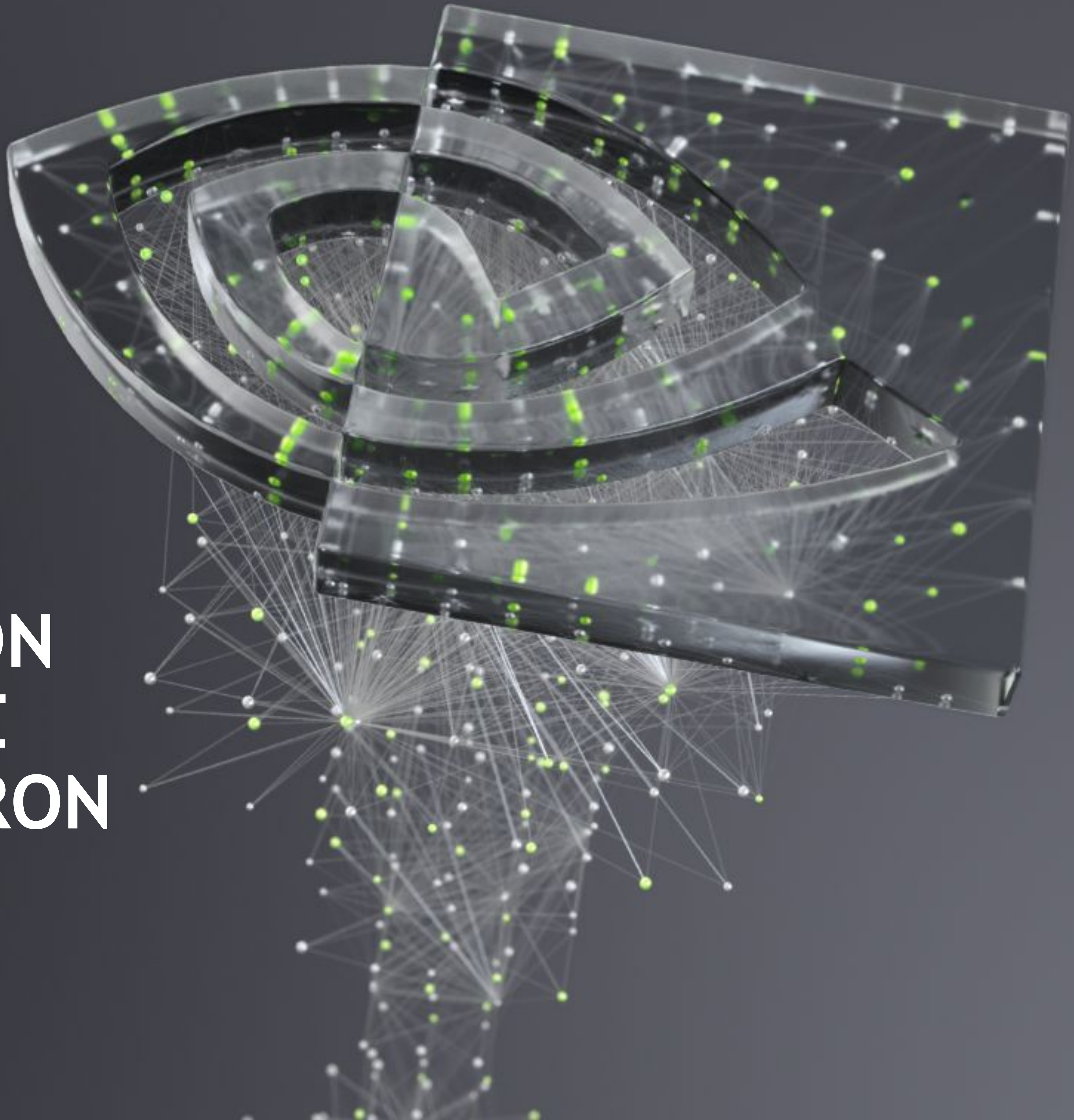
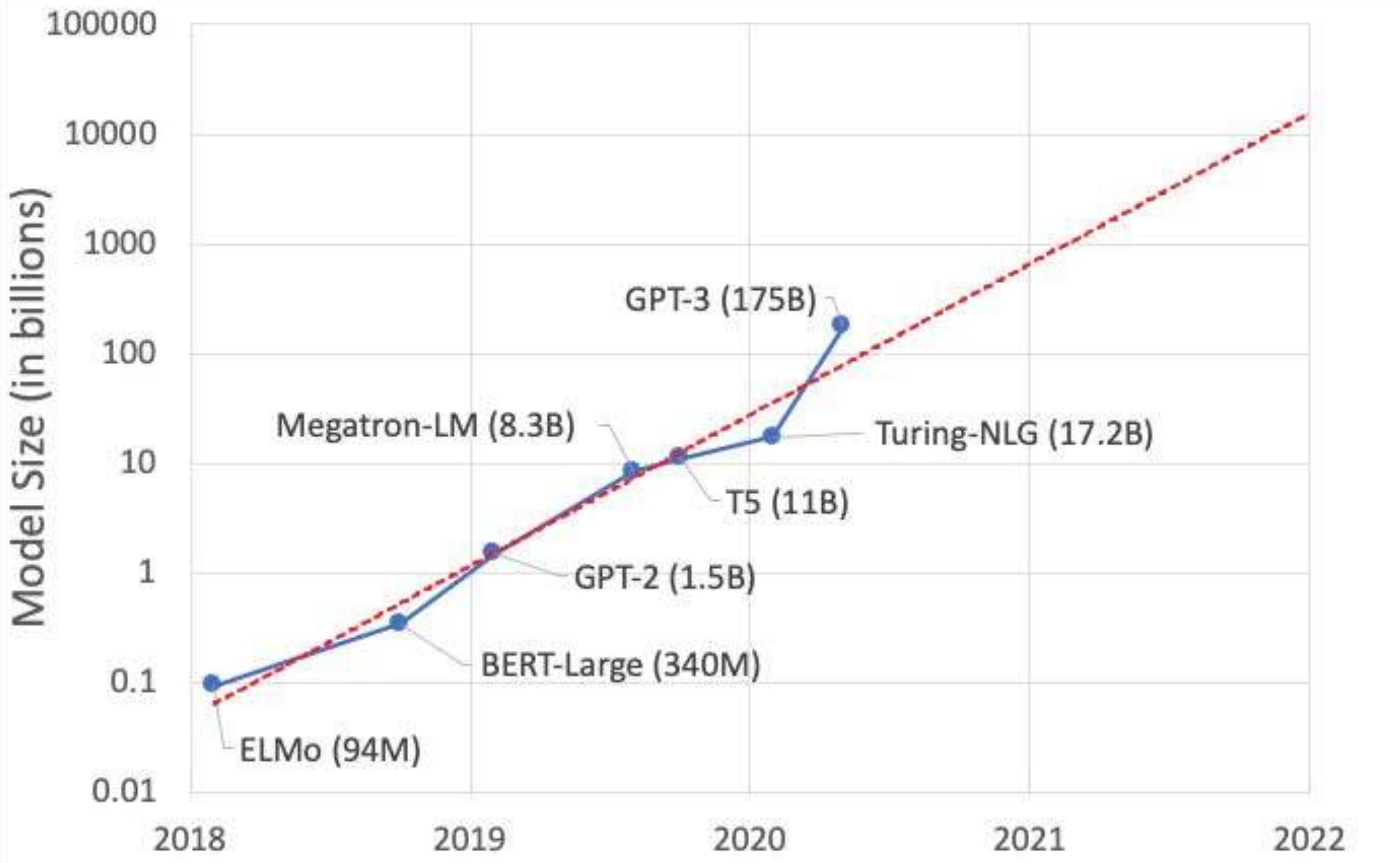
TRAINING MULTI-BILLION PARAMETER LANGUAGE MODELS USING MEGATRON

Mohammad Shoeybi
Hot Chips Conf., Aug. 16, 2020



MODEL SIZE TREND IN NLP

- Training the **largest transformer-based** language model has recently been one of the best ways to advance the state-of-the-art in NLP applications
- NLP model size increases by almost **an order of magnitude** every year

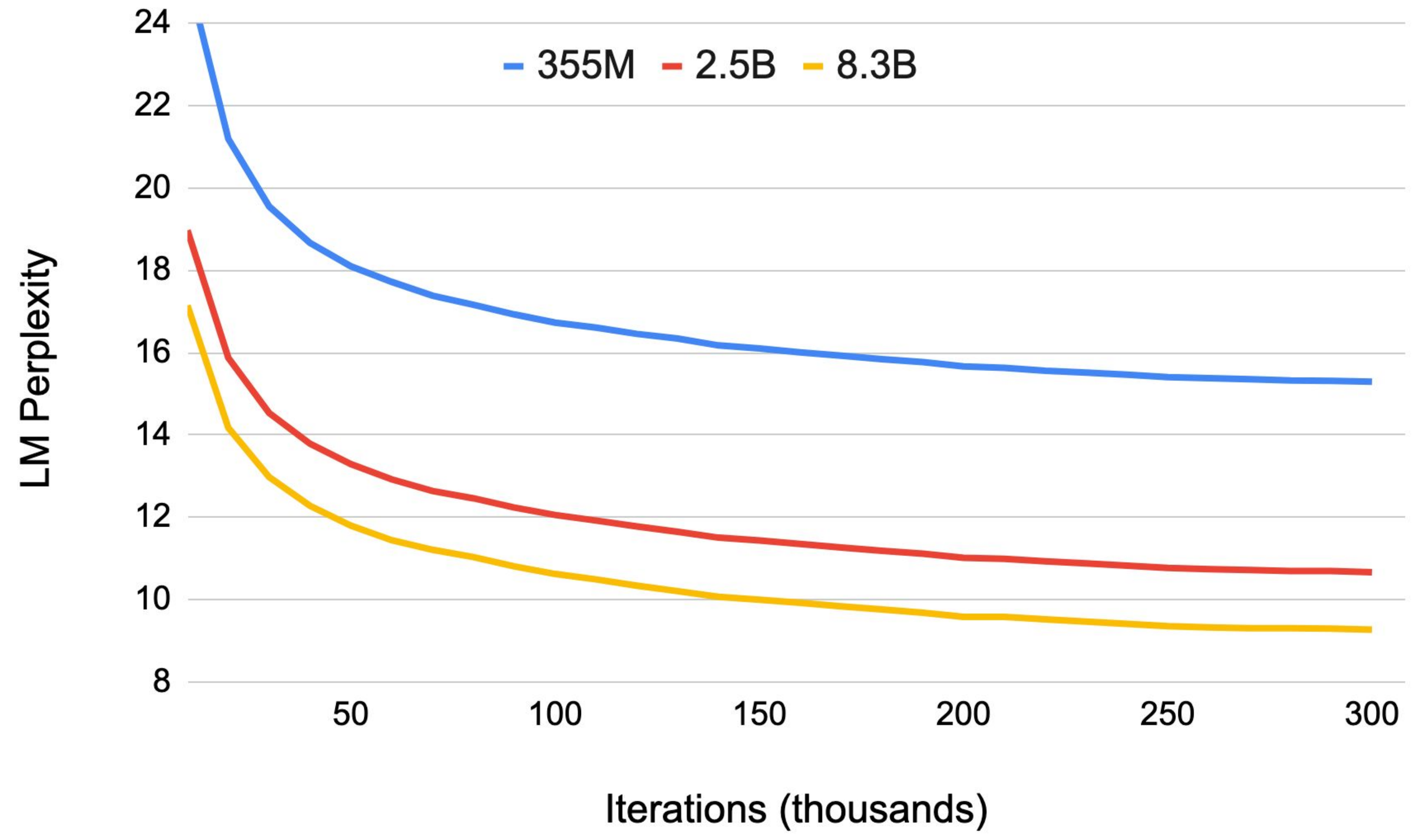


WHY LARGE NLP MODELS?

- Imagenet moment of NLP
- Unsupervised pretraining on large text corpora has eliminated training dataset size issues
- Lots of downstream NLP applications have benefited from recent advancements
- Training larger models with more data results in better accuracy in almost all cases

EXAMPLE 1: LEFT-TO-RIGHT LANGUAGE MODEL

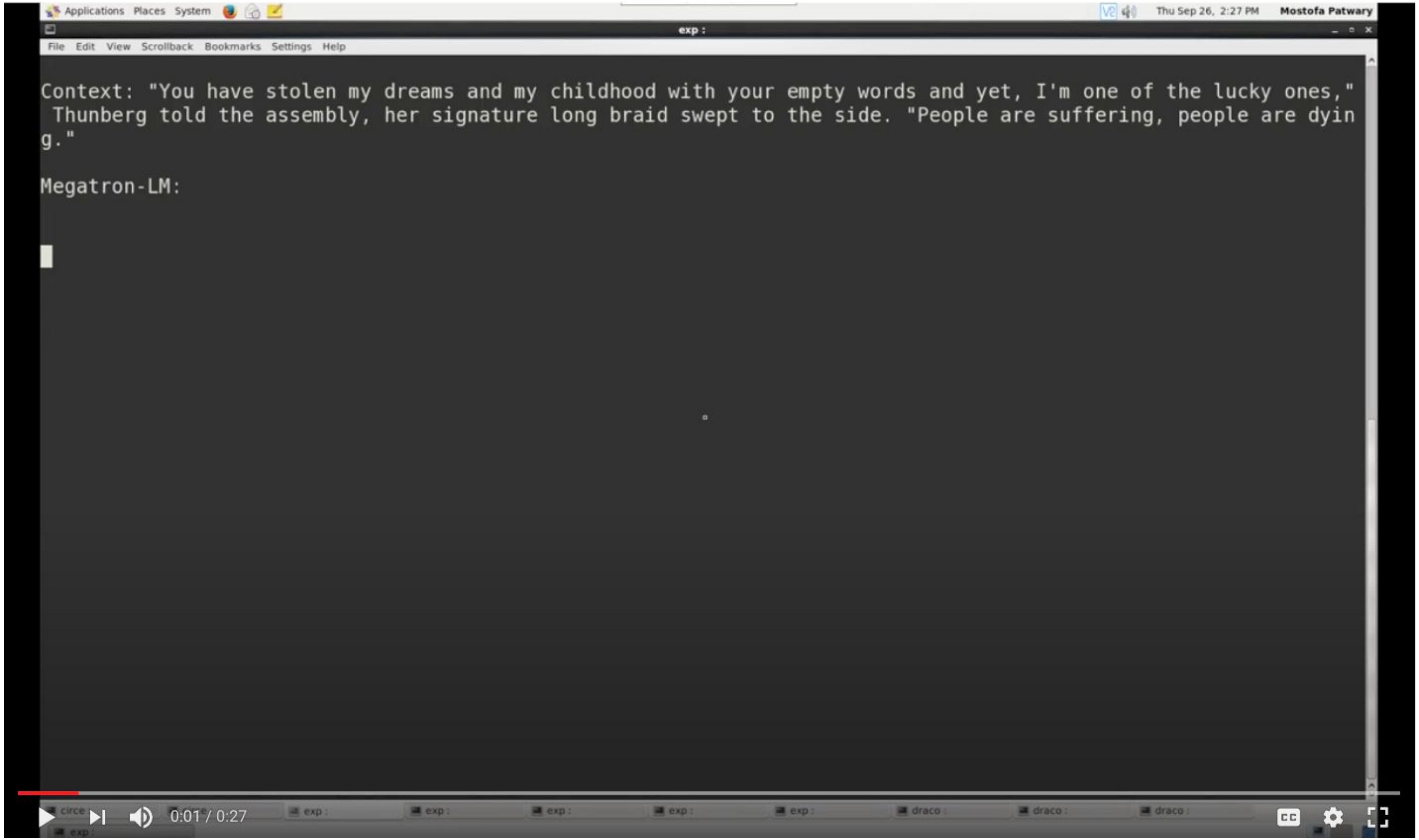
- GPT2-based language model
- Training data:
  174 GB WebText / CC-Stories / Wikipedia / RealNews



Model Size	Wikitext-103 (Perplexity ↓)
355 M	19.22
2.5 B	12.68
8.3 B	**10.81**
Previous SOTA	16.43*

* Dynamic Evaluation of Transformer Language Models, Krause et. al., 2019

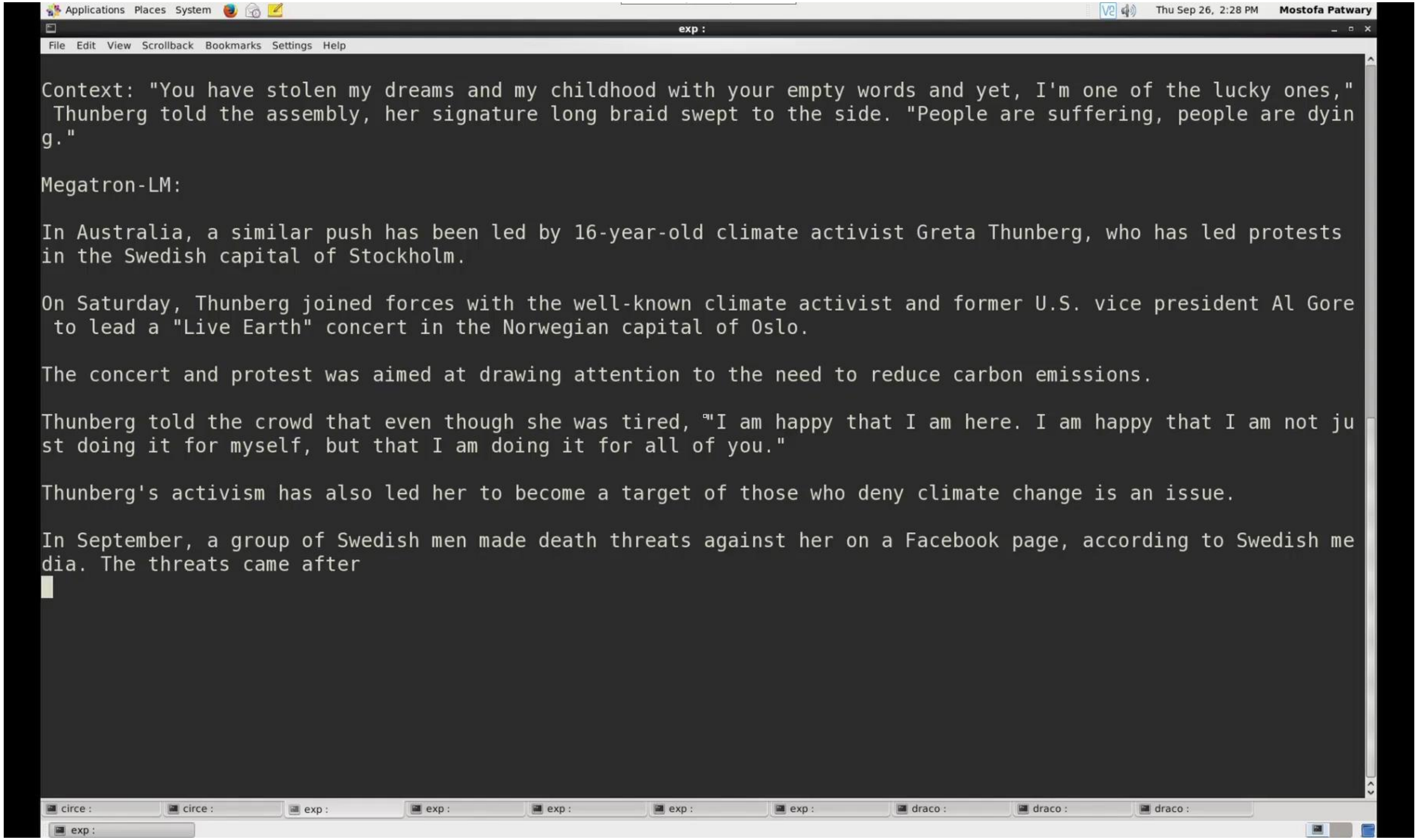
POWERFUL GENERATIONS



Applications Places System
Thu Sep 26, 2:27 PM
Mostofa Patwary
exp :
File Edit View Scrollback Bookmarks Settings Help
0:01 / 0:27

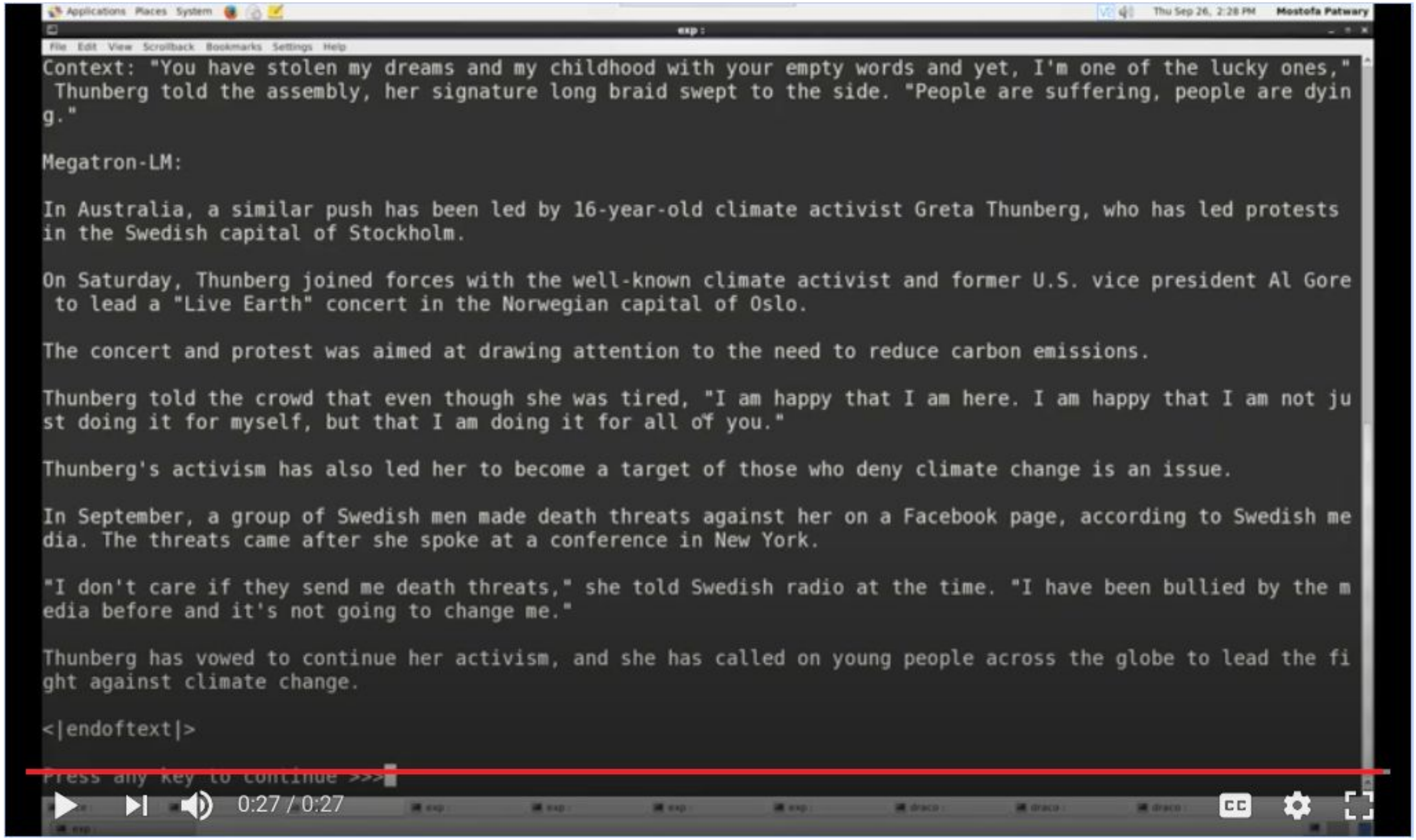
```
Context: "You have stolen my dreams and my childhood with your empty words and yet, I'm one of the lucky ones,"
 Thunberg told the assembly, her signature long braid swept to the side. "People are suffering, people are dyin
g."

Megatron-LM:
```

POWERFUL GENERATIONS



```
Context: "You have stolen my dreams and my childhood with your empty words and yet, I'm one of the lucky ones,"
 Thunberg told the assembly, her signature long braid swept to the side. "People are suffering, people are dyin
g."

Megatron-LM:

In Australia, a similar push has been led by 16-year-old climate activist Greta Thunberg, who has led protests
in the Swedish capital of Stockholm.

On Saturday, Thunberg joined forces with the well-known climate activist and former U.S. vice president Al Gore
 to lead a "Live Earth" concert in the Norwegian capital of Oslo.

The concert and protest was aimed at drawing attention to the need to reduce carbon emissions.

Thunberg told the crowd that even though she was tired, "I am happy that I am here. I am happy that I am not ju
st doing it for myself, but that I am doing it for all of you."

Thunberg's activism has also led her to become a target of those who deny climate change is an issue.

In September, a group of Swedish men made death threats against her on a Facebook page, according to Swedish me
dia. The threats came after
```

POWERFUL GENERATIONS

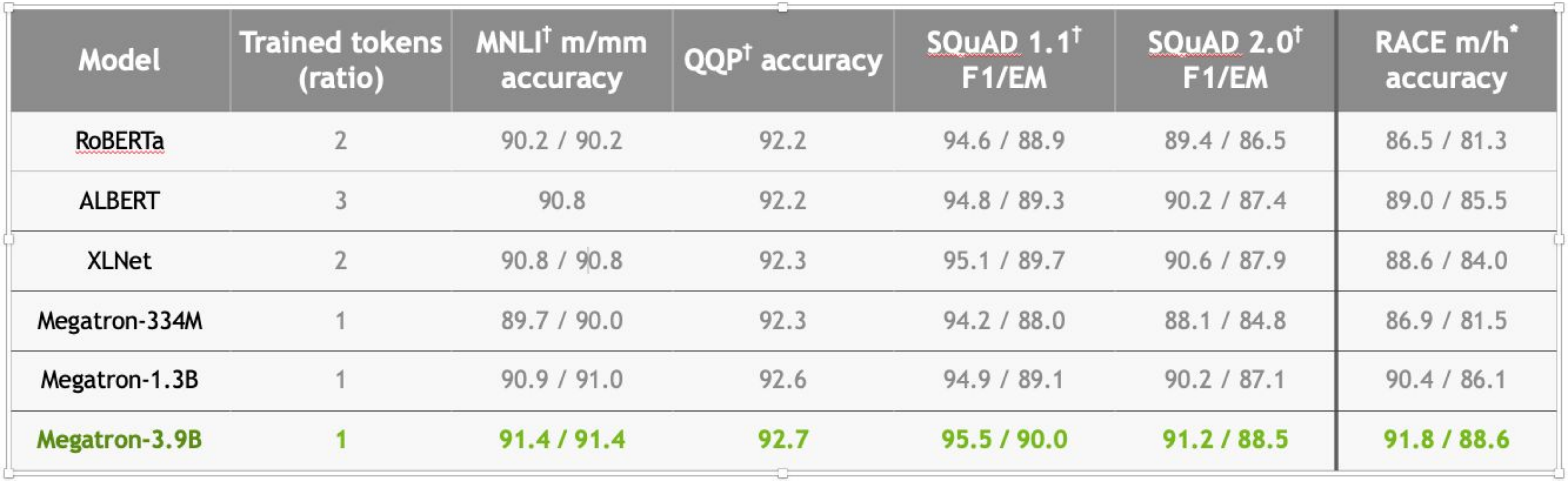


Applications Places System
Thu Sep 26, 2:28 PM
Mostofa Patwary
exp :
File Edit View Scrollback Bookmarks Settings Help
0:27 / 0:27

```
Context: "You have stolen my dreams and my childhood with your empty words and yet, I'm one of the lucky ones,"
 Thunberg told the assembly, her signature long braid swept to the side. "People are suffering, people are dyin
g."

Megatron-LM:

In Australia, a similar push has been led by 16-year-old climate activist Greta Thunberg, who has led protests
in the Swedish capital of Stockholm.

On Saturday, Thunberg joined forces with the well-known climate activist and former U.S. vice president Al Gore
 to lead a "Live Earth" concert in the Norwegian capital of Oslo.

The concert and protest was aimed at drawing attention to the need to reduce carbon emissions.

Thunberg told the crowd that even though she was tired, "I am happy that I am here. I am happy that I am not ju
st doing it for myself, but that I am doing it for all of you."

Thunberg's activism has also led her to become a target of those who deny climate change is an issue.

In September, a group of Swedish men made death threats against her on a Facebook page, according to Swedish me
dia. The threats came after she spoke at a conference in New York.

"I don't care if they send me death threats," she told Swedish radio at the time. "I have been bullied by the m
edia before and it's not going to change me."

Thunberg has vowed to continue her activism, and she has called on young people across the globe to lead the fi
ght against climate change.

<|endoftext|>

Press any key to continue >>>
```

EXAMPLE 2: BERT

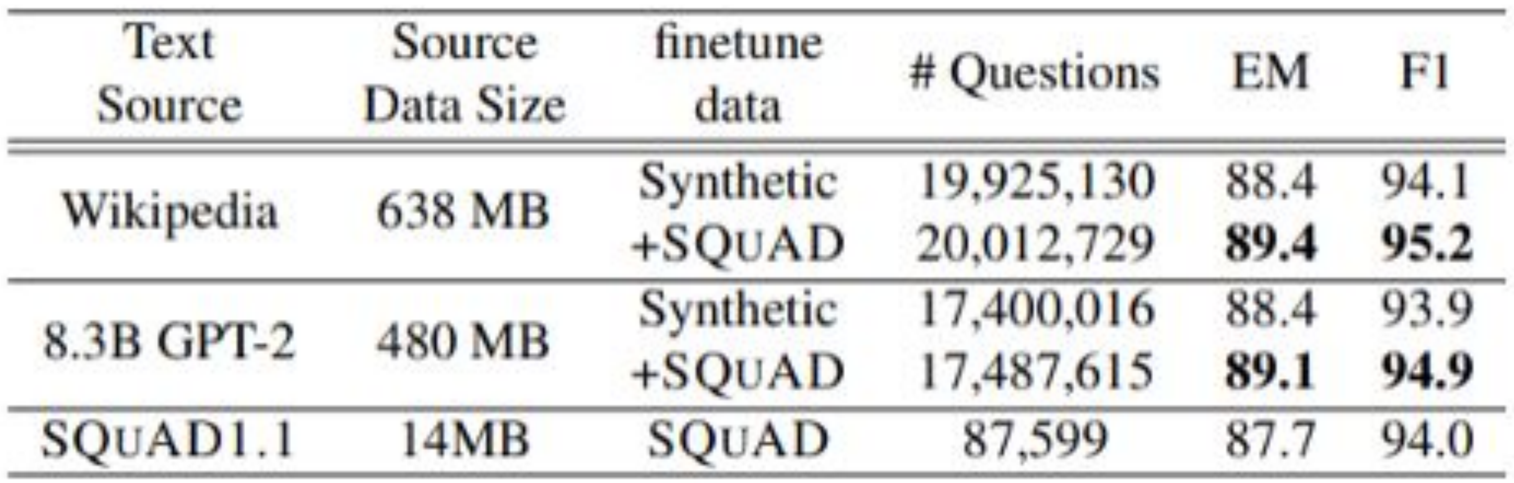
- Canonical downstream tasks
- Larger model trained on fewer number of tokens produces better results



Model	Trained tokens (ratio)	MNLI† m/mm accuracy	QQP† accuracy	SQuAD 1.1† F1/EM	SQuAD 2.0† F1/EM	RACE m/h* accuracy
RoBERTa	2	90.2 / 90.2	92.2	94.6 / 88.9	89.4 / 86.5	86.5 / 81.3
ALBERT	3	90.8	92.2	94.8 / 89.3	90.2 / 87.4	89.0 / 85.5
XLNet	2	90.8 / 90.8	92.3	95.1 / 89.7	90.6 / 87.9	88.6 / 84.0
Megatron-334M	1	89.7 / 90.0	92.3	94.2 / 88.0	88.1 / 84.8	86.9 / 81.5
Megatron-1.3B	1	90.9 / 91.0	92.6	94.9 / 89.1	90.2 / 87.1	90.4 / 86.1
Megatron-3.9B	**1**	**91.4 / 91.4**	**92.7**	**95.5 / 90.0**	**91.2 / 88.5**	**91.8 / 88.6**

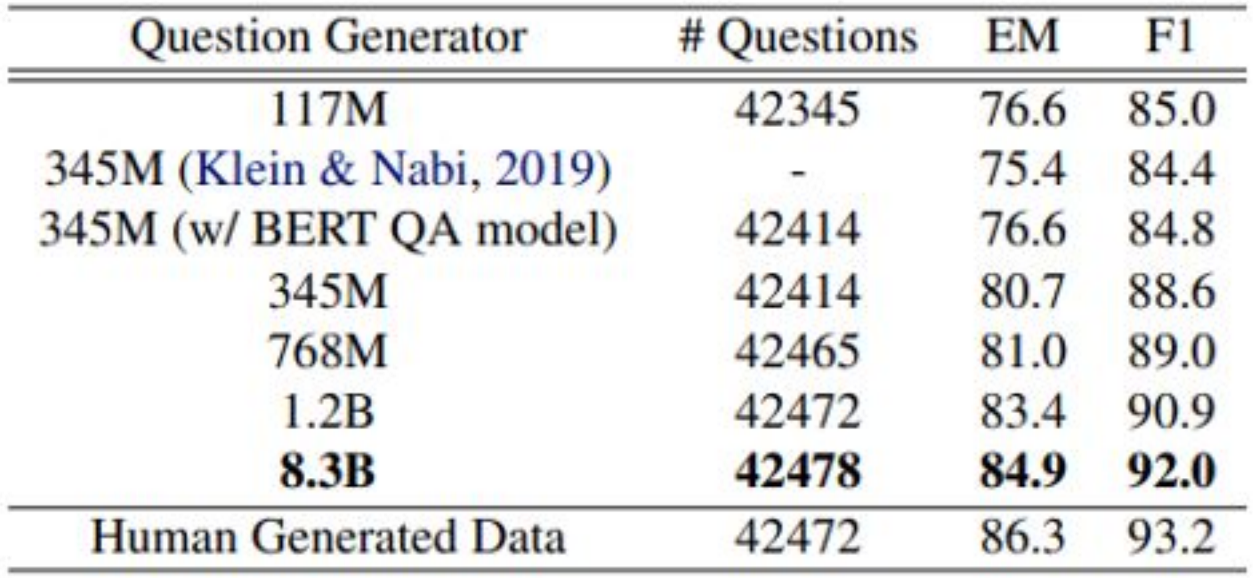
EXAMPLE 3: TEACHING MODELS TO GENERATE QUESTIONS AND ANSWERS

- Using synthetic data can beat using only human labeled data

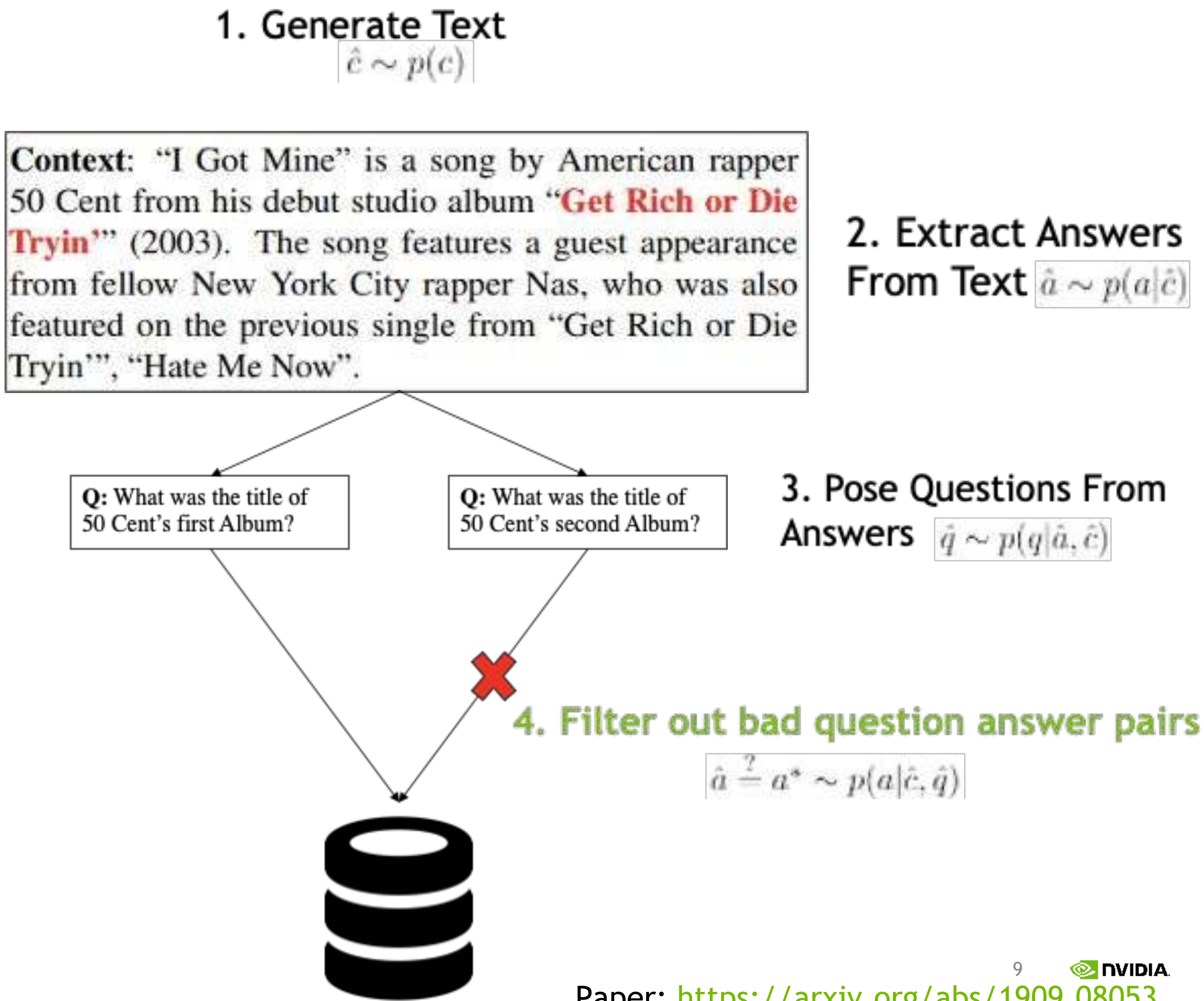


Text Source	Source Data Size	finetune data	# Questions	EM	F1
Wikipedia	638 MB	Synthetic	19,925,130	88.4	94.1
		+SQuAD	20,012,729	**89.4**	**95.2**
8.3B GPT-2	480 MB	Synthetic	17,400,016	88.4	93.9
		+SQuAD	17,487,615	**89.1**	**94.9**
SQuAD1.1	14MB	SQuAD	87,599	87.7	94.0

- Scale is the key in improving results



Question Generator	# Questions	EM	F1
117M	42345	76.6	85.0
345M (Klein & Nabi, 2019)	-	75.4	84.4
345M (w/ BERT QA model)	42414	76.6	84.8
345M	42414	80.7	88.6
768M	42465	81.0	89.0
1.2B	42472	83.4	90.9
8.3B	**42478**	**84.9**	**92.0**
Human Generated Data	42472	86.3	93.2



1. Generate Text $\hat{c} \sim p(c)$

Context: "I Got Mine" is a song by American rapper 50 Cent from his debut studio album "**Get Rich or Die Tryin'**" (2003). The song features a guest appearance from fellow New York City rapper Nas, who was also featured on the previous single from "Get Rich or Die Tryin'", "Hate Me Now".

2. Extract Answers From Text $\hat{a} \sim p(a|\hat{c})$

Q: What was the title of 50 Cent's first Album?

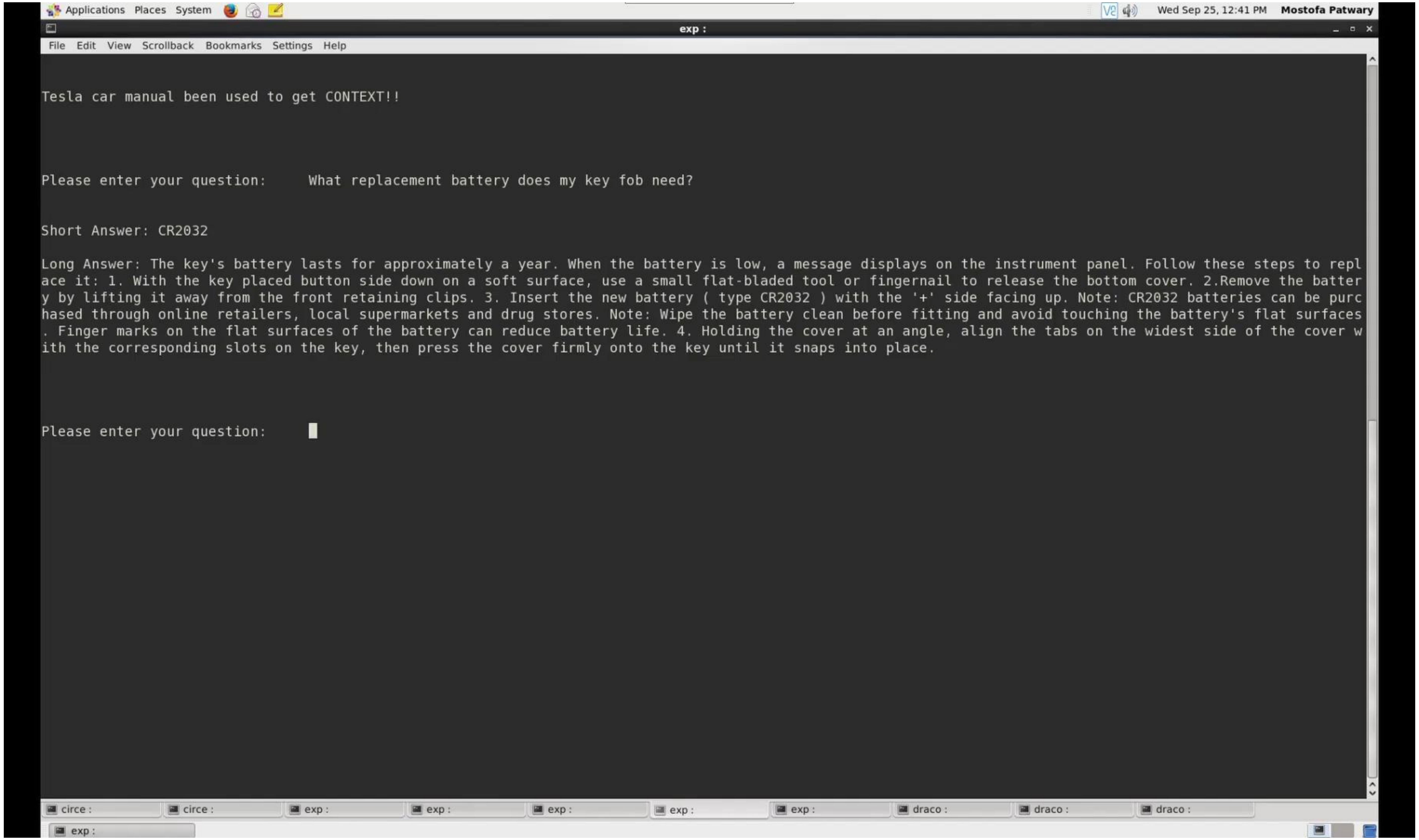
Q: What was the title of 50 Cent's second Album?

3. Pose Questions From Answers $\hat{q} \sim p(q|\hat{a},\hat{c})$

4. Filter out bad question answer pairs $\hat{a} \overset{?}{=} a^* \sim p(a|\hat{c},\hat{q})$

Paper: https://arxiv.org/abs/1909.08053

EXAMPLE 3: Q&A OVER TESLA CAR MANUAL



Applications Places System
Wed Sep 25, 12:41 PM Mostofa Patway
exp :
File Edit View Scrollback Bookmarks Settings Help
circe : circe : exp : exp : exp : exp : exp : draco : draco : draco :
exp :

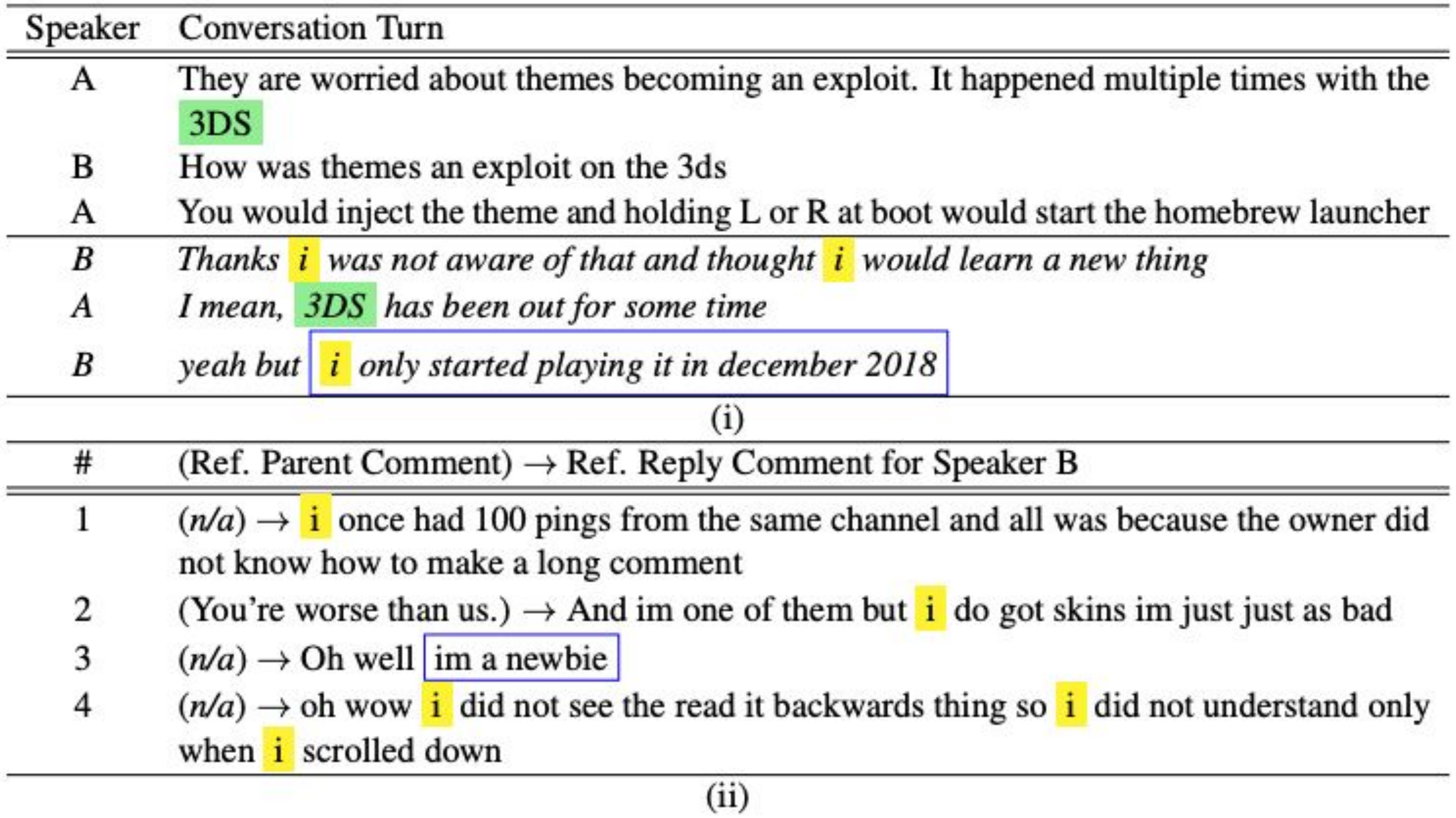
```
Tesla car manual been used to get CONTEXT!!



Please enter your question:     What replacement battery does my key fob need?


Short Answer: CR2032

Long Answer: The key's battery lasts for approximately a year. When the battery is low, a message displays on the instrument panel. Follow these steps to repl
ace it: 1. With the key placed button side down on a soft surface, use a small flat-bladed tool or fingernail to release the bottom cover. 2.Remove the batter
y by lifting it away from the front retaining clips. 3. Insert the new battery ( type CR2032 ) with the '+' side facing up. Note: CR2032 batteries can be purc
hased through online retailers, local supermarkets and drug stores. Note: Wipe the battery clean before fitting and avoid touching the battery's flat surfaces
. Finger marks on the flat surfaces of the battery can reduce battery life. 4. Holding the cover at an angle, align the tabs on the widest side of the cover w
ith the corresponding slots on the key, then press the cover firmly onto the key until it snaps into place.



Please enter your question:
```

EXAMPLE 4: CHATBOTS

- Generative Conversation Control (GCC)
  - Paper: https://www.aclweb.org/anthology/2020.acl-main.8/

- Persona control by conditioning on prior conversations

	# Parameters	Data Source	Persona Control
DLGNet	345 Million	Movie Triples / Ubuntu Dialogue Corpus	👎
DialoGPT	768 Million	Reddit (2005-2017)	👎
Meena	2.6 Billion	Public Social Media	👎
GCC	8.3 Billion	Reddit (2019)	👍



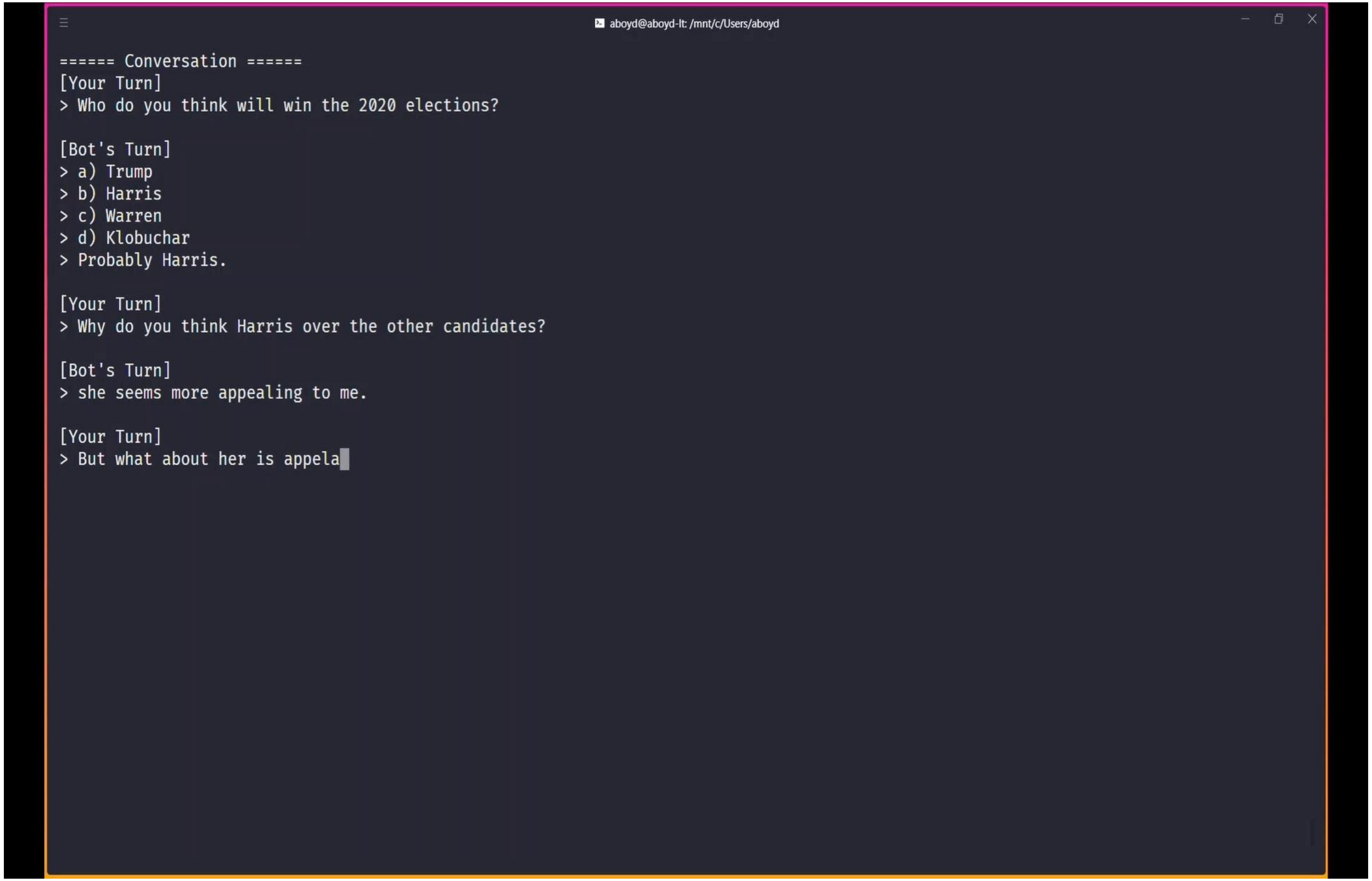
Speaker	Conversation Turn
A	They are worried about themes becoming an exploit. It happened multiple times with the 3DS
B	How was themes an exploit on the 3ds
A	You would inject the theme and holding L or R at boot would start the homebrew launcher
B	*Thanks i was not aware of that and thought i would learn a new thing*
A	*I mean, 3DS has been out for some time*
B	*yeah but i only started playing it in december 2018*

(i)

#	(Ref. Parent Comment) → Ref. Reply Comment for Speaker B
1	*(n/a)* → i once had 100 pings from the same channel and all was because the owner did not know how to make a long comment
2	(You're worse than us.) → And im one of them but i do got skins im just just as bad
3	*(n/a)* → Oh well im a newbie
4	*(n/a)* → oh wow i did not see the read it backwards thing so i did not understand only when i scrolled down

(ii)

EXAMPLE 4: CHATBOTS



```
====== Conversation ======
[Your Turn]
> Who do you think will win the 2020 elections?

[Bot's Turn]
> a) Trump
> b) Harris
> c) Warren
> d) Klobuchar
> Probably Harris.

[Your Turn]
> Why do you think Harris over the other candidates?

[Bot's Turn]
> she seems more appealing to me.

[Your Turn]
> But what about her is appela
```

RECAP: WHY LARGE NLP MODELS?

- Imagenet moment of NLP
- Unsupervised pretraining on large text corpora has eliminated training dataset size issues
- Lots of downstream NLP application have benefited from recent advancements
- Training larger models with more data results in better accuracy in all cases

We need software and hardware that can efficiently and easily enable researchers to train large NLP models

MEGATRON

NVIDIA's framework for efficiently training the world's largest transformer-based language models.

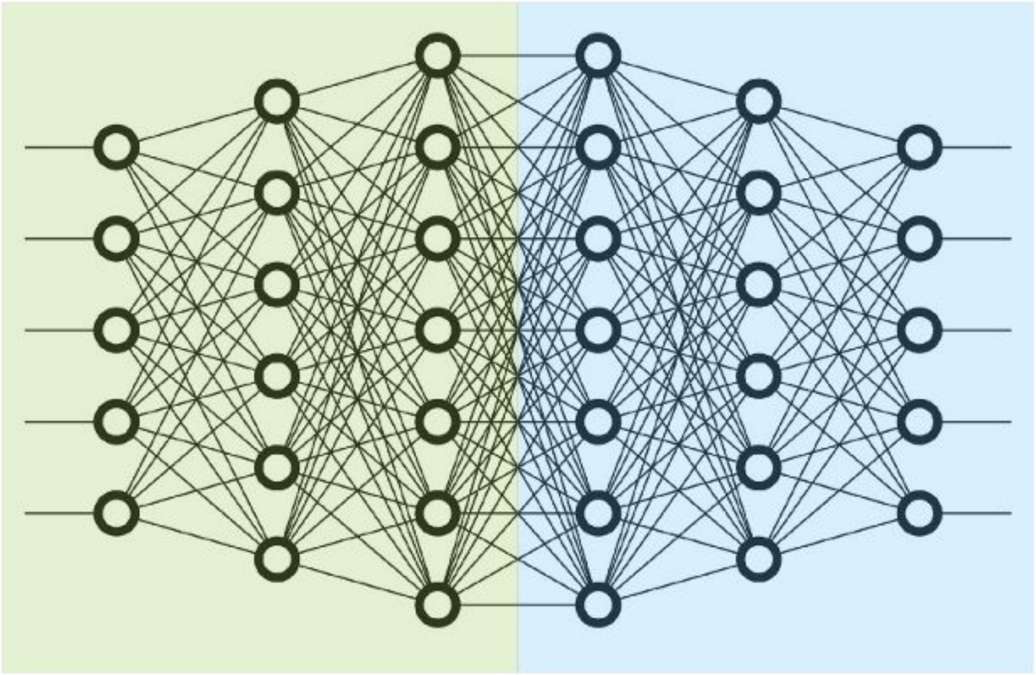
Paper: https://arxiv.org/abs/1909.08053
Repo: https://github.com/NVIDIA/Megatron-LM

- All the aforementioned cases are trained using Megatron
- Megatron has been directly used to train Turing-NLG (17.2B)
- Megatron has inspired other work such as BlenderBot by Facebook

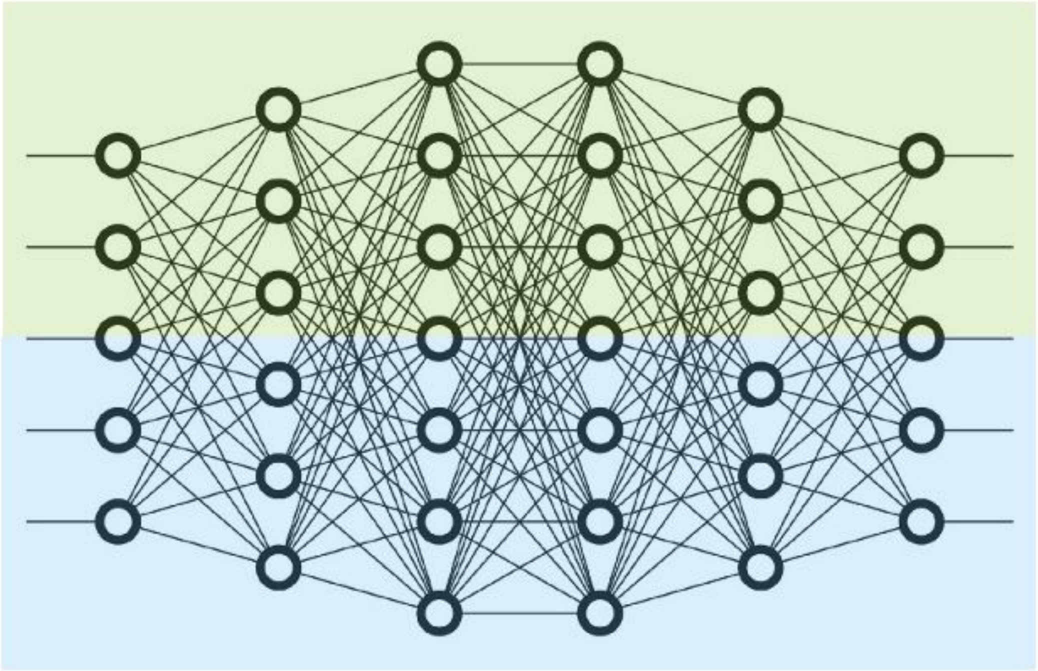
GOALS FOR MEGATRON

- Low barrier to entry
    - Ability to build on top our existing code base (no rewriting, …)
    - Devising simple methods that require minimal changes
- Training of transformer-based language models with billions of parameters
    - Requires model parallelism to fit in GPU memory
    - Only support transformer-based model
- Achieving high utilization and scaling up to thousands of GPUs

MODEL PARALLELISM



- Inter-Layer (Pipeline) Parallelism
    - Split sets of layers across multiple devices
    - Layer 0,1,2 and layer 3,4,5 are on difference devices



- Intra-Layer (Tensor) Parallelism
    - Split individual layers across multiple devices
    - Both devices compute difference parts of Layer 0,1,2,3,4,5

WHY INTRA-LAYER MODEL PARALLELISM

- Tensor parallelism is much simpler to implement
- Easier to load-balance
- Less restrictive on the batch-size (bubble issue in pipelining)
    - Intra-layer model parallelism is orthogonal to pipeline parallelism: very large models such as GPT-3 use both.
- Transformers have large GEMMs
    - Tensor parallelism works well for large matrices
- NVIDIA DGX-A100 boxes with nvlink
    - DGX-A100 has 600 GB/sec GPU-to-GPU bidirectional bandwidth

CHALLENGES WITH INTRA-LAYER MODEL PARALLELISM

- Tensor splitting results in lower math intensity
  - This approach is not suitable for strong scaling
- We should make sure math intensity stays above that of the processor
  - A100 math intensity = 312 teraFLOPs/1555 GB/sec = 200
- Intra-layer model parallelism requires more communication.
  - DGX-A100 with nvlink mitigates this issue.

Parallel

Operation: $Y_{n\times(n/p)} = X_{n\times n}A_{n\times(n/p)}$

Flops: $2n^3/p$

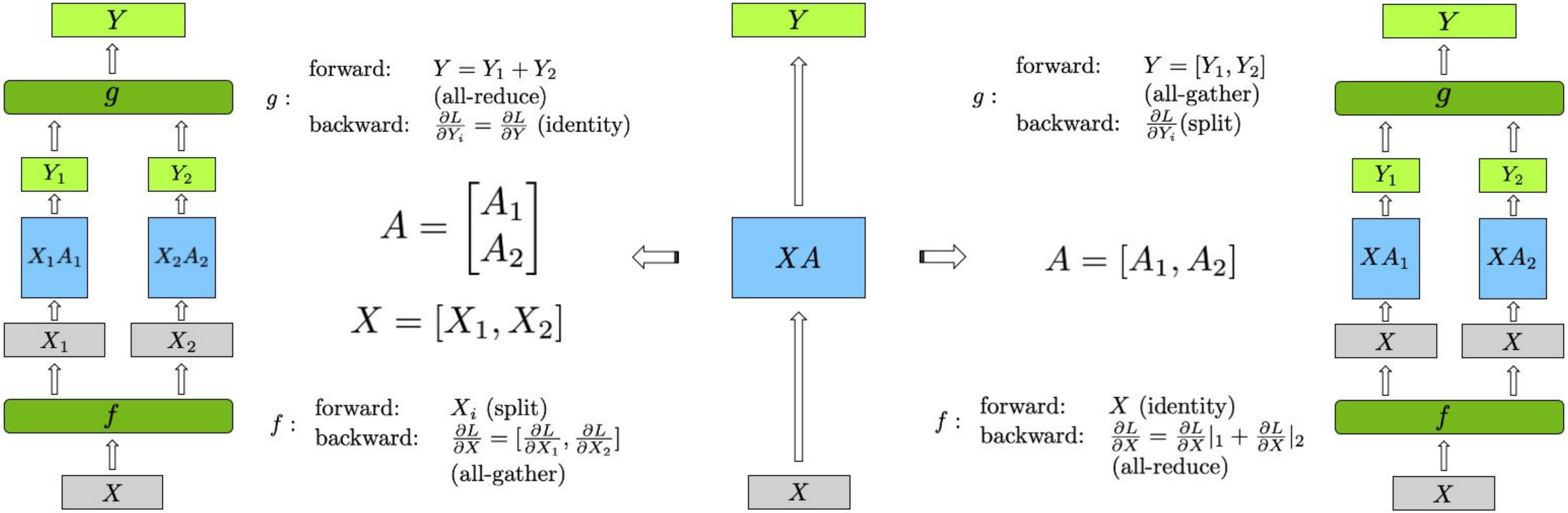
Bandwidth: $2n^2(1+2/p)$

Intensity: $\dfrac{1}{2+p}n$

SIMPLE EXAMPLE OF TENSOR PARALLELISM

Row Parallel Linear Layer

Column Parallel Linear Layer



$g:$ forward: $Y = Y_1 + Y_2$ (all-reduce)
backward: $\frac{\partial L}{\partial Y_i} = \frac{\partial L}{\partial Y}$ (identity)

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

$$X = [X_1, X_2]$$

$f:$ forward: X_i (split)
backward: $\frac{\partial L}{\partial X} = [\frac{\partial L}{\partial X_1}, \frac{\partial L}{\partial X_2}]$ (all-gather)

$g:$ forward: $Y = [Y_1, Y_2]$ (all-gather)
backward: $\frac{\partial L}{\partial Y_i}$(split)

$$A = [A_1, A_2]$$

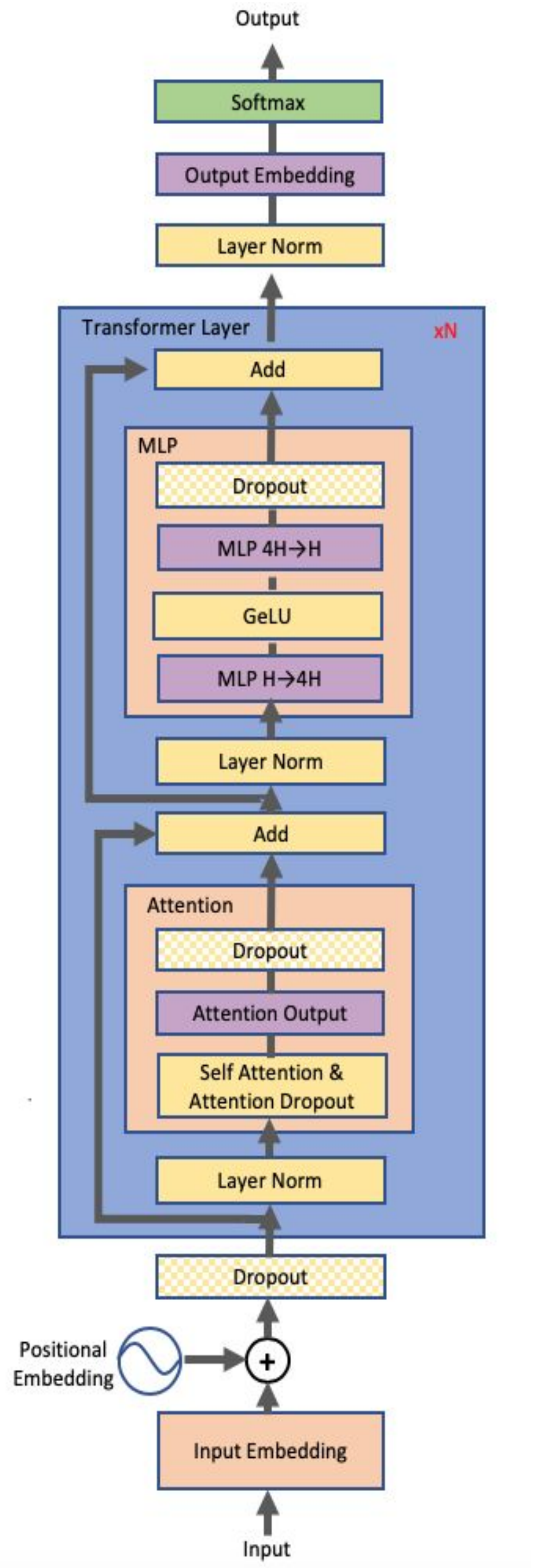
$f:$ forward: X (identity)
backward: $\frac{\partial L}{\partial X} = \frac{\partial L}{\partial X}|_1 + \frac{\partial L}{\partial X}|_2$ (all-reduce)

APPROACH FOR TRANSFORMER MODELS

- Develop an approach that can be fully implemented with insertion of few simple collectives

- Rely on pre-existing NCCL/PyTorch operations for a native PyTorch implementation

- Group math heavy operations (such as GEMMs) before parallel communication point



PARTITIONING MLP

- MLP:

$$Y = \mathrm{GeLU}(XA)$$
$$Z = \mathrm{Dropout}(YB)$$

- Approach 1: split X column-wise and A row-wise:

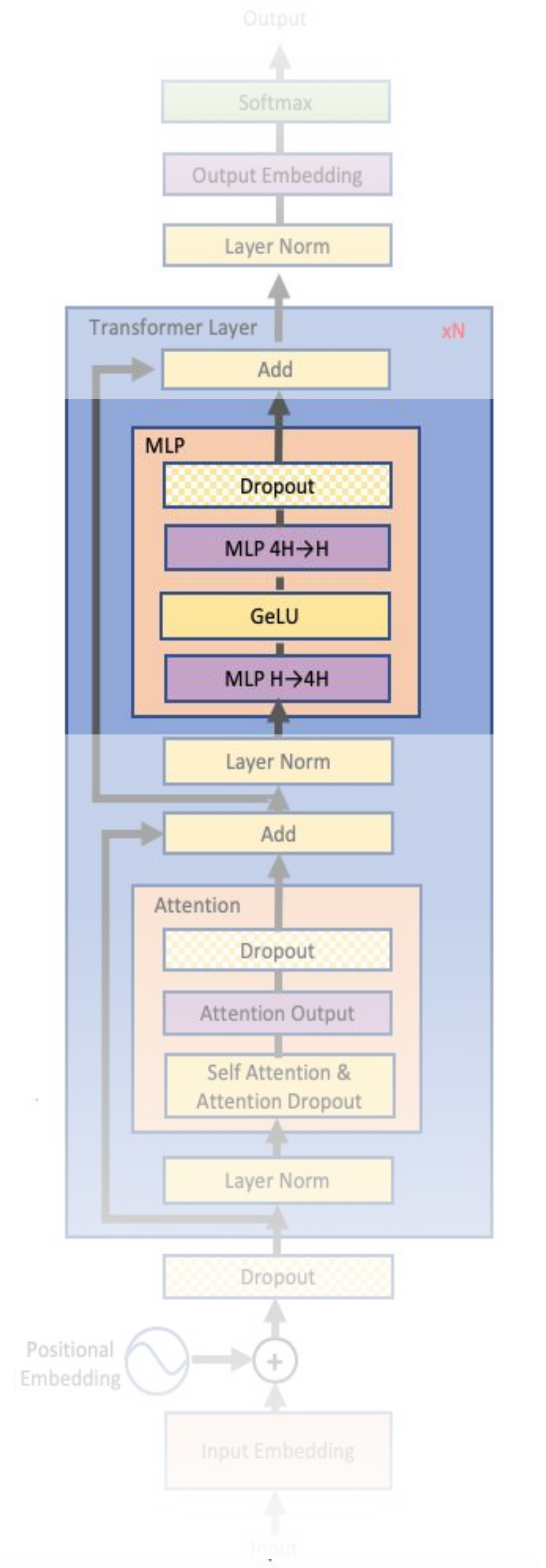
$$X = [X_1, X_2] \quad A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \longrightarrow Y = \mathrm{GeLU}(X_1A_1 + X_2A_2)$$

  - Before GeLU, we will need a communication point

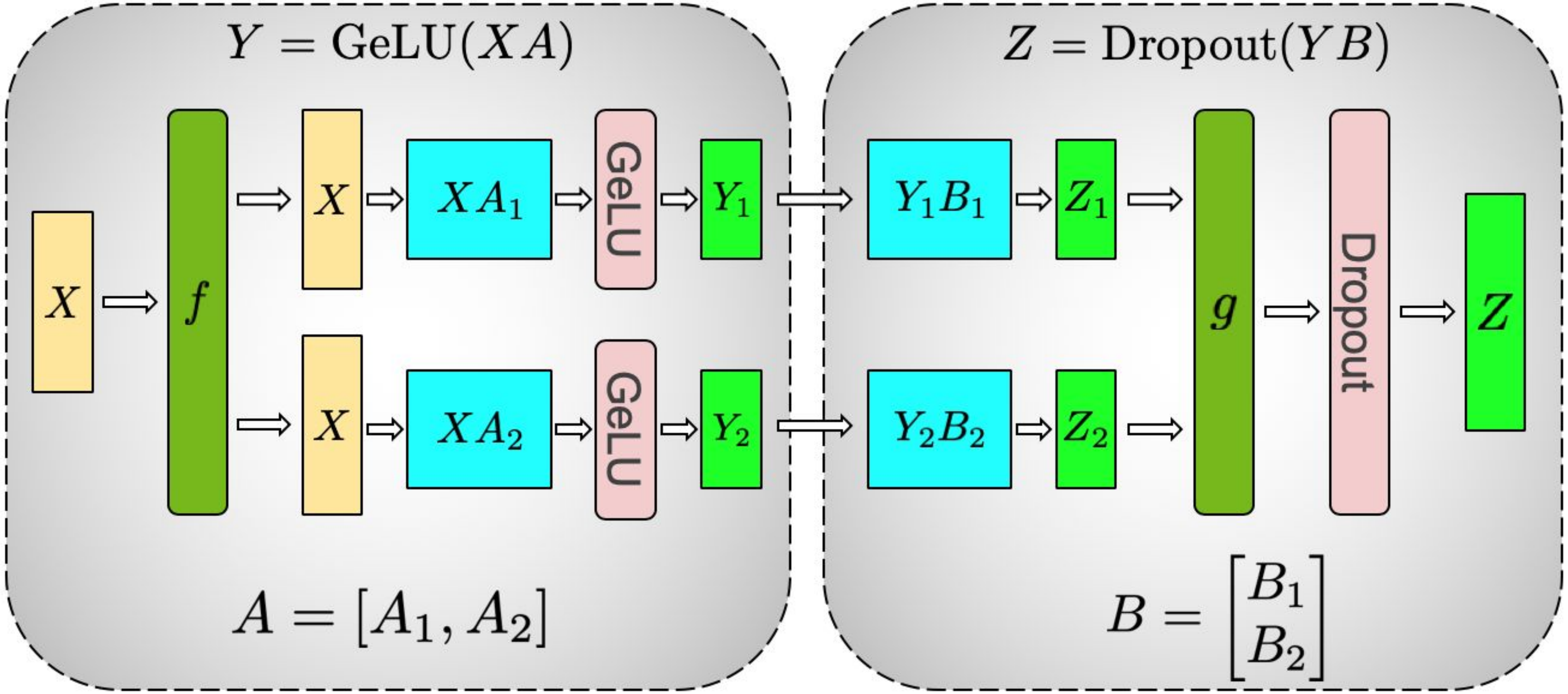
- Approach 2: split A column-wise:

$$A = [A_1, A_2] \longrightarrow [Y_1, Y_2] = [\mathrm{GeLU}(XA_1), \mathrm{GeLU}(XA_2)]$$

  - no communication is required

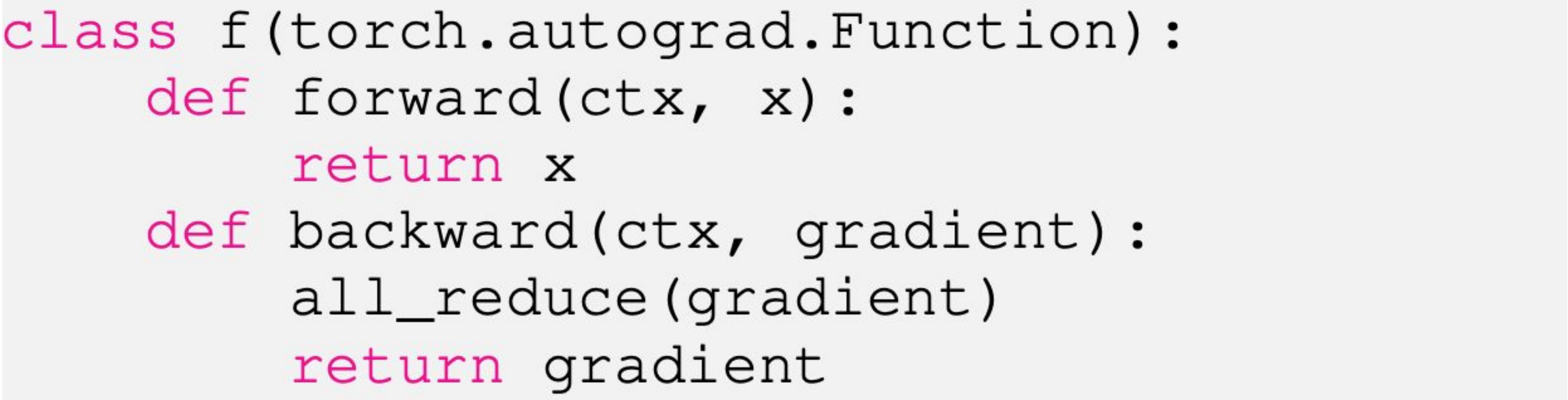


MLP



f and ***g*** are **conjugate**, ***f*** is **identity** operator in the forward pass and **all-reduce** in the backward pass while ***g*** is **all-reduce** in forward and **identity** in backward.

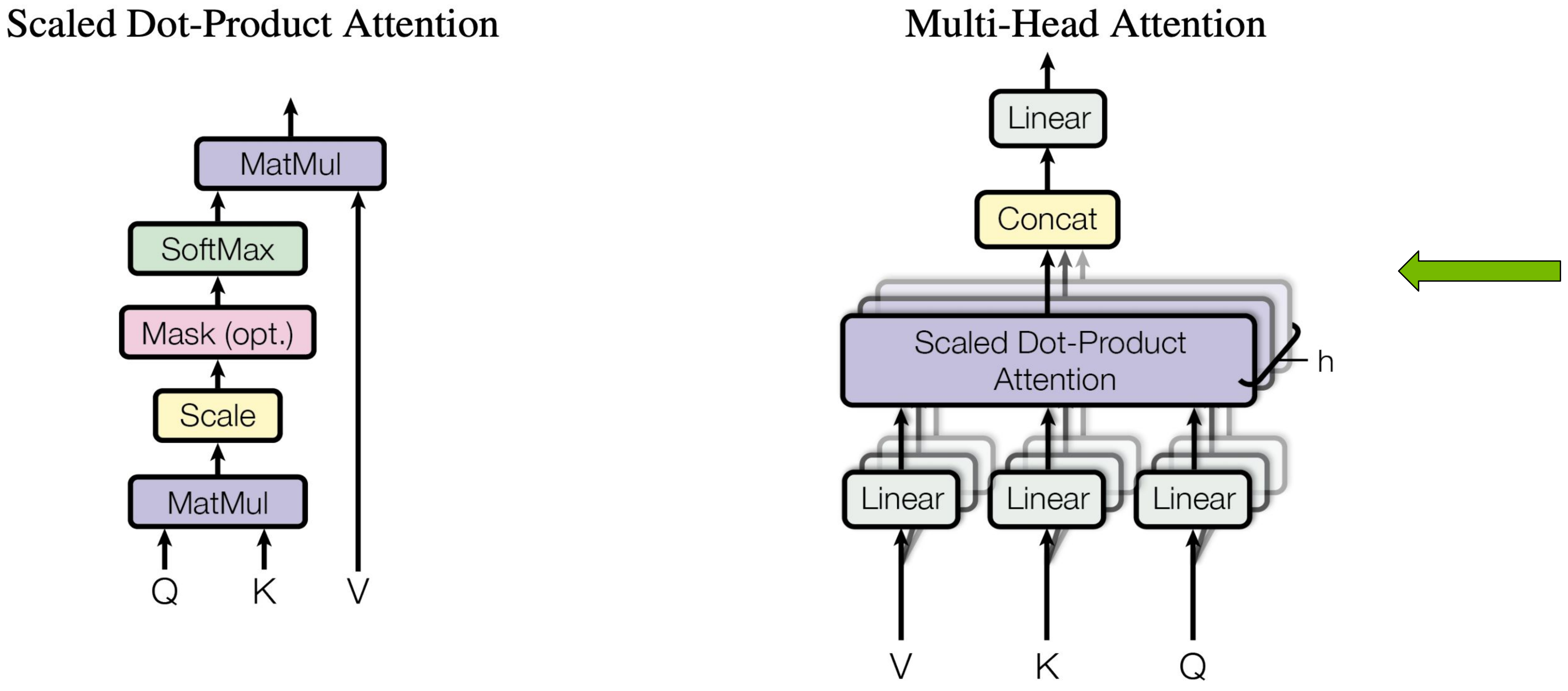
f AND g ARE SIMPLE



```
class f(torch.autograd.Function):
    def forward(ctx, x):
        return x
    def backward(ctx, gradient):
        all_reduce(gradient)
        return gradient
```

PARTITIONING: SELF-ATTENTION

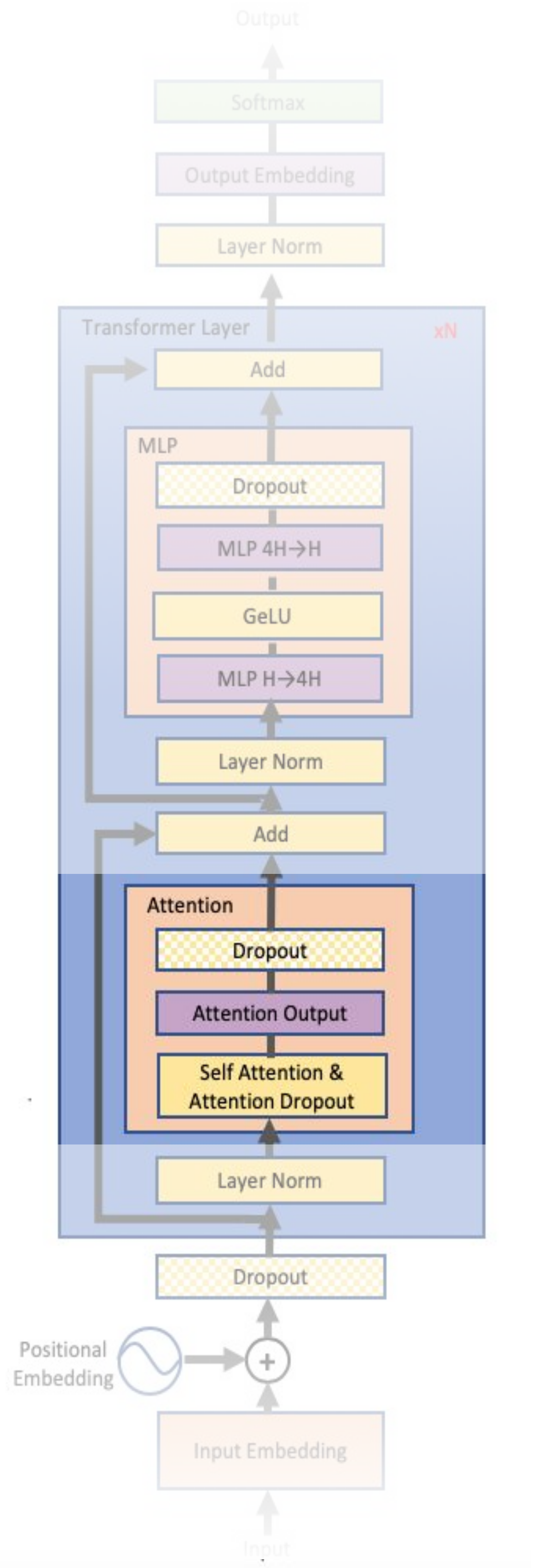
- Self-attention is more complex than MLP



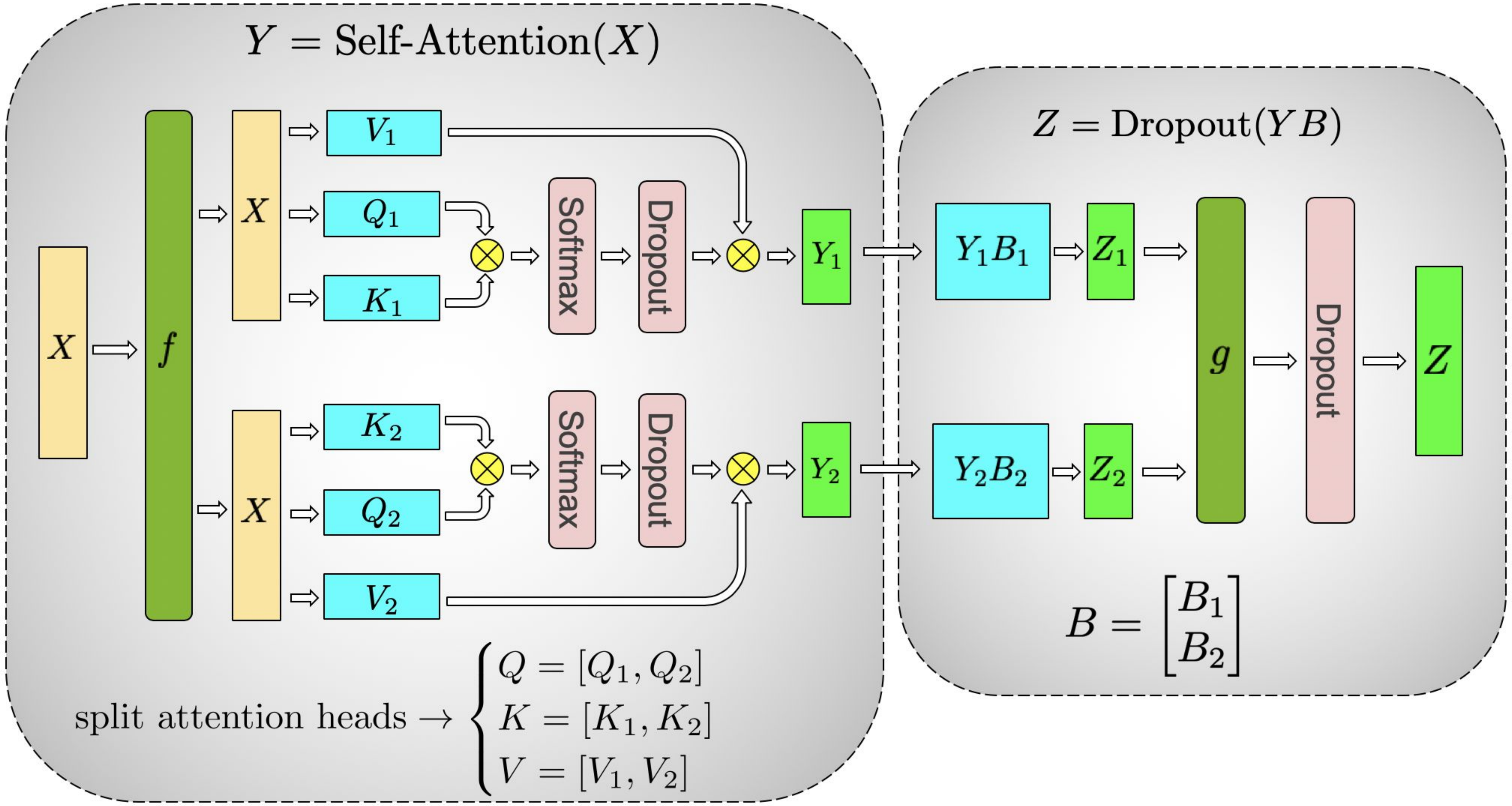
Scaled Dot-Product Attention

Multi-Head Attention

Figure courtesy of Vaswani et al. 2017

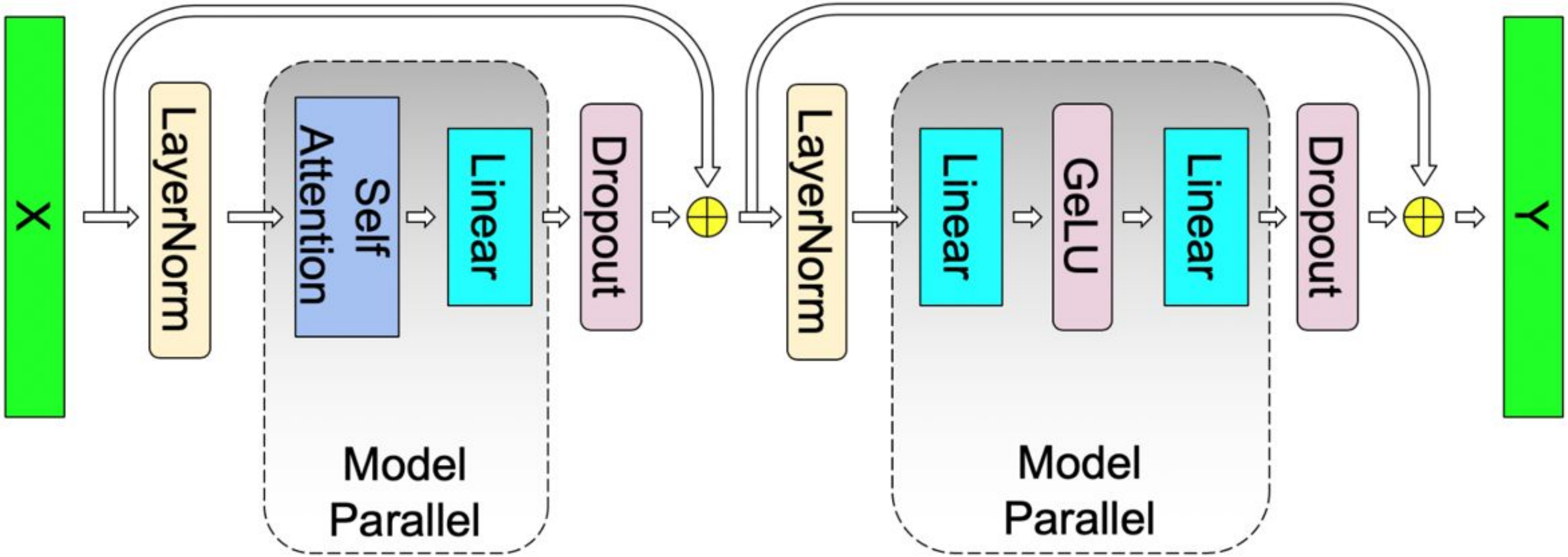


SELF-ATTENTION



f and ***g*** are **conjugate**, ***f*** is **identity** operator in the forward pass and **all-reduce** in the backward pass while ***g*** is **all-reduce** in forward and **identity** in backward.

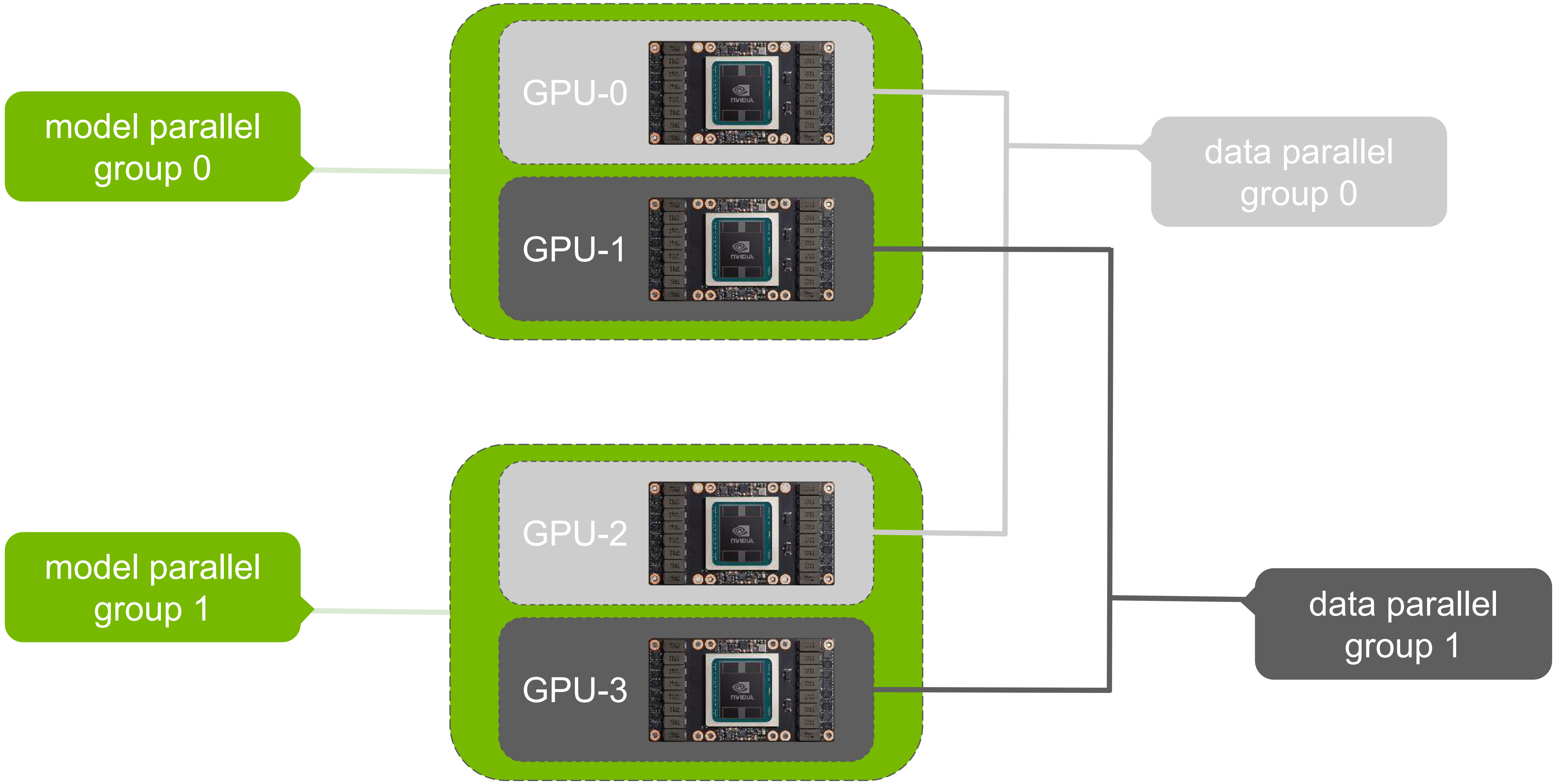
PARALLEL TRANSFORMER LAYER



2 All-Reduce
(forward + backward)

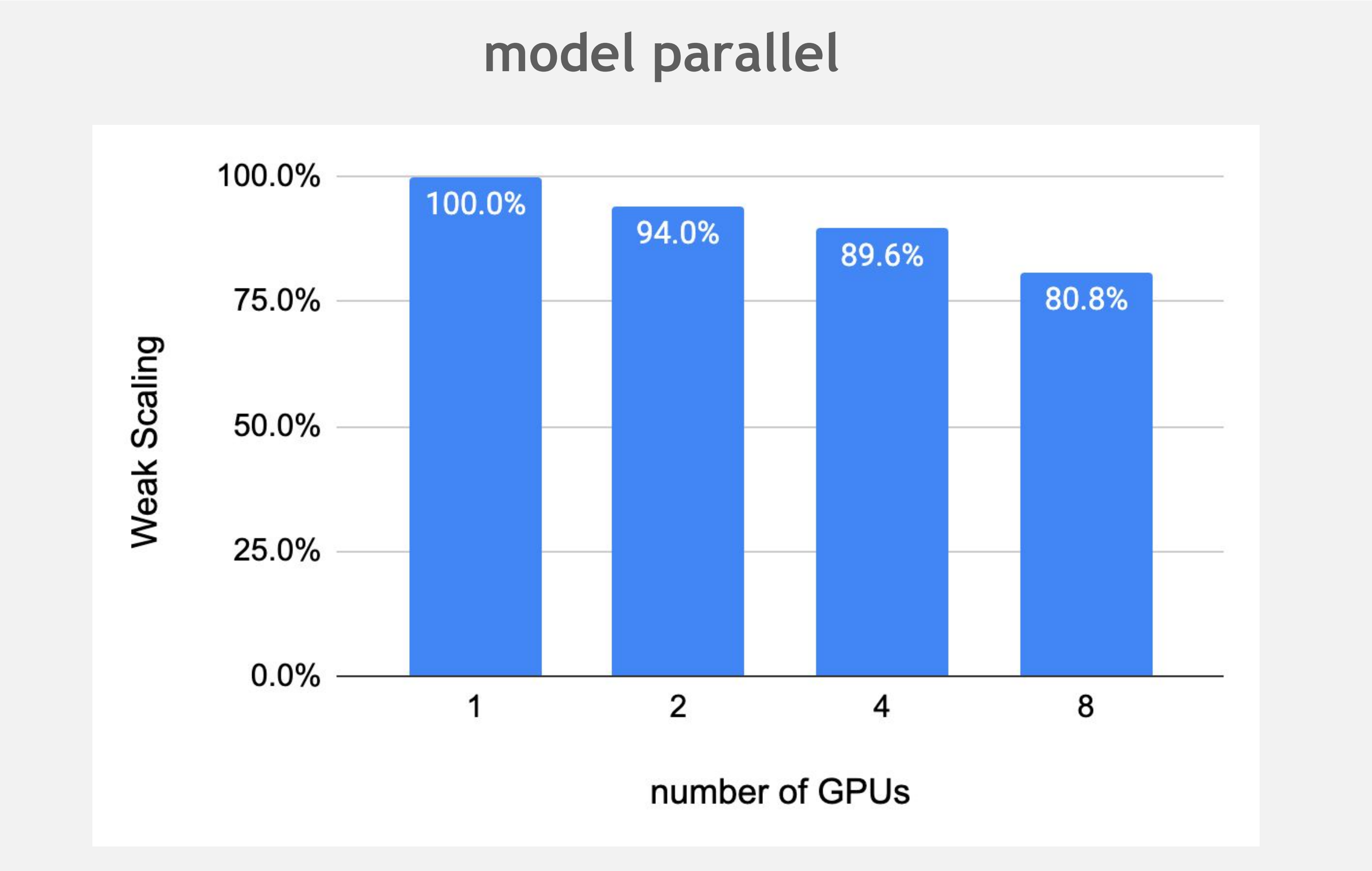
2 All-Reduce
(forward + backward)

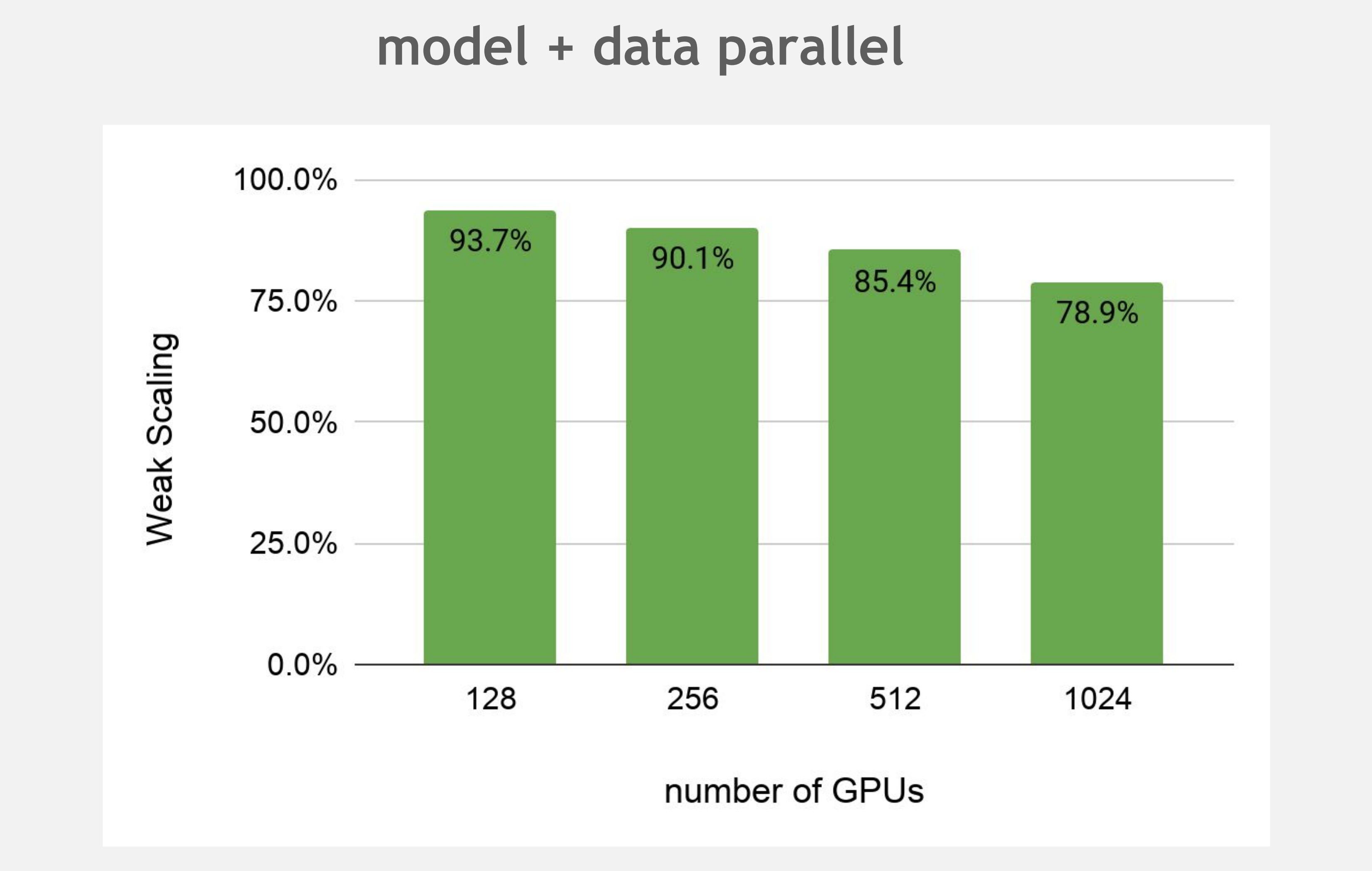
HYBRID MODEL+DATA PARALLELISM



WEAK SCALING EFFICIENCY ON SELENE

case	hidden size	attention heads	num layers	num parameters (billions)	model parallel size	model+data parallel size
1B	1920	15	24	1.2	1	128
2B	2304	18	30	2.0	2	256
4B	3072	24	36	4.2	4	512
8B	4096	32	42	8.7	8	1024



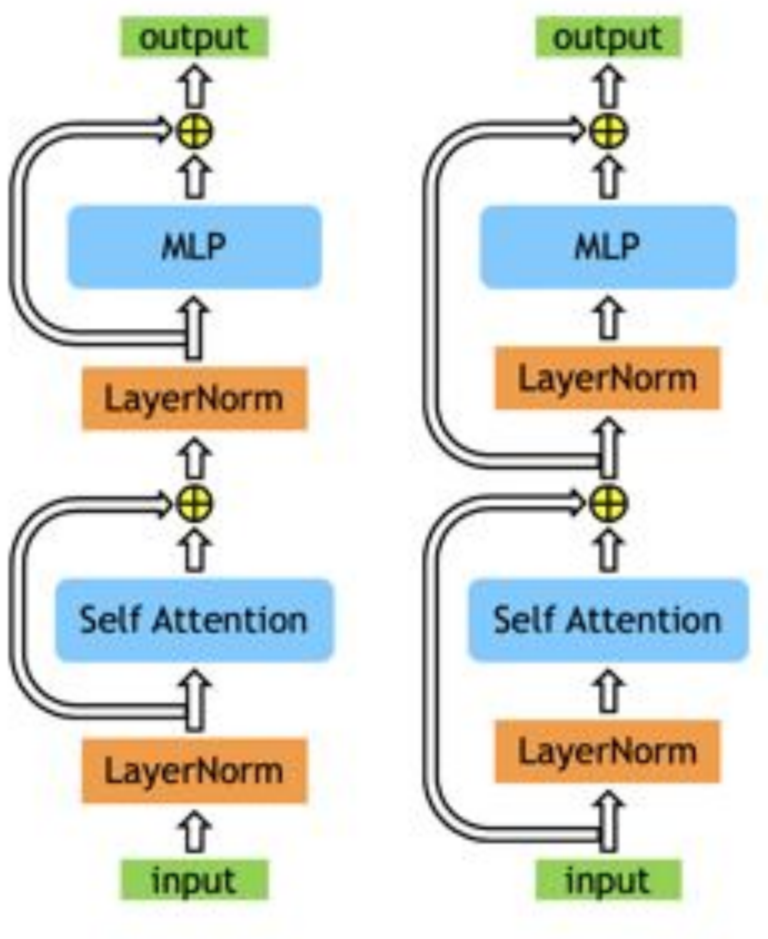


Baseline (1.2B parameters on a single GPU) sustains **118 teraFLOPs/sec on an A100 GPU** during the entire training process.

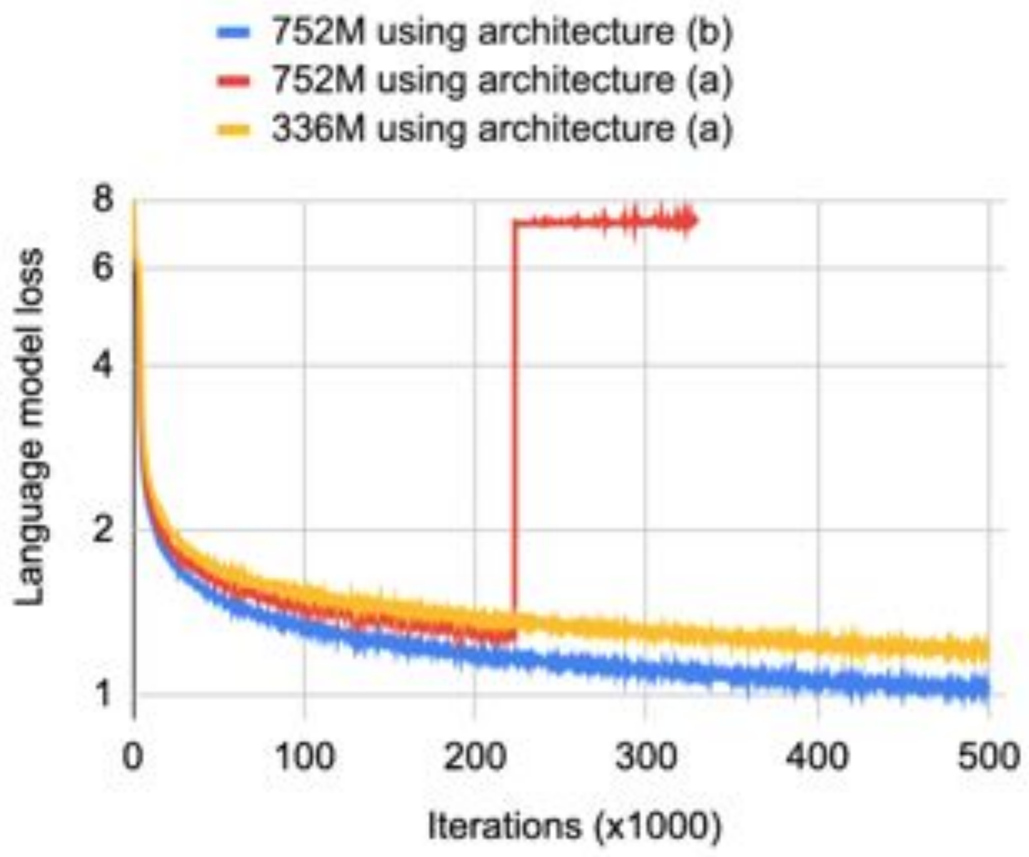
SIDE CHALLENGES OF TRAINING LARGE MODELS

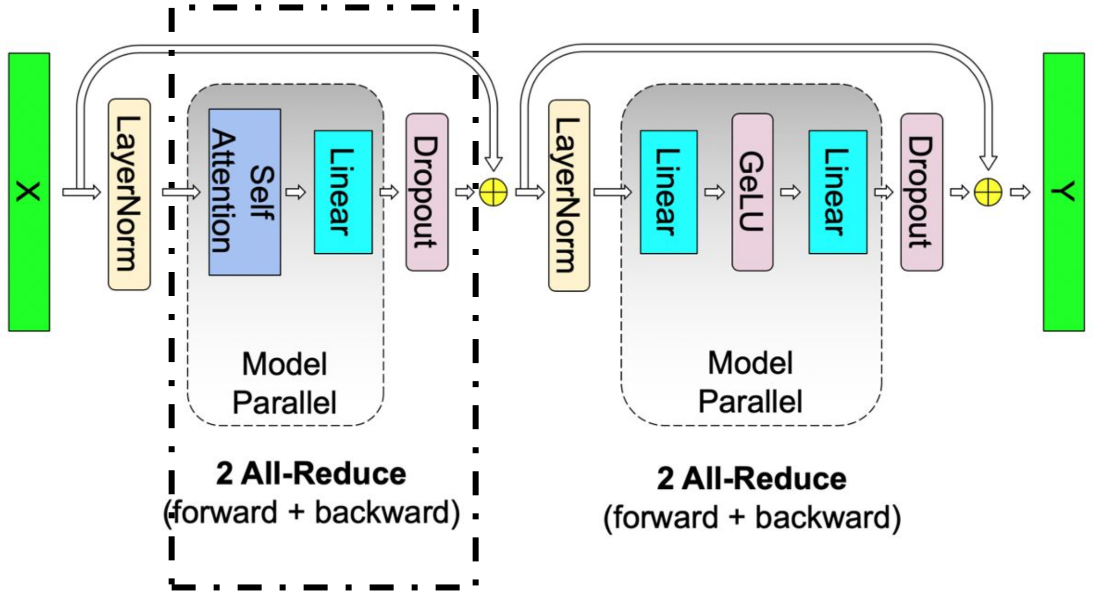
CHANGES IN MODEL STRUCTURE AND INITIALIZATION

- Scaling BERT model as presented in the original work beyond 336M parameters results in instabilities
- Rearranging the residual connection to allow for the direct flow of gradients is necessary
- However, the weights right before the residual connection need to be initialized with a much lower variance



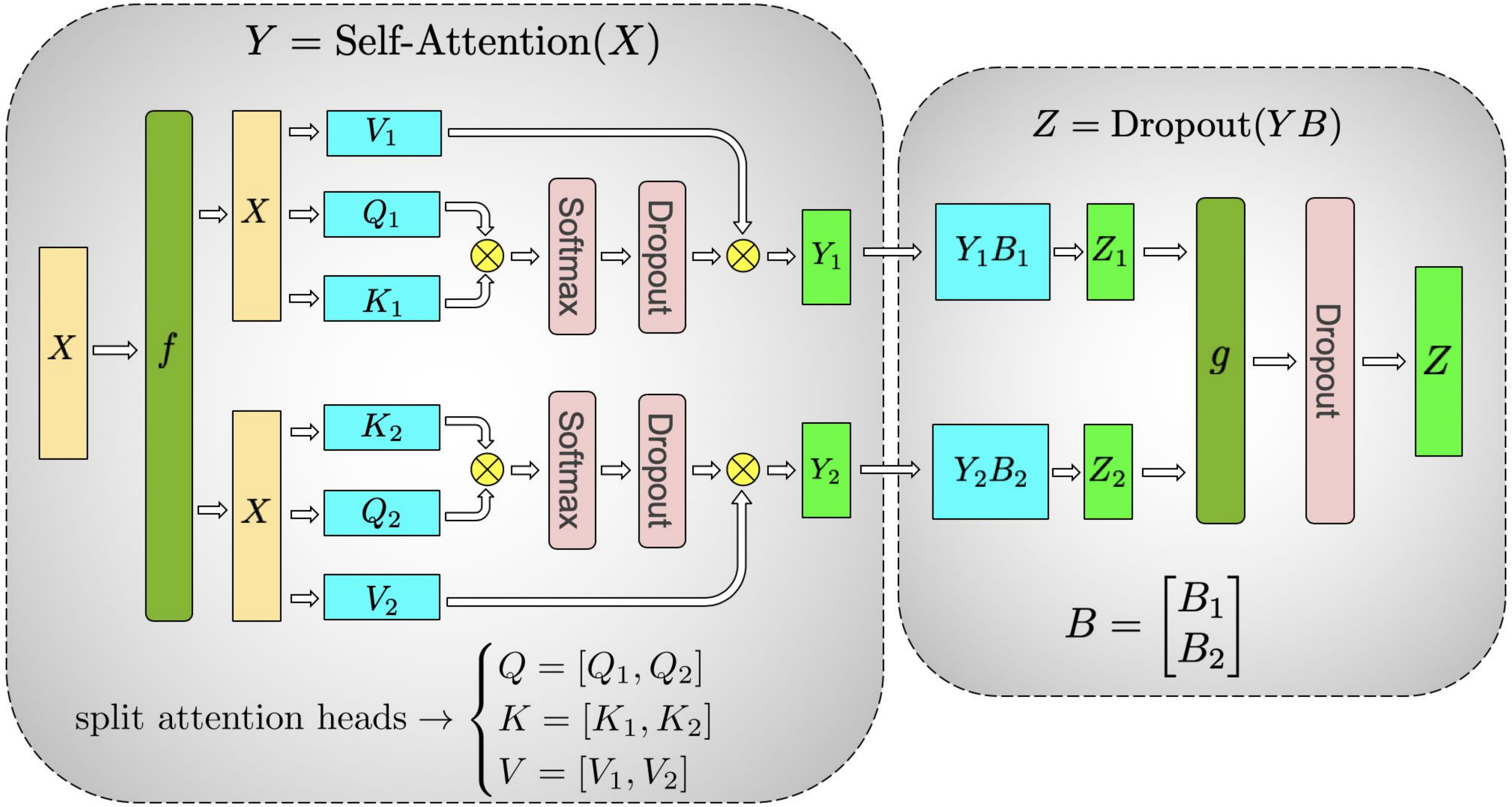
(a) (b)





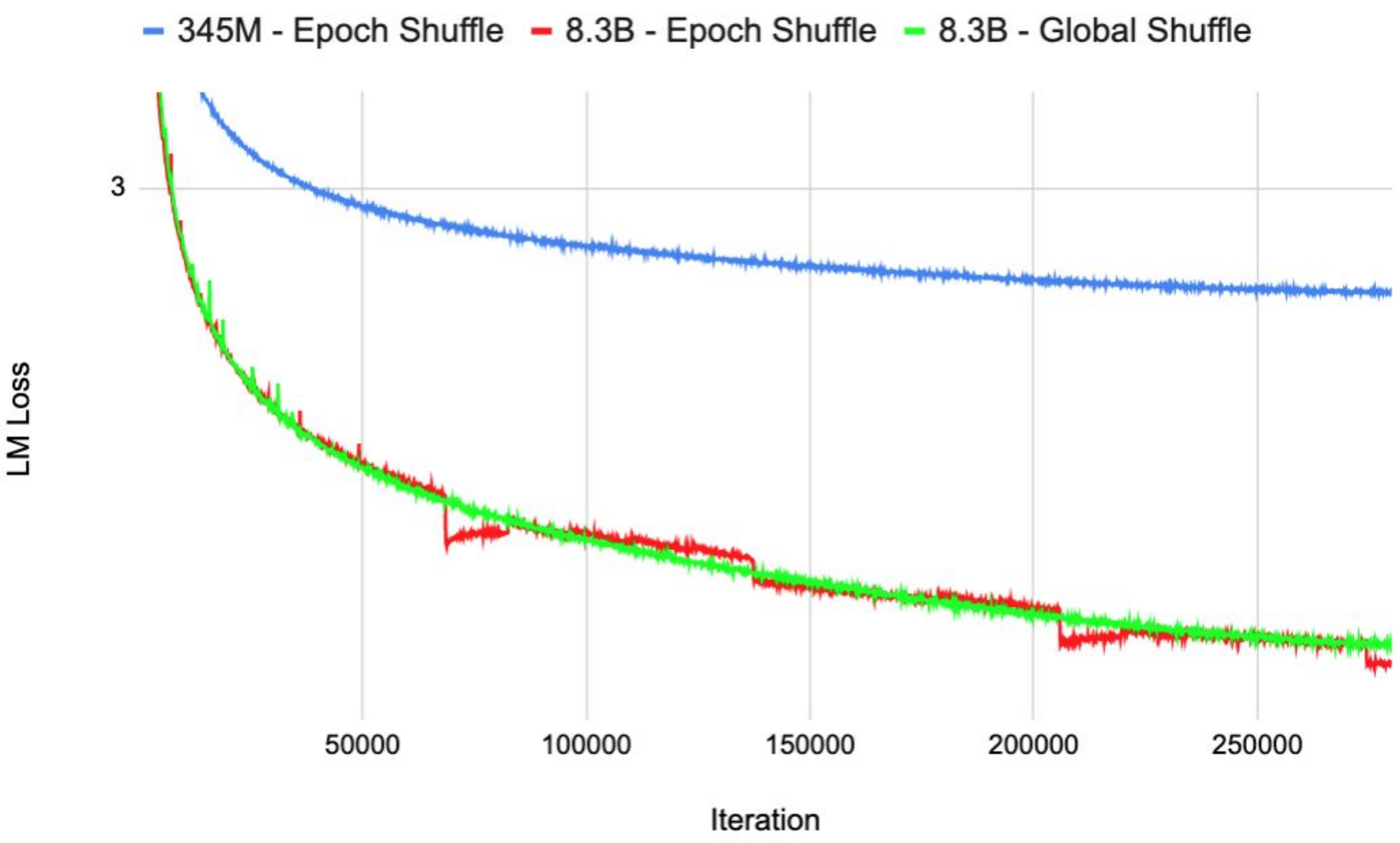
RANDOM NUMBERS IN SELF-ATTENTION

- Tensor splitting results in both serial and parallel regions
- For a model instance
  - serial regions need to have the same RNG sate
  - parallel regions need to have different RNG states
- As a result we need to track two sets of RNG states


$Y = \text{Self-Attention}(X)$
X
f
X
V_1
Q_1
K_1
Softmax
Dropout
Y_1
X
K_2
Q_2
V_2
Softmax
Dropout
Y_2
split attention heads → $\begin{cases} Q = [Q_1, Q_2] \\ K = [K_1, K_2] \\ V = [V_1, V_2] \end{cases}$
$Z = \text{Dropout}(YB)$
Y_1B_1
Z_1
Y_2B_2
Z_2
g
Dropout
Z
$B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$

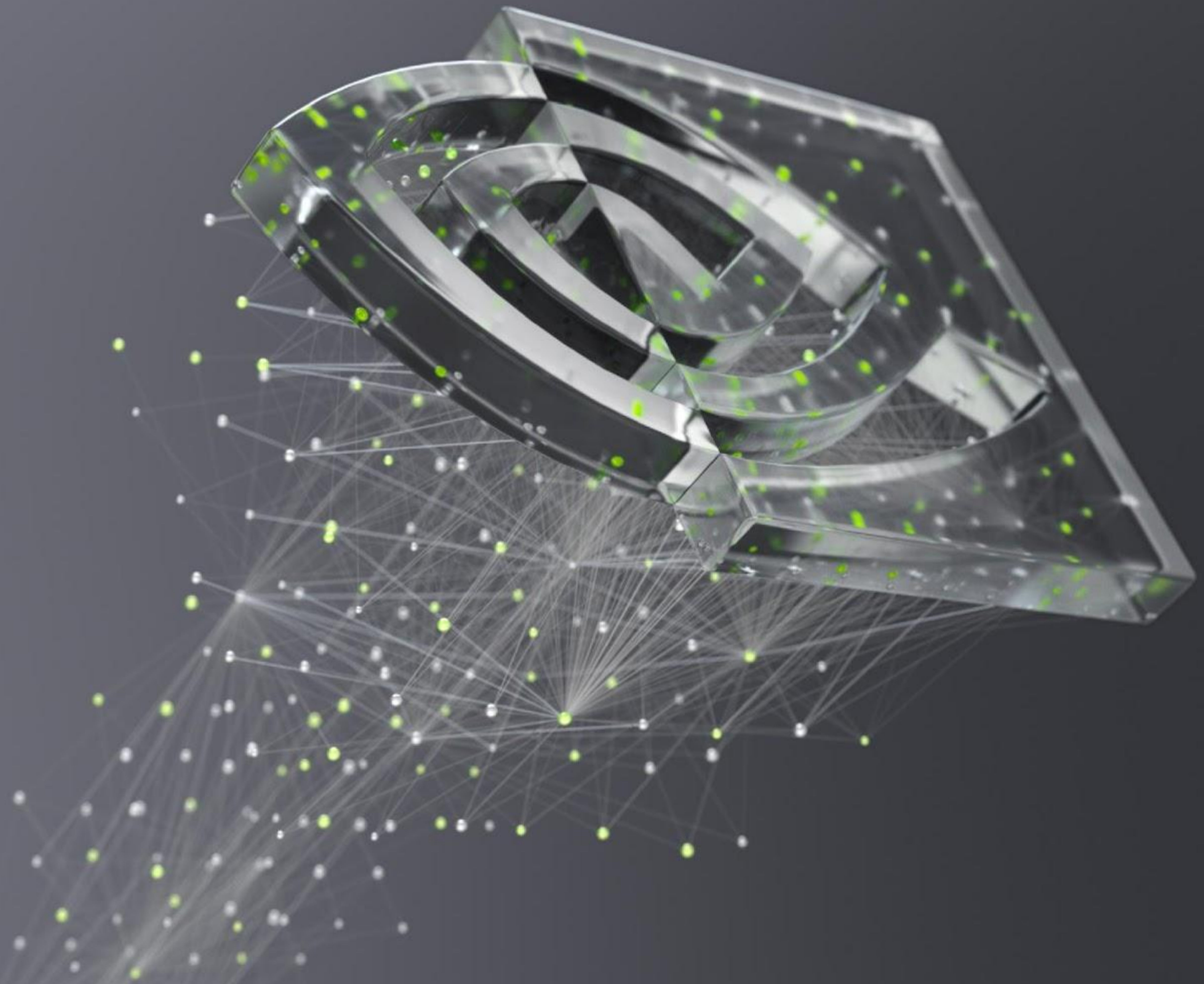
LARGE MODELS ARE MORE SENSITIVE TO DATA SHUFFLING

- Small models are insensitive to the dataset order
- Larger models have much higher power of memorization and as a result more sensitive to shuffling scheme
- Special attention to dataloaders is required for large models



SUMMARY

- Ability to train large language models is essential for today’s NLP application.
- Intra-layer model parallelism is powerful yet simple for NLP models.
- We can scale models efficiently to billions of parameters on thousands of GPUs.
- When scaling models, careful attention to model structure and data loaders is required



NVIDIA