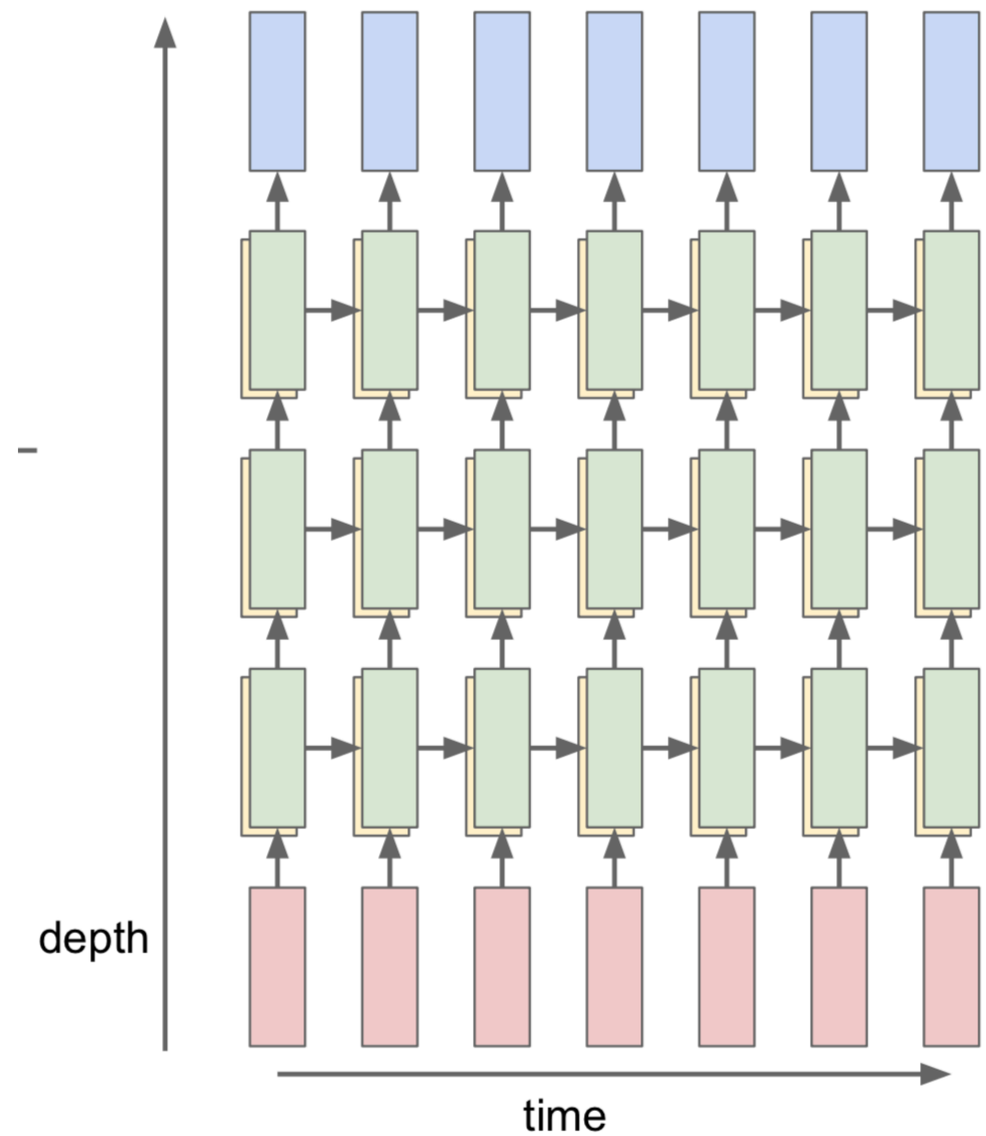
Lec 22: Attention

Yaoliang Yu

Jul 23, 2020

UNIVERSITY OF WATERLOO

Pros and Cons of RNN



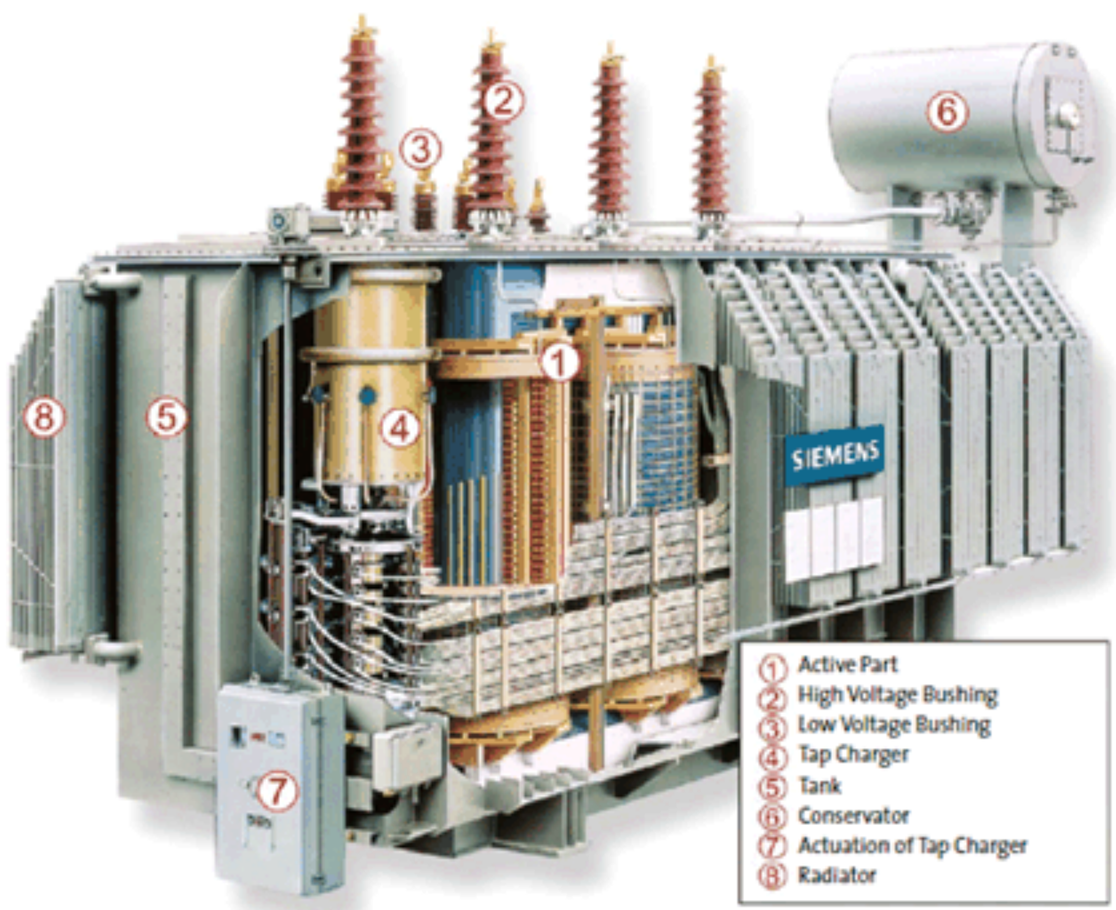
- Sequential in both directions; slow and expensive!

- Trade depth for time

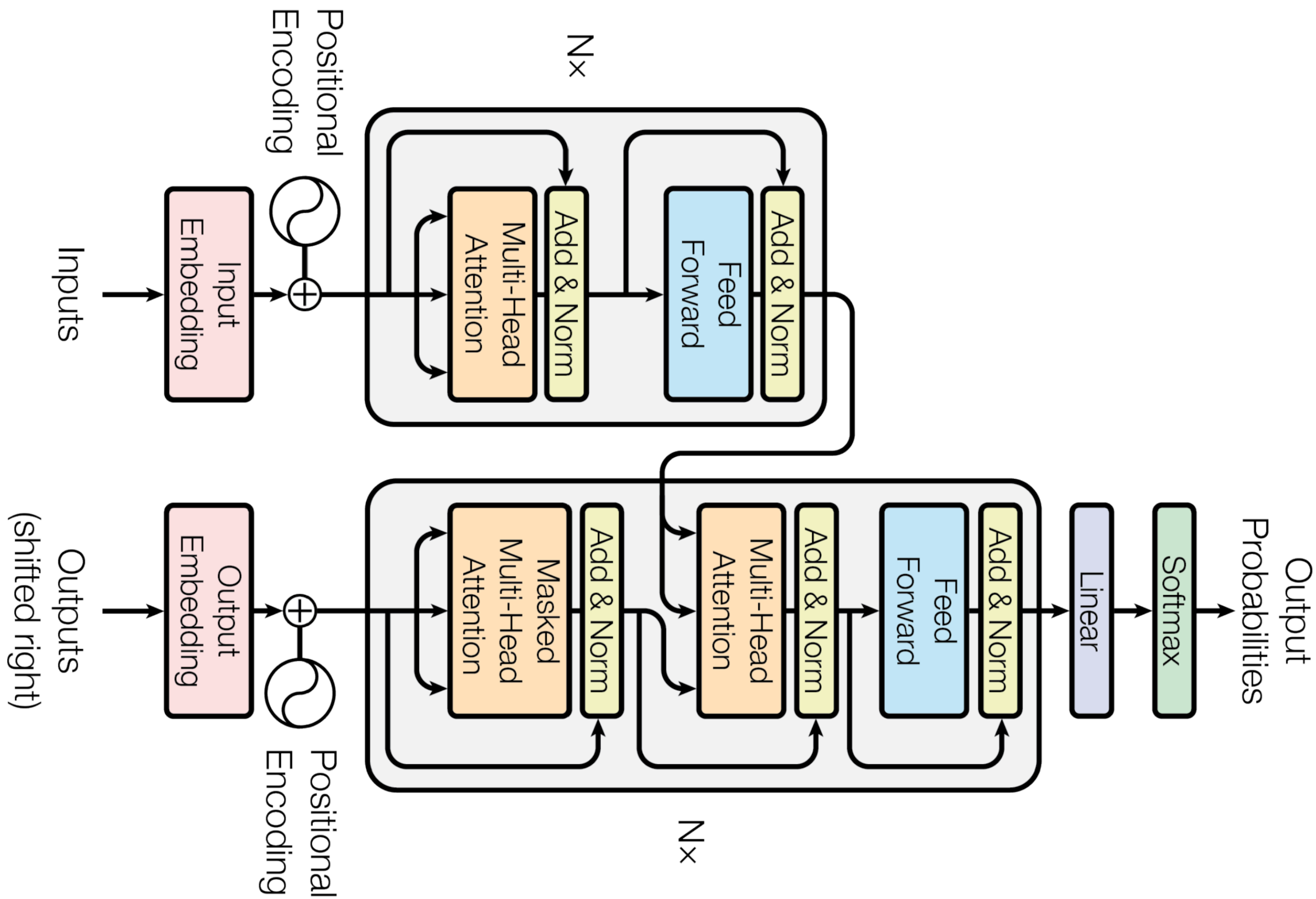
Transformer

https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Transformers (I know)



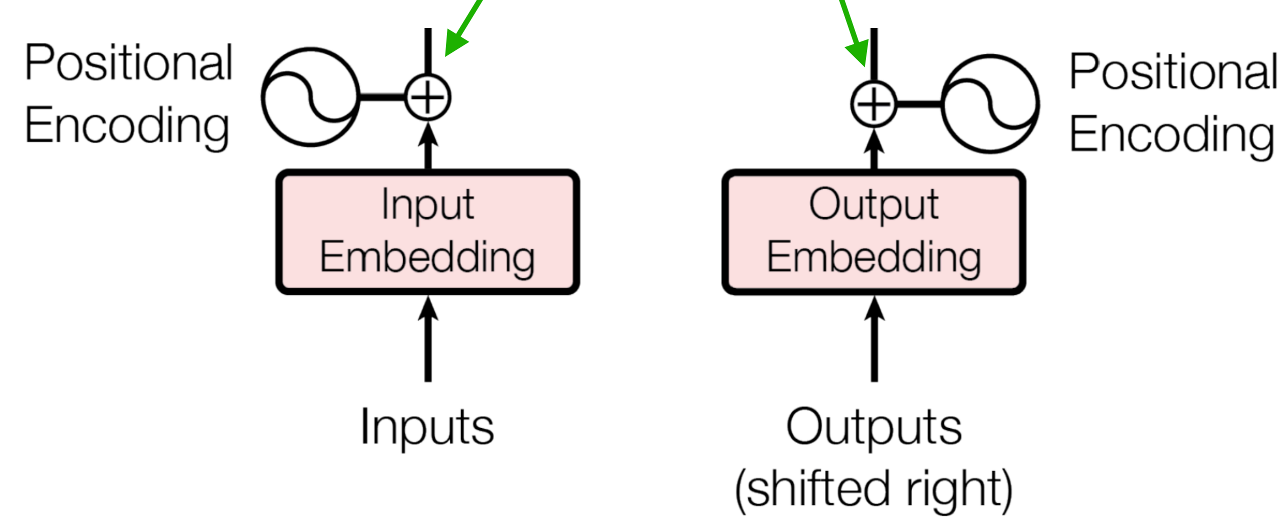
Transformer in 1 Fig



x and y

- Input sequence $x = (x_1, x_2, \ldots, x_m)^T$, $x_j \in \mathbb{R}^p$ one-hot
- Output sequence $y = (y_1, y_2, \ldots, y_l)^T$, $y_j \in \mathbb{R}^p$ one-hot
- Embedding: xW_e and yW_e, $W_e \in \mathbb{R}^{p \times d}$
- Positional encoding:

$$\boxed{\mathfrak{p}_{t,2i} = \sin\left(t/10000^{2i/d}\right), \quad \mathfrak{p}_{t,2i+1} = \cos\left(t/10000^{2i/d}\right), \quad i = 0, \ldots, \frac{d}{2} - 1.}$$



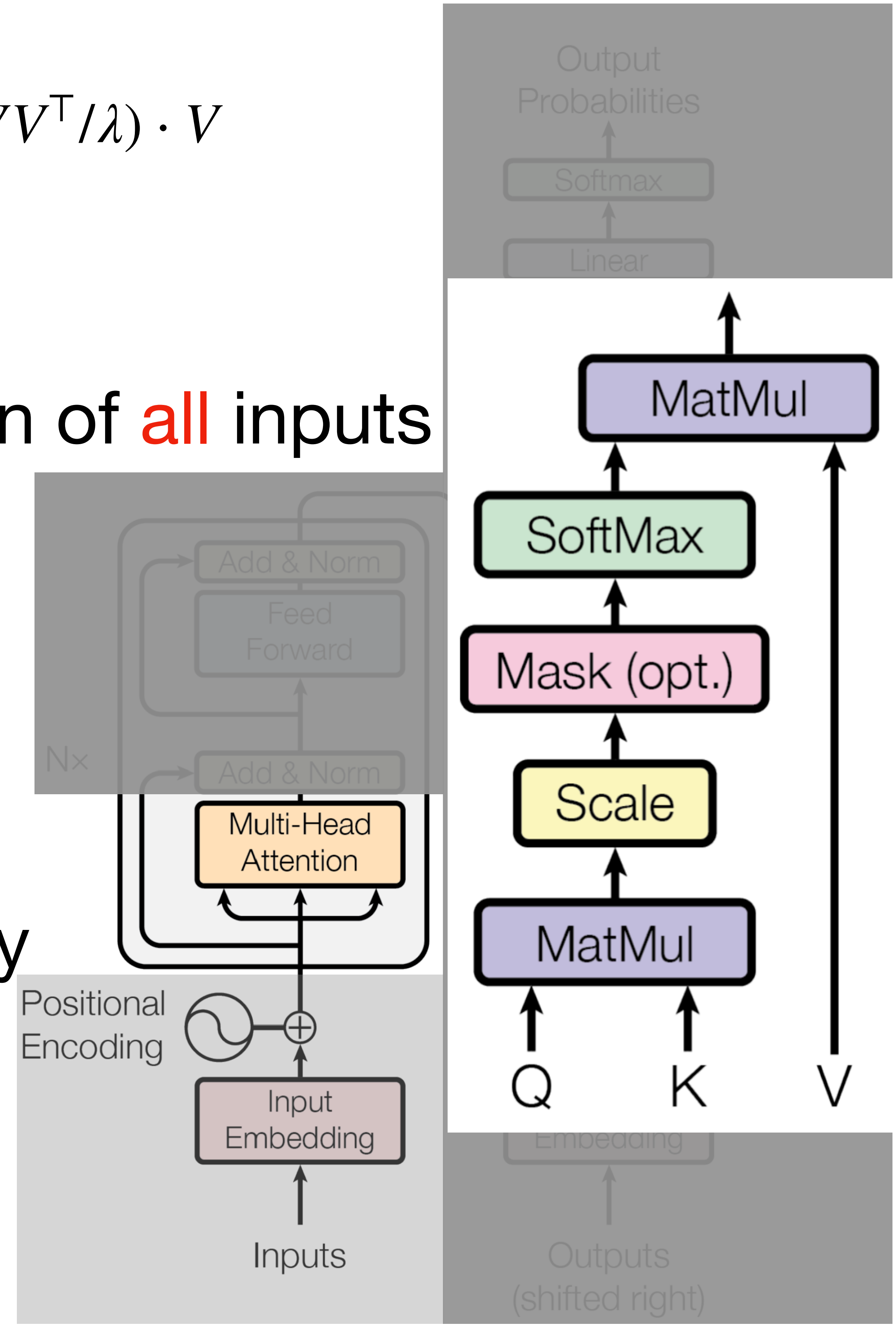
d = 512

Self-attention

$$V \leftarrow VW^V$$

$$V \leftarrow A_\lambda(V;V) = \mathtt{softmax}(VV^\top/\lambda)\cdot V$$

- Replacement of recurrence
- Each output is a convex combination of all inputs
- Matrix product highly parallelizable
- Softmax is dense
- Dot product $\mathbf{v}_i^\top \mathbf{v}_j$ measures similarity
  - *More similar, more contribution*



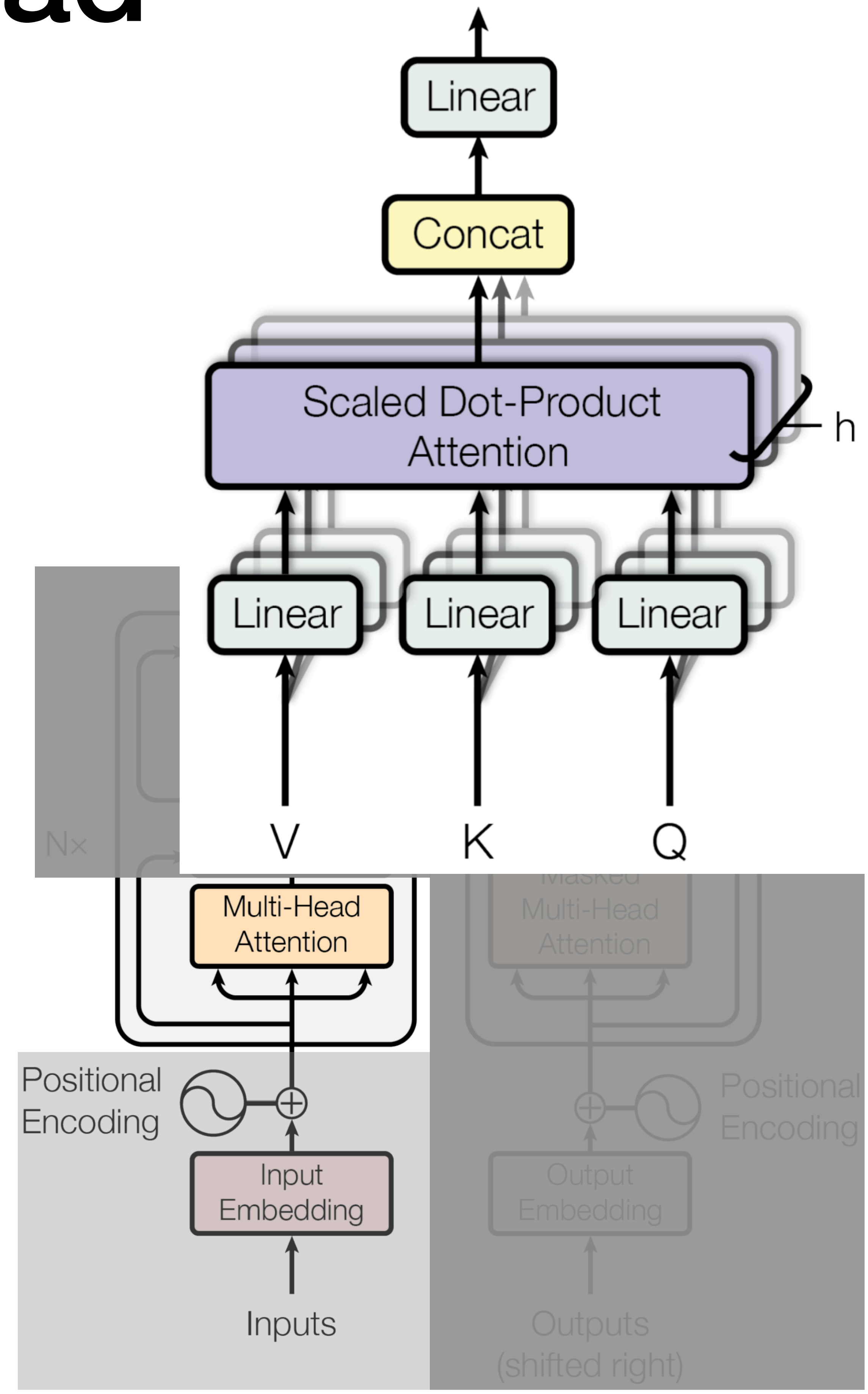
Multi-head

For $i = 1, \ldots, h$

$$V_i \leftarrow V W_i^V$$

$$V_i \leftarrow A_\lambda(V_i; V_i) = \texttt{softmax}(V_i V_i^\top / \lambda) \cdot V_i$$

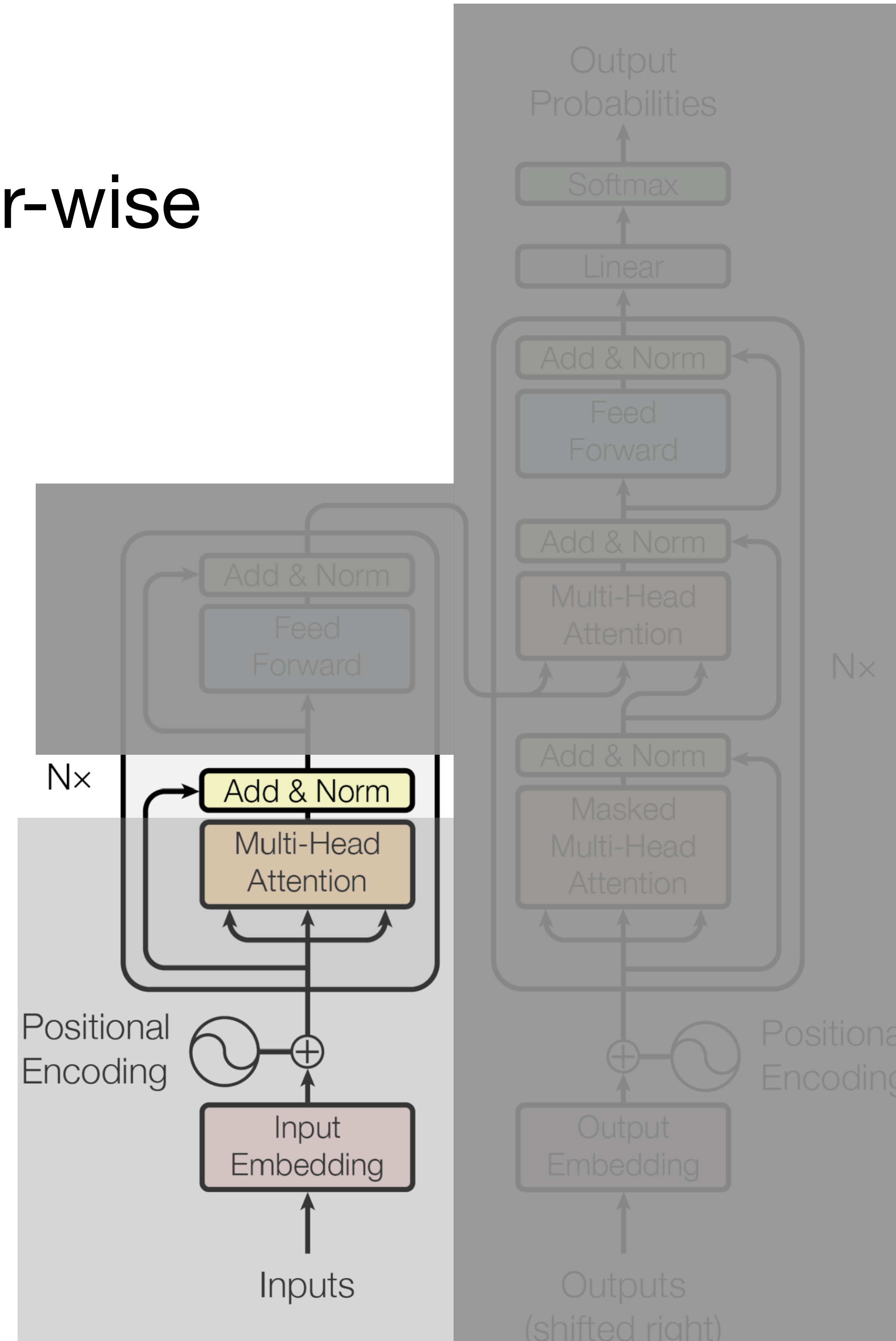
$$V \leftarrow [V_1, \ldots, V_h] W$$



H = 8

Residual & Layer Normalization

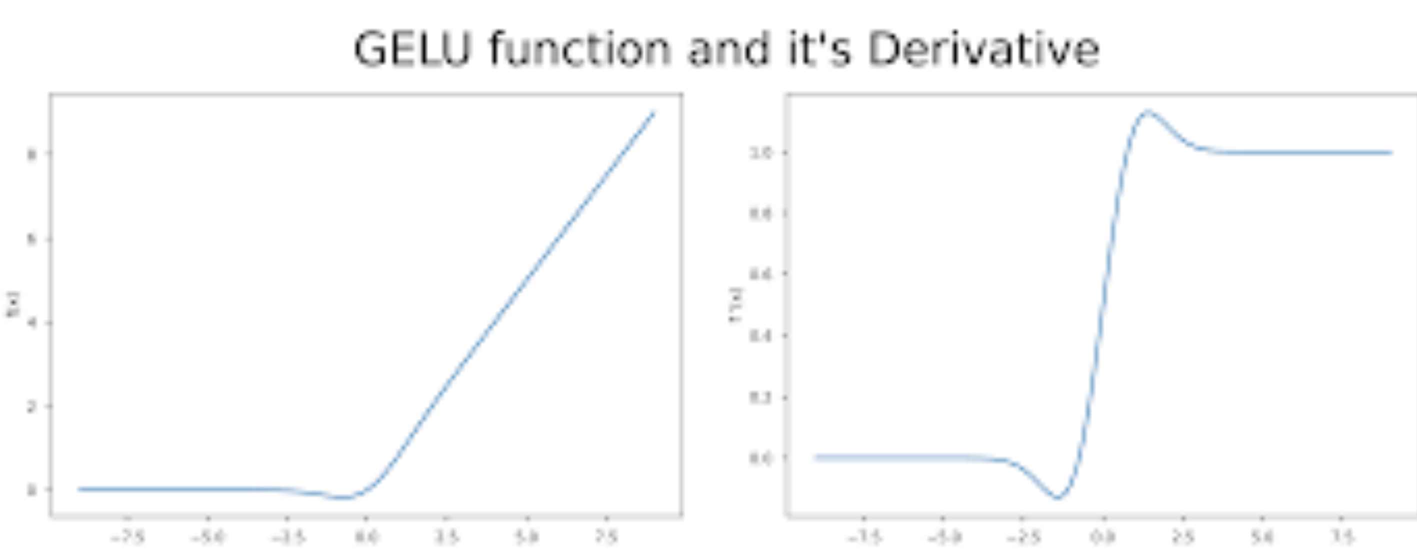
- Add residual connection and layer-wise normalization to ease training



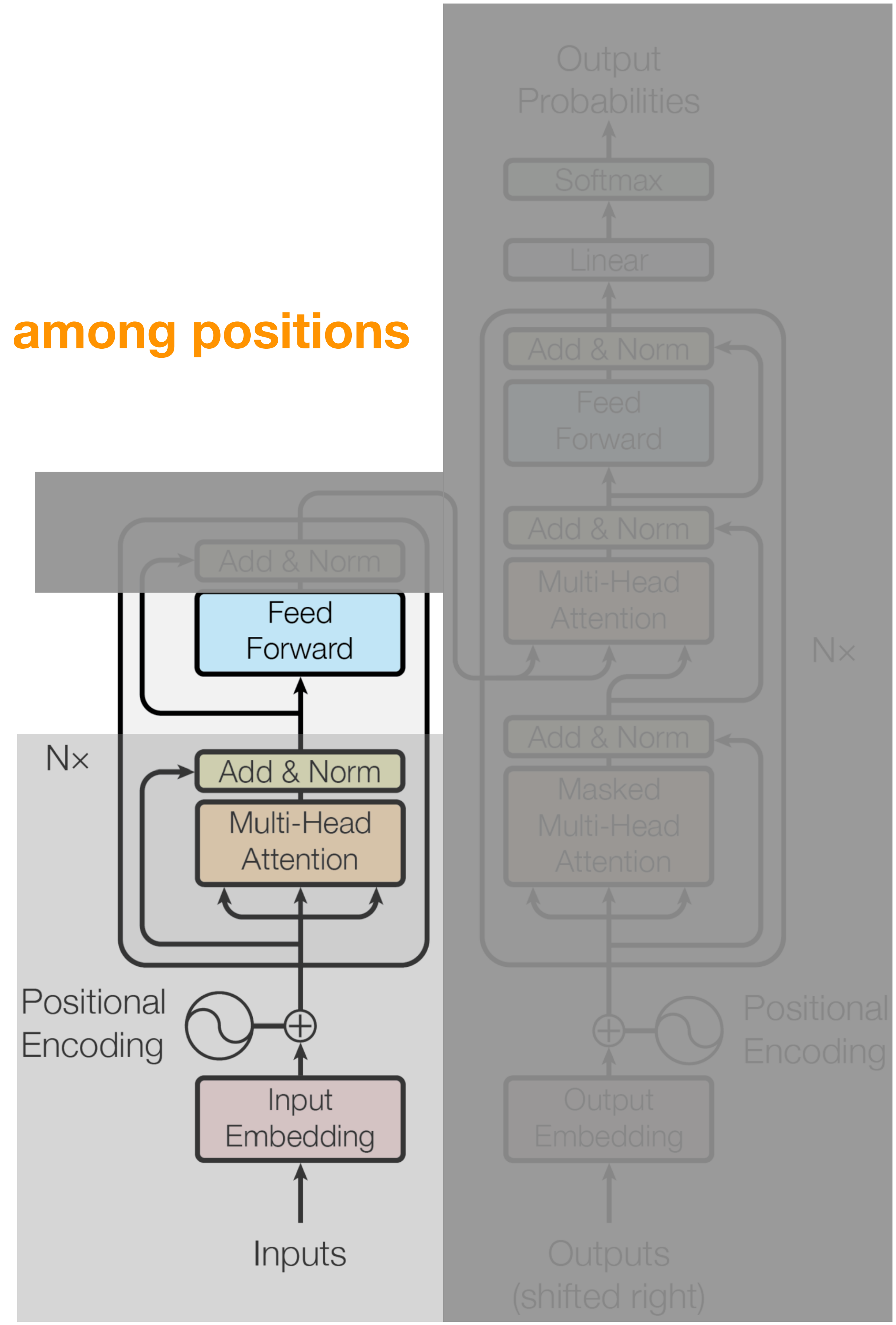
Feed-forward

$$\mathtt{FFN}(\mathbf{v}_t) = \sigma(\mathbf{v}_t W_1) W_2$$

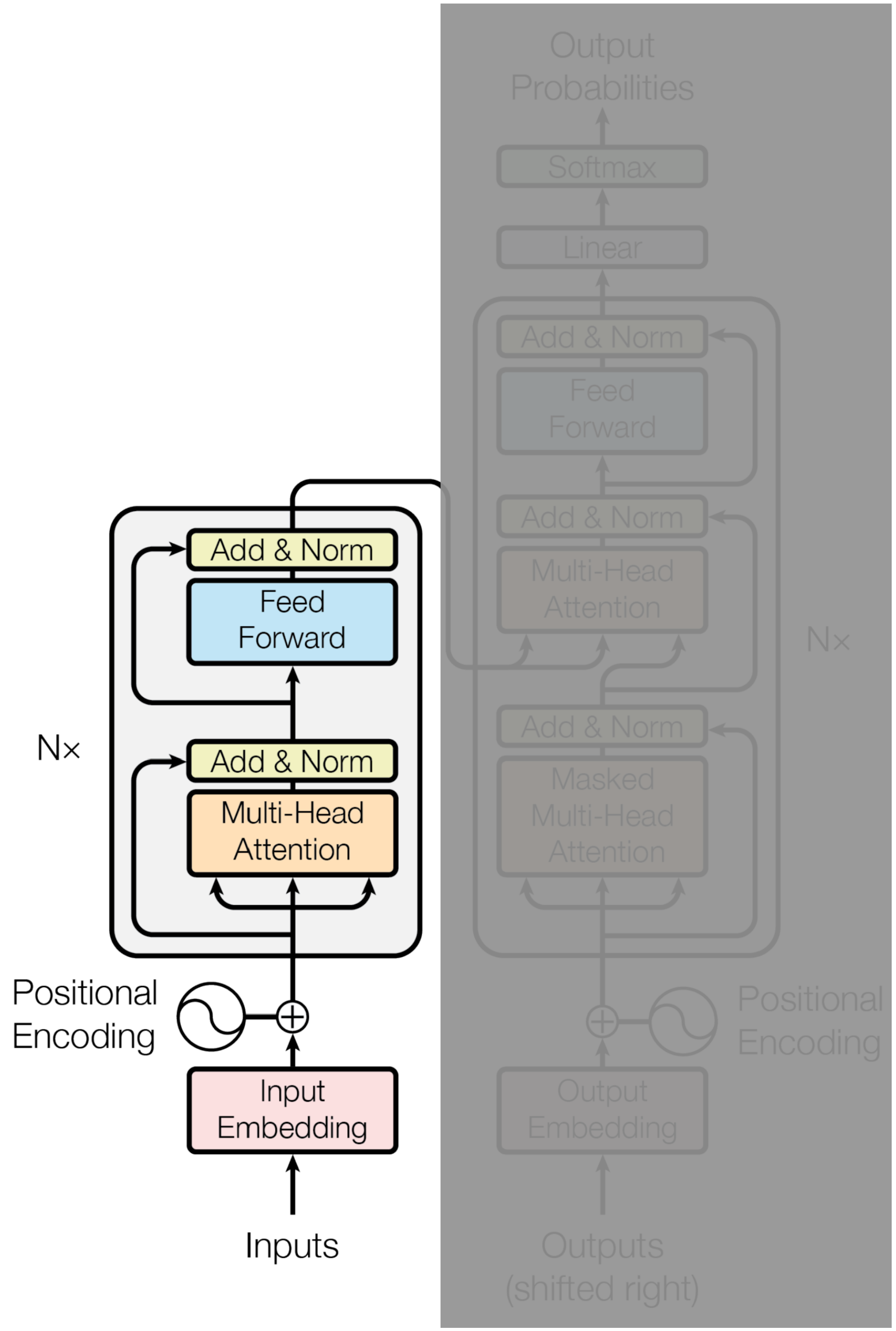
shared among positions



$$\sigma(x) = x\Phi(x)$$

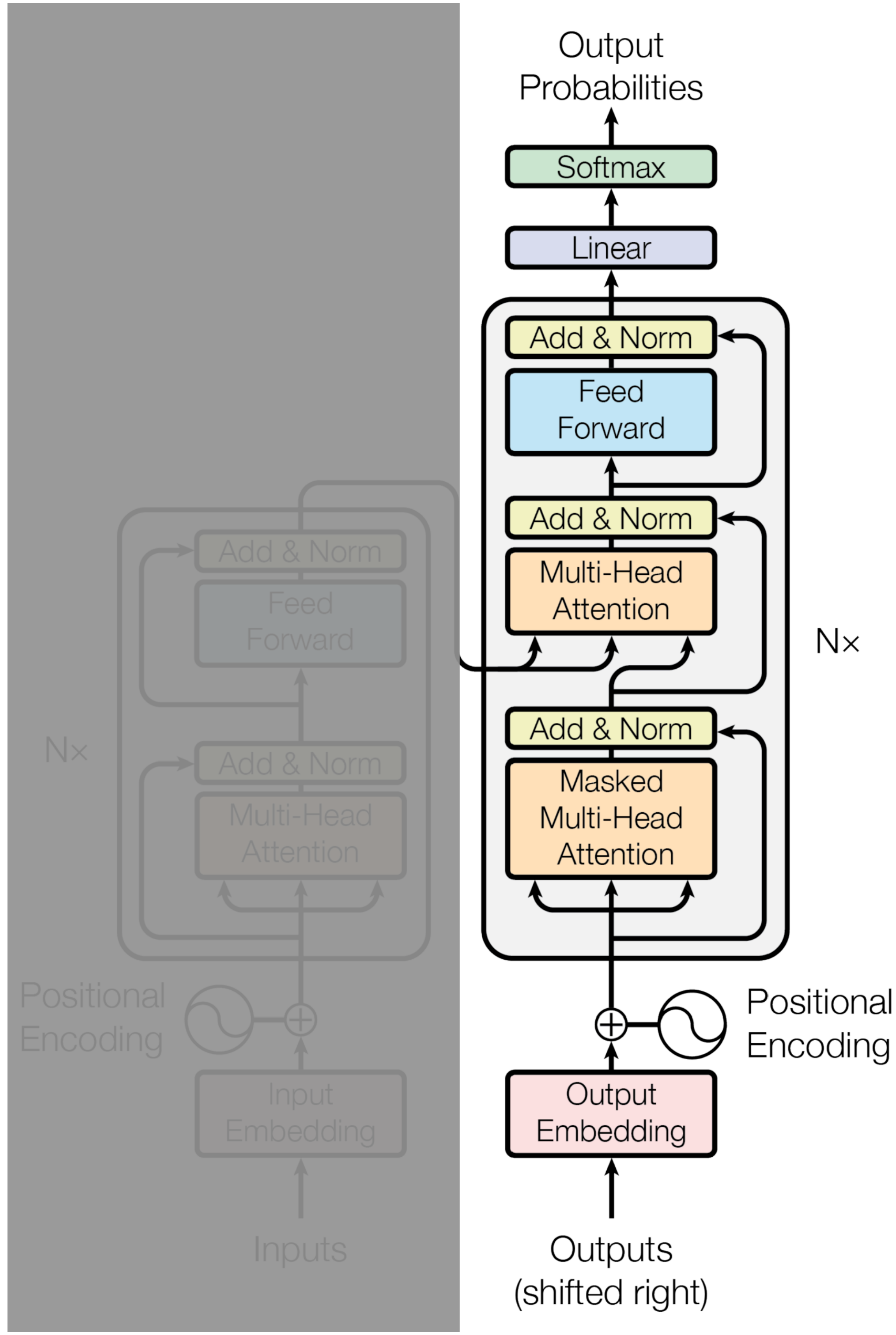


The Encoder



Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Positional
Encoding
Output
Embedding
Outputs
(shifted right)
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Input
Embedding
Inputs

The Decoder



Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
N×
Positional Encoding
Output Embedding
Outputs (shifted right)
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
N×
Positional Encoding
Input Embedding
Inputs

Masked Self-attention

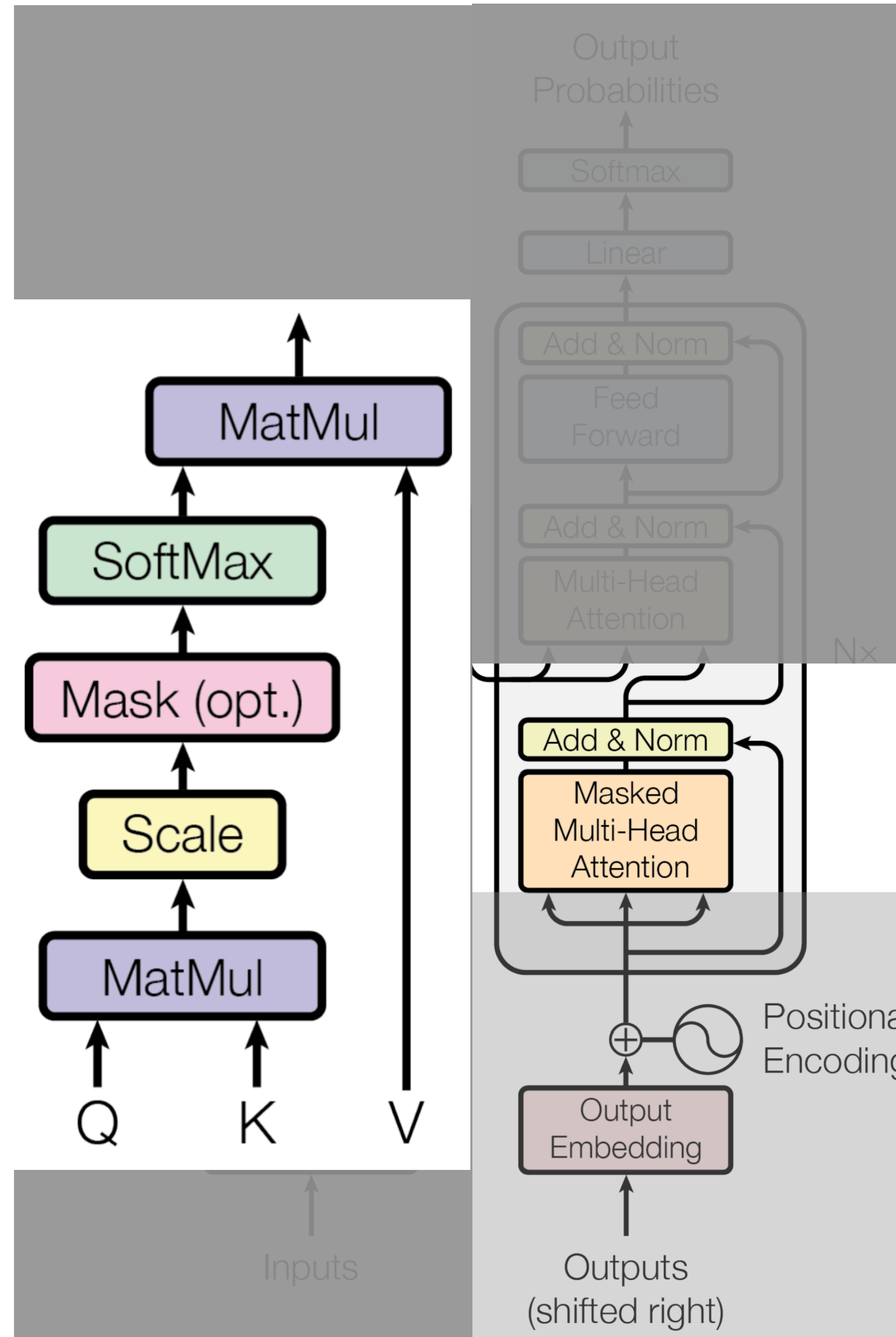
$$Q \leftarrow QW_Q$$

$$Q \leftarrow A_\lambda(Q;Q) = \texttt{softmax}(Q^\top Q/\lambda)\cdot Q$$

- Causal: Any output can only depend on previous outputs

- Reset $\mathbf{q}_i^\top \mathbf{q}_j = -\infty$ for all $i < j$

- Apply multi-head



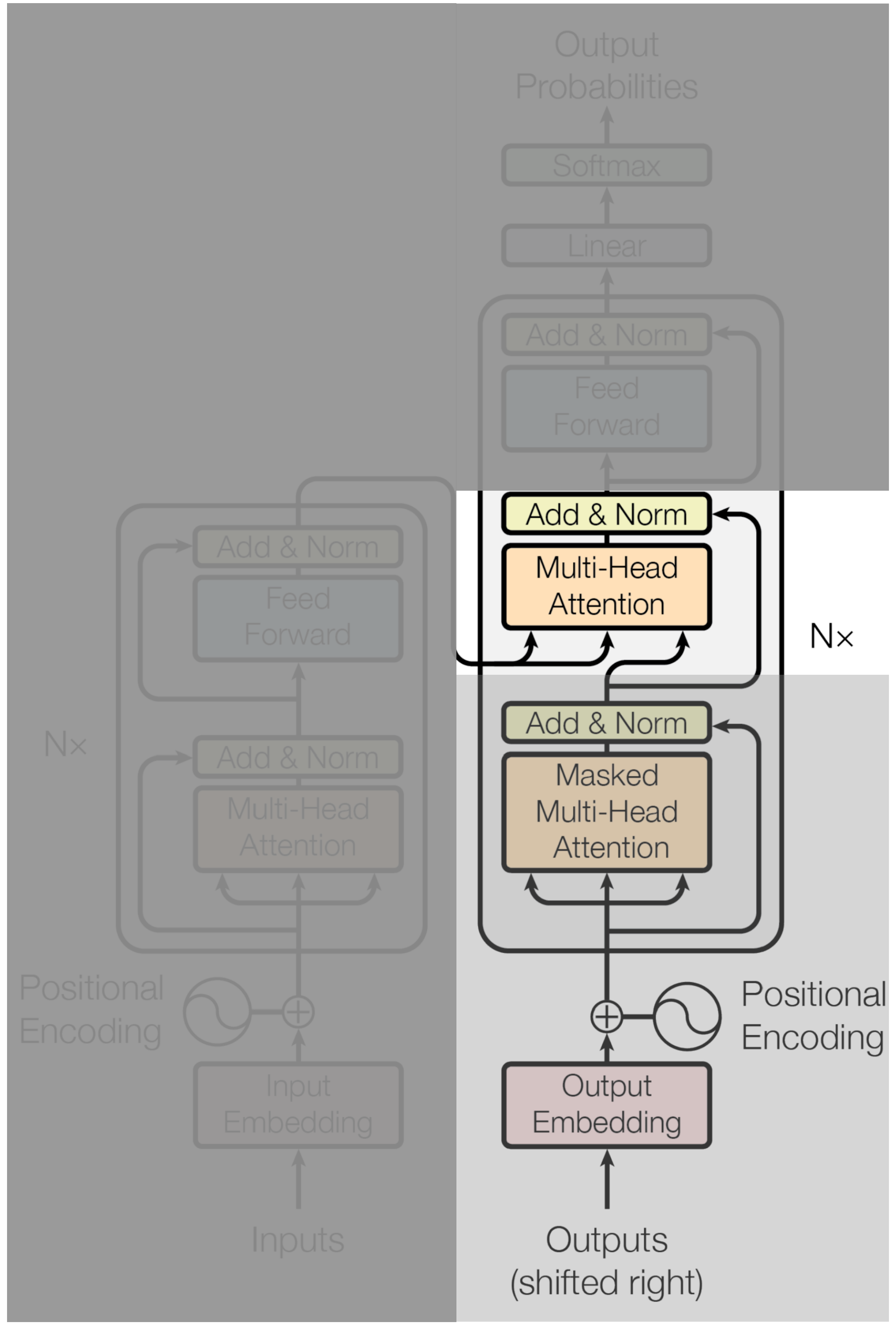
Context Attention

$$Q \leftarrow QW_Q$$

$$V \leftarrow VW_V$$

$$Q \leftarrow A_\lambda(Q; V) = \texttt{softmax}(Q^\top V)V$$

- V comes from encoder output
- Q comes from decoder
- Apply multi-head

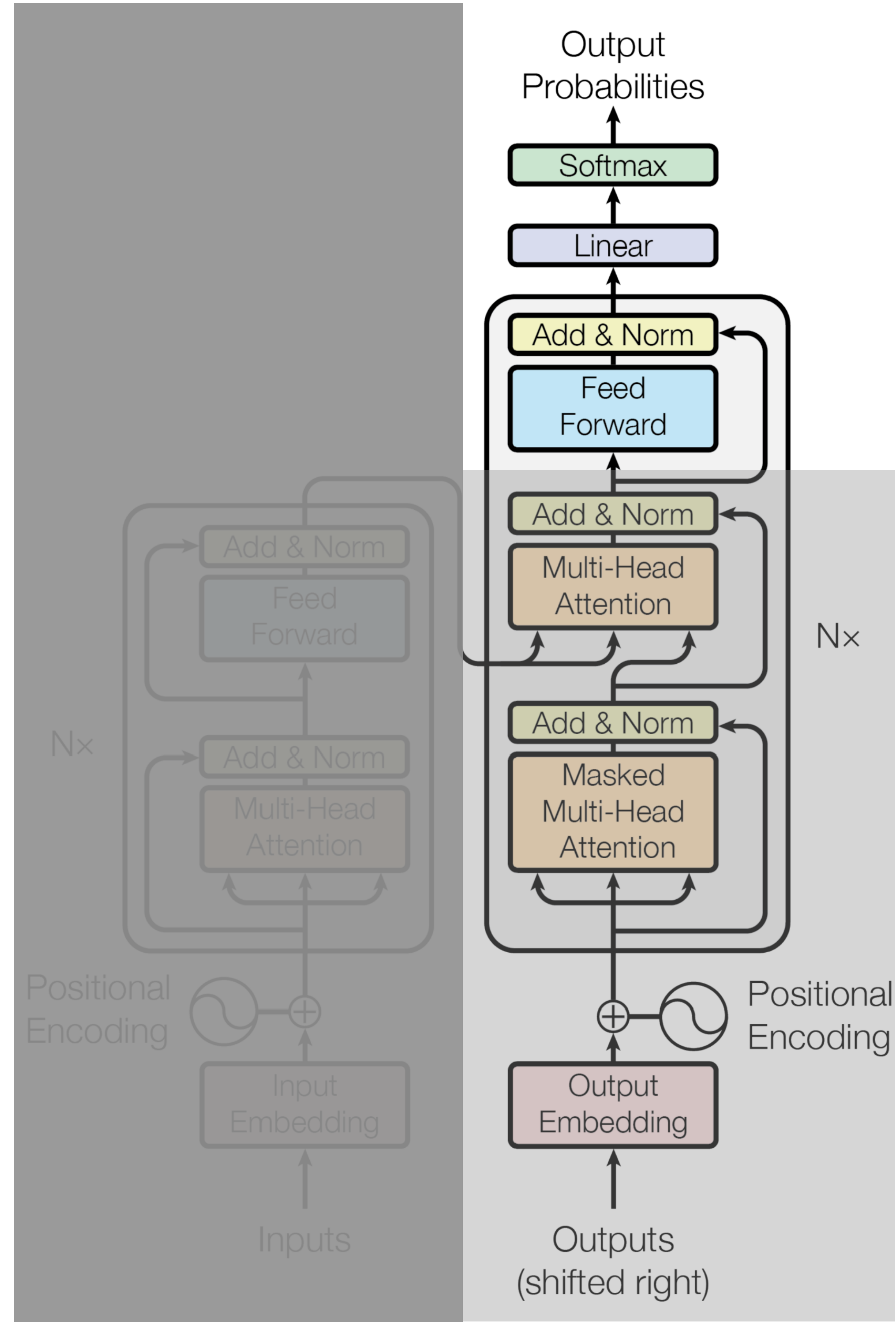


Softmax

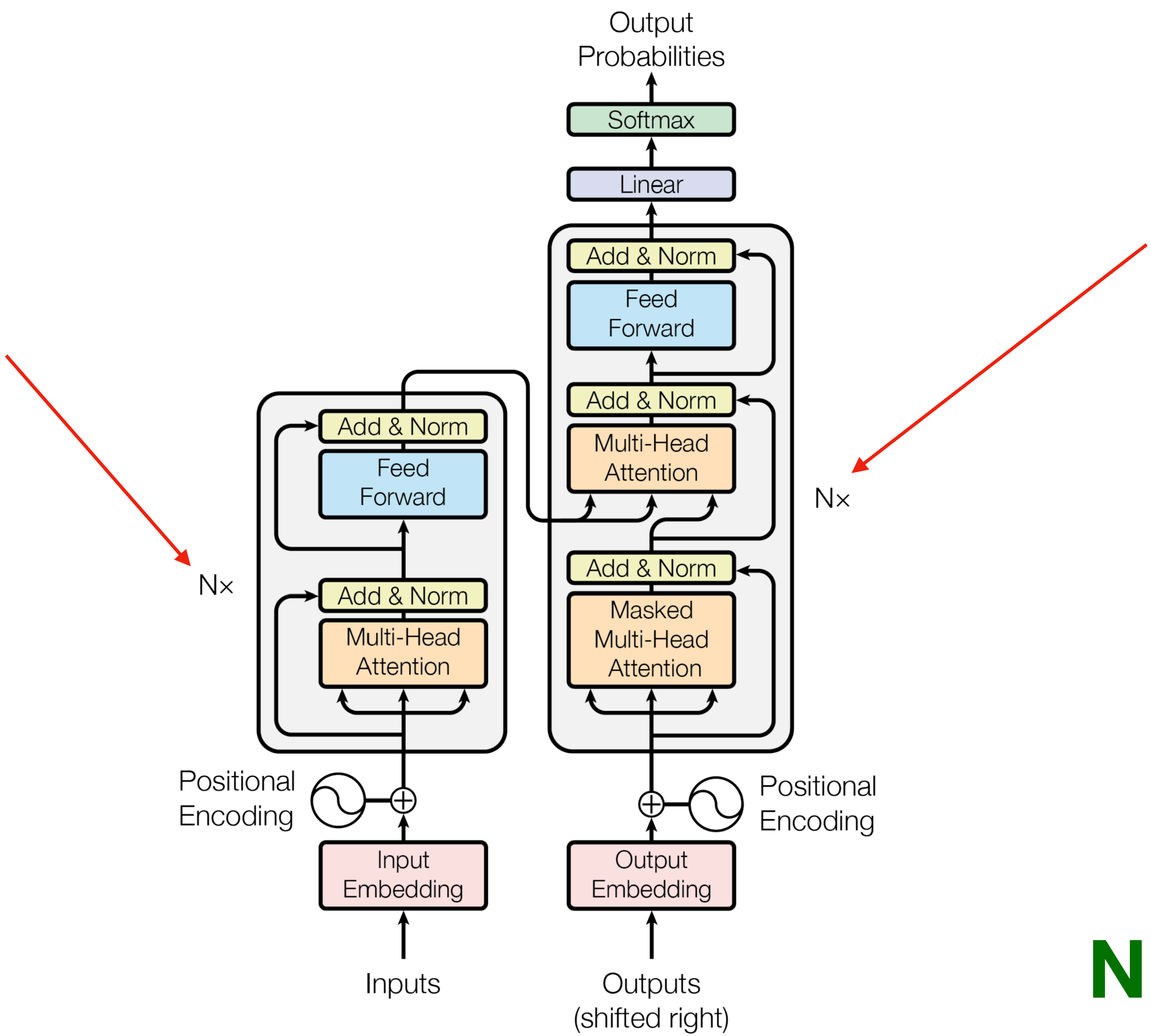
$$\hat{\mathbf{y}} = \texttt{softmax}(\mathbf{q}W_e^\top)$$

$$\min \ \hat{\mathbb{E}}\left(\sum_{j=1}^{l} -\mathbf{y}_j^\top \log \hat{\mathbf{y}}_j\right)$$

$$W_e, W_i^V, W_i, W_{1,i}, W_{2,i}, W_i^Q$$



Going Deep



N = 6

Does It Work?

Layer type	per-layer complexity	sequential operations	max path length
Self-attention	$O(m^2 d)$	$O(1)$	$O(1)$
Recurrent	$O(md^2)$	$O(m)$	$O(m)$
Convolution	$O(kmd^2)$	$O(1)$	$O(\log_k m)$
Self-attention (restricted)	$O(rmd)$	$O(1)$	$O(m/r)$

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

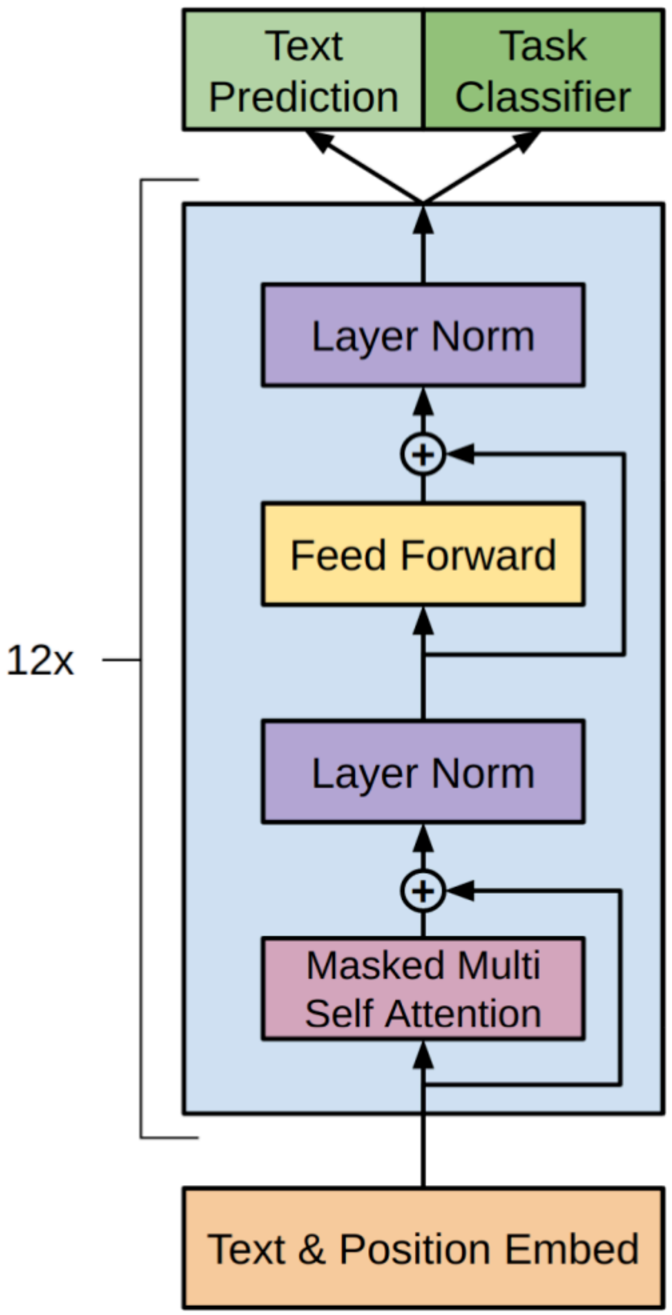
Supervised Learning

GPT-1

Pre-training
Fine-tuning

https://openai.com/blog/language-unsupervised/

Generative Pre-Training (GPT)



- Unsupervised pre-training

$$\min_{\Theta} \quad -\hat{\mathsf{E}} \log \mathsf{p}(X|\Theta), \quad \text{where} \quad \mathsf{p}(X|\Theta) = \prod_{j=1}^{m} \mathsf{p}(\mathbf{x}_j|\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}; \Theta).$$

$$H^{(0)} = XW_e + W_p$$

$$H^{(\ell)} = \texttt{transformer\_decoder\_block}(H^{(\ell-1)}), \quad \ell = 1, \ldots, L$$

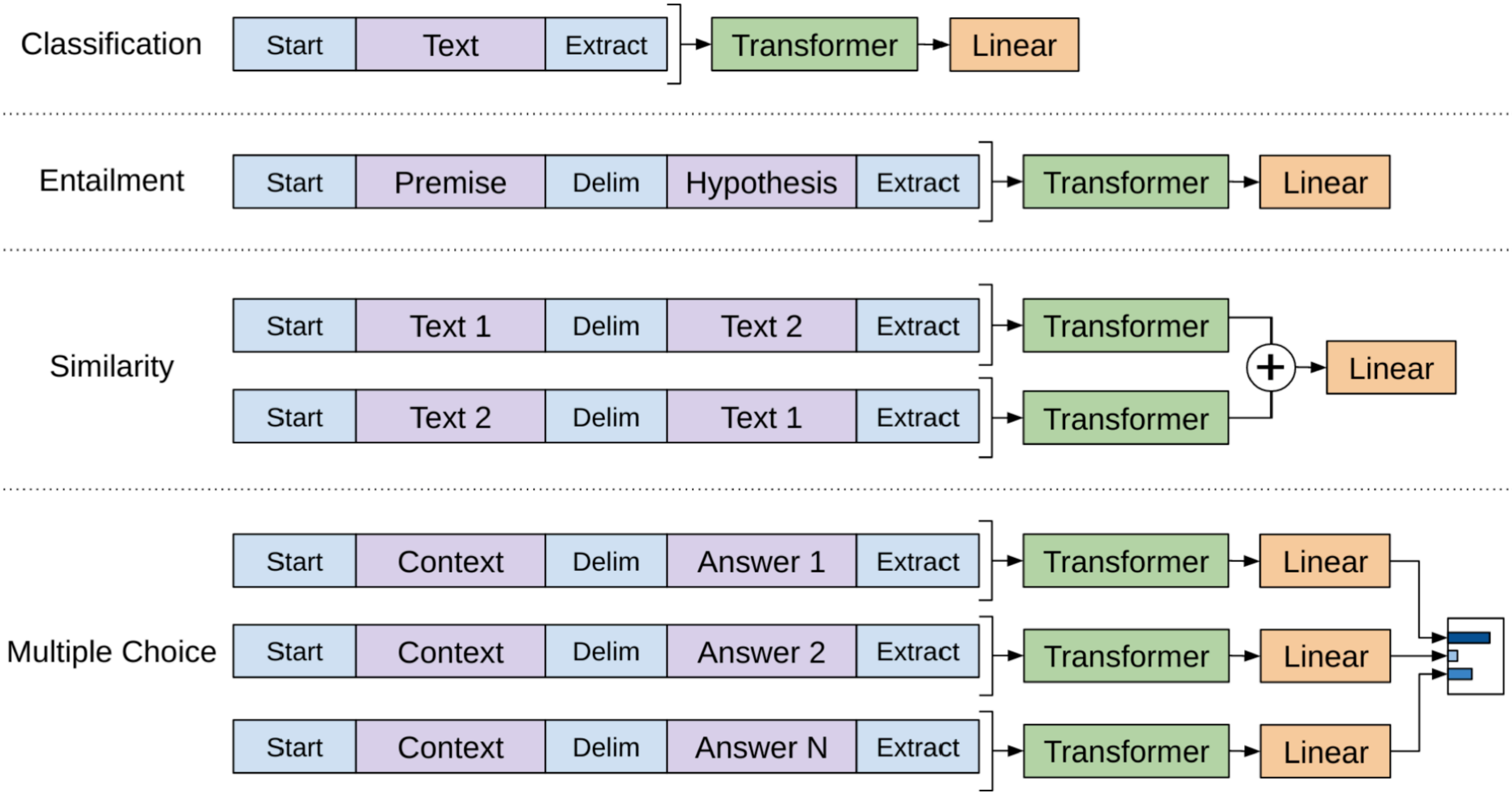
$$\mathsf{p}(\mathbf{x}_j|\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}; \Theta) = \texttt{softmax}(\mathbf{h}_j^{(L)} W_e^\top).$$

- Supervised fine-tuning

$$\min_{W_y} \min_{\Theta} \quad -\hat{\mathsf{E}} \log \mathsf{p}(\mathbf{y}|X, \Theta) - \lambda \cdot \hat{\mathsf{E}} \log \mathsf{p}(X|\Theta),$$

$$\text{where} \quad \mathsf{p}(\mathbf{y}|X, \Theta) = \left\langle \mathbf{y}, \texttt{softmax}(\mathbf{h}_m^{(L)} W_y) \right\rangle$$

Input Transformations



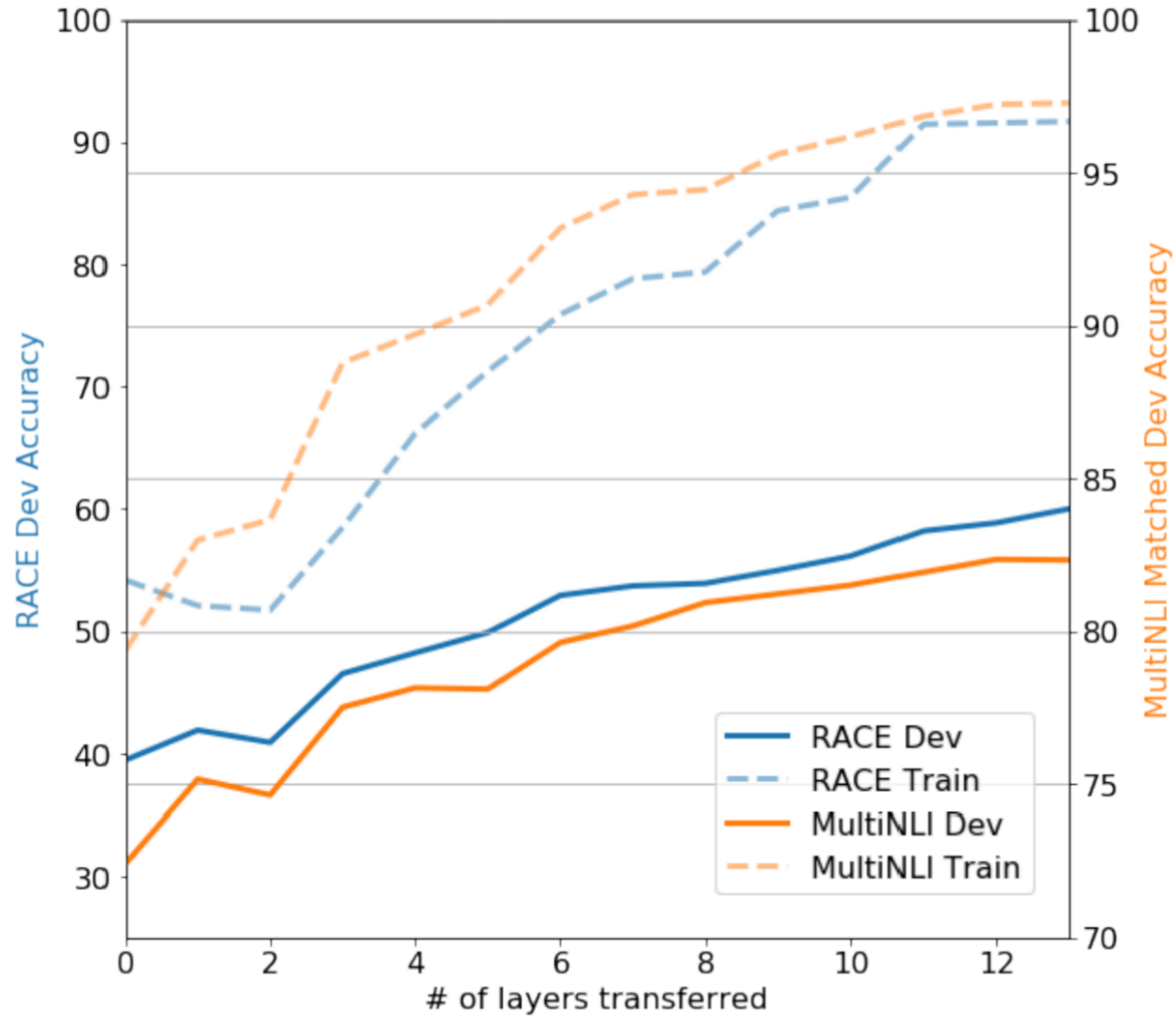
Fine-tuning Results

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

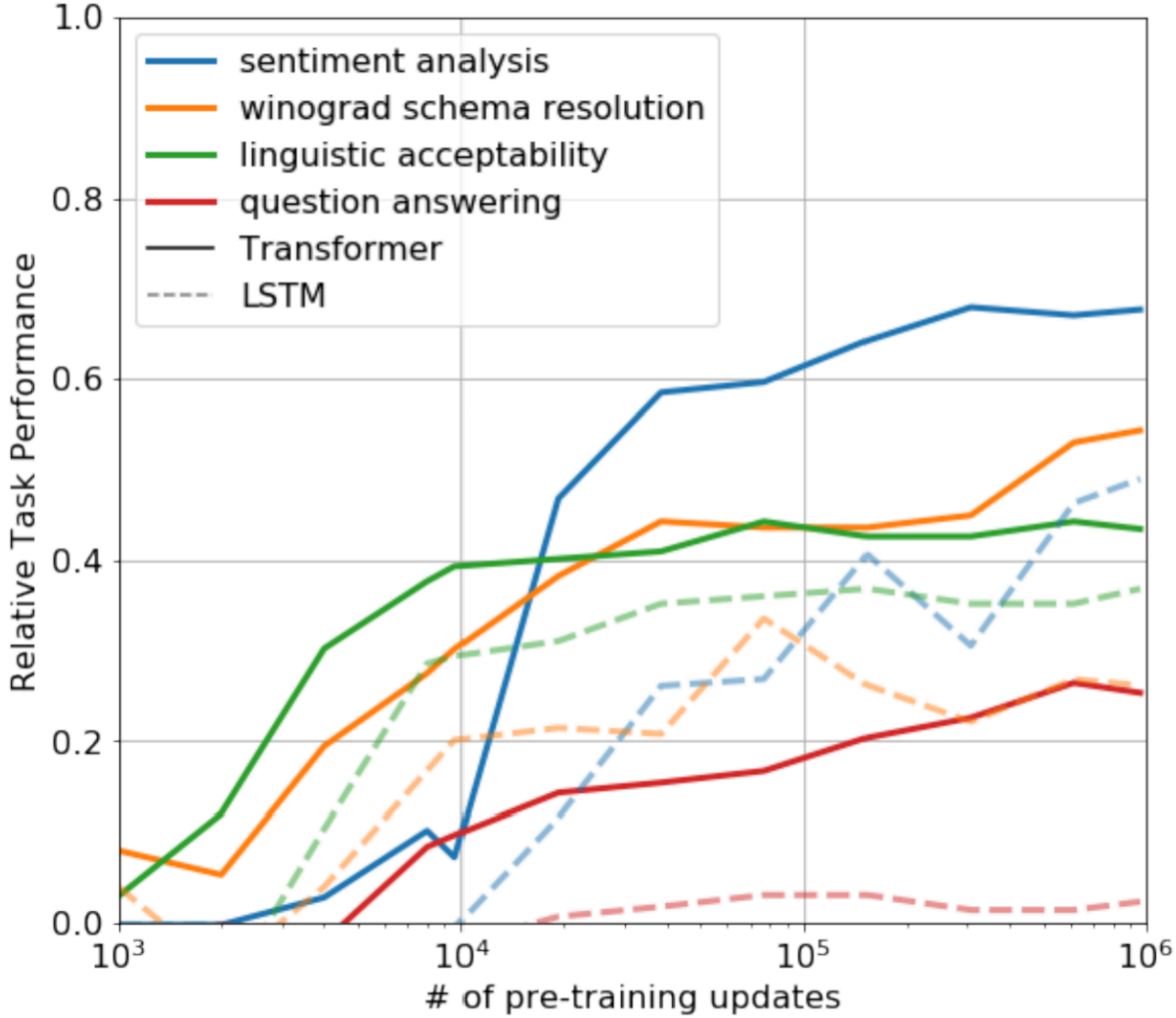
Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

Method	**Classification**		**Semantic Similarity**			**GLUE**
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

We Need to Go Deep?



Zero-Shot Relative Perf



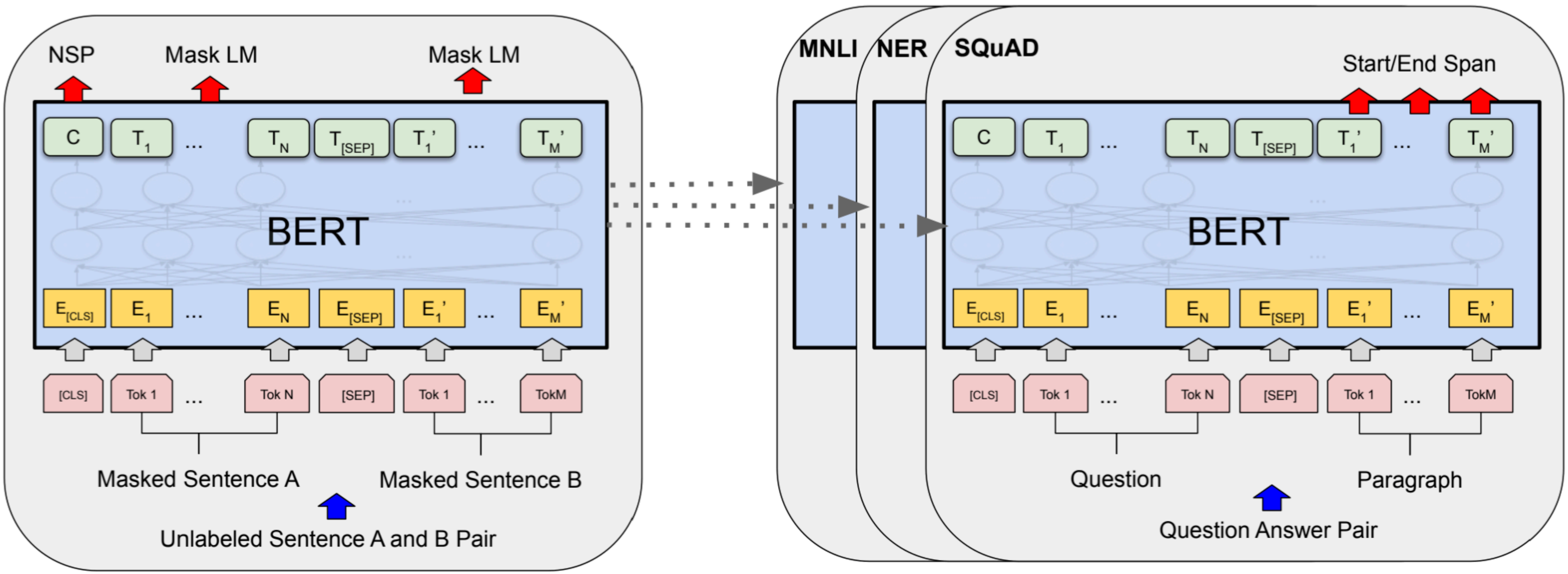
Pre-training Helps

Method	Avg. Score	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	MNLI (acc)	QNLI (acc)	RTE (acc)
Transformer w/ aux LM (full)	74.7	45.4	91.3	82.3	82.0	**70.3**	**81.8**	**88.1**	**56.0**
Transformer w/o pre-training	59.9	18.9	84.0	79.4	30.9	65.5	75.7	71.2	53.8
Transformer w/o aux LM	**75.0**	**47.9**	**92.0**	**84.9**	**83.2**	69.8	81.1	86.9	54.4
LSTM w/ aux LM	69.1	30.3	90.5	83.2	71.8	68.1	73.7	81.1	54.6

- Transformer is better than LSTM
- Auxiliary LM loss improves performance on larger datasets
- Pre-training helps a lot on certain datasets

https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

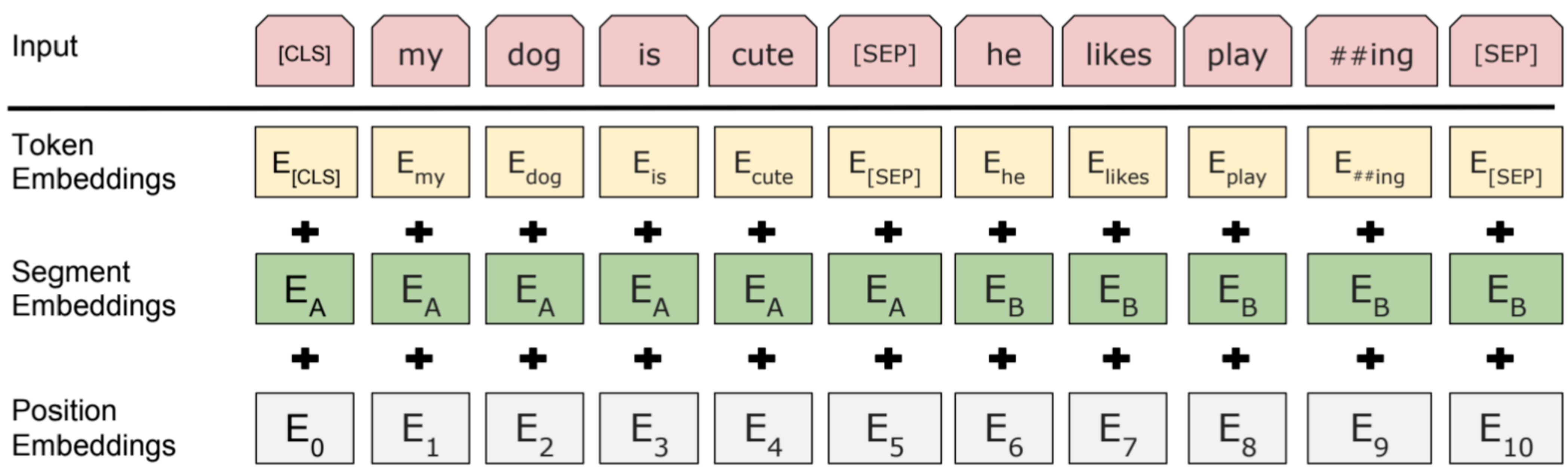
BERT in 1 Fig



Pre-training

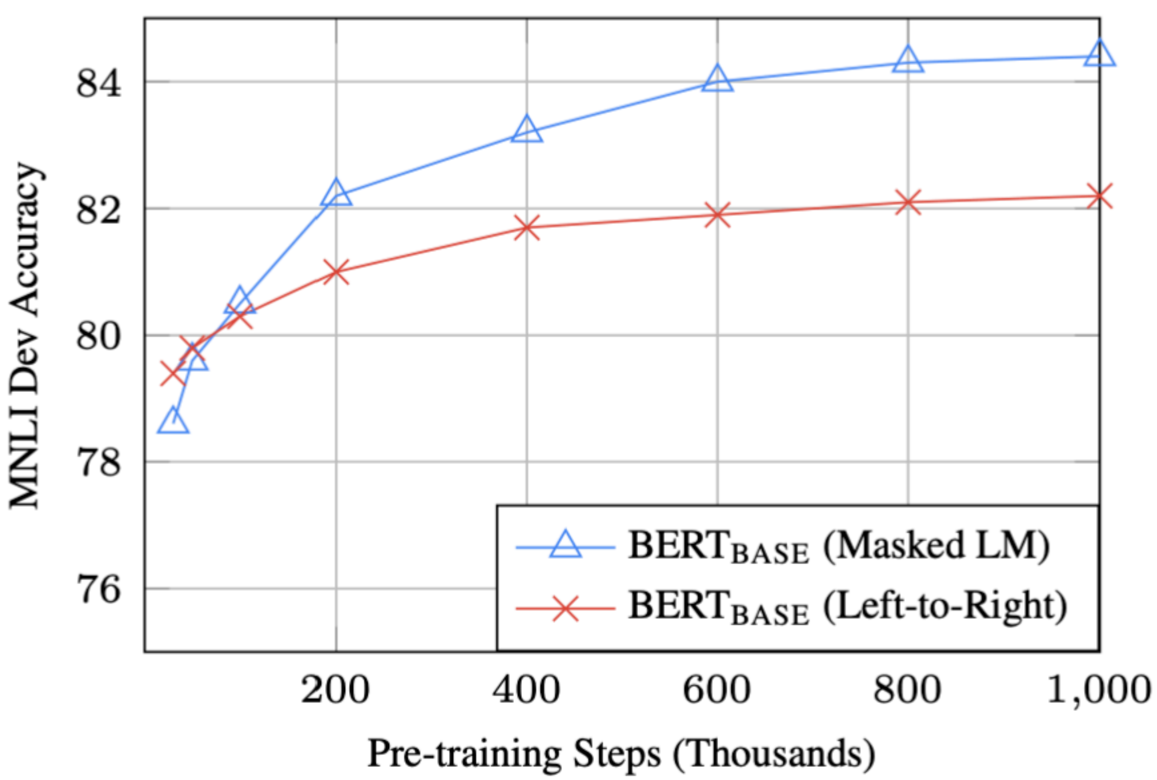
Fine-Tuning

Input Transformations



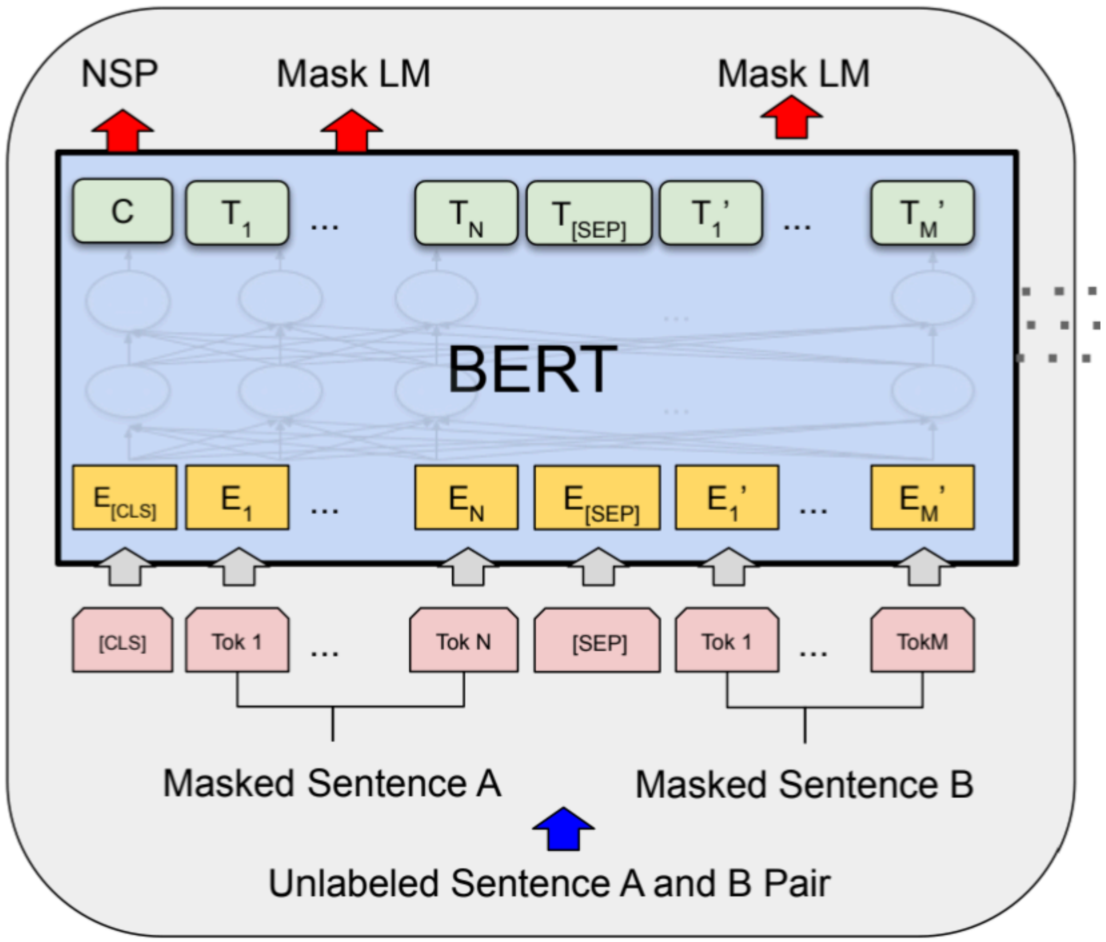
Mask Language Model

- Randomly select 15% input tokens, change to [Mask]
- Add softmax to predict the [Mask] tokens
- Actually 12% replaced with [Mask], 1.5% with random



Masking Rates			Dev Set Results		
MASK	SAME	RND	MNLI Fine-tune	NER Fine-tune	NER Feature-based
80%	10%	10%	84.2	95.4	94.9
100%	0%	0%	84.3	94.9	94.0
80%	0%	20%	84.1	95.2	94.6
80%	20%	0%	84.4	95.2	94.7
0%	20%	80%	83.7	94.8	94.6
0%	0%	100%	83.6	94.9	94.6

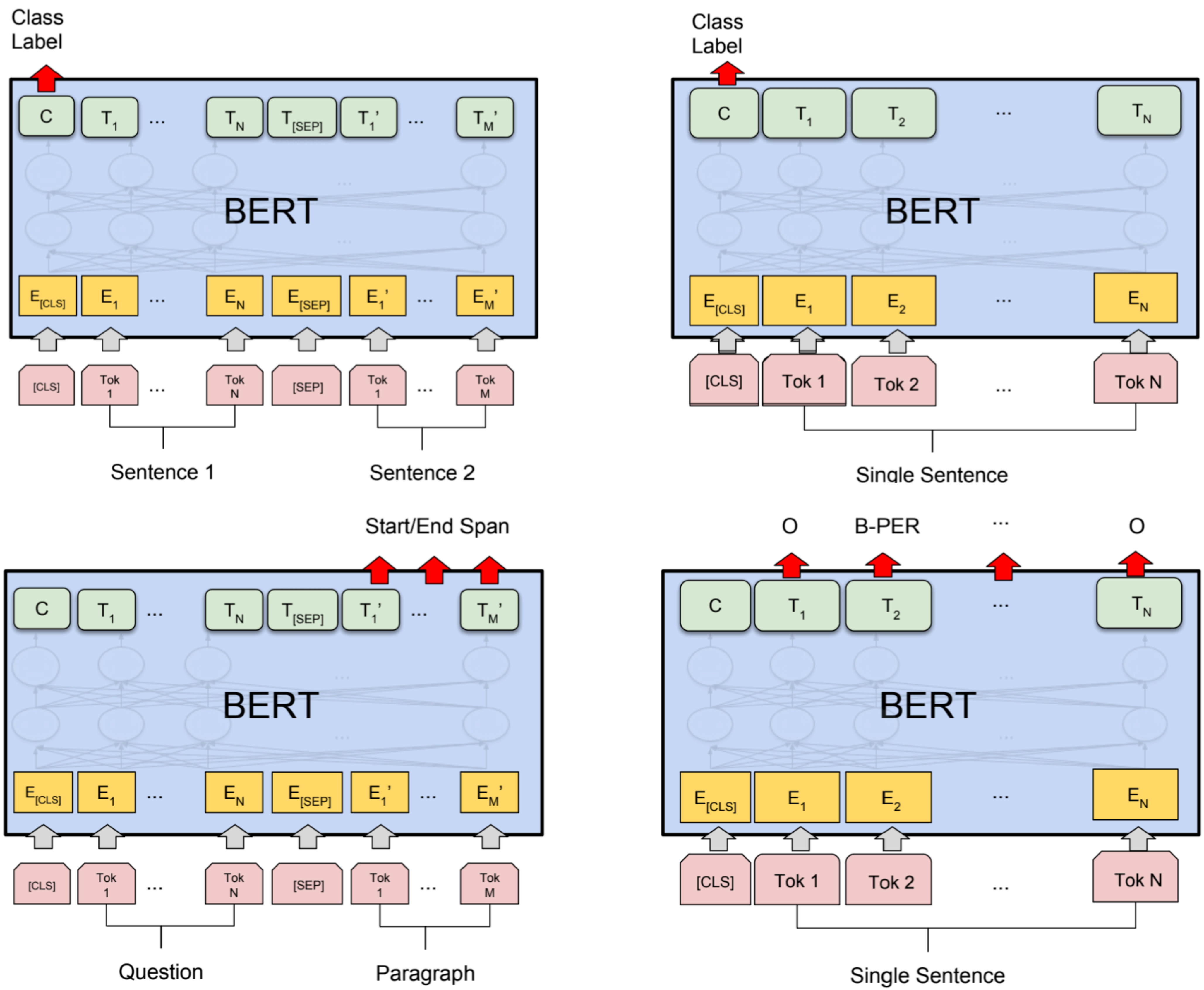
Next Sentence Prediction



- 50% of time, B follows A
- 50% of time, B is randomly chosen
- Binary classification to train sentence-level representation on [CLS]
- Training objective: MLM + NSP

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Task Transformation



Fine-tuning vs Two-stage

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

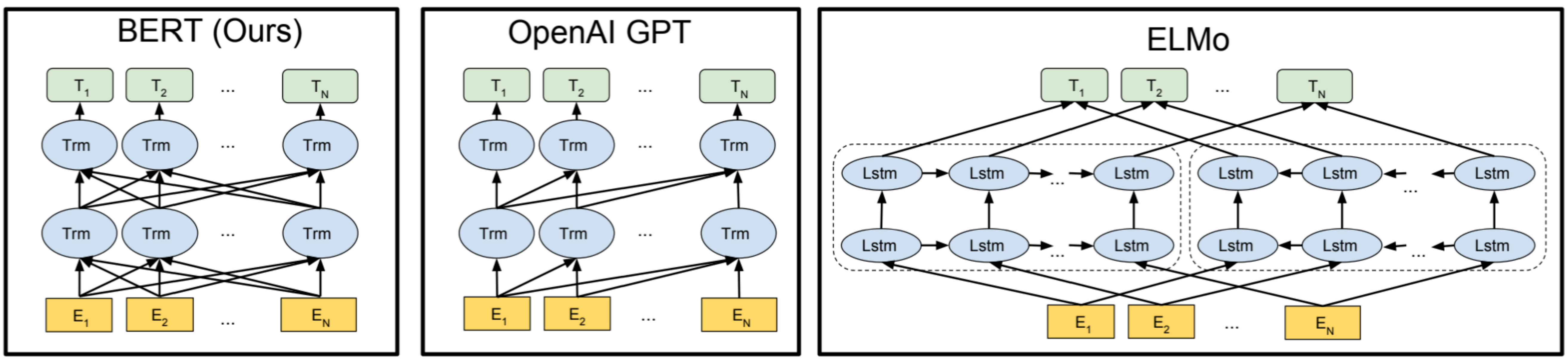
System	Dev F1	Test F1
ELMo (Peters et al., 2018a)	95.7	92.2
CVT (Clark et al., 2018)	-	92.6
CSE (Akbik et al., 2018)	-	**93.1**
Fine-tuning approach		
$\text{BERT}_{\text{LARGE}}$	96.6	92.8
$\text{BERT}_{\text{BASE}}$	96.4	92.4
Feature-based approach ($\text{BERT}_{\text{BASE}}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

The Bigger, The Better?

Hyperparams				Dev Set Accuracy		
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

Table 6: Ablation over BERT model size. #L = the number of layers; #H = hidden size; #A = number of attention heads. "LM (ppl)" is the masked LM perplexity of held-out training data.

Comparison



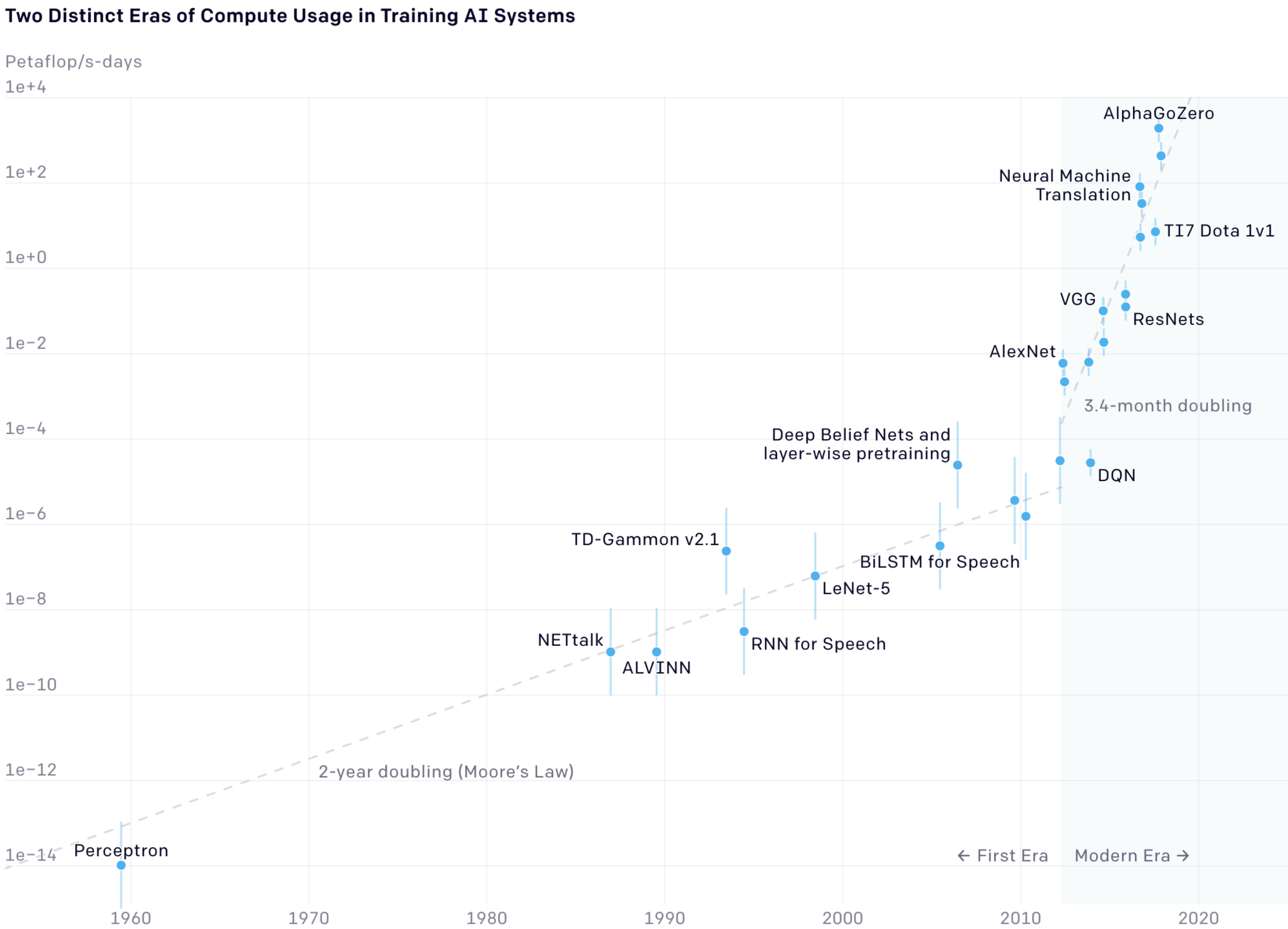
Transformer Encoder

Transformer Decoder

Bidirectional LSTM

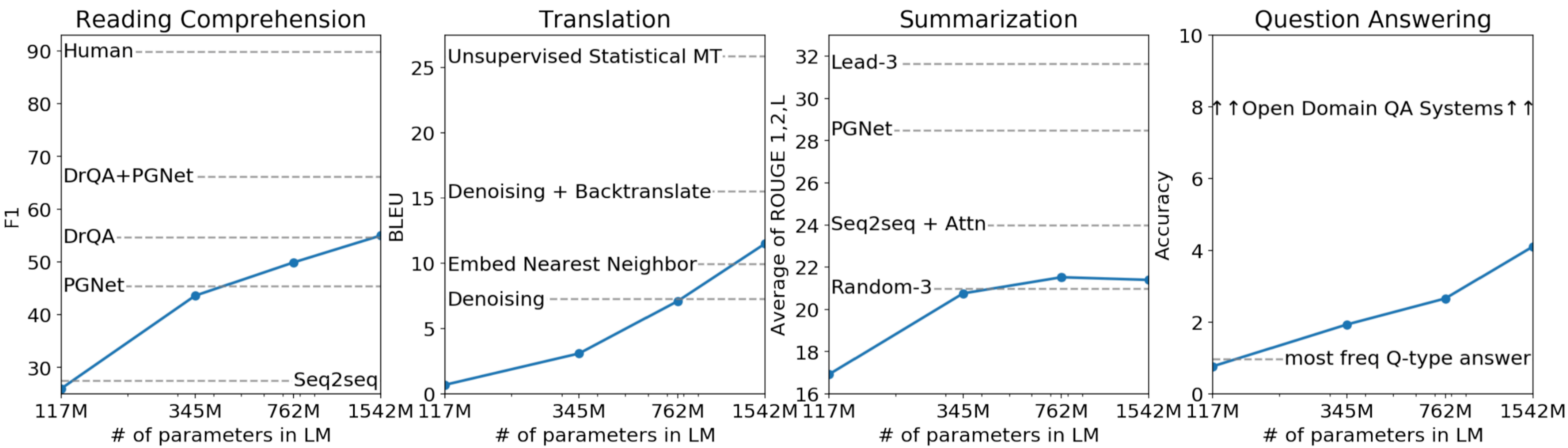
GPT-2

https://openai.com/blog/better-language-models/



https://openai.com/blog/ai-and-compute/

GPT-2 in 1 Fig



- Exclusively on zero-shot
- The bigger, the better?

A glance on training set

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”

“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

Children’s book test

Context:
1 So they had to fall a long way .
2 So they got their tails fast in their mouths .
3 So they couldn't get them out again .
4 That 's all .
5 `` Thank you , " said Alice , `` it 's very interesting .
6 I never knew so much about a whiting before . "
7 `` I can tell you more than that , if you like , " said the Gryphon .
8 `` Do you know why it 's called a whiting ? "
9 `` I never thought about it , " said Alice .
10 `` Why ? "
11 `` IT DOES THE BOOTS AND SHOES . '
12 the Gryphon replied very solemnly .
13 Alice was thoroughly puzzled .
14 `` Does the boots and shoes ! "
15 she repeated in a wondering tone .
16 `` Why , what are YOUR shoes done with ? "
17 said the Gryphon .
18 `` I mean , what makes them so shiny ? "
19 Alice looked down at them , and considered a little before she gave her answer .
20 `` They 're done with blacking , I believe . "

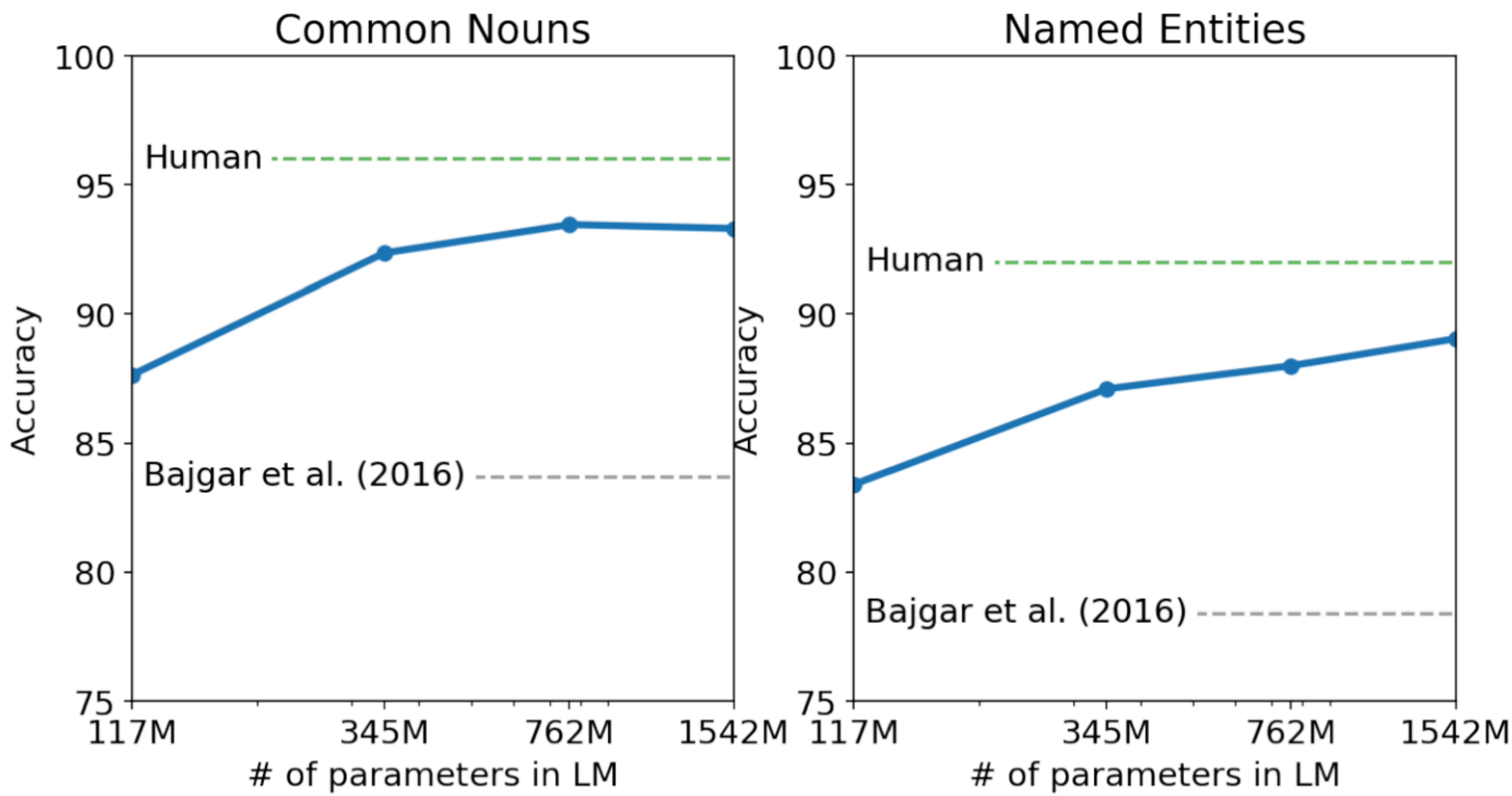
Query: `` Boots and shoes under the sea , " the XXXXX went on in a deep voice , `` are done with a whiting ".

Candidates: Alice|BOOTS|Gryphon|SHOES|answer|fall|mouths|tone|way|whiting

Answer: gryphon

https://research.fb.com/downloads/babi/

Children's book test



- Exclusively on zero-shot
- The bigger, the better?

Translation

- context: English sentence = French sentence
- generation: English sentence =

- EN->FR: 5 BLEU
- FR->EN: 11.5 BLEU
- SOTA unsupervised: 33.5 BLEU

Question Answering

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Contamination

	PTB	WikiText-2	enwik8	text8	Wikitext-103	1BW
Dataset train	**2.67%**	0.66%	**7.50%**	2.34%	**9.09%**	**13.19%**
WebText train	0.88%	**1.63%**	6.31%	**3.94%**	2.42%	3.75%

- 3.3% reported on CIFAR-10 (Barz & Denzler, 2020)

Text Generation

Context (WebText test)

Xmonad, Xresources, and Fonts

Xresources can be an absolute pain (they were for me). So, this a bit of help:

First of all, the way that fonts are generally handled in Xmonad is through XFT . Xft is the X FreeType Interface library and is a library that handles the actual rendering of fonts in Xmonad/urxvt (Xresources specifies a bunch of resources to things that launch under the X window manager. Settings for individual programs are specified in Xresources like so: Urxvt.background:background settins here).

Anyway, fonts are set using this syntax in Xresources :

urxvt*font:xft:your font name goes here:size=the size of the font goes here,xft:fall-back font 1 goes here

So, thats the most basic part of fonts in urxvt specified via Xresources.

Fixing /.ssh/ Permissions in Cygwin

In Cygwin on Windows I found that I could not set the permissions of my /.ssh/ folder to be 0600, as is required for ssh to allow you to use keys. The symptom I had was that I no matter what I did, it always modfied the owner and the group

Small model completion	**GPT-2 completion**
was set to /.ssh/ , which was the way things were. This could not be fixed completely. I found I could make that much more workable with just typing the username with your mouse then trying to set permissions to be "0600" for the group, which is what my system is now configured to do. However, as of Cygwin 7 on Linux I have found that doing so is now "unnecessary" for my system! I also installed a package called chroot. It does that by allowing you to install/upgrade to your system directory. So, as of Windows 8 and up, installing that can require chroot, so I tried and tried and found that it wasn't so bad when installing the package on Linux and then uninstalling it. As of my last write-up on Cygwin, the script just does its job (the sudo commands do the rest), if you try it you will have to log the change. If that wasn't confusing enough, this is what I tried and just went back to my /.ssh/ directory and the chroot script. I found I was finally running out of permissions (because i forgot to create a symlink to ssh from the file system and	, making these files read-only. I tried just starting ssh and using a program like passwd which modifies the owner/group to 0644 permissions, but it didn't work. It was annoying but not nearly as bad as it could be, just annoying to have to fiddle with permissions. So I decided to do this as the reverse: For each .ssh folder in your Cygwin directory, set the permission of each file and directory to 2644. Then start ssh with: ssh-add /.ssh/ The first time you connect, your permissions of the /.ssh folder that you just had in Cygwin will be set to 0600. Then every time you connect again, your permissions will change to 0644. Makefile for Xresources as used in /.xsessionrc These are made in my .xsessionrc : .xprofile: .xsession-errors: color="" For example: .xprofile: color=blue As the colors are stored in a separate file, the color in your /.xsession-errors will look like cyan instead of blue. This is mostly because, as of now, only the color is assigned to your

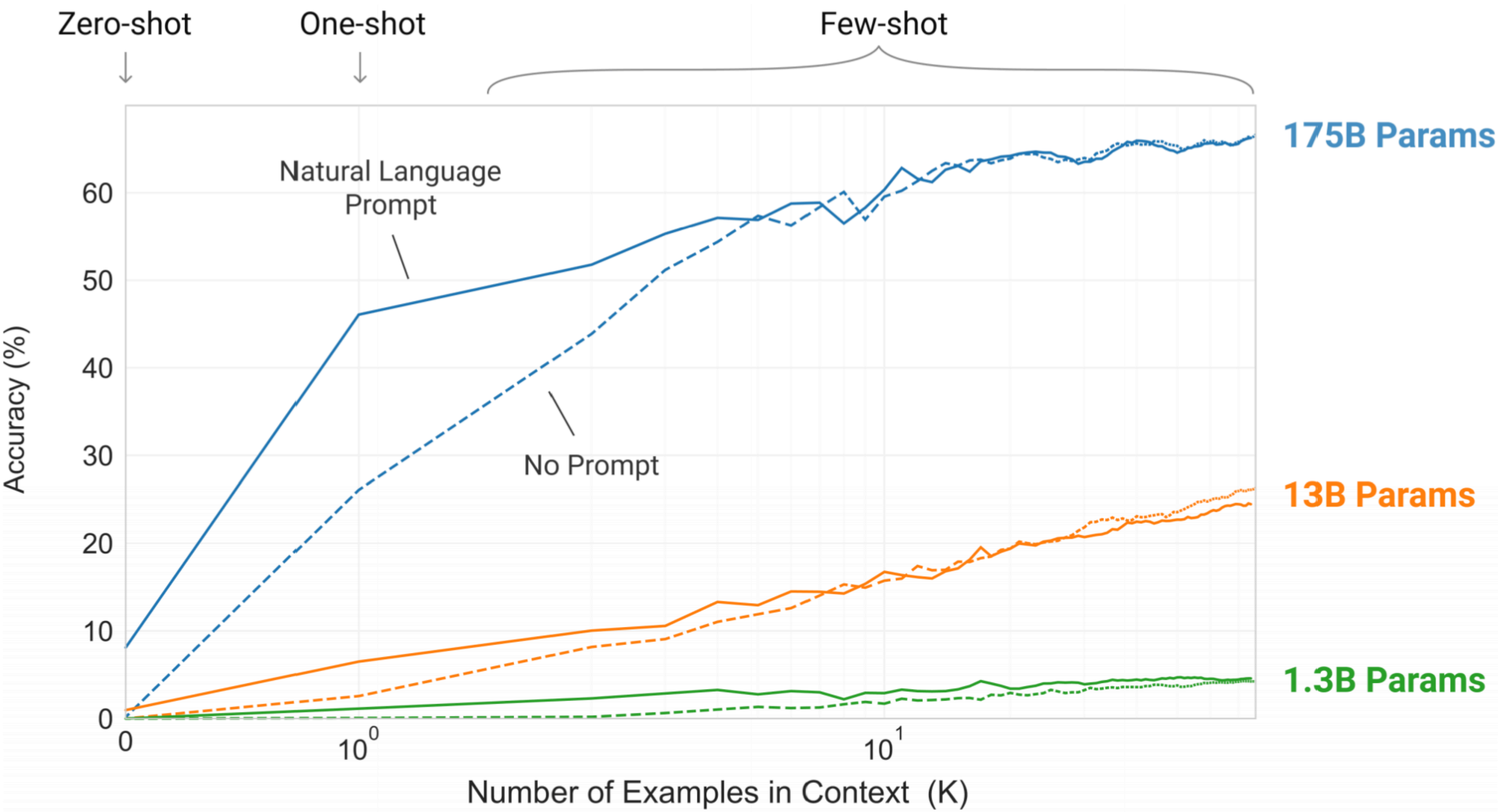
Closing out

DATASET	METRIC	OUR RESULT	PREVIOUS RECORD	HUMAN
Winograd Schema Challenge	accuracy (+)	**70.70%**	63.7%	92%+
LAMBADA	accuracy (+)	**63.24%**	59.23%	95%+
LAMBADA	perplexity (–)	**8.6**	99	~1-2
Children's Book Test Common Nouns (validation accuracy)	accuracy (+)	**93.30%**	85.7%	96%
Children's Book Test Named Entities (validation accuracy)	accuracy (+)	**89.05%**	82.3%	92%
Penn Tree Bank	perplexity (–)	**35.76**	46.54	unknown
WikiText-2	perplexity (–)	**18.34**	39.14	unknown
enwik8	bits per character (–)	**0.93**	0.99	unknown
text8	bits per character (–)	**0.98**	1.08	unknown
WikiText-103	perplexity (–)	**17.48**	18.3	unknown

GPT-3

https://arxiv.org/abs/2005.14165

GPT-3 in 1 Fig



Zero-shot
One-shot
Few-shot
Natural Language Prompt
No Prompt
175B Params
13B Params
1.3B Params
Accuracy (%)
60
50
40
30
20
10
0
0
10^0
10^1
Number of Examples in Context (K)

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:      ← task description
cheese =>                         ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:      ← task description
sea otter => loutre de mer        ← example
cheese =>                         ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:      ← task description
sea otter => loutre de mer        ← examples
peppermint => menthe poivrée      ←
plush girafe => girafe peluche    ←
cheese =>                         ← prompt
```

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
sea otter => loutre de mer        ← example #1
```

↓

gradient update

↓

```
peppermint => menthe poivrée      ← example #2
```

↓

gradient update

↓

• • •

↓

```
plush giraffe => girafe peluche   ← example #N
```

gradient update

```
cheese =>                         ← prompt
```

GPT-3 Family

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

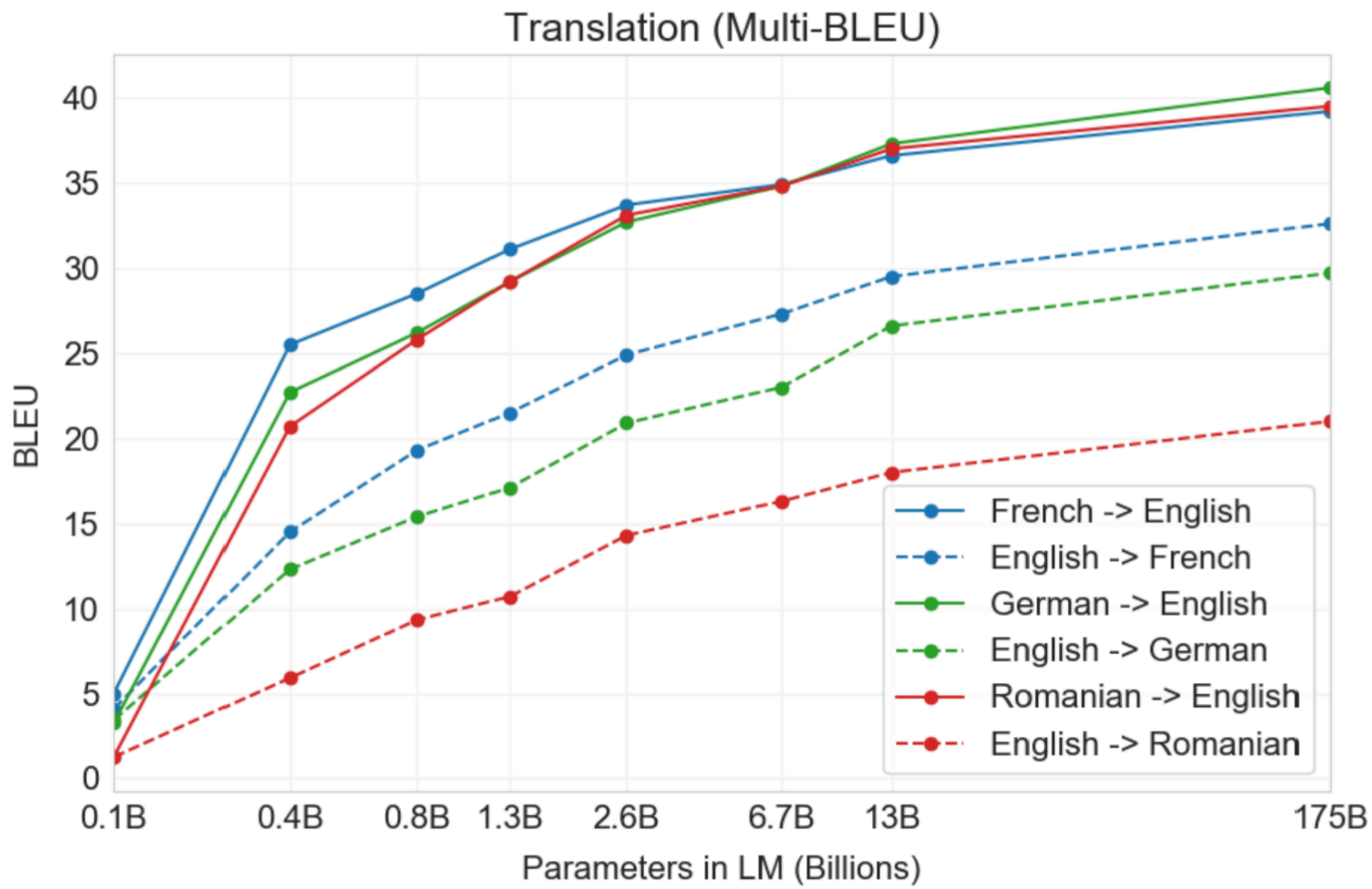
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

How Costly is GPT-3?

"Unfortunately, a bug in the filtering caused us to ignore some overlaps, and due to the cost of training it was not feasible to retrain the model."

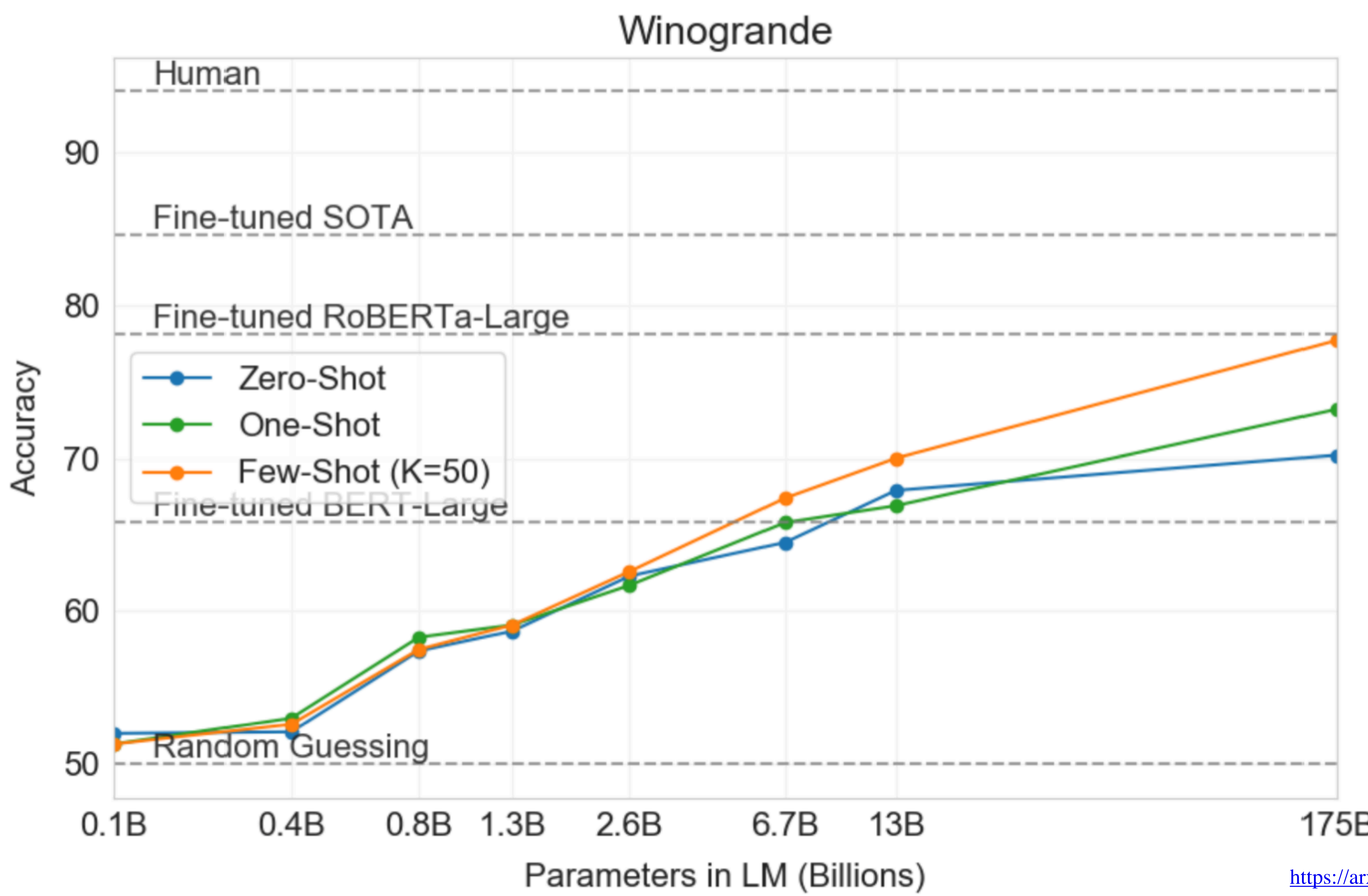
Translation

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5



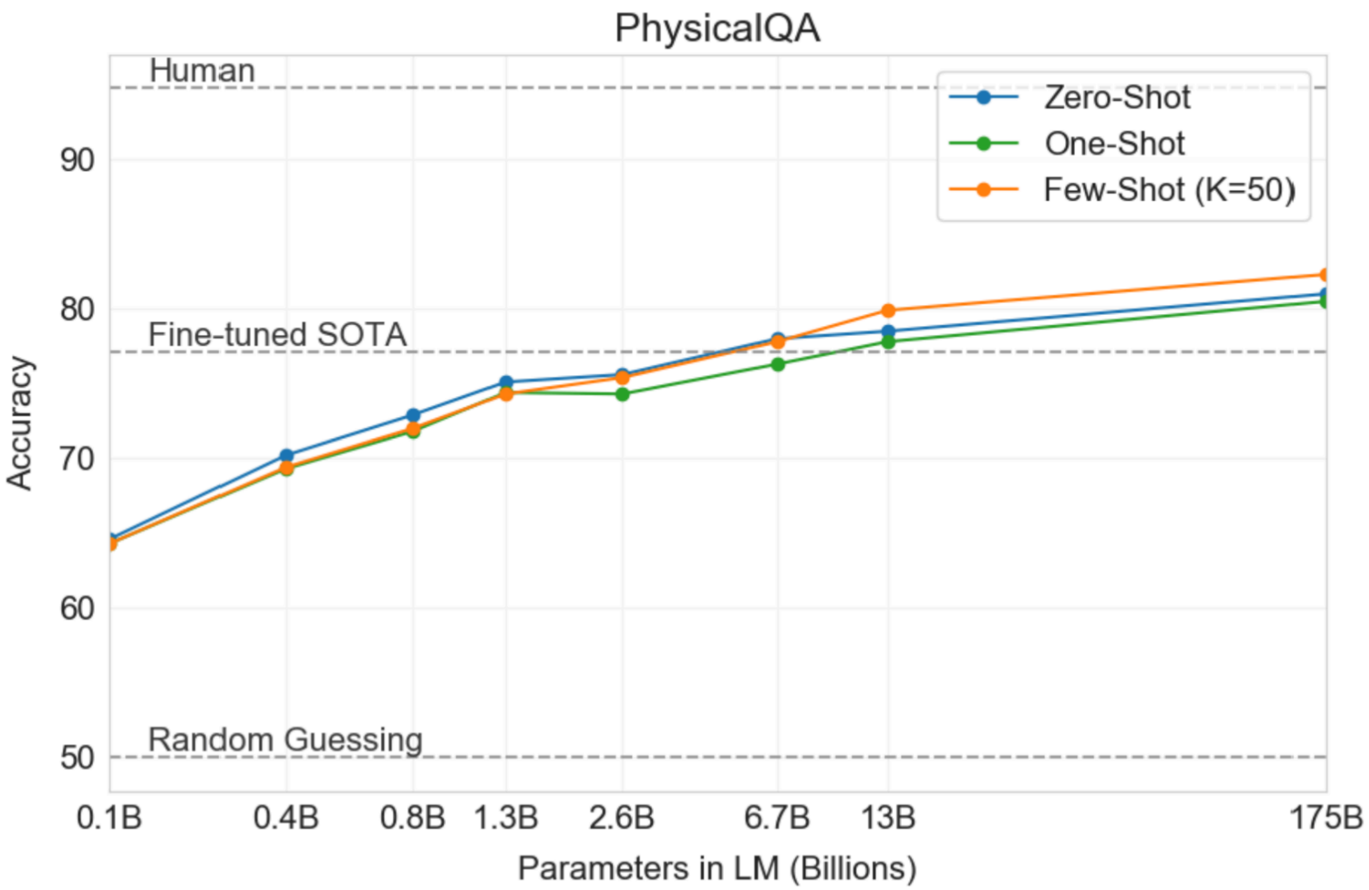
Winogrande

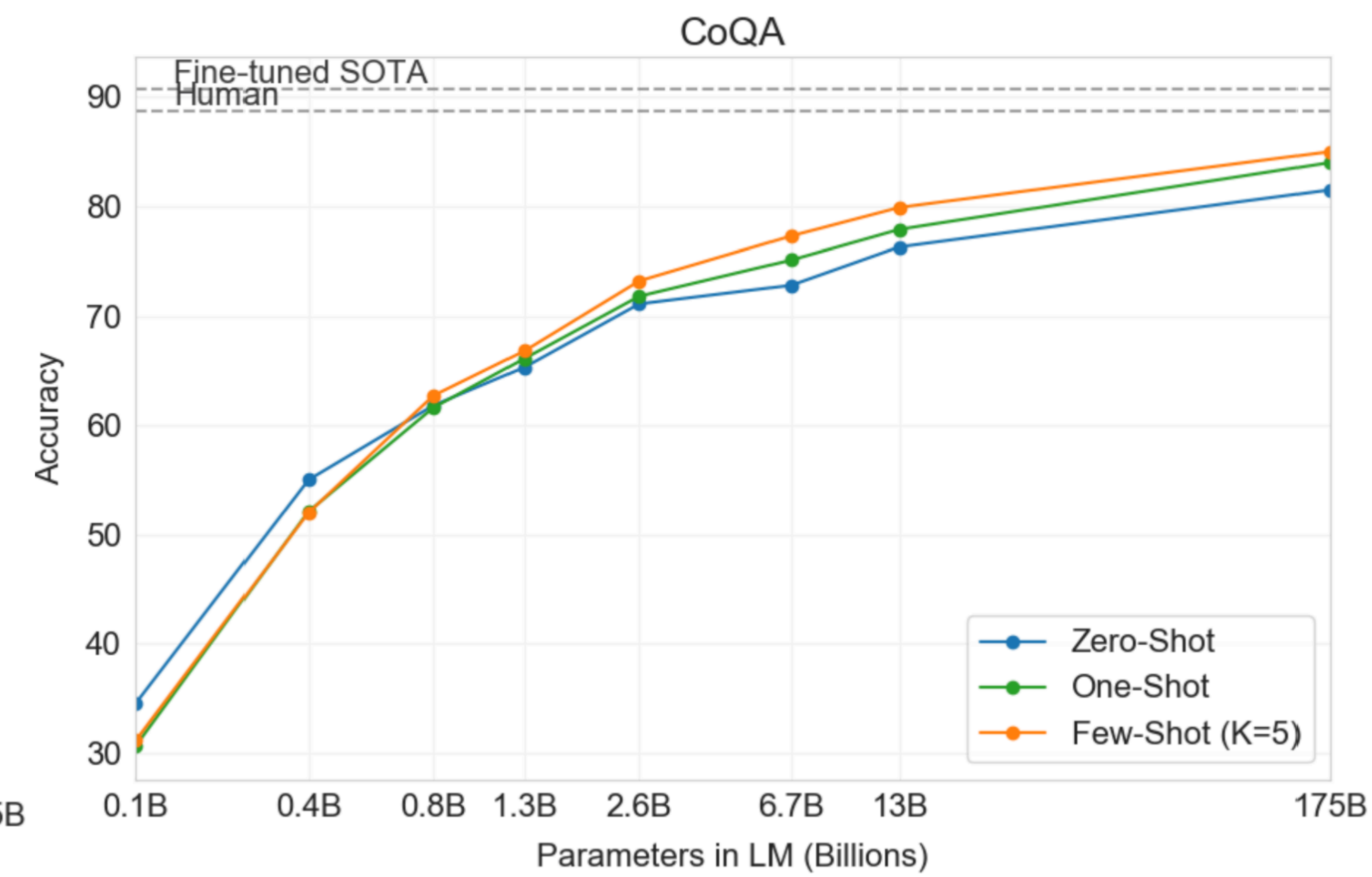
“The city councilmen refused the demonstrators a permit because they [feared/advocated] violence”



https://arxiv.org/abs/1907.10641

Question Answering





Q: To separate egg whites from the yolk using a water bottle, you should…

(a) Squeeze the water bottle and press it against the yolk. Release, which creates suction and lifts the yolk.

(b) Place the water bottle and press it against the yolk. Keep pushing, which creates suction and lifts the yolk.

https://arxiv.org/abs/1911.11641

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q1: Who had a birthday?
A1: Jessica
R1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

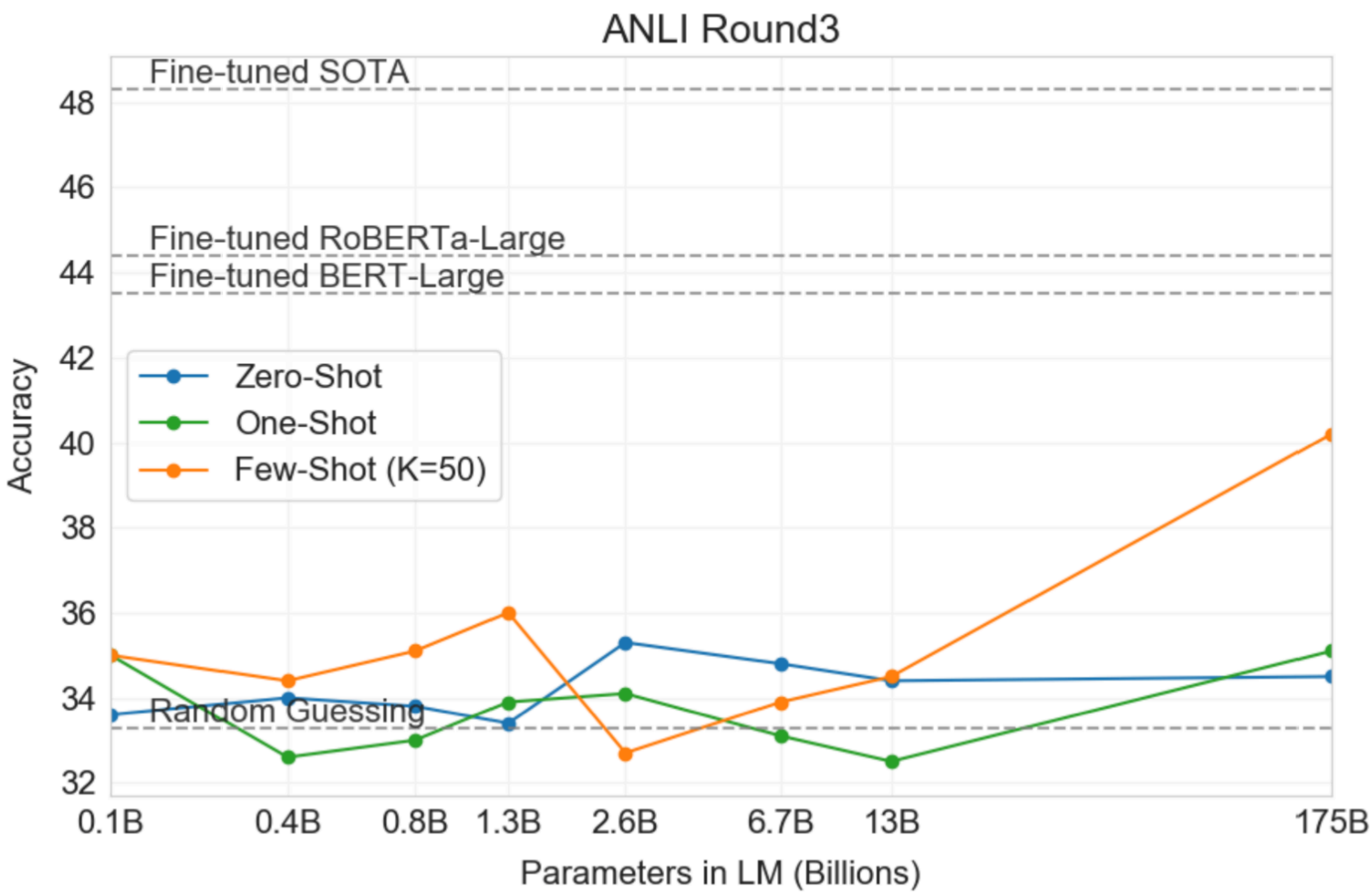
https://www.aclweb.org/anthology/Q19-1016/

Adversarial Natural Language Inference

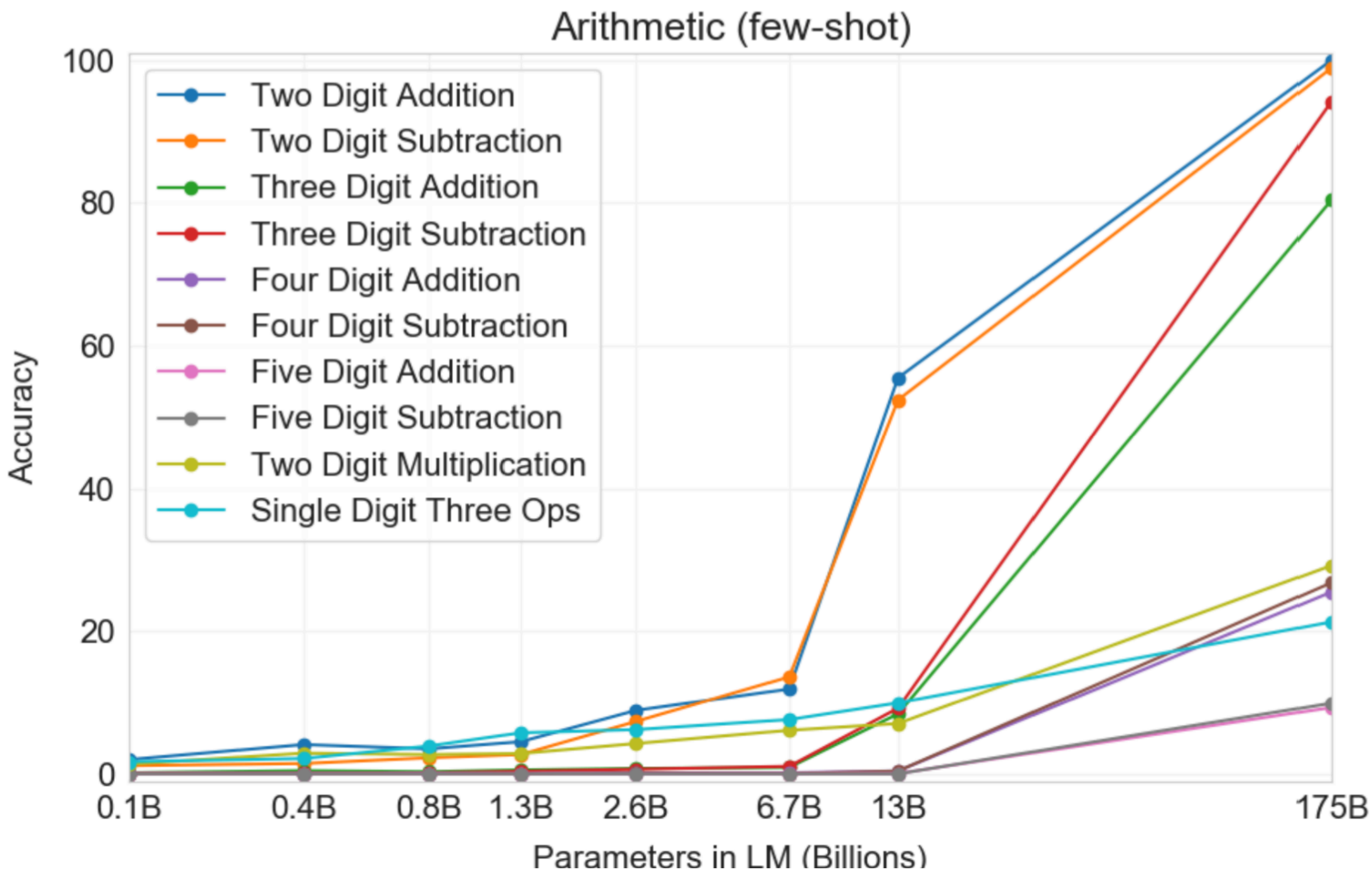
Context	Hypothesis	Reason	Round	Labels orig.	Labels pred.	Labels valid.	Annotations
Roberto Javier Mora García (c. 1962 – 16 March 2004) was a Mexican journalist and editorial director of "El Mañana", a newspaper based in Nuevo Laredo, Tamaulipas, Mexico. He worked for a number of media outlets in Mexico, including the "El Norte" and "El Diario de Monterrey", prior to his assassination.	Another individual laid waste to Roberto Javier Mora Garcia.	The context states that Roberto Javier Mora Garcia was assassinated, so another person had to have "laid waste to him." The system most likely had a hard time figuring this out due to it not recognizing the phrase "laid waste."	A1 (Wiki)	E	N	E E	Lexical (assassination, laid waste), Tricky (Presupposition), Standard (Idiom)
A melee weapon is any weapon used in direct hand-to-hand combat; by contrast with ranged weapons which act at a distance. The term "melee" originates in the 1640s from the French word "mělée", which refers to hand-to-hand combat, a close quarters battle, a brawl, a confused fight, etc. Melee weapons can be broadly divided into three categories	Melee weapons are good for ranged and hand-to-hand combat.	Melee weapons are good for hand to hand combat, but NOT ranged.	A2 (Wiki)	C	E	C N C	Standard (Conjunction), Tricky (Exhaustification), Reasoning (Facts)
If you can dream it, you can achieve it—unless you're a goose trying to play a very human game of rugby. In the video above, one bold bird took a chance when it ran onto a rugby field mid-play. Things got dicey when it got into a tussle with another player, but it shook it off and kept right on running. After the play ended, the players escorted the feisty goose off the pitch. It was a risky move, but the crowd chanting its name was well worth it.	The crowd believed they knew the name of the goose running on the field.	Because the crowd was chanting its name, the crowd must have believed they knew the goose's name. The word "believe" may have made the system think this was an ambiguous statement.	A3 (News)	E	N	E E	Reasoning (Facts), Reference (Coreference)

https://arxiv.org/abs/1910.14599

Adversarial Natural Language Inference

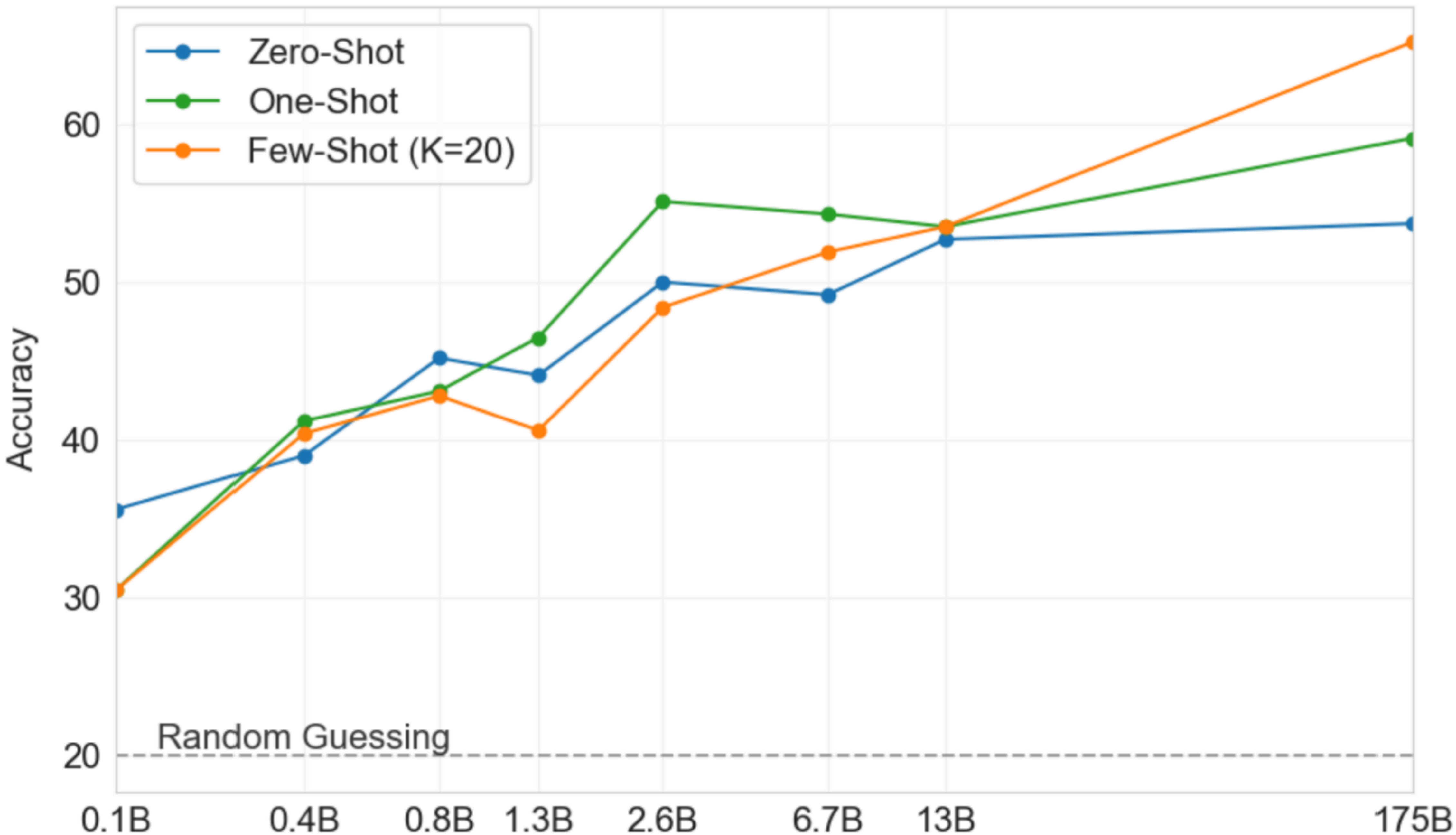


Arithmetic



Setting	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3 Zero-shot	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3 One-shot	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3 Few-shot	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3

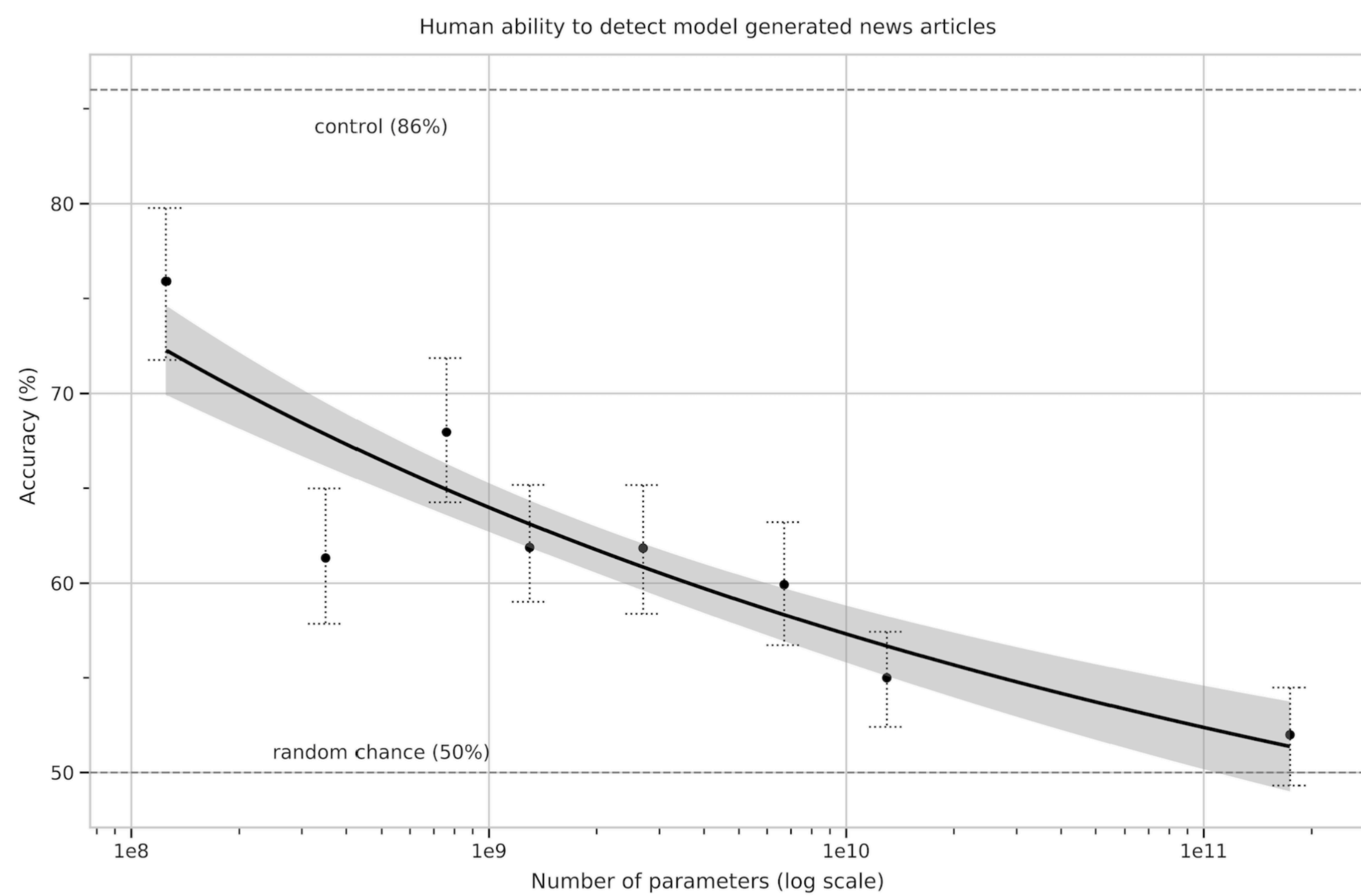
SAT Analogies



Audacious is to boldness as (a) sanctimonious is to hypocrisy, (b) anonymous is to identity, (c) remorseful is to misdeed, (d) deleterious is to result, (e) impressionable is to temptation

News Article Generation

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	"I don't know" assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

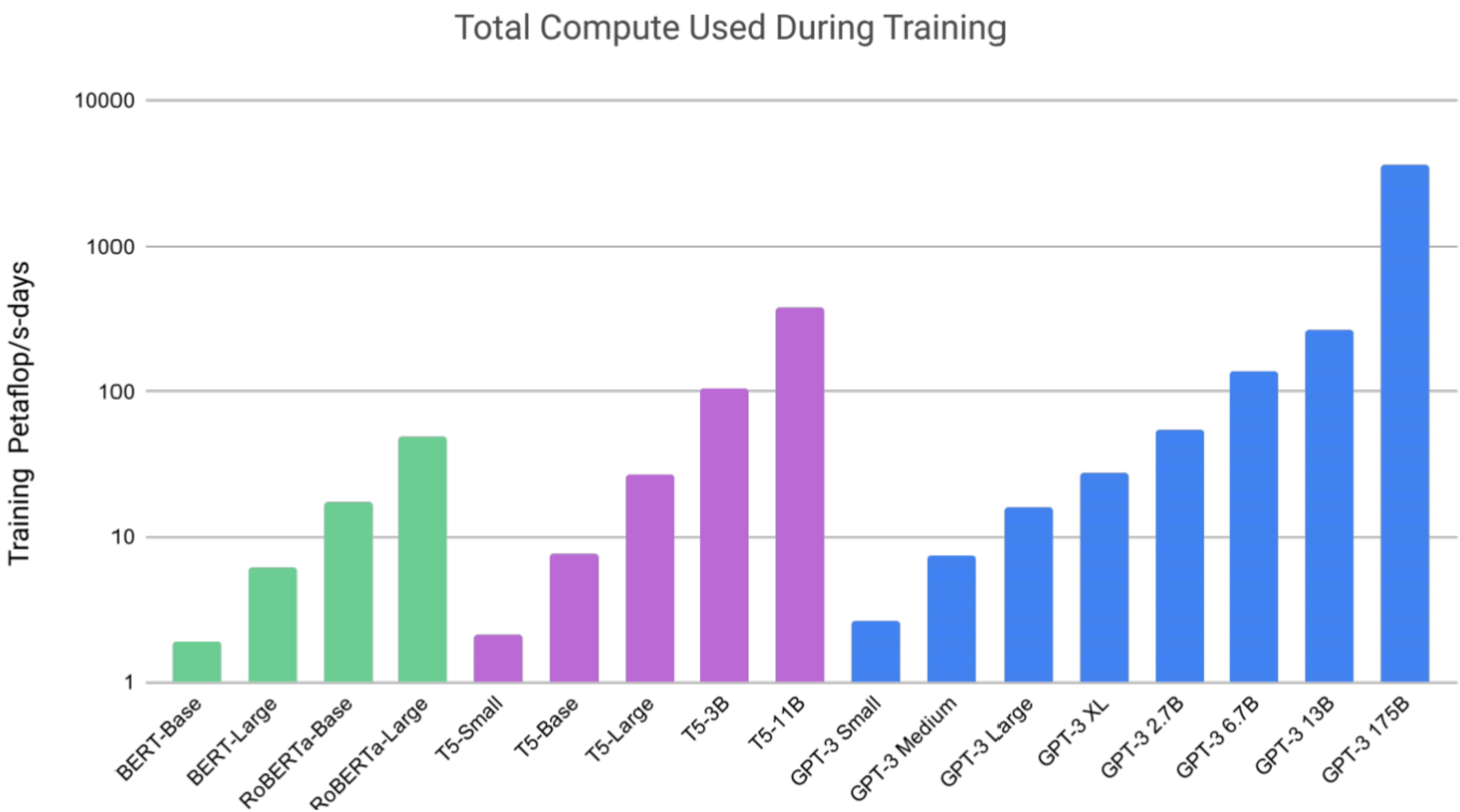


News Article Example

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Compute



https://arxiv.org/abs/2001.08361