Hindi-Small Language Model @Gemma 270M

Utpal Anand, 3rd Year BS-MS

Motivation

- We don’t need Large Language Model every time.
- LLMs require huge computation and High-end GPUs.

QUESTION: But can we have the same performance from models with fewer parameters , mainly trained for specific task?

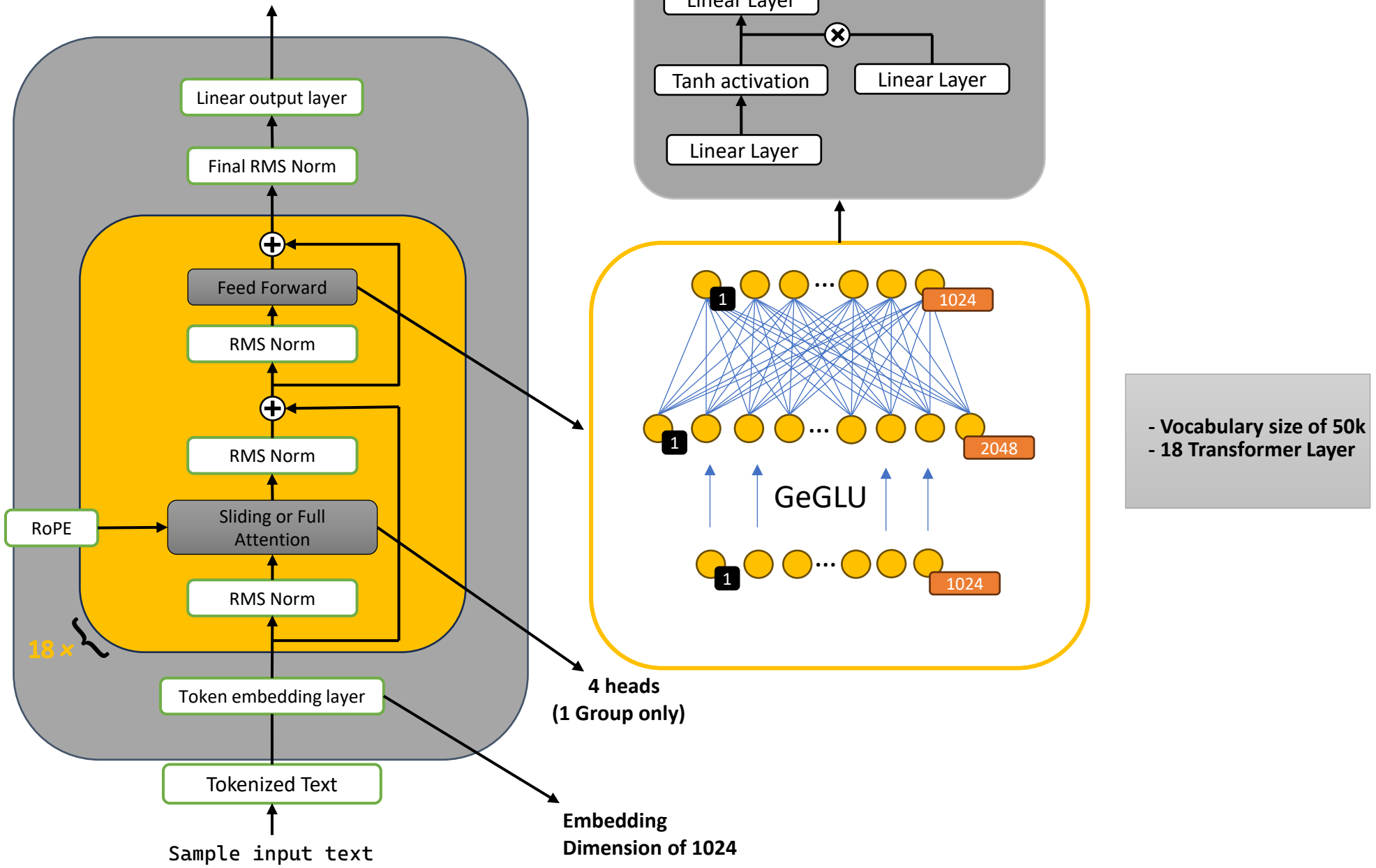
ANSWER: SLMs (≲1–12B params) are not only sufficient but often superior for agentic workloads (e.g. RAG)

NVIDIA: SLMs are the future of agentic AI and edge inference, emphasizing cost/latency/energy advantages and the role of guided decoding.

References

1. Small Language Models for Agentic Systems: A Survey of Architectures, Capabilities, and Deployment Trade-offs (https://arxiv.org/pdf/2510.03847)
2. Small Language Models are the Future of Agentic AI (https://arxiv.org/pdf/2506.02153)

Gemma 270M@Hindi SLM



$$\begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1d} \\ q_{21} & q_{22} & \cdots & q_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{Nd} \end{pmatrix} \times \begin{pmatrix} k_{11} & k_{21} & \cdots & k_{N1} \\ k_{12} & k_{22} & \cdots & k_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1d} & k_{2d} & \cdots & k_{Nd} \end{pmatrix} \longrightarrow \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1N} \\ s_{21} & s_{22} & \cdots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NN} \end{pmatrix}$$

$Q\ (N \times d)$ $K^T\ (d \times N)$ $QK^T\ (N \times N)$

$$s_{11} = q_{11}k_{11} + q_{12}k_{12} + \cdots + q_{1d}k_{1d}$$

Attention score between Query and Key of Token 1

$$\begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1d} \\ q_{21} & q_{22} & \cdots & q_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{Nd} \end{pmatrix} \times \begin{pmatrix} k_{11} & k_{21} & \cdots & k_{N1} \\ k_{12} & k_{22} & \cdots & k_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1d} & k_{2d} & \cdots & k_{Nd} \end{pmatrix} \longrightarrow \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1N} \\ s_{21} & s_{22} & \cdots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NN} \end{pmatrix}$$

$Q\ (N \times d)$ $K^T\ (d \times N)$ $QK^T\ (N \times N)$

$$s_{12} = q_{11}k_{21} + q_{12}k_{22} + \cdots + q_{1d}k_{2d}$$

Attention score between Query of Token 1 and Key of Token 2

$$\begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1d} \\ q_{21} & q_{22} & \cdots & q_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{Nd} \end{pmatrix} \times \begin{pmatrix} k_{11} & k_{21} & \cdots & k_{N1} \\ k_{12} & k_{22} & \cdots & k_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1d} & k_{2d} & \cdots & k_{Nd} \end{pmatrix} \longrightarrow \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1N} \\ s_{21} & s_{22} & \cdots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NN} \end{pmatrix}$$

$Q\ (N \times d)$ $K^T\ (d \times N)$ $QK^T\ (N \times N)$

$$s_{21} = q_{21}k_{11} + q_{22}k_{12} + \cdots + q_{2d}k_{1d}$$

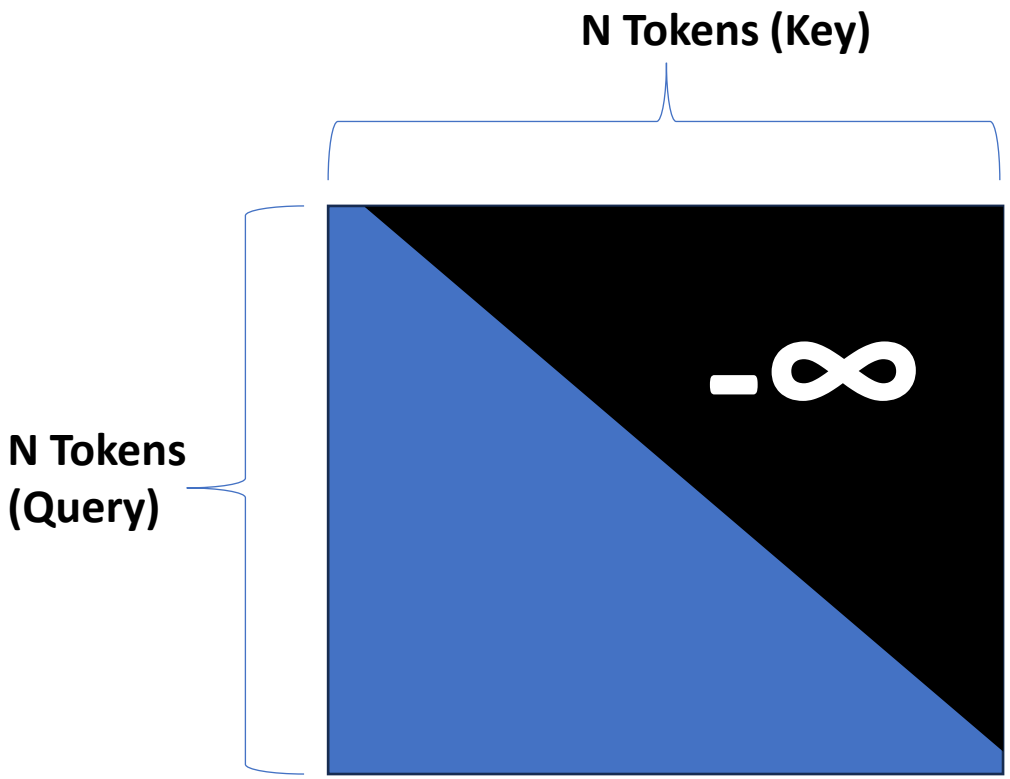
Attention score between Query of Token 2 and Key of Token 1

	Key1	Key2	Key j	KeyN
Query1	s_{11}	s_{12}	$\cdots$	s_{1N}
Query2	s_{21}	s_{22}	$\cdots$	s_{2N}
Query i	$\vdots$	$\vdots$	s_{ij}	$\vdots$
QueryN	s_{N1}	s_{N2}	$\cdots$	s_{NN}

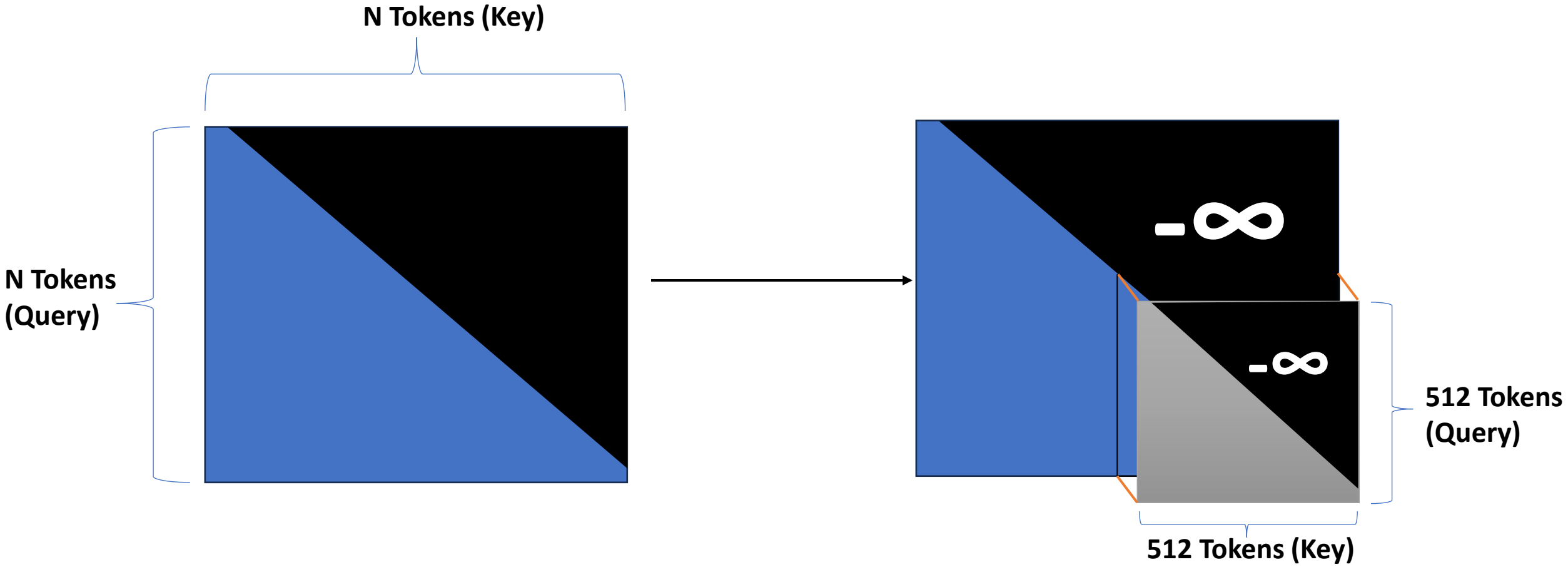
Full Attention Score (QK^T matrix) for one head only

Sliding window attention (Masked Attention Score)

Each token only pays attention to the 512 tokens immediately before it.

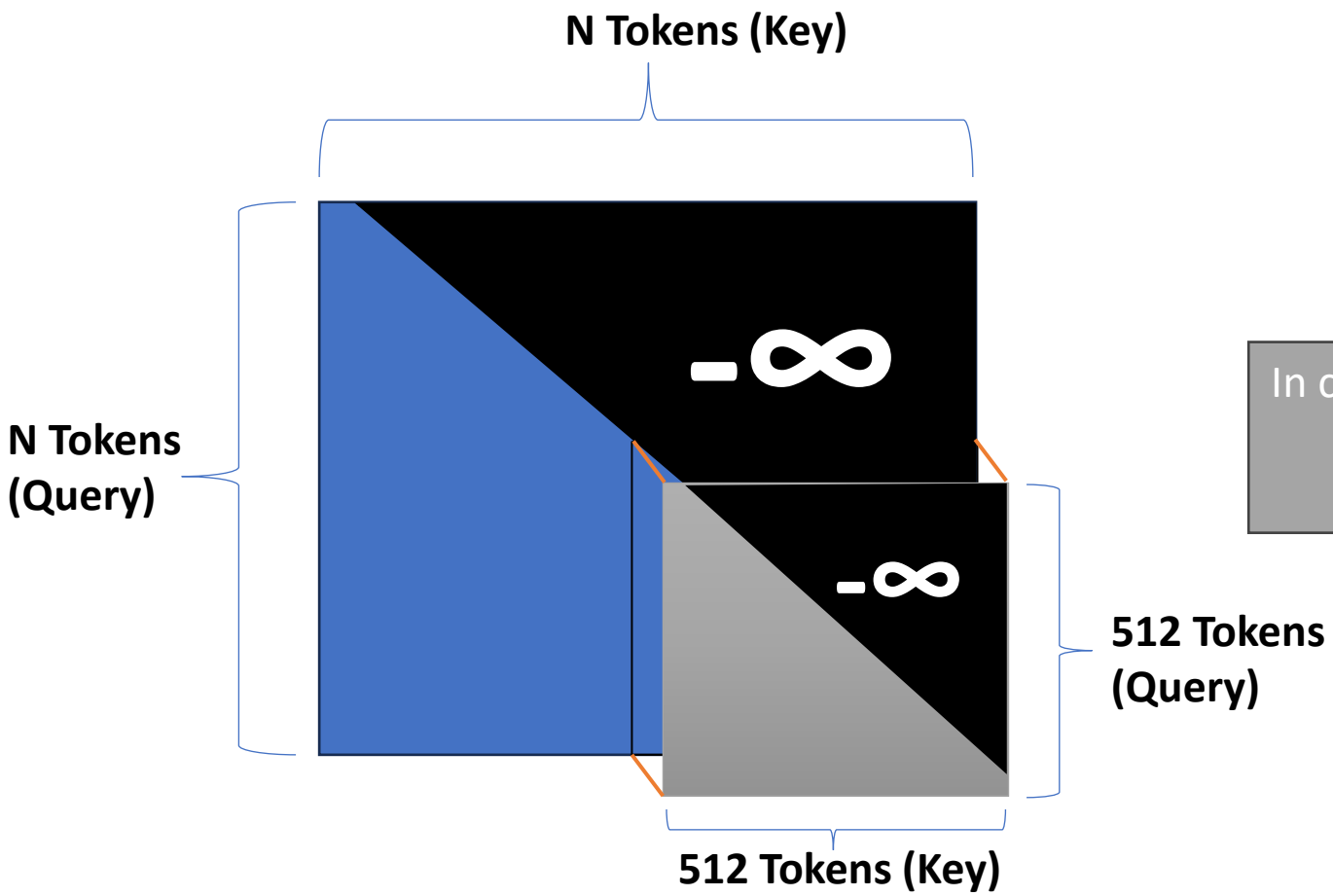


In case of N less than 512,
it will consider all token.
It behaves like Full Attention!


N Tokens (Key)
N Tokens (Query)
-∞
-∞
512 Tokens (Query)
512 Tokens (Key)

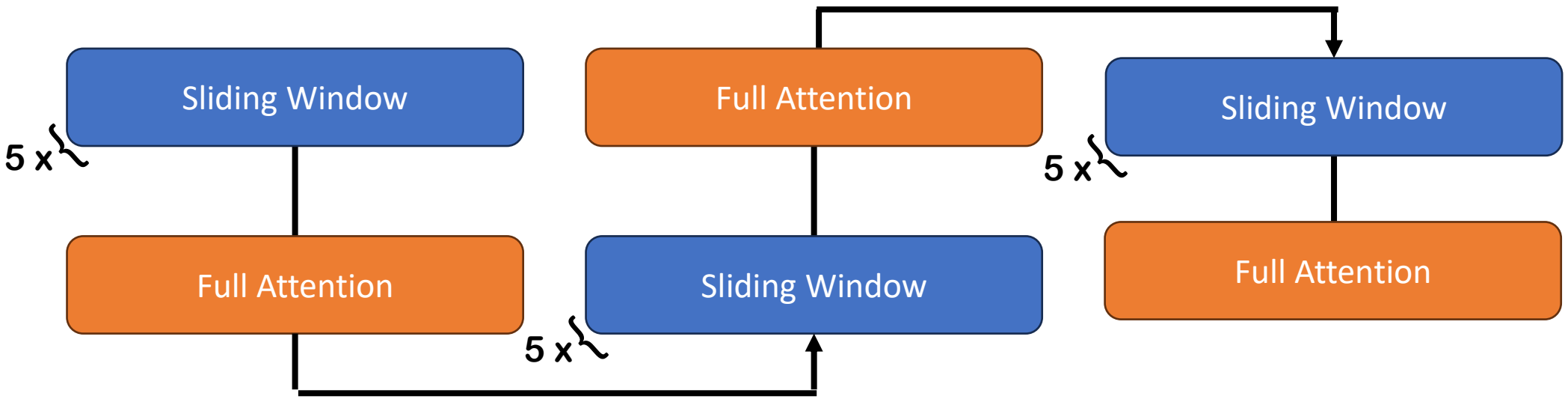
Sliding window attention (Masked Attention Score)

Each token only pays attention to the 512 tokens immediately before it.



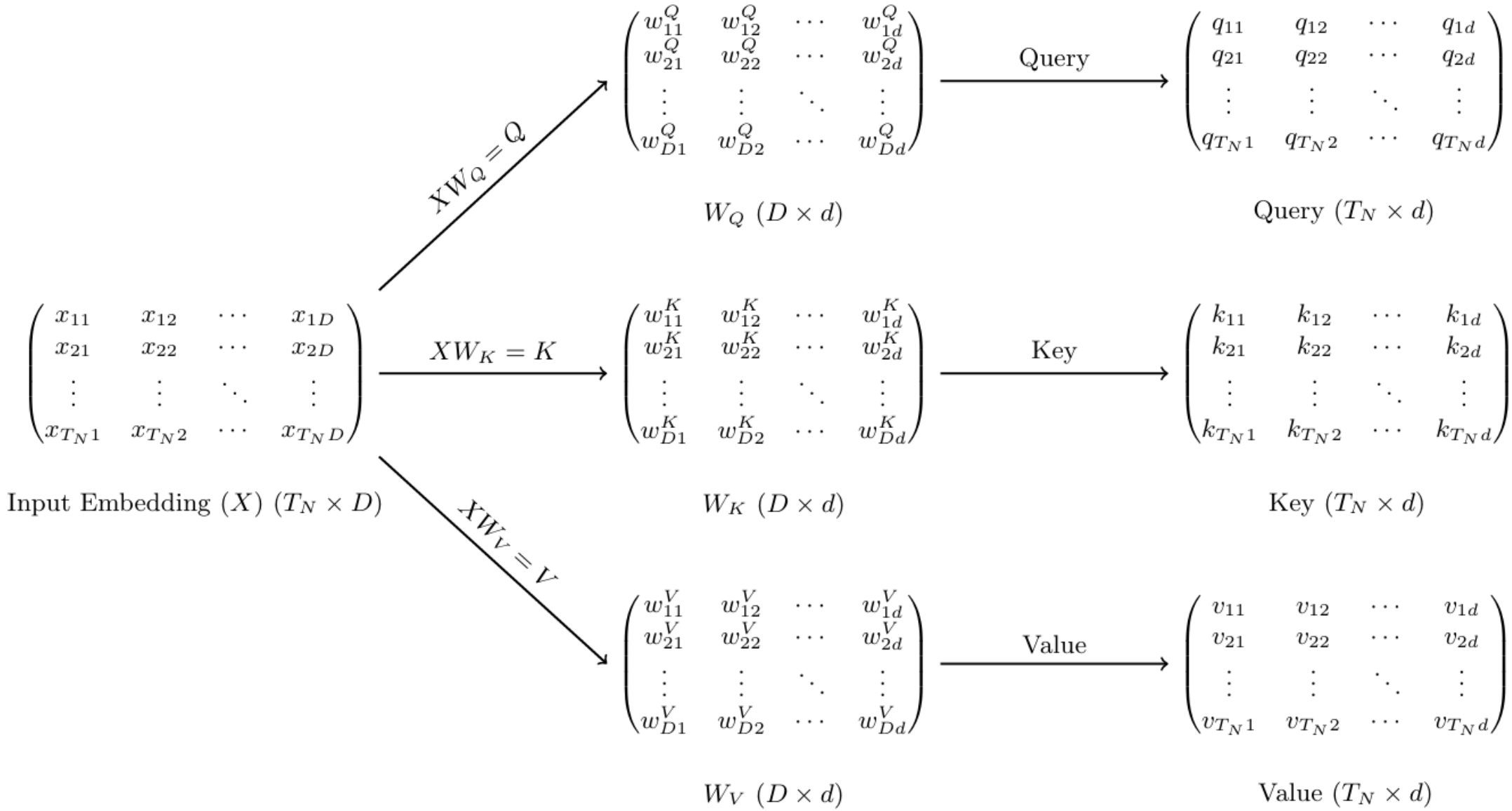
In case of N more than 512, it will consider only previous 512 tokens.

Pipeline of 18 Transformer Stack



ROTARY POSITION EMBEDDING (RoPE)

But before this……..



If we have n embedding dimension…..

We will make pairs of two dimensions like:

1 & 2

3 & 4

5 & 6

.

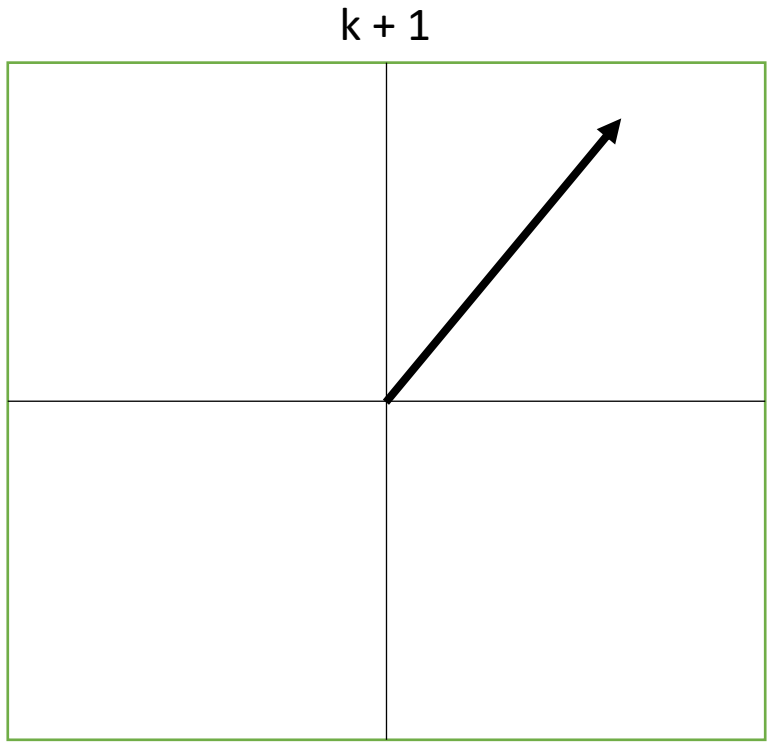
.

.

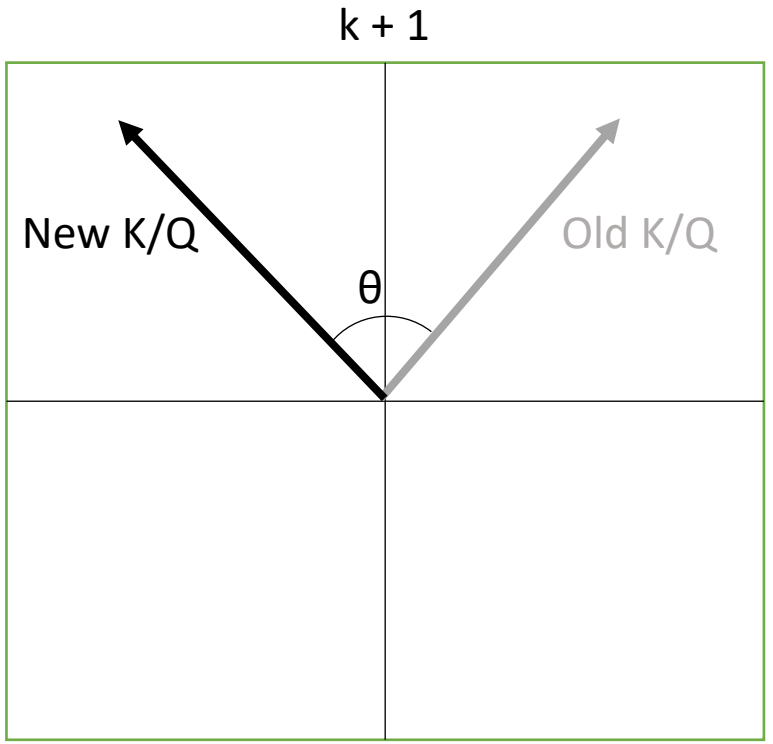
n-1 & n

Let’s take one pair : k & k+1 → We will rotate Query and Key by some angle with respect to position and dimension.

For query and key both:



k $\rightarrow R =$ $\begin{bmatrix} \cos\theta_{p,k} & -\sin\theta_{p,k} \\ \sin\theta_{p,k} & \cos\theta_{p,k} \end{bmatrix}$ $\rightarrow$



$$\text{For each pair } (k, k+1): \quad \theta_{p,k} = \frac{p}{\theta_{\text{base}}^{k/d_{\text{head}}}}$$

For sliding attention: $\theta_{\text{base}} = 10^4 = 10{,}000$

For full attention: $\theta_{\text{base}} = 10^6 = 1{,}000{,}000$

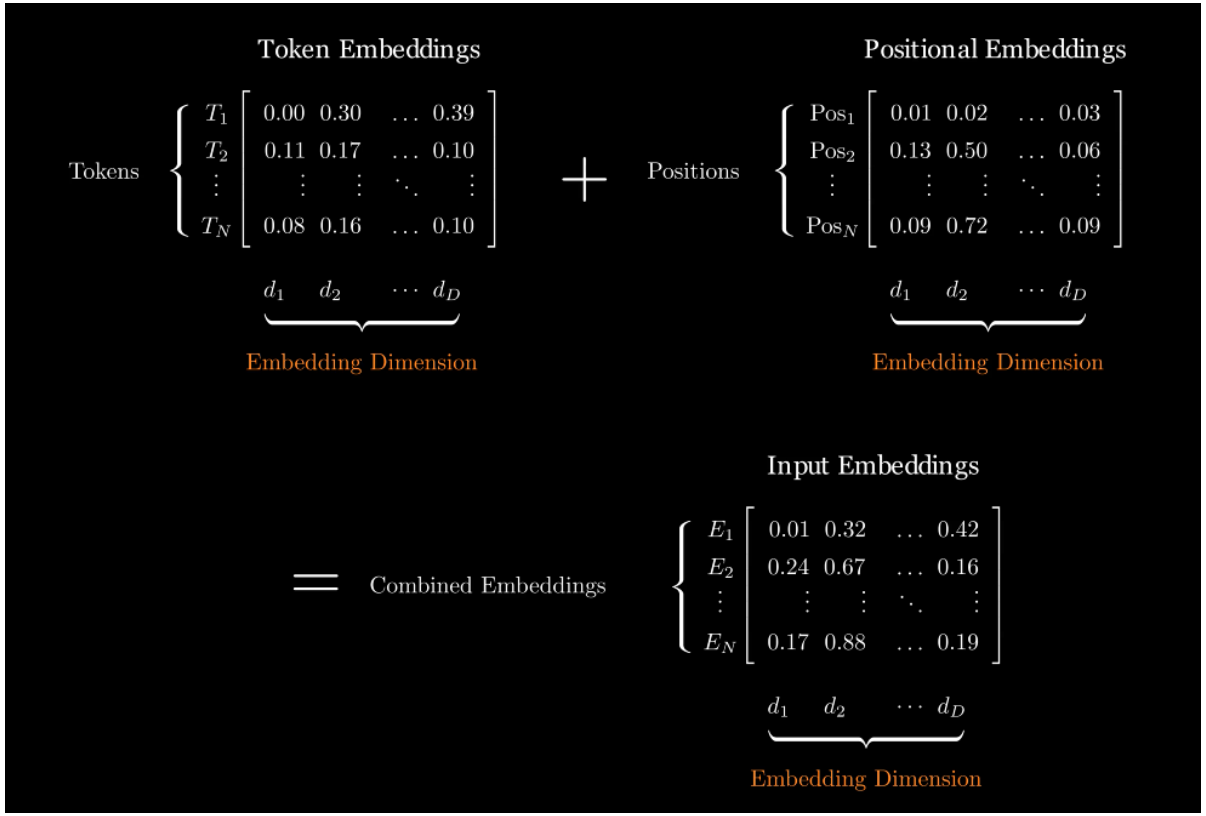
While in Vaswani 2017:

Since our model contains no recurrence and no convolution, in order for the model to make use of the order of the sequence, we must inject some information about the relative or absolute position of the tokens in the sequence. To this end, we add "positional encodings" to the input embeddings at the bottoms of the encoder and decoder stacks. The positional encodings have the same dimension d_{model} as the embeddings, so that the two can be summed. There are many choices of positional encodings, learned and fixed [9].

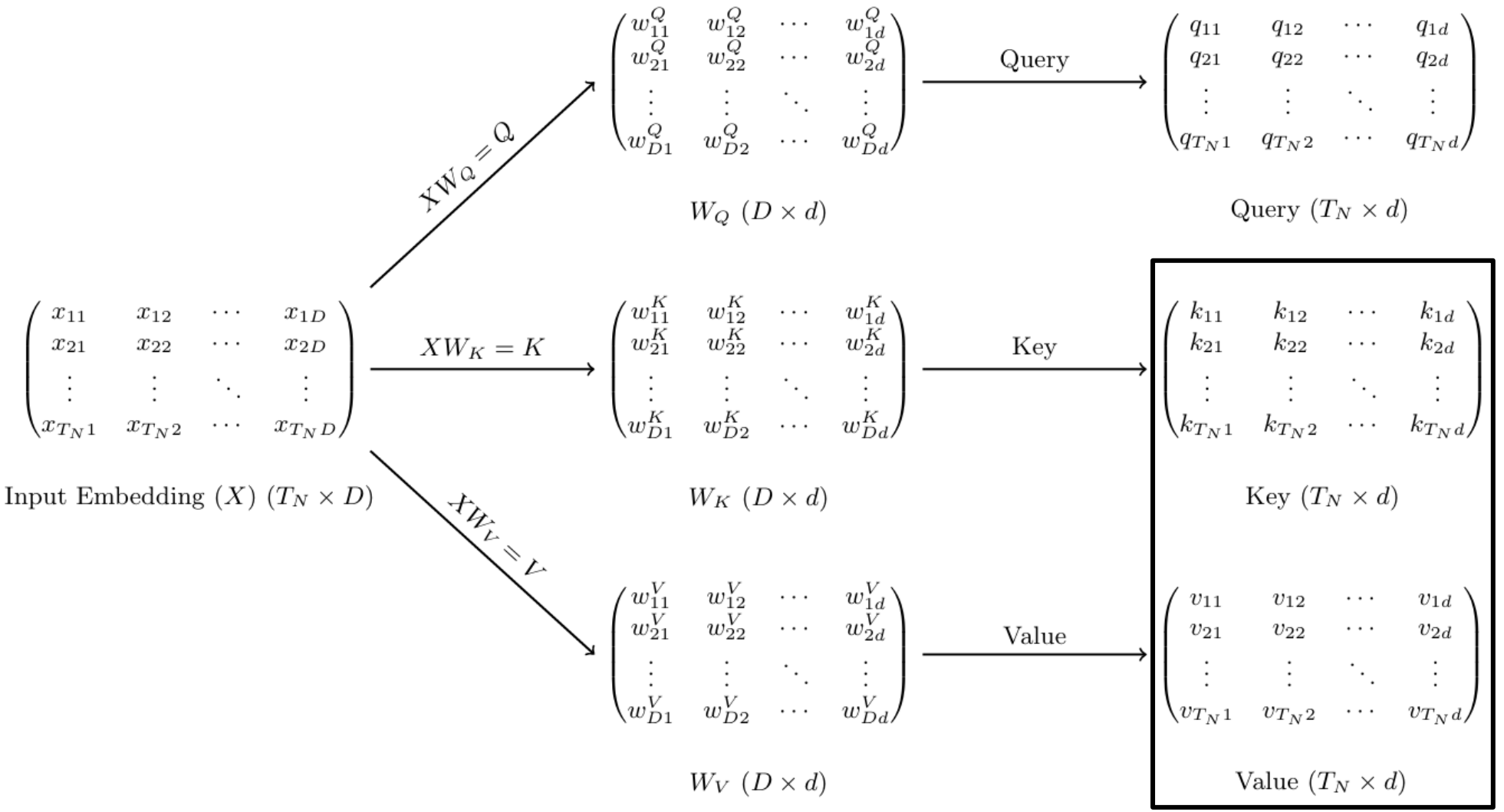
In this work, we use sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

where pos is the position and i is the dimension. That is, each dimension of the positional encoding corresponds to a sinusoid. The wavelengths form a geometric progression from 2π to $10000 \cdot 2\pi$. We chose this function because we hypothesized it would allow the model to easily learn to attend by relative positions, since for any fixed offset k, PE_{pos+k} can be represented as a linear function of PE_{pos}.

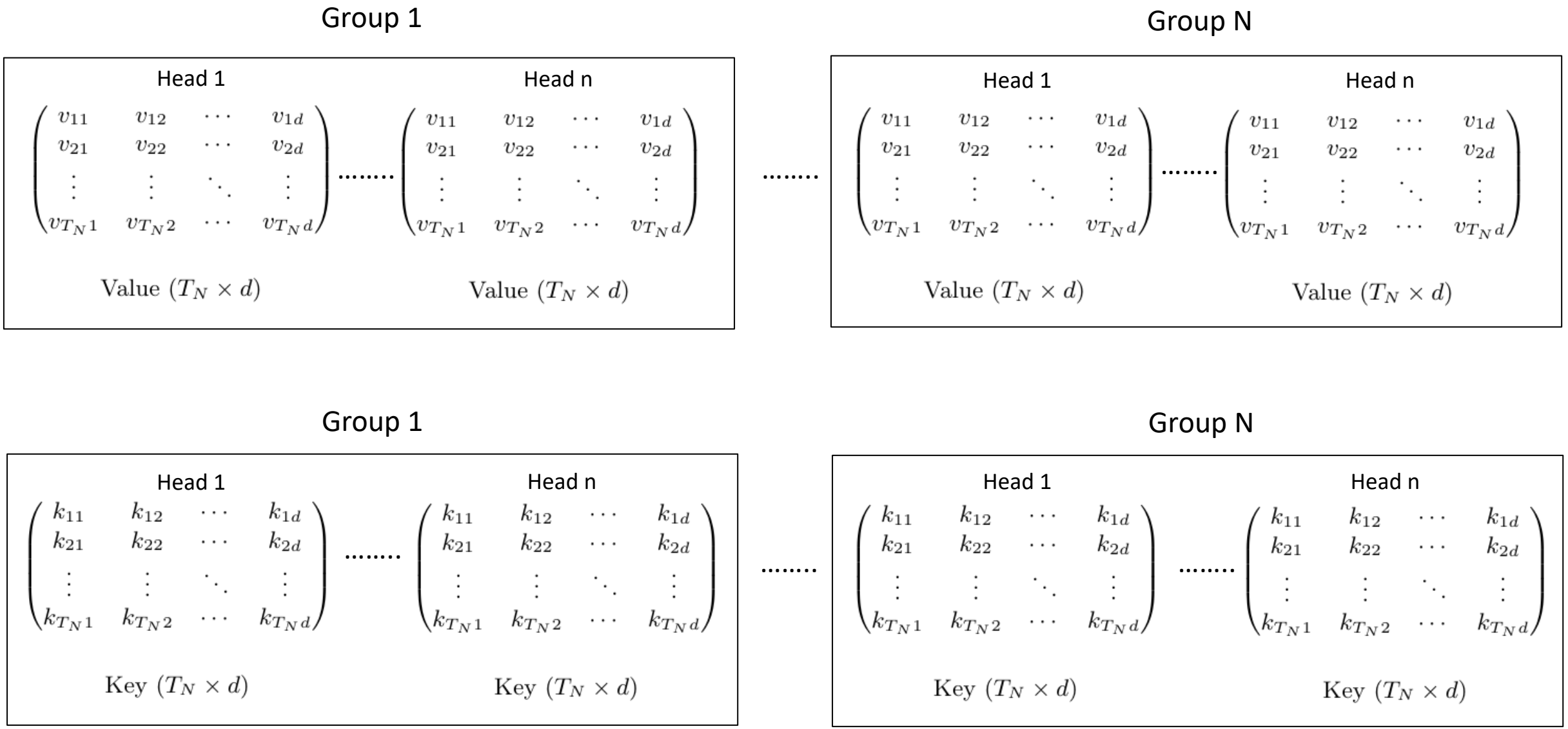


Multi/Grouped Query Attention



This is common across heads.

For KV cache Reduction: Heads in a group share the same weight (matrix).



Gemma 270M

-32k(32768) context window
-vocab size = 262144
-140 languages
-Trained over 6 trillion tokens

Gemma 270M @Hindi

-1024 context window
-vocab size = 50000
-Hindi only
-Trained over 1.2 billion tokens

Same across both

```
"num_attention_heads": 4,
"num_hidden_layers": 18,
"num_key_value_heads": 1,
```

Block size is 128 because it has seen maximum only 128 token at once while training!
[effective context per sample]

First Training:
With the small dataset of around 50 million tokens.
Iterations = 200k (≈ 262 epochs)

-Batch size = 16

-Block size = 128(context size)

-32 Gradient Accumulation steps

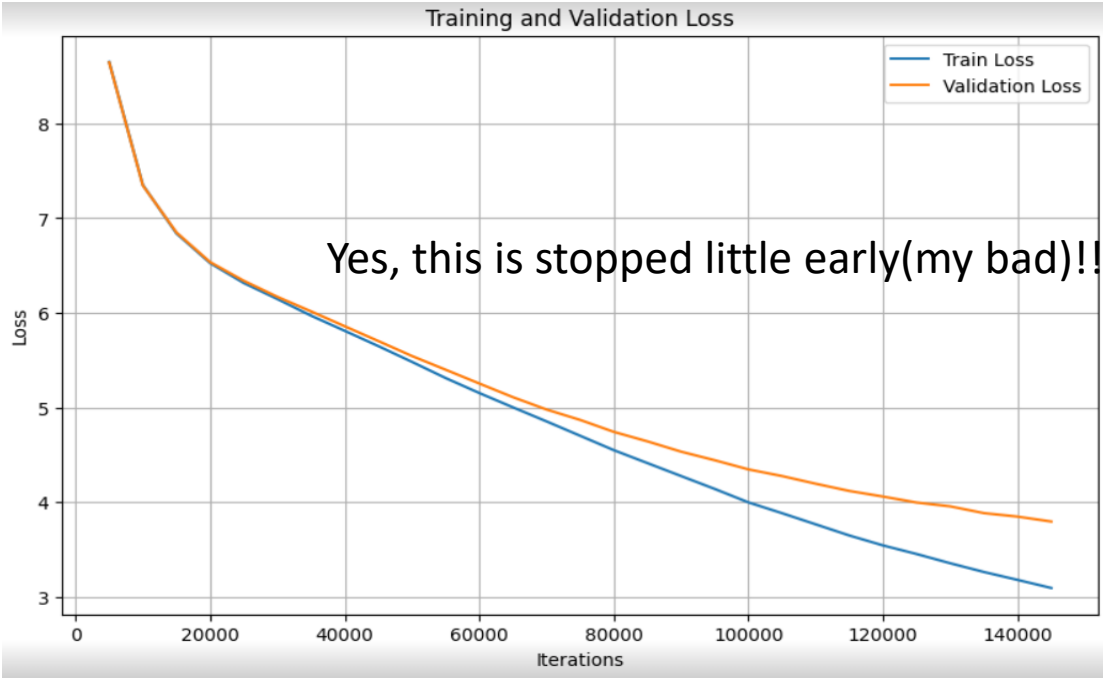
-Learning rate = 1e-4

-optimizer = torch.optim.AdamW(model.parameters(), lr=learning_rate, betas=(0.9, 0.95), weight_decay=0.1, eps=1e-9)

Tokenizer trained on this dataset only!
With 50k Vocabulary size!

Dataset: https://www.kaggle.com/datasets/disisbig/hindi-wikipedia-articles-172k

Let's see some inference----



temperature=0.7,
top_k=20

It is repeating same tokens!

भारत में बारिश कब होगी ? ? ? ? ? ? ? क्या ? क्या ? ? ? क्या ? ? ? ? ? ? ? ? ? ? ? ? क्या ? क्या ? ? क्या ? ? ? ? ? ? क्या ? ? क्या ? ? ? ? ? ? ? क्या ? ? क्या ? ? ? ? ? ? ? क्या ? ? ? ? ? ? मेरा जन्म सिद्ध अधिकार ? ? ? ? ? ? क्या ? ? ? क्या क्या तुम मुझसे सात दिन ? क्या कहै : क्या तुम मुझसे मत पूछो। क्या ? ? ? ? ? ? क्या तुम्हारी प्रार्थना क्या ? क्या ? क्या तुम मुझसे पूछा जाय, यह कैसे ? ? क्या तुम कौन ? क्या , हम किसी ने ? ? क्या तुम माँग सकता है? क्या तुम मुझसे पूछा जाय, यह कैसे ? क्या बात ? क्या तुम मुझसे कहा ? क्या तुम्हारी कविता क्यों ? क्या क्या ? क्या तुम्हारी कविता ? क्या क्या ? ? क्या यह तो क्या ? मैंने उसे ? क्या

temperature=0.7, top_k=20

भारत में बारिश कब होगी । पुलिस ने बताया कि सरकार ने कहा कि वह अपनी ही में नहीं दिया । वहीं , ' जब तक ' की मौत - Amar U j ob : राहुल और राज्य सरकार के बीच भी हैं । वे यह भी एक नए विंडो में एक और कहा है कि यह भी कोई बात नहीं हो रहा है , क्योंकि यह भी नहीं है । इसके अलावा , एक ही नहीं । लेकिन अगर आप सभी को भी कुछ नहीं कर सकता । यह भी अपने एक बार फिर से भी एक ही नहीं कर रहे हैं । इस बारे में तो वो एक बार फिर भी हो सकता है । वो मेरे घर में कोई भी कुछ और भी नहीं । एक ऐसी बात तो आपको कुछ नहीं है , लेकिन मैं भी नहीं हूं । वह भी नहीं होता । मैं क्या आप अपने अपने नाम पर मैं आपको बता दें हूँ और मेरे साथ ही नहीं । जब भी कोई ही नहीं नहीं है । तो नहीं तो आप इस पर आप अपने साथ ही नहीं करते समय से । यह बात है कि ये भी एक ही है । यह भी कुछ भी नहीं है

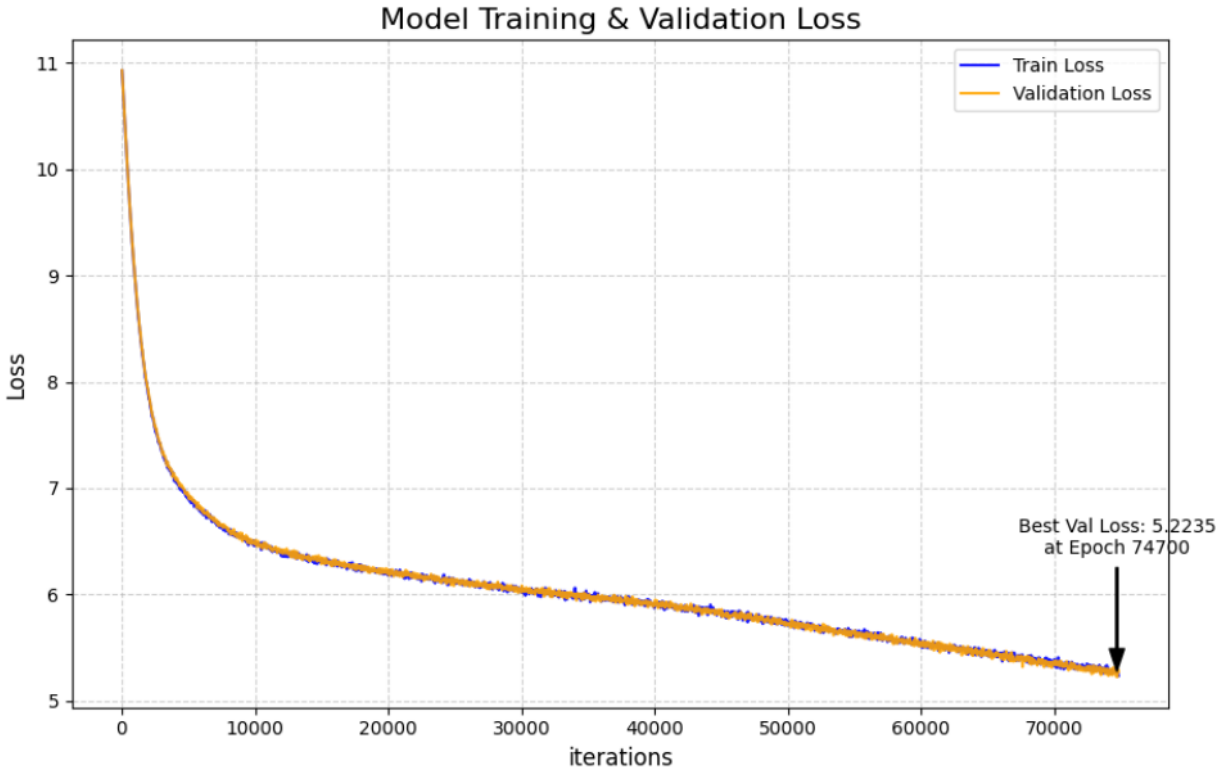
temperature=0.7,
top_p=0.4,
repetition_penalty=1.2

भारत में बारिश कब होगी । इस बात को लेकर इस तरह से पहले ही नहीं किया गया है । यह एक साथ ही समय पर कुछ ऐसा हो सकता है कि वह अपने घर के लिए भी आपको पता चला है , लेकिन यह कोई नहीं होगा । इससे पहले की वजह से अधिक जानकारी दी गई थी । अगर आप भी पढ़ें : क्या आपके लिए तो ये हैं तो आप हमें फेसबुक और भी कम करने का भी होता है । इसके बाद अब तक हम सब लोग अपनी पत्नी को भी काम कर सकते हैं । आज हम सभी को दूर करें । आपका स्वागत है । यह बहुत ज्यादा समय ही एक नए विंडो में है जो एक ही है , जब मैं अपने घर में नहीं है , वो अपने घर में यह है कि उन्हें उनकी मदद करता है । मुझे भी किसी भी तरह से नहीं है । मैंने कहा जाता है कि वे ही कुछ भी नहीं है । उसके बाद उस दिन पहले ही ही अपना न करें । आप जानते हैं , लेकिन इन दिनों से दूर हो रहा था । इसे एक ही एक ही सही है । इस दौरान कई बार फिर

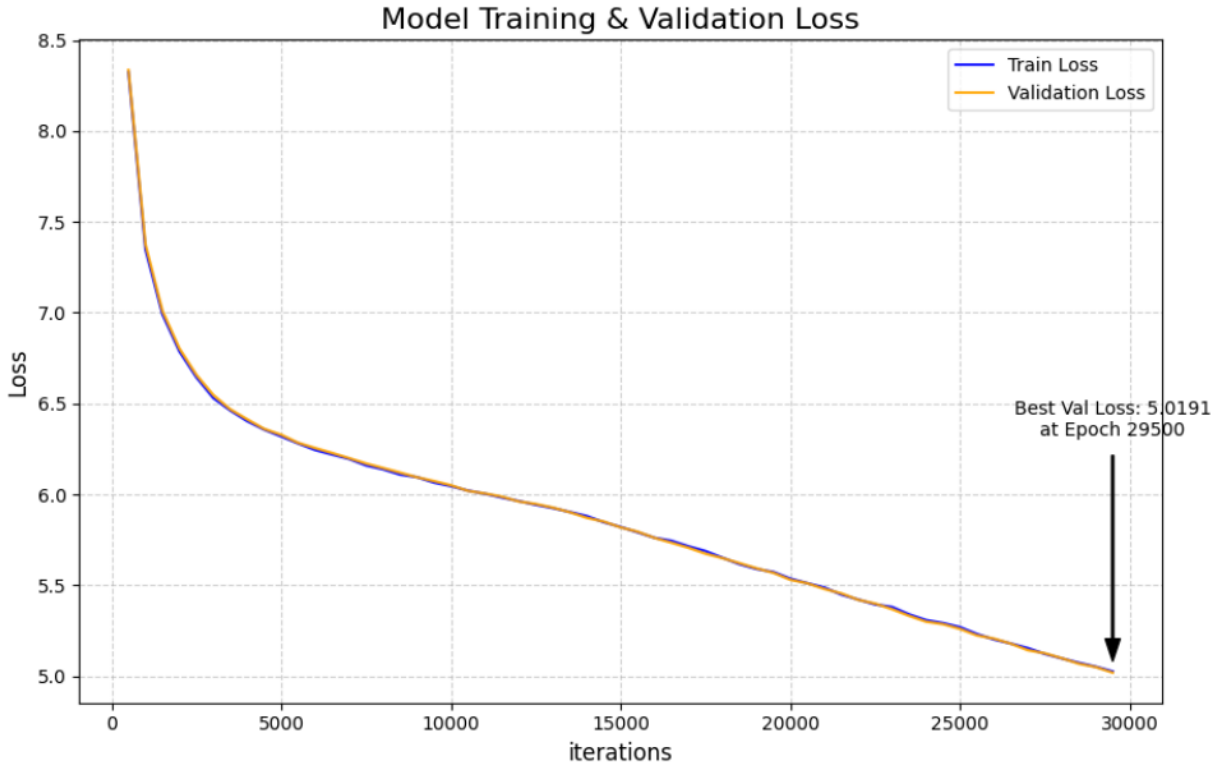
Second phase of training: This time, it is on CC100 hindi devanagri dataset. (Contains more than 1 Billion tokens)

Same learning rate = 1e-4!

World Size: 4, Batch per GPU: 48, Effective Batch Size: 192
max_iters: 84000
warmup_steps: 8000
Gradient Accumulation = 8
≈ 13 hours on 4 NVIDIA A30



World Size: 12, Batch per GPU: 48, Effective Batch Size: 576
max_iters: 30000
warmup_steps: 8000
Gradient Accumulation = 4
≈ 4.5 hours on 12 NVIDIA A30



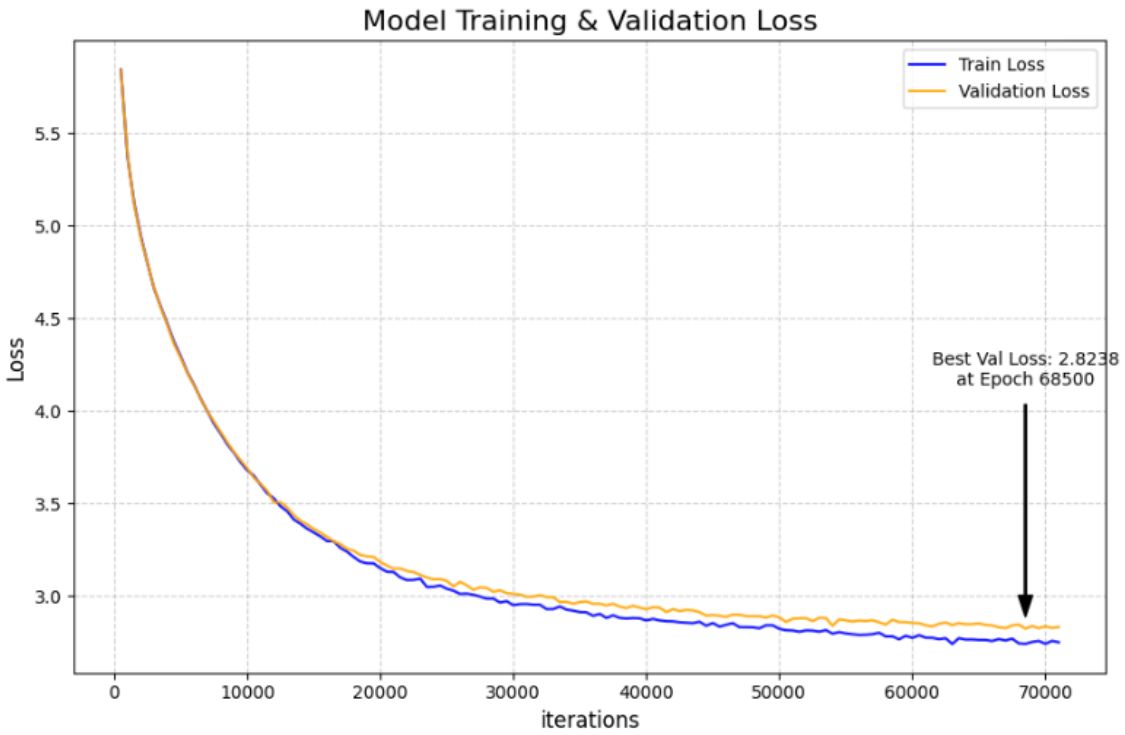
```
temperature=0.8,
top_p=0.5,
repetition_penalty=1.2
```

-It is not repeating because we added a repetition penalty.

भारत में बारिश कब होगी ? ? - नवम्बर को व्याकुल हो गया, लेकिन क्या करना है? वह तो एक सामान्य आदमी है, जो खुद को दोहराता है। वे उसे अपनी कहानी बताते हैं और महसूस करते हुए भी उन दोनों की हिम्मत न करने लगें । इसके बाद जब यह बात पता चलती है तो उसने सच्चाई का परिचय देते हुए कहा कि अगर वह अपने से कम ख़र्च नहीं होती तो ऐसी हालत है कि उसका दिल टूट जायेगा। अगर तुम लोग कुछ समझ कर ही जाते हो, तो फिर वह झूठ तूफ़ान आ गई होगी। इसलिए उन्होंने देखा कि जैसे वह लोग मारे जायेंगे ! उसके पिता ने अपनी गलती का गलत ब पूछा तो बस यही था के लिए मुझे माफ कर दिया गया। मैंने तुझे ढूं ढ़ ते नजर आये " और नल द गी बोली, 'हे मेरे पास मेरी बीवी है।' ... ’ स्मिथ की कमी और थोड़ा दूर रहना चाहिए "। तेरे नाम के इस नियम को देखकर राजा ने पूछा, 'ऐ लाने लगे - क्या तुमने मुझसे कहा ?" कहने पर मैं तुम्हें कहता हूं कि तुम्हारा ? मेरा कोई मतलब ना होगा ? ' मैं ने मुझसे पूछ ो। मैं क्षमा चाहता हूँ और अधिक रू

Third phase of training:

World Size: 14, Batch per GPU: 48, Effective Batch Size: 672
max_iters: 170000 ← But this was not completed !
warmup_steps: 8000
Gradient Accumlation = 2
New learning rate = 3e-4 ←3 times more than earlier
For 110k iteration : **≈ 14-15 hours on 14 NVIDIA A30**



I saved models because of checkpointing and renamed one model to save (because it was overwriting same file)!
The program was crashed around 110k Iterations. Due to no force flushing of logs, I didn't receive loss log after 71k iterations!

Saved in between one by renaming!

temperature=0.7, top_k=20

भारत में बारिश कब होगी । यह एक बार फिर से चीन में एक बड़ी समस्या है । चीन के एक और खेल की तरह है । भारत और ईरान के बीच भारत के साथ भारत - पाकिस्तान , चीन , जापान और ब्रिटेन के बीच की एक शानदार प्रदर्शन का हिस्सा , भारत के लिए सबसे बड़ा बदलाव , पाकिस्तान के साथ - साथ ऐसा किया जा रहा है कि भारत के साथ भारत में भी भारत का भविष्य निधि नहीं हुआ है । भारत और पाकिस्तान के बीच भी कई देशों में ही है । भारत की सबसे बड़ी जीत , बांग्लादेश में भारत को लगा ओ यू . पी . ने कहा कि भारत के साथ काम करते हुए एक और दूसरे से बातचीत कर रहे हैं । चीन का एक नया मोड़ पर है । भारत के लिए पाकिस्तान में एक और भारतीय खिलाड़ी बनने के लिए बेहद आसान है । भारत के लिए भारत में सबसे ज्यादा मुनाफा एक और रिकॉर्ड बनाने के लिए सबसे ज्यादा नुकसान है । भारत और पाकिस्तान के बीच है । इस समय भारतीय मूल के सबसे अमीर , चीन का भारत में सबसे अधिक है भारत की अर्थव्यवस्था के लिए सबसे ज्यादा नुकसान होगा

temperature=0.7,
top_p=0.4,
repetition_penalty=1.2

भारत में बारिश कब होगी ? (1) एक बार फिर से भारत का पहला मुकाबला है , जब भी वह अपने देश के लिए और सबसे अच्छा विकल्प है । 3 . भारत की अर्थव्यवस्था को किस तरह से समाप्त करने वाले हैं ? (2) भारत सरकार ने विश्व कप में भारत के साथ मिलकर काम किया था ? (4) भारत के संविधान में कौन - सा भार : (5) भारत के किस प्रकार की जनसंख्या को विकसित करना होगा कि भारत में कुल 6 प्रतिशत तक नहीं होता है ? (7) पाकिस्तान में चीन की सीमा पर भारत के प्रधानमंत्री नरेंद्र मोदी ने कहा कि यह पहला स्थान हासिल कर रहा है क्योंकि उन्होंने कहा कि हम लोगों को भारत के विकास के लिए भारत में सबसे ज्यादा पढ़े गए हैं । इस दौरान कई राज्यों में दुनिया भर में एक दूसरे के प्रति संवेदना हो गया है । लेकिन वे किसी भी राजनीतिक दलों के बीच होने वाली थी । ये लोग इस बात की जानकारी मिली है कि भारत के पास भारत में भी भारत में निवेश करेगी । (3) भारत में सबसे अधिक है । भारत के राष्ट्रपति चुनाव

Not so good, I think loss is still high!

This is something good!

temperature=0.7, top_k=20

भारत में बारिश कब होगी इसकी संभावना कम ही है । यहां का मौसम शुष्क रहता है । यहां की नमी से लोग परेशान रहते हैं । मौसम शुष्क रहने के कारण यहां का अधिकतम तापमान 35 डिग्री तक पहुंच रहा है । इसके चलते यहां का अधिकतम तापमान 32 डिग्री से नीचे पहुंच गया है । यहां का न्यूनतम तापमान 1 . 5 डिग्री से नीचे दर्ज किया गया । यहां का न्यूनतम तापमान 21 डिग्री से नीचे दर्ज किया गया । वहीं अधिकतम तापमान 35 डिग्री से नीचे दर्ज किया गया है । - मौसम विभाग ने कहा - अगले 24 घंटे में कुछ बू ंदा बां दी हो सकती है । - मौसम विभाग ने कहा है कि अगले 24 घंटों में यहां बारिश होने की संभावना है । - वहीं , बारिश के चलते यहां की स्थिति खराब हो सकती है । - वहीं , मौसम विभाग के अनुसार अगले 24 घंटों में यहां का अधिकतम तापमान 38 डिग्री से ऊपर रहने की संभावना है । - अगर अभी मौसम साफ होता है तो , अगले 24 घंटों में यहां तापमान में बढ़ोतरी हो सकती है । - वहीं , अगले 24 घंटों में मौसम साफ रहेगा । बारिश की संभावना

temperature=0.7,
top_p=0.6,
repetition_penalty=1.2

भारत में बारिश कब होगी ? – जुलाई , अगस्त , सितंबर , अक्टूबर , दिसंबर , फरवरी और मार्च के महीने में मानसून का आगमन होता है । इस समय वर्षा की मात्रा बहुत अधिक होती है इसलिए मानसून की स्थिति को देखते हुए इस दौरान बारिश होने से बचने के लिए सावधानी बर तें । इसके अलावा मानसून के मौसम में जहां - त हां बारिश हो रही हो तो सावधान रहें क्योंकि इन दिनों बरसात के साथ ही ओले गिरने लगते हैं जिससे बारिश का पानी खेतों तक पहुंच जाता है । इस कारण किसानों को भारी नुकसान उठाना पड़ता है । अगर आप भी इस समस्या से परेशान रहते हैं तो हम आपको बता रहे हैं कि कैसे आप इस समस्या से निजात पा सकते हैं । इस लेख में हम आपको बता येंगे कि कैसे आप इस समस्या से छुटकारा पा सकते हैं । 1 . सबसे पहले अपने घर के अंदर या बाहर निकलने वाले कूड़े दानों पर ध्यान दें । 2 . इस बात का खास ख्याल रखें कि कूड़े दान में जो कूड़ा डाला गया है उसे न फे कें । 3 . अगर आप इस साइट पर कोई अपडेट चेक करें तो तुरंत बदल लें ।

At least the context is same!!!
To make it learn facts , it is little hard!

Another one!

temperature=0.7, top_k=20

दिनेश कार्तिक और कुल दीप यादव को आराम दिए जाने और टीम प्रबंधन को लेकर भी सवाल उठाए . कुल दीप यादव की बात करें तो भुवनेश्वर कुमार ने भी पिछले मैच में शानदार गेंदबाजी की थी . उन्होंने भुवनेश्वर कुमार की जगह उमेश यादव को मौका दिया था . वहीं दूसरी तरफ बांग्लादेश की गेंदबाजी में भी भुवनेश्वर कुमार और जस प्रीत बु मरा ह की स्पिन जोड़ी के खिलाफ अच्छा प्रदर्शन किया था . वहीं , गेंदबाजी में भी भुवनेश्वर कुमार ने कमाल दिखाया था . भुवनेश्वर कुमार ने भी पिछले मैच में कुल दीप यादव की जगह भुवनेश्वर कुमार को मौका दिया था . वहीं वेस्टइंडीज की गेंदबाजी में भी भुवनेश्वर कुमार ने कमाल किया था . उन्होंने वेस्टइंडीज के खिलाफ दूसरे टी - 20 मैच में दो शानदार छक्के लगाए थे . उन्होंने इस मैच में दो छक्के लगाए थे . वहीं रोहित शर्मा ने टी20 में अपनी गेंदबाजी में भी कमाल किया था . रोहित शर्मा ने पिछले मैच में भी शानदार गेंदबाजी की थी . इस मैच में रोहित शर्मा ने अपने दूसरे मैच में भी दो छक्के लगाए थे . वहीं तीसरे टी - 20 मैच में कुल दीप यादव की जगह भुवनेश्वर कुमार को

temperature=0.7,
top_p=0.6,
repetition_penalty=1.2

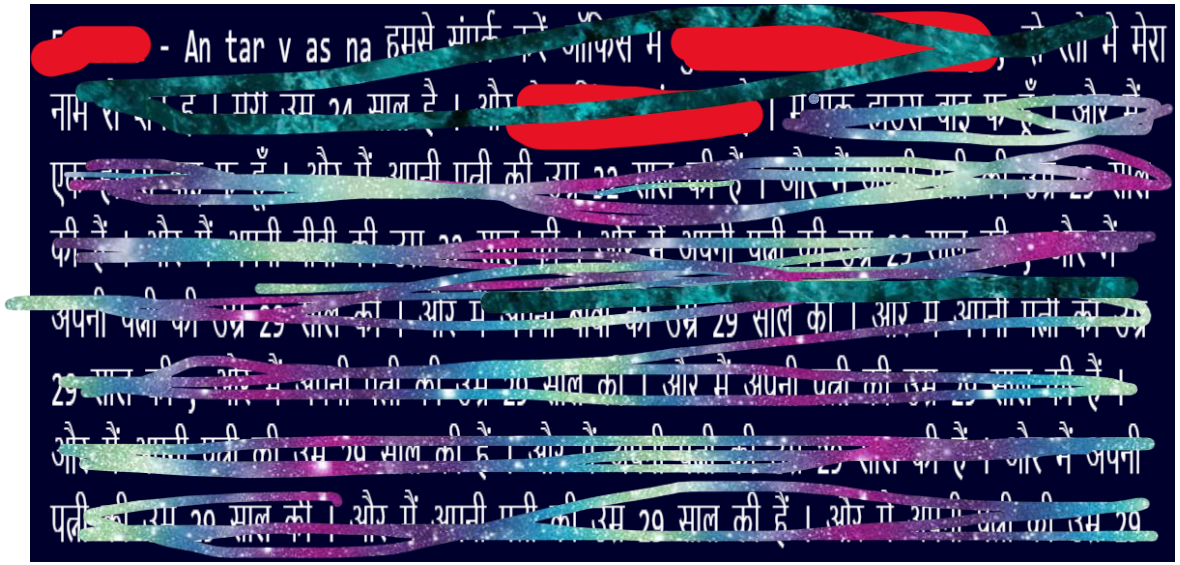
दिनेश कार्तिक की कप्तानी में भारत ने रविवार को खेले गए एशिया कप के सुपर फोर मुकाबले में बांग्लादेश को 8 विकेट से हराकर टूर्नामेंट का अपना पहला मैच जीत लिया . इसी के साथ ही भारतीय टीम एशिया कप के फाइनल में पहुंच गई है . अब उसका सामना श्रीलंका और पाकिस्तान से होगा . रोहित शर्मा , शिखर धवन , विराट कोहली (कप्तान), मनीष पांडे , एमएस धोनी (विकेटकीपर), हार्दिक पंड्या , भुवनेश्वर कुमार , कुल दीप यादव , युज वेंद्र चहल , जस प्रीत बु मरा ह , उमेश यादव , शार् दुल ठाकुर , मोहम्मद शमी . Pre v ious article अब इस क्रिकेटर पर बनेगी बायो पिक , खुद भी कर चुके हैं एक्टिंग N ext article इंग्लैंड के खिलाफ वनडे सीरीज के लिए ऑस्ट्रेलियाई टीम घोषित , ये दिग्गज खिलाड़ी हुआ बाहर ... बिहार : हत्या के शक में भीड़ ने महिला को पीटा ; एक शख्स को पैंट में दर्द एससी / एसटी एक्ट पर मायावती फॉर्मूला - ‘ अगर B J P जीती तो संविधान बदलने वाला पीएम मोदी आएगा ’ क्रिकेट से संन्यास लेने के बाद बोले चे तेश्वर पु जारा , टीम इंडिया के नाम किया है सम्मानित

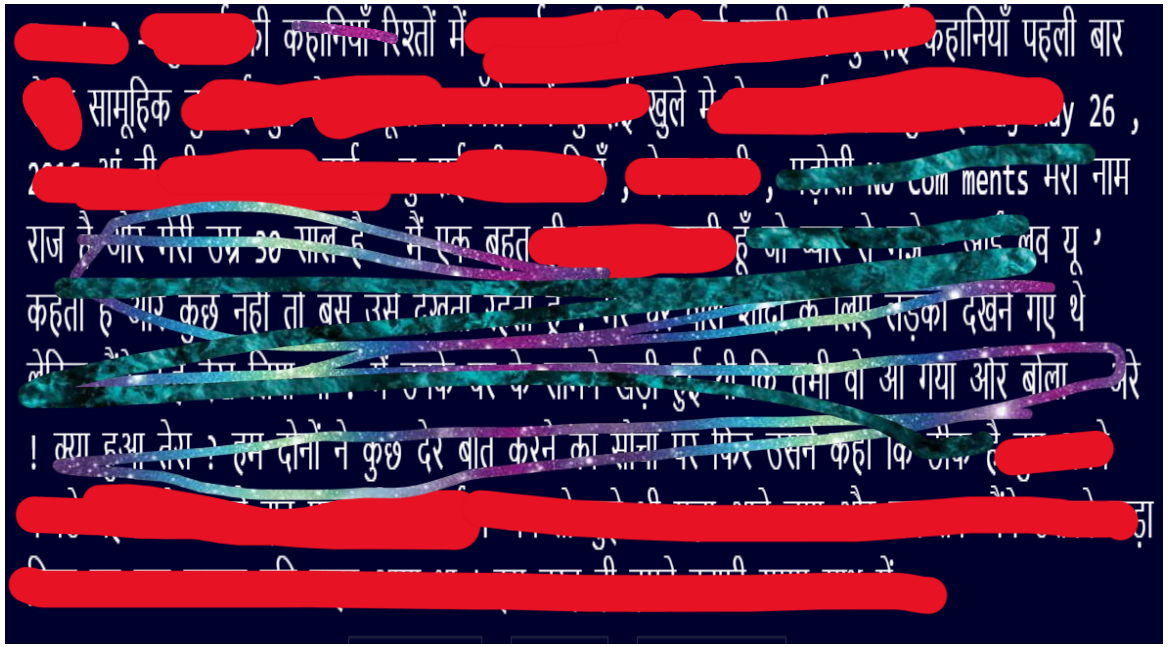
Gemma270 has done:
Data Preprocessing like:

•CSAM Filtering: Rigorous CSAM (Child Sexual Abuse Material) filtering was applied at multiple stages in the data preparation process to ensure the exclusion of harmful and illegal content.

•Sensitive Data Filtering: As part of making Gemma pre-trained models safe and reliable, automated techniques were used to filter out certain personal information and other sensitive data from training sets.

Why is it important? Obvious answer!
Because your model will start using bad words!





Model Improvement:

1. Should use cleaned dataset!
2. Addition of some Hinglish and Hindi dataset
3. Hindi and Hinglish Tokenizer (a better one)!
4. Can increase head numbers!
5. Weight tying between the input and output embedding!
6. Try to do Supervised Fine Tuning for better conversational answers!

Configuration about Training:

- Mixed precision training using Automatic Mixed Precision (AMP (torch.amp.autocast)

-For the initialisation , i have used nn.Linear in GroupedQueryAttention, Feedforward and Gemma3Model.out_head and it uses Kaiming (He) uniform initialization.

-For the nn.Embedding (the self.tok_emb layer) --> normal distribution of mean 0 and standard 1.

-For RMS Norm: self.scale --> initialized to all zeros using nn.Parameter(torch.zeros(emb_dim))

-torch.manual_seed(42+rank) --> default initialization reproducible

-While no other calls custom initialization functions were called: (like _init_weights or nn.init calls)

Performance Improvements!!!

Yes , some parts were implemented very unoptimized.

-In manual attention, attn_scores was also stored as intermediate tensor(materialized tensor), now to be written to and read from the GPU's main VRAM(HBM).

-Batch* Heads* Seq_Len* Seq_Len - It is very big number!

-Moving this giant tensor from VRAM to SRAM for the SoftMax is a huge performance bottleneck.

-In case **of F.scaled_dot_product_attention**: In this case, it is now going to use fused kernel and it will use algorithm like FlashAttention(where it does things in block).

-F.cross_entropy (fused kernel for loss calculation)

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{e^{z_{i,y_i}}}{\sum_{k=1}^{C} e^{z_{i,k}}}\right)$$

Thanks