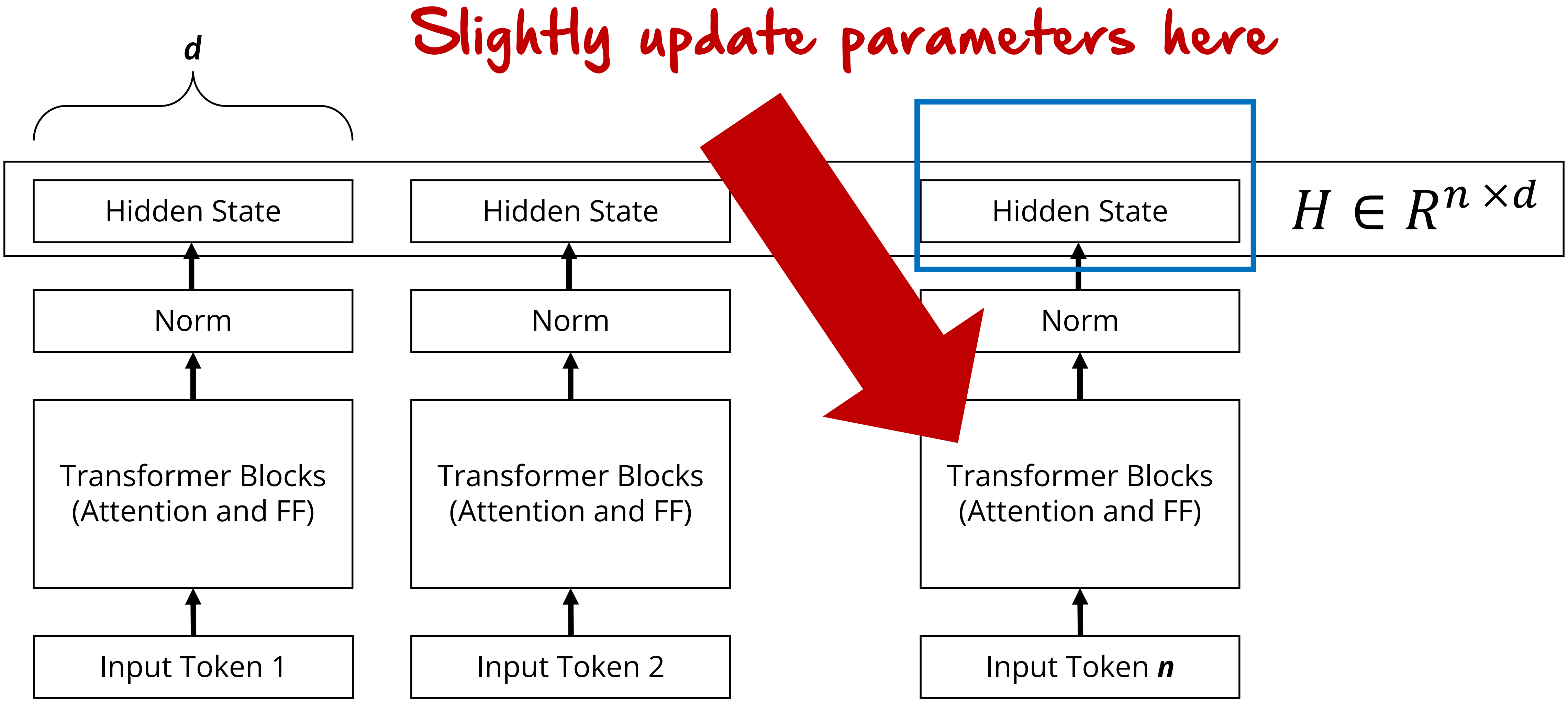
EOS-Last Token Pooling: **H[n, :]**



Mean Pooling: mean(H[1,:], H[2,:], ..., H[n,:])



Fine-tuning


Slightly update parameters here
d
Hidden State
Hidden State
Hidden State
$H \in R^{n \times d}$
Norm
Norm
Norm
Transformer Blocks
(Attention and FF)
Transformer Blocks
(Attention and FF)
Transformer Blocks
(Attention and FF)
Input Token 1
Input Token 2
Input Token ***n***

Fine-tuning via contrastive loss: Basic Idea

- Pull positive pairs (e.g., related query-document) closer
- Push negative pairs (e.g., irrelevant query-document) apart

{**"user_query":** "How to use Microsoft Power BI for data analysis",
"positive_document": "Microsoft Power BI is a sophisticated tool that requires time and practice to master. In this tutorial, we'll show you how to navigate Power BI … *(omitted)* ",
“hard_negative_document”: “Excel is an incredibly powerful tool for managing and analyzing large amounts of data. Our tutorial series focuses on how you…*(omitted)*” }

Contrastive loss: InfoNCE loss

- Pull positive pairs (e.g., related query-document) closer
- Push negative pairs (e.g., irrelevant query-document) apart
- InfoNCE loss: Information Noise Contrastive Estimation loss

$$\boxed{\min \quad \mathbb{L} = -\log \frac{\phi(q_{\text{inst}}^{+}, d^{+})}{\phi(q_{\text{inst}}^{+}, d^{+}) + \sum_{n_i \in \mathbb{N}} (\phi(q_{\text{inst}}^{+}, n_i))}}$$

Contrastive loss: InfoNCE loss

- Pull positive pairs (e.g., related query-document) closer
- Push negative pairs (e.g., irrelevant query-document) apart
- InfoNCE loss: Information Noise Contrastive Estimation loss

$$\min \quad \mathbb{L} = -\boxed{\log \frac{\phi(q_{\text{inst}}^{+}, d^{+})}{\phi(q_{\text{inst}}^{+}, d^{+}) + \sum_{n_i \in \mathbb{N}} (\phi(q_{\text{inst}}^{+}, n_i))}}$$

This whole log will be maximized

Contrastive loss: InfoNCE loss

- Pull positive pairs (e.g., related query-document) closer
- Push negative pairs (e.g., irrelevant query-document) apart
- InfoNCE loss: Information Noise Contrastive Estimation loss

$$\min \ \mathbb{L} = -\log \frac{\boxed{\phi(q_{\text{inst}}^{+}, d^{+})}}{\boxed{\phi(q_{\text{inst}}^{+}, d^{+})} + \boxed{\sum_{n_i \in \mathbb{N}} (\phi(q_{\text{inst}}^{+}, n_i))}}$$

Positive similarity will be maximized

Negative similarity will be minimized

Extract training examples from ChatGPT

Task group: long-short matching
Task definition: Identifying severity level of customer complaints in support tickets
Generated data: { "input_text": "I am writing to express my intense dissatisfaction with one of your products, a TV that has stopped functioning only a month after purchase. This situation yields less satisfaction to me and speaks voluminously about your quality control procedures in assembly lines. I hope this troubling issue etches into your improvement list for invoking earnest attention.", "label": "High Severity", "misleading_label": "Low Severity" }

Task group: short-short matching
Task definition: Provided a movie quote, find the movie title in which it is said.
Generated data: { "input": "I’m going to make him an offer he can’t refuse.", "positive_document": "The Godfather" }

Benefits of fine-tuning

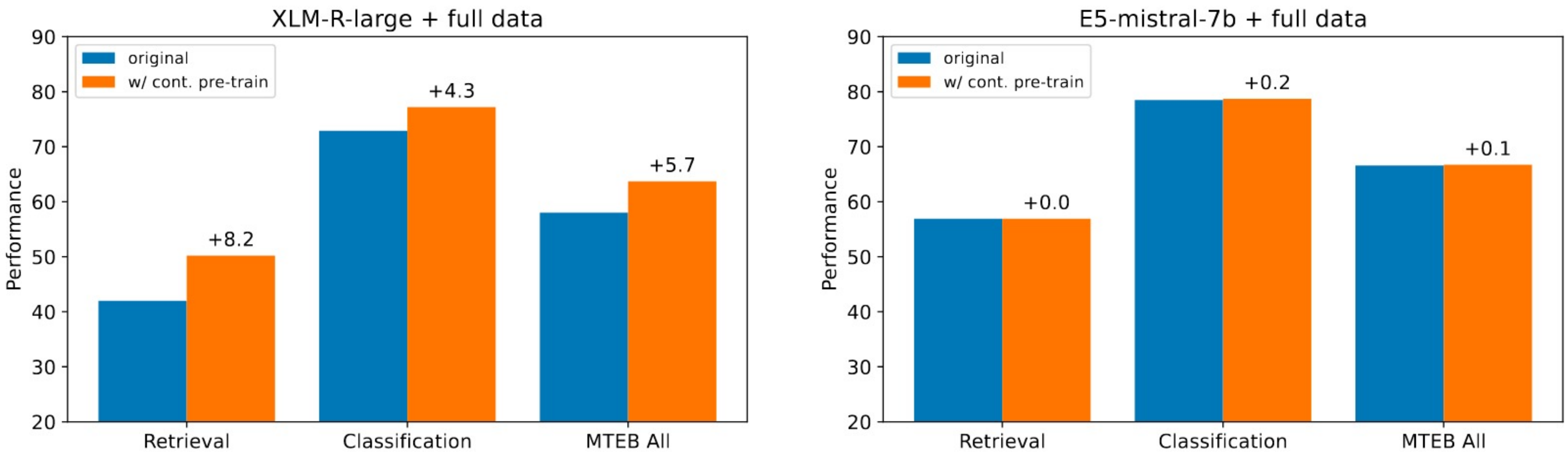


Figure 3: Effects of contrastive pre-training. Detailed numbers are in Appendix Table 7.

Summary

- Large language models can capture semantics accurately
- Can use their internal states as an embedding of the whole document
- Multiple ways to extra internal states
  - EOS-last token, Mean-pooling, Fine-tuning
- Fine-tuning offers more advantage for smaller models

Questions?