

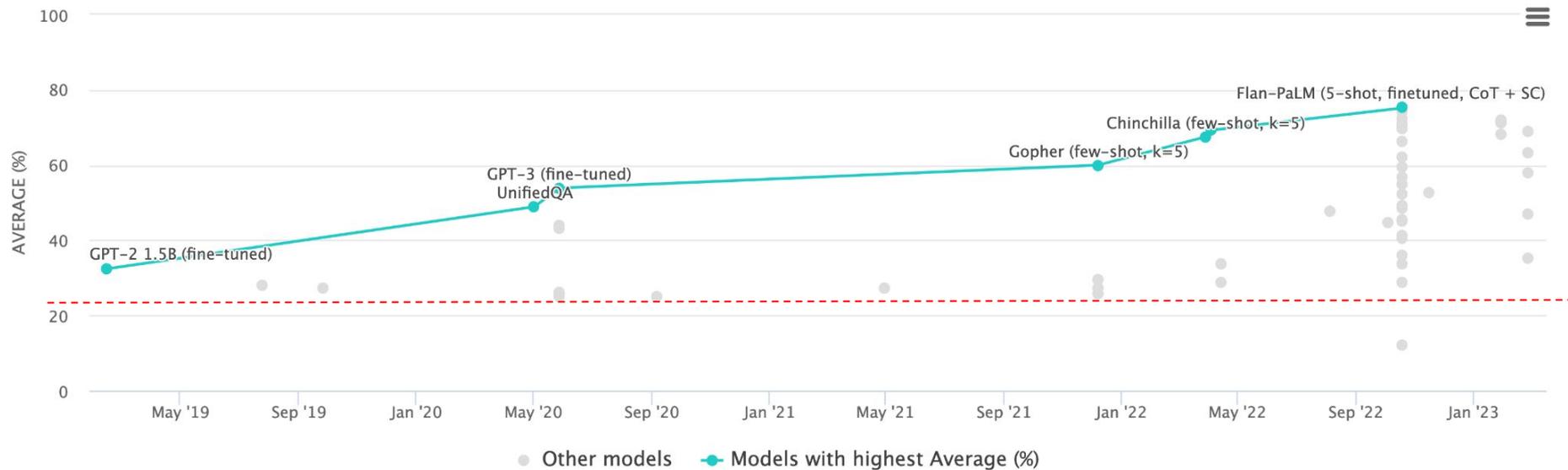
<https://tinyurl.com/mlhc2023>

Sign-in sheet!

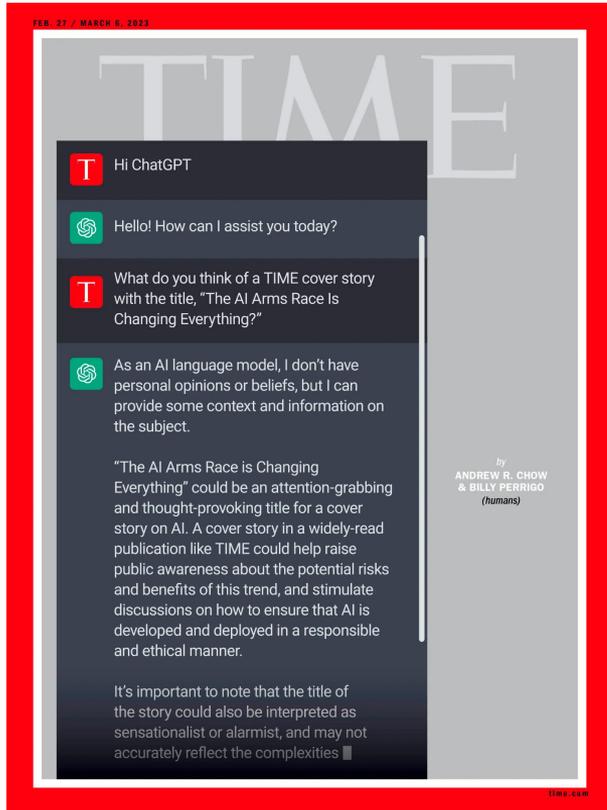
# Clinical NLP (Mostly NLP)

Eric Lehman, 3/7/23

# NLP has moved... Fast



# NLP has moved... Fast



# ChatGPT is very very good...

About 212 results (0.22 seconds)

 HealthITAnalytics

## ChatGPT Passes US Medical Licensing Exam Without Clinician Input

According to an MGH research spotlight describing the findings, ChatGPT is an advanced AI chatbot developed by OpenAI and released to the public...

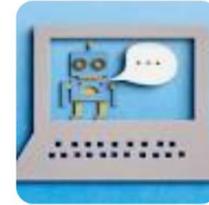
2 weeks ago

 The Daily Beast

## ChatGPT Passed the U.S. Medical Licensing Exam. Will It Be ...

The researchers tested ChatGPT on 350 of 376 publicly available questions from the June 2022 USMLE. The chatbot was able to score between 52.4...

3 weeks ago



ChatGPT is very very good...

## The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model

David M Levine, Rudraksh Tuwani, Benjamin Kompa, Amita Varma, Samuel G. Finlayson, Ateev Mehrotra, Andrew Beam

**doi:** <https://doi.org/10.1101/2023.01.30.23285067>

**Results** Among all cases, GPT-3 replied with the correct diagnosis in its top 3 for 88% (95% CI, 75% to 94%) of cases, compared to 54% (95% CI, 53% to 55%) for lay individuals ( $p < 0.001$ ) and 96% (95% CI, 94% to 97%) for physicians ( $p = 0.0354$ ). GPT-3 triaged (71% correct; 95% CI, 57% to 82%) similarly to lay individuals (74%; 95% CI, 73% to 75%;  $p = 0.73$ ); both were significantly worse than physicians (91%; 95% CI, 89% to 93%;  $p < 0.001$ ). As measured by the Brier score, GPT-3 confidence in its top prediction was reasonably well-calibrated for diagnosis (Brier score = 0.18) and triage (Brier score = 0.22).

However...

However...

LE

Who is Pete Szolovits?

# However...

LE

Who is Pete Szolovits?



Pete Szolovits is a computer scientist and professor at the Massachusetts Institute of Technology (MIT). He was born on October 8, 1946, in Budapest, Hungary. He received his undergraduate degree from Brown University in 1968 and his PhD from Stanford University in 1974.

However...

# However...



Can you give me some references on digital epidemiology?



Sure, here are some references on digital epidemiology that you might find useful:



- Fox, J., & Dugas, M. (2018). Digital epidemiology: A new paradigm for infectious disease surveillance. *Annual Review of Public Health*, 39(1), 171-186.
- Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., & Ferguson, N. M. (2010). Discriminating between epidemic and non-epidemic situations using high-resolution surveillance data. *Epidemics*, 2(3), 87-95.
- Salathé, M., & Khandelwal, S. (2017). Digital epidemiology. *Annual Review of Public Health*, 38, 105-120.
- Scarpino, S. V., & Allard, A. (2017). Big data in epidemiology: Opportunities and challenges. *Journal of the Royal Society Interface*, 14(130), 20170408.
- Tariq, M. H., & Davis, C. (2019). Digital epidemiology: An emerging discipline for studying infectious diseases in the digital age. *Infection, Disease & Health*, 24(3), 153-157.
- Vespignani, A. (2012). Modelling dynamic and multi-scale phenomena with computational social science. *Nature Physics*, 8(9), 32-39.

# However...



Hiccups, a common affliction that most individuals experience at some point in their lives, are often benign and resolve spontaneously. However, persistent hiccups, defined as hiccups lasting for more than 48 hours, can lead to significant morbidity, including malnutrition, weight loss, sleep disturbance, and rare complications such as pneumothorax or esophageal rupture. Treatment options for persistent hiccups are varied and often unsuccessful, with few proven effective therapies.

Water submersion therapy, also known as the "dunking" or "sink" method, has emerged as a promising treatment option for persistent hiccups. This therapy involves immersing the head in a container of water, with the goal of interrupting the hiccup reflex arc by activating the vagus and phrenic nerves.

A recent randomized controlled trial conducted by our group sought to evaluate the effectiveness of water submersion therapy in the treatment of persistent hiccups. A total of 50 patients with persistent hiccups were enrolled in the study and randomly assigned to receive either water submersion therapy or standard medical therapy. The primary outcome measure was resolution of hiccups within 24 hours of treatment initiation.

And even worse...



**Welcome to the new  
Bing**

Your AI-powered  
answer engine

Ask anything

And even worse...



ALL

IMAGES

VIDEOS

MAPS

NEWS

**Welcome to the new  
Bing**

Your AI-powered  
answer engine

Ask anything

And even worse...



is it safe to boil a baby

ALL

IMAGES

VIDEOS

MAPS

NEWS

About 86,100,000 results

Date ▾

Welcome to the new  
Bing

Your AI-powered  
answer engine

Ask anything

Yes

According to 2 sources

# Main Questions

- So you're probably wondering... how did we get here?

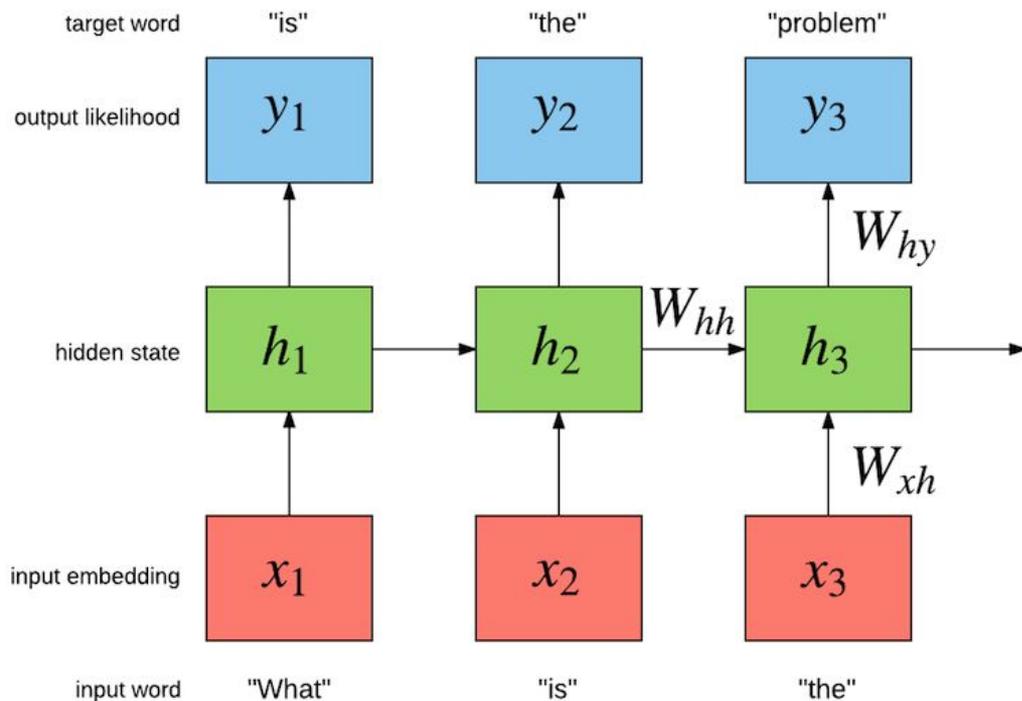
# Main Questions

- So you're probably wondering... how did we get here?
- Why has the field moved so fast?

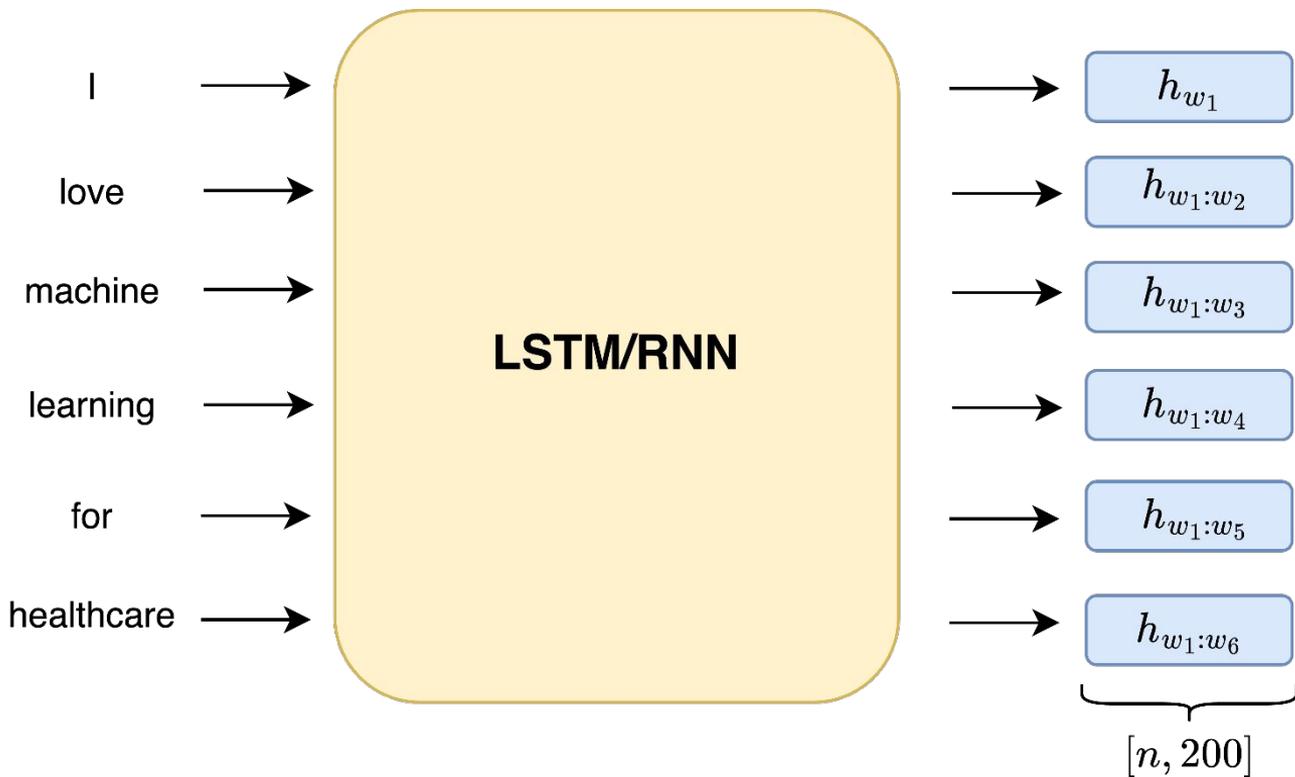
# Main Questions

- So you're probably wondering... how did we get here?
- Why has the field moved so fast?
- What does it mean for healthcare?

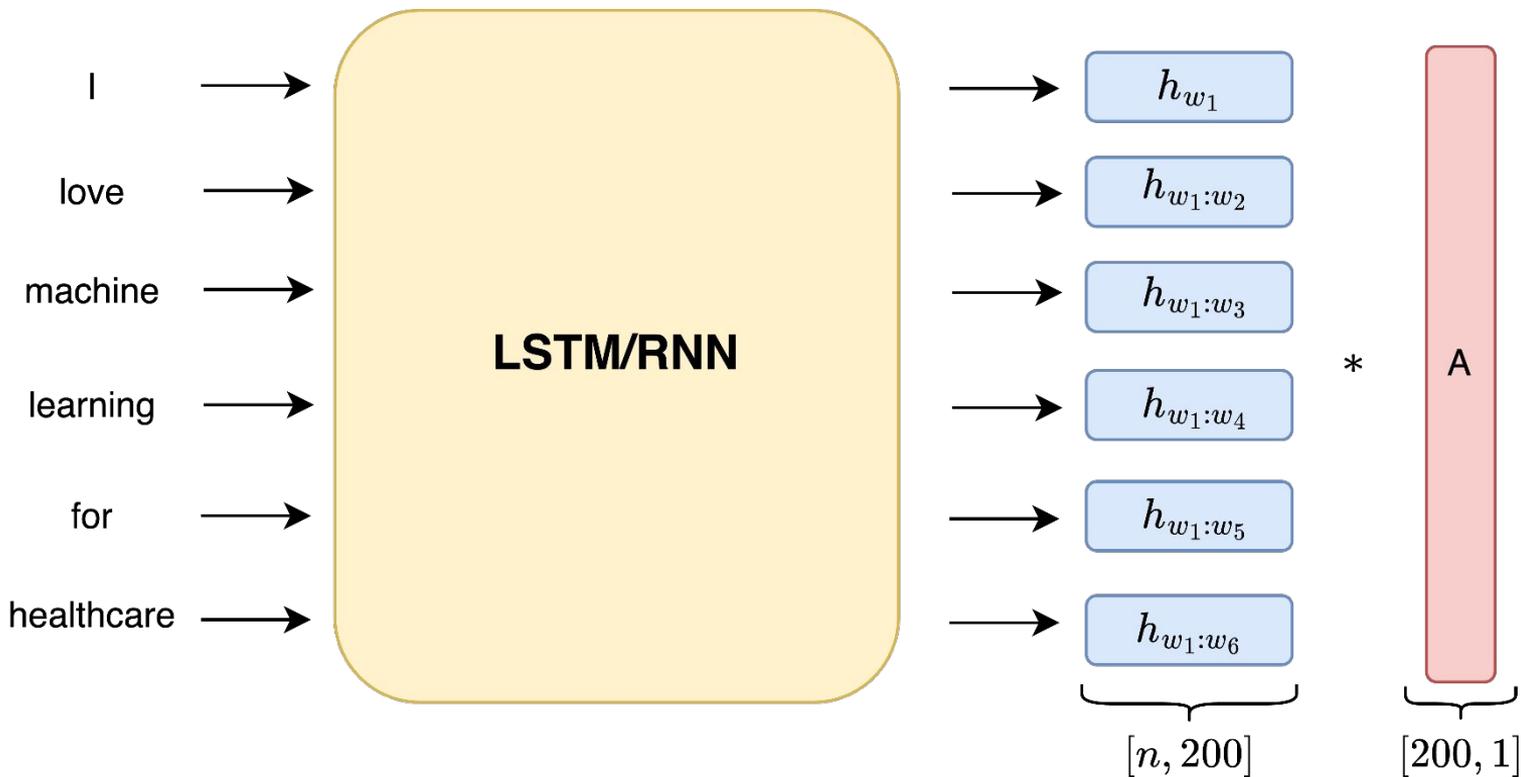
# Review: Recurrent Neural Networks



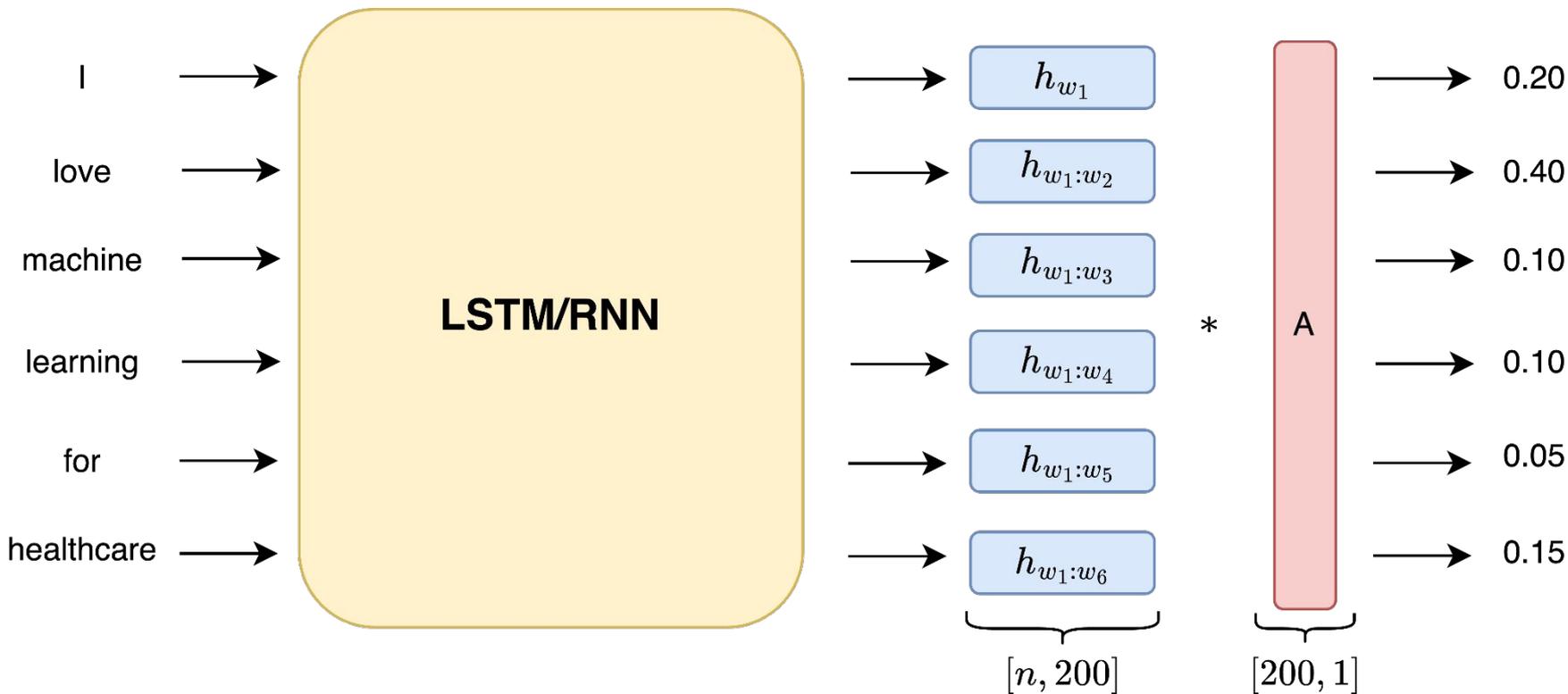
# Attention Mechanism (2015)



# Attention Mechanism (2015)



# Attention Mechanism (2015)



# Attention Mechanism (2015)

# Attention Mechanism (2015)

$$0.2 * h_{w_1:w_1}$$

+

$$0.4 * h_{w_1:w_2}$$

+

$$0.1 * h_{w_1:w_3}$$

+

...

+

$$0.15 * h_{w_1:w_n}$$

# Attention Mechanism (2015)

$$0.2 * h_{w_1:w_1}$$

+

$$0.4 * h_{w_1:w_2}$$

+

$$0.1 * h_{w_1:w_3}$$

+

...

+

$$0.15 * h_{w_1:w_n}$$

**Idea:** How important is **this** word?

# Attention Heatmaps: Predicting Number of Stars

*Task: Hotel location*

you get what you pay for . not the **cleanest rooms** but bed was **clean** and so was **bathroom** . bring your own towels though as very **thin** . service was **excellent** , let us book in at 8:30am ! for **location and price** , **this ca n't be beaten** , but it is **cheap** for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

*Task: Hotel cleanliness*

you get what you pay for . **not the cleanest rooms but bed was clean and so was bathroom** . bring your own towels though as very **thin** . service was **excellent** , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is **cheap** for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

*Task: Hotel service*

you get what you pay for . not the cleanest rooms but bed was **clean** and so was **bathroom** . bring your own **towels** though as very **thin** . **service was excellent** , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is **cheap** for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

# Attention is Not Explanation (2018)

*after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore*

original  $\alpha$

$$f(x|\alpha, \theta) = 0.01$$

## Attention is Not Explanation (2018)

*after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore*

original  $\alpha$

$$f(x|\alpha, \theta) = 0.01$$

*after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore*

adversarial  $\tilde{\alpha}$

$$f(x|\tilde{\alpha}, \theta) = 0.01$$

# Next Word Prediction/Language Modeling

# Next Word Prediction/Language Modeling

**Text:** Second Law of Robotics: A robot must obey the orders given it by human beings

# Next Word Prediction/Language Modeling

**Text:** Second Law of Robotics: A robot must obey the orders given it by human beings



**Generated training examples**

**Example #**

**Input (features)**

**Correct output (labels)**

1

Second law of robotics :

a

# Next Word Prediction/Language Modeling

**Text:** Second Law of Robotics: A robot must obey the orders given it by human beings



**Generated training examples**

**Example #**

**Input (features)**

**Correct output (labels)**

1

Second law of robotics :

a

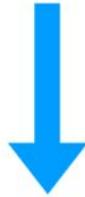
2

Second law of robotics : a

robot

# Next Word Prediction/Language Modeling

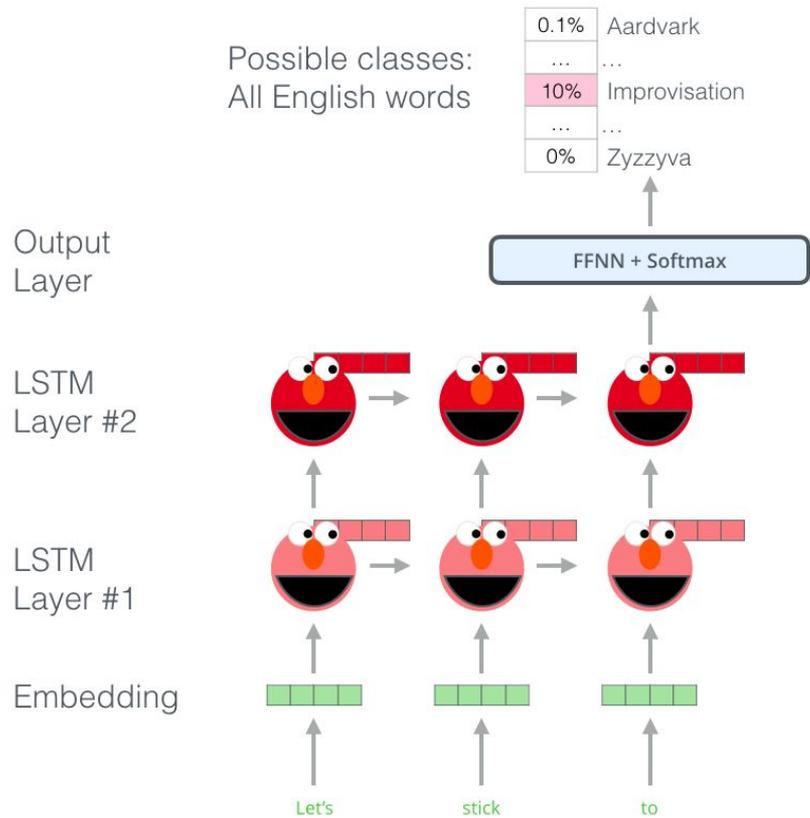
**Text:** Second Law of Robotics: A robot must obey the orders given it by human beings



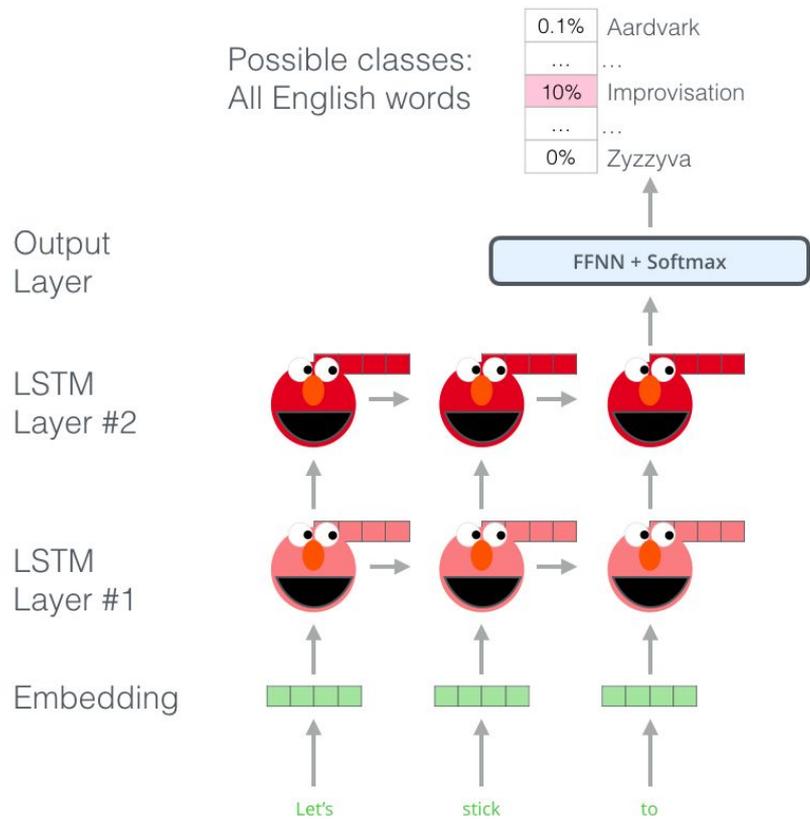
**Generated training examples**

Example #	Input (features)	Correct output (labels)
1	Second law of robotics :	a
2	Second law of robotics : a	robot
3	Second law of robotics : a robot	must
...		

# ELMO Pretraining – Language Modeling

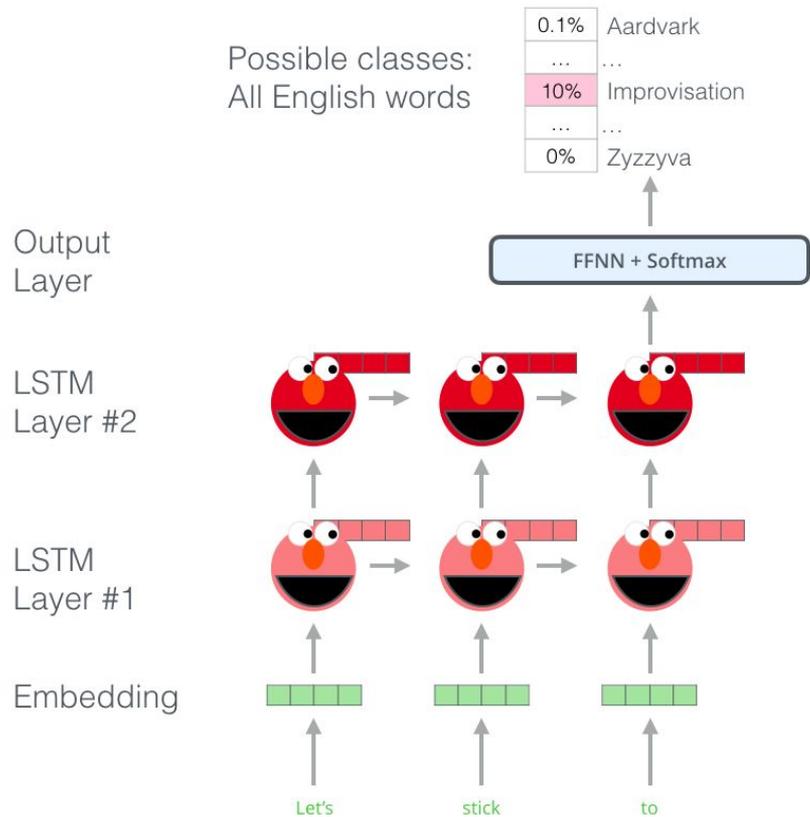


# ELMO Pretraining – Language Modeling



Q: How is learning to predict the next word relevant to doing deidentification, predicting sepsis from nursing notes, etc?

# ELMO Pretraining – Language Modeling



Q: How is learning to predict the next word relevant to doing deidentification, predicting sepsis from nursing notes, etc?

A: It allows us learn how to represent words & sentences FOR FREE.

How Would You Use ELMO For a Clinical Task?

## How Would You Use ELMO For a Clinical Task?

1. Load the model that has been trained to predict:

$$P(w_i | w_1, w_2, \dots, w_{i-1})$$

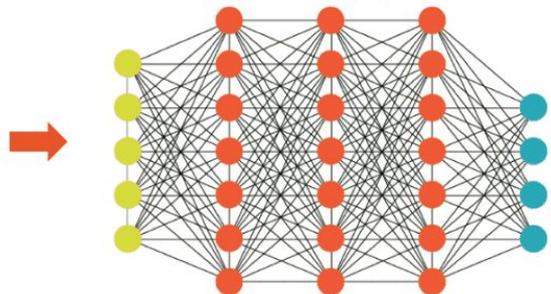
## How Would You Use ELMO For a Clinical Task?

1. Load the model that has been trained to predict:

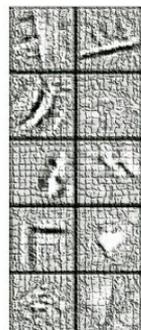
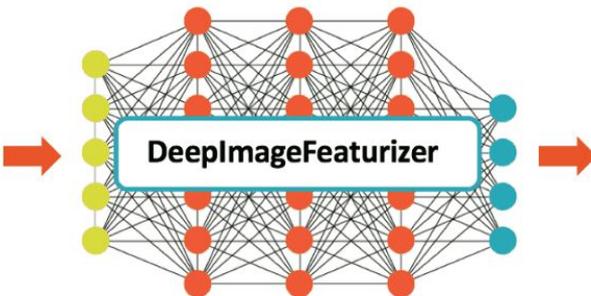
$$P(w_i | w_1, w_2, \dots, w_{i-1})$$

2. Use the model to encode your sentence and train it on **YOUR** task.

# Transfer Learning/Pretraining



GIANT PANDA 0.9  
RED PANDA 0.05  
RACCOON 0.01  
...



Chihuahua

Any Questions Thus Far?

This is essential for the rest of the lecture!

# ELMO (2018)

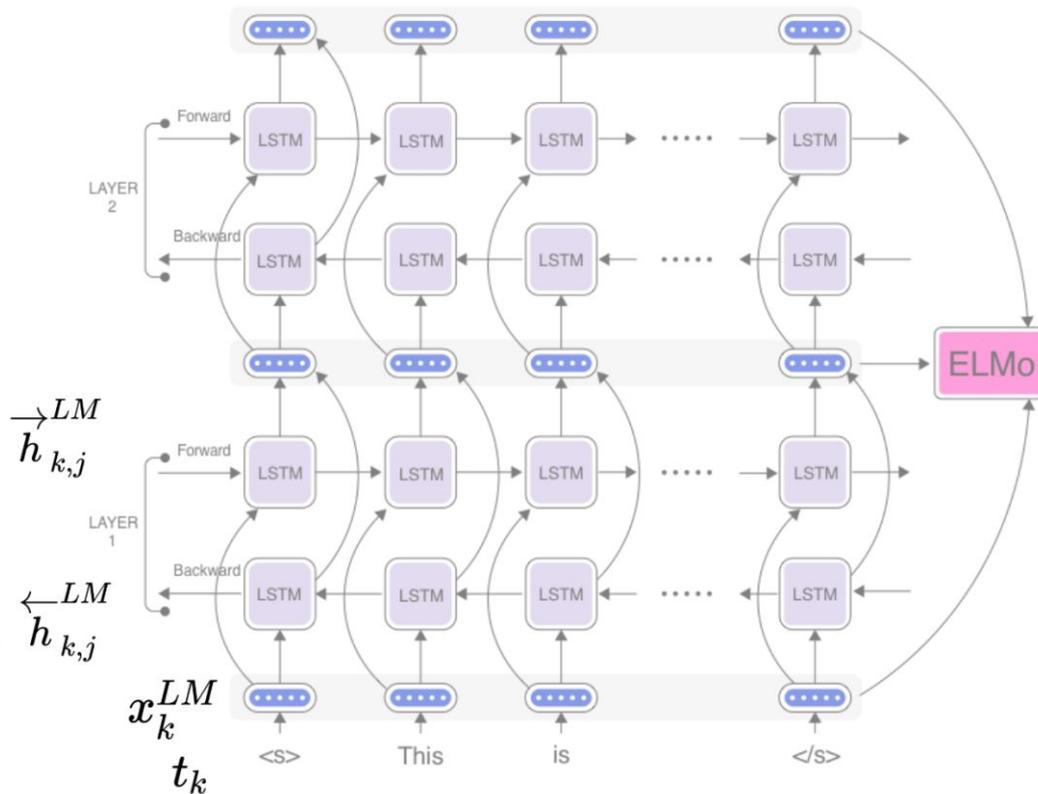
## Structure

Each token  $t_k$

L-layer biLM  
computes  $2L+1$   
representations

k is the k-th token

j is the j-th biLM layer



## ELMO Summary (2018)

- Allows for contextual representations of words

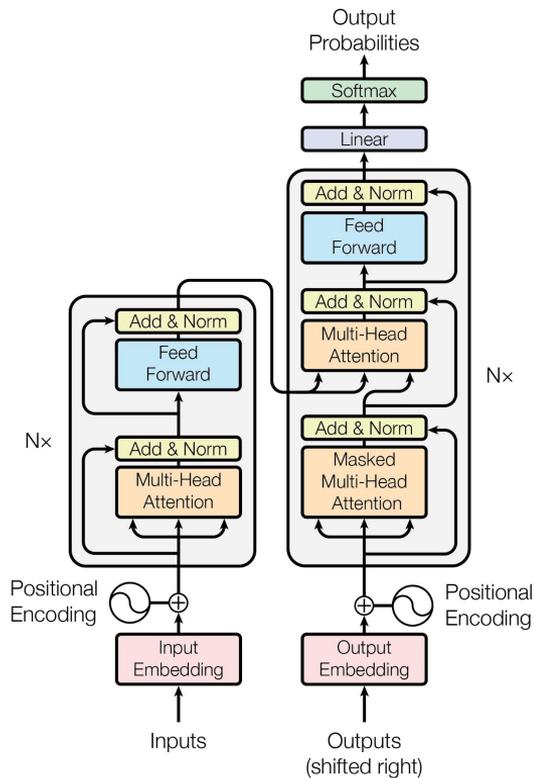
## ELMO Summary (2018)

- Allows for contextual representations of words
- Similar to Word2Vec, this is unsupervised (i.e., pretraining)

## ELMO Summary (2018)

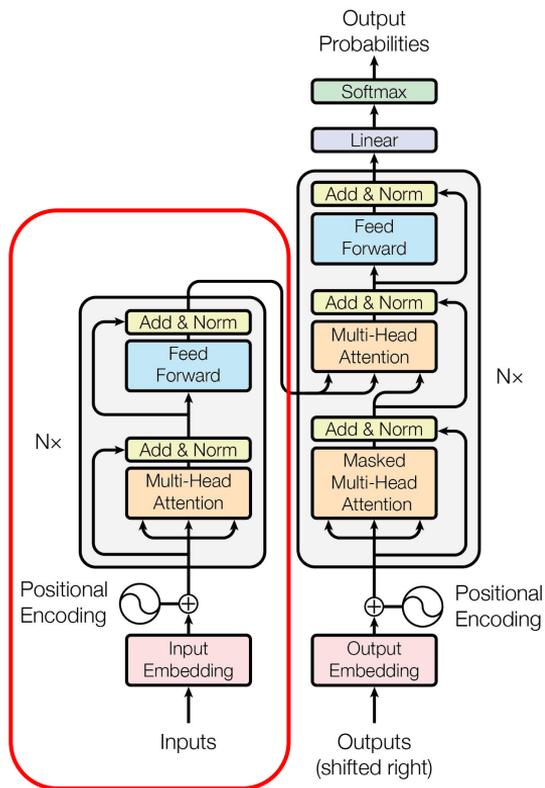
- Allows for contextual representations of words
- Similar to Word2Vec, this is unsupervised (i.e., pretraining)
- SUPER SLOW

# Transformer (2017)

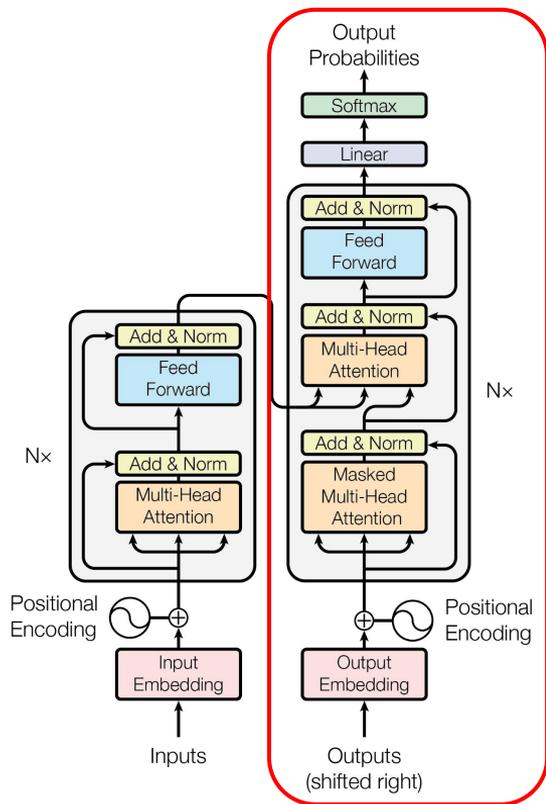


# Transformer (2017)

**Encoder:** reads the entire sequence all at once.



# Transformer (2017)



**Encoder:** reads the entire sequence all at once.

**Decoder:** reads left to right (but parallelized)

# Self-Attention

**Step 1:** Take each word in the sequence and create a word embedding for it.

# Self-Attention

**Step 1:** Take each word in the sequence and create a word embedding for it.

*“Anthony Hopkins admired Michael Bay as a great director”*



# Self-Attention

**Step 1:** Take each word in the sequence and create a word embedding for it.

*“Anthony Hopkins admired Michael Bay as a great director”*



Anthony

Hopkins

admired

Michael

Bay

as

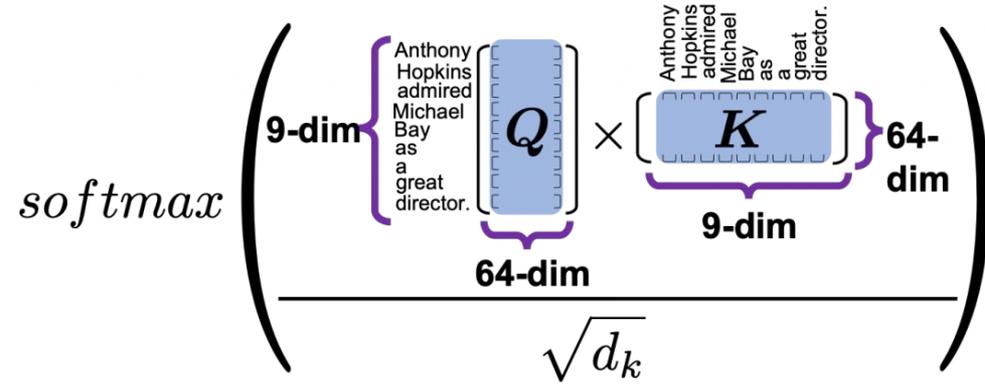
a

great

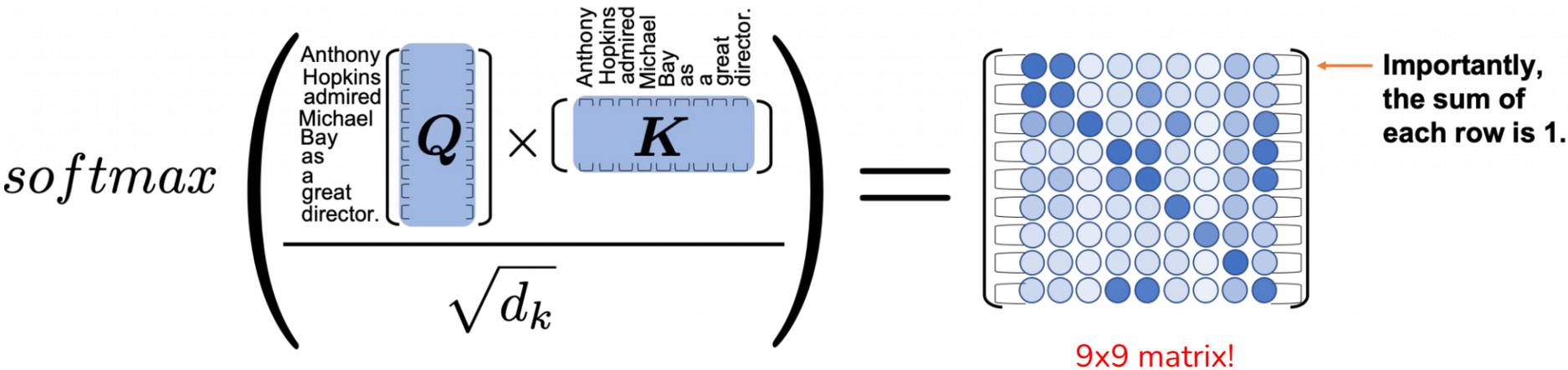
director

[1, 64]

# Self-Attention



# Self-Attention



# Self-Attention

for `word1` in `sentence`:

```
row = []
```

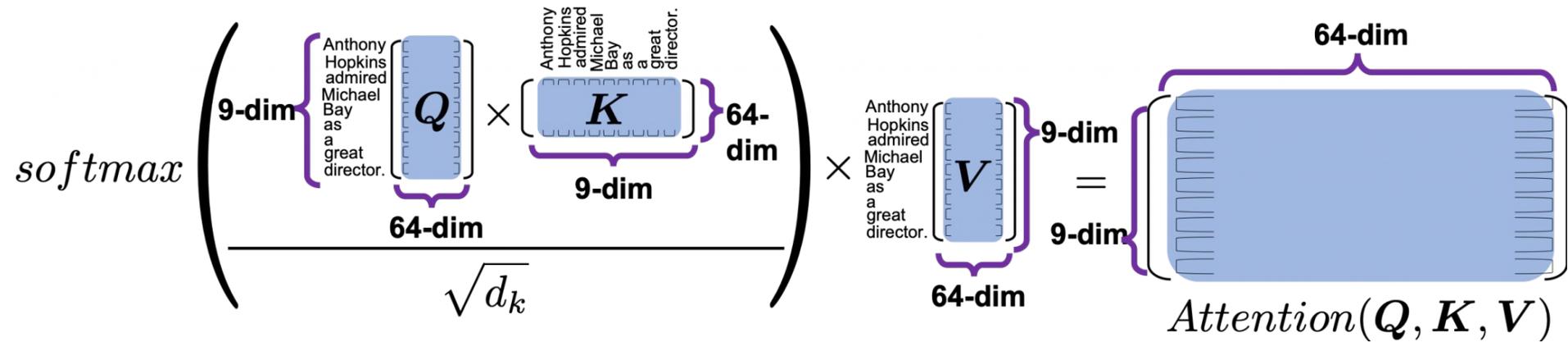
for `word2` in `sentence`:

```
score = word1 * word2
```

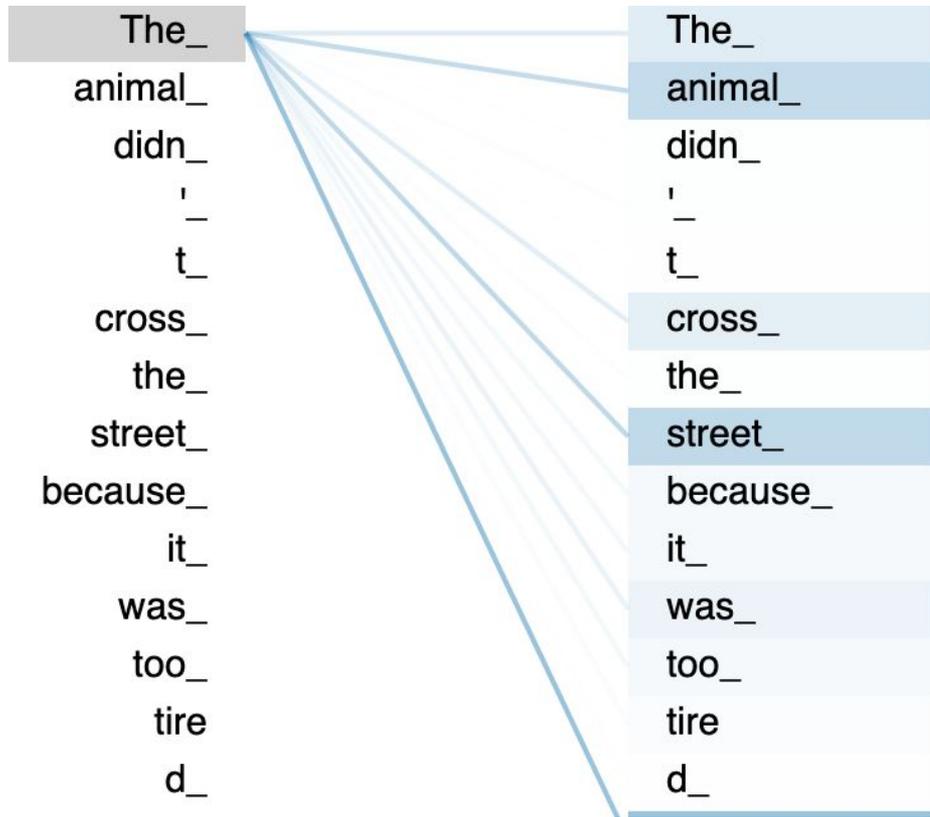
```
row.append(score)
```

```
softmax(row)
```

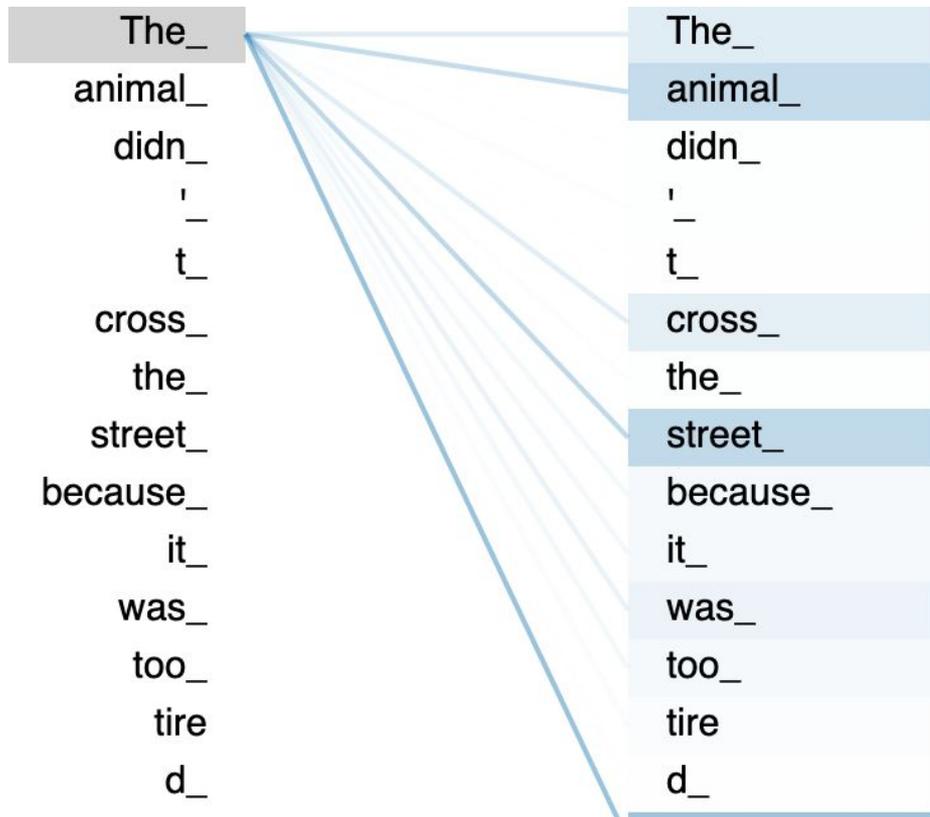
# Self-Attention



# Transformer (2017)

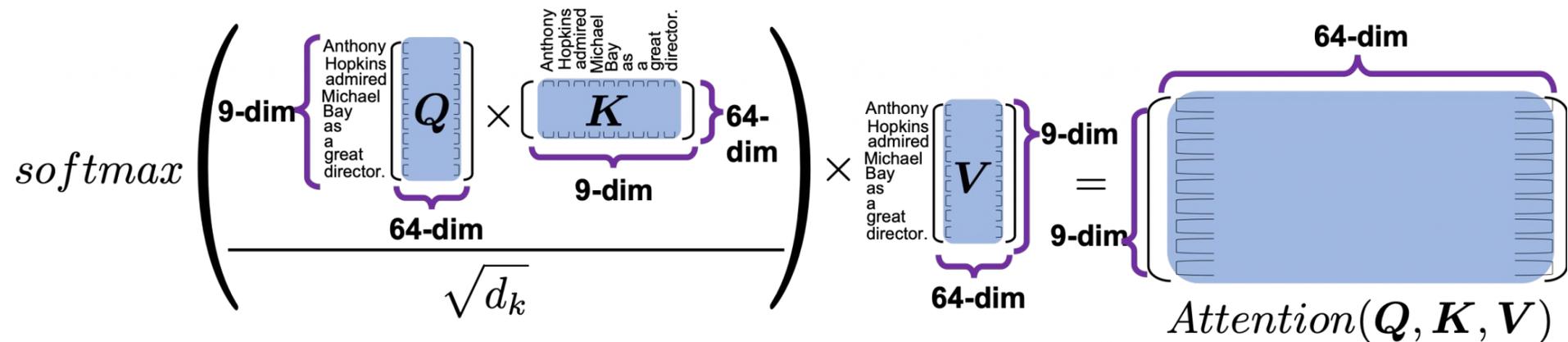


# Transformer (2017)



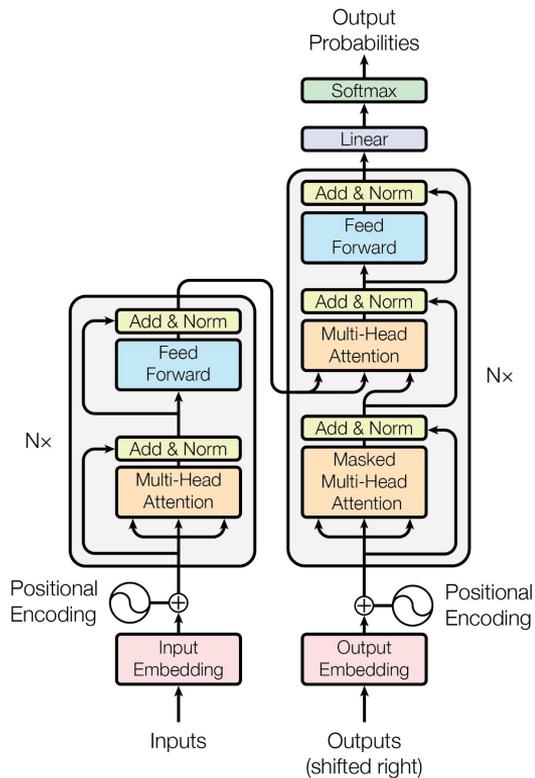
**Idea:** How important is *this word*, with respect to ALL other words?

# Transformer (2017) – Position Embeddings

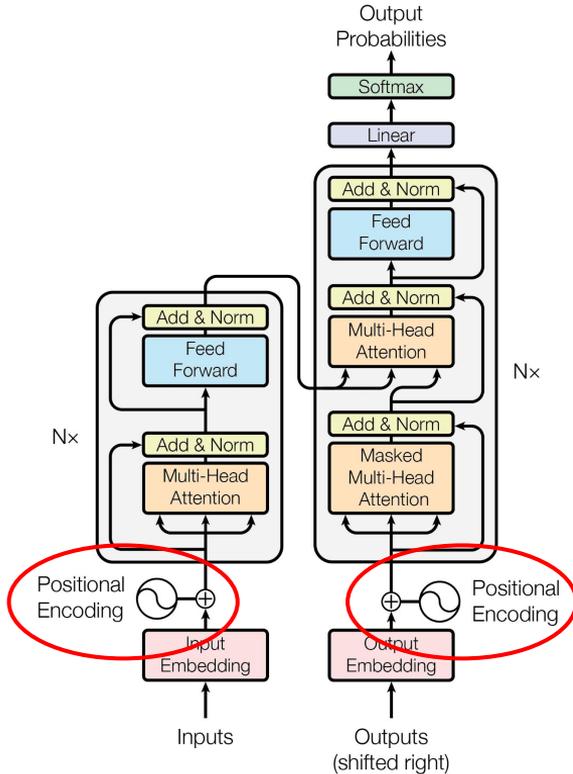


Where is the ordering? Isn't this just a big bag of words???

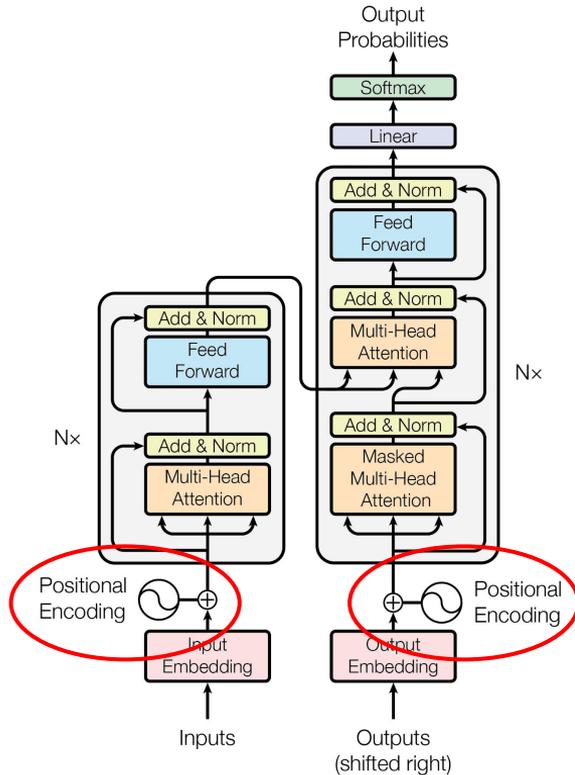
# Transformer (2017) – Position Embeddings



# Transformer (2017) – Position Embeddings

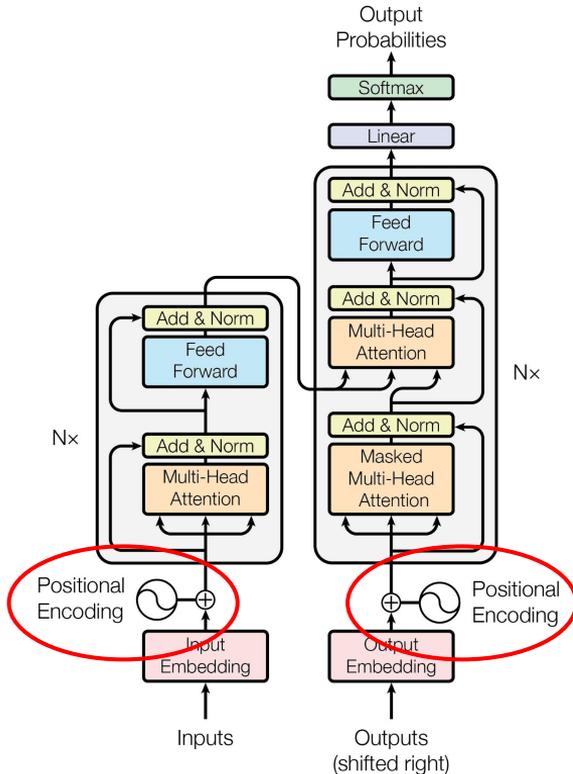


# Transformer (2017) – Position Embeddings



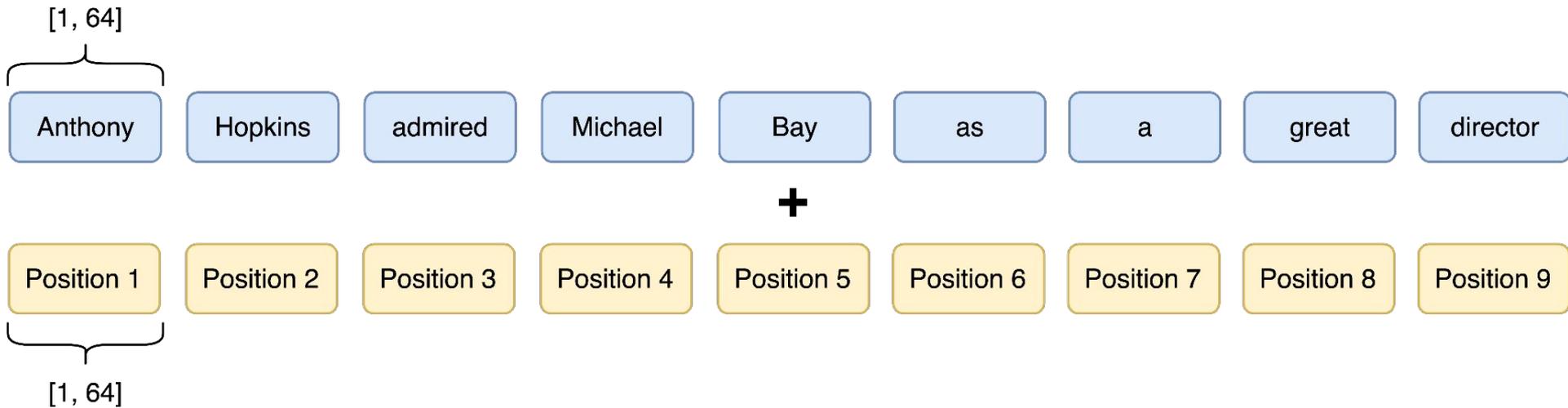
- Every position (i.e., the first, second, third... word) has a unique vector that represents its position in the sentence/paragraph

# Transformer (2017) – Position Embeddings



- Every position (i.e., the first, second, third... word) has a unique vector that represents its position in the sentence/paragraph
- These are randomly initialized and learned by the model!

# Transformer (2017) – Position Embeddings



# Transformer Performance

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

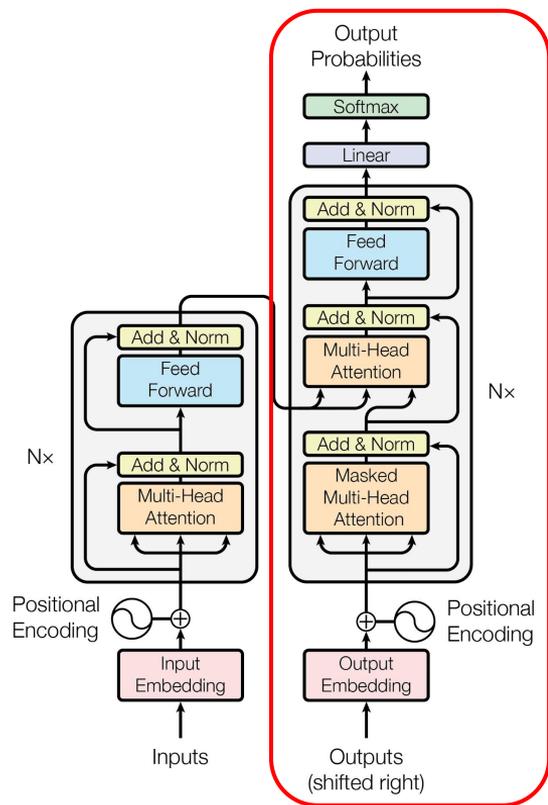
Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	<b>41.29</b>	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	<b><math>3.3 \cdot 10^{18}</math></b>	
Transformer (big)	<b>28.4</b>	<b>41.8</b>	$2.3 \cdot 10^{19}$	

# Transformer Speed

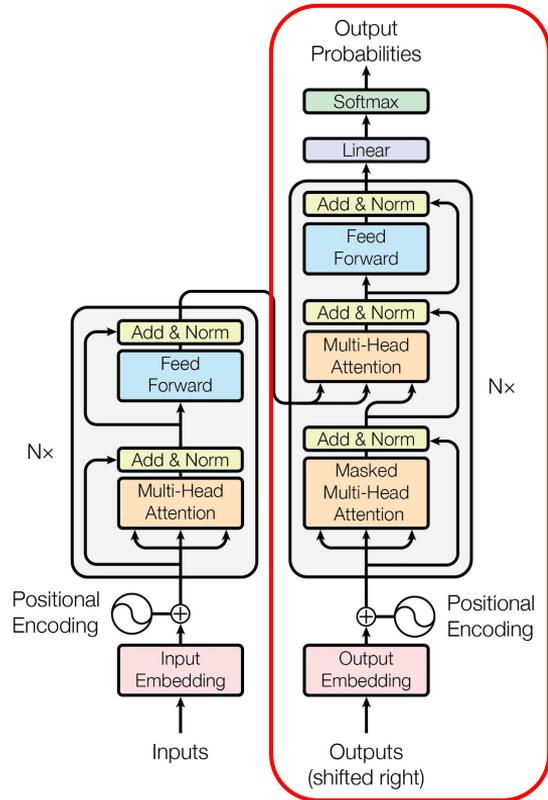
Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types.  $n$  is the sequence length,  $d$  is the representation dimension,  $k$  is the kernel size of convolutions and  $r$  the size of the neighborhood in restricted self-attention.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(\log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

# Generative Pretrained Transformer (GPT), 2018

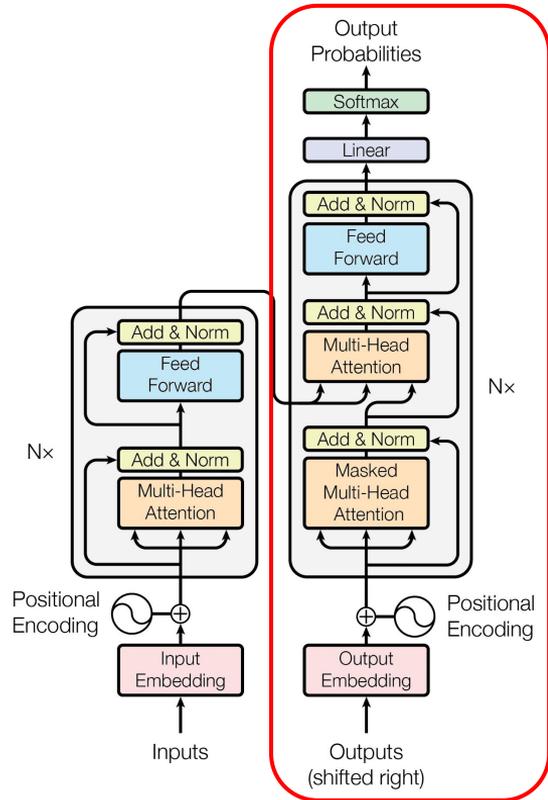


# Generative Pretrained Transformer (GPT), 2018



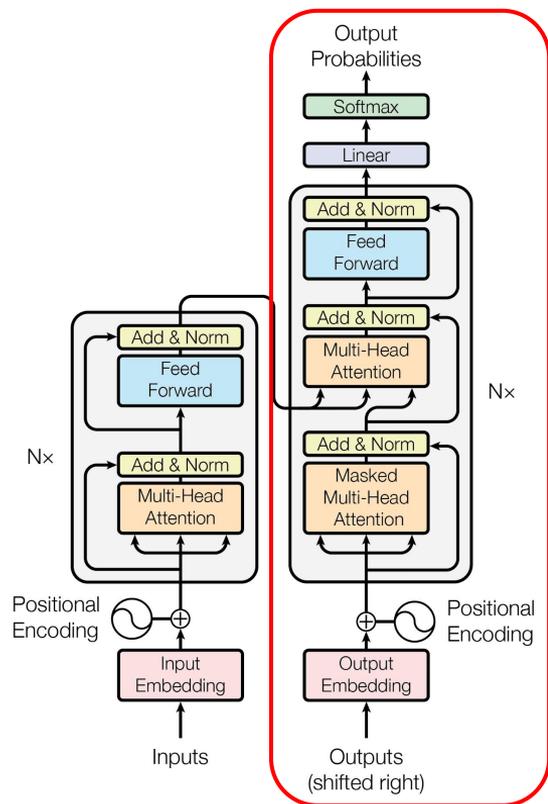
- Pretraining using next word prediction on 7000 books (1B words)

# Generative Pretrained Transformer (GPT), 2018



- Pretraining using next word prediction on 7000 books (1B words)
- Autoregressive: Reads left-to-right

# Generative Pretrained Transformer (GPT), 2018

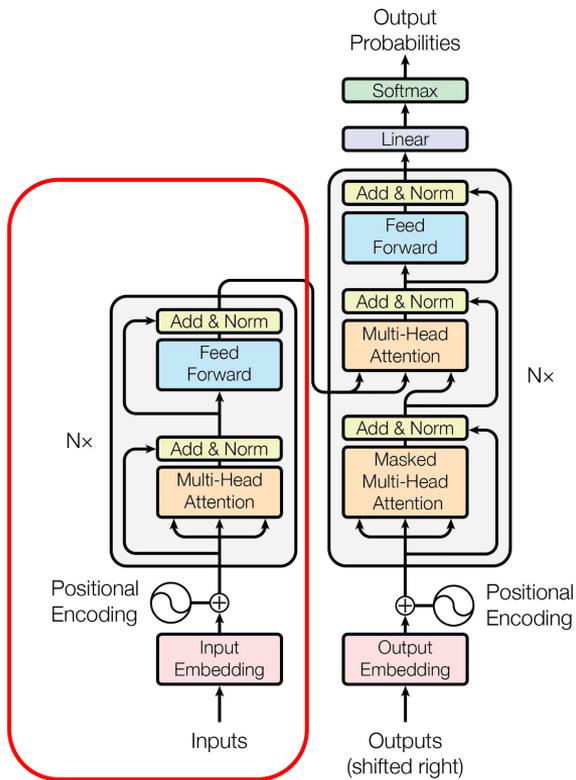


- Pretraining using next word prediction on 7000 books (1B words)
- Autoregressive: Reads left-to-right
- 117M parameters

# Generative Pretrained Transformer (GPT), 2018

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

# Bidirectional Encoder Representations from Transformers (BERT)

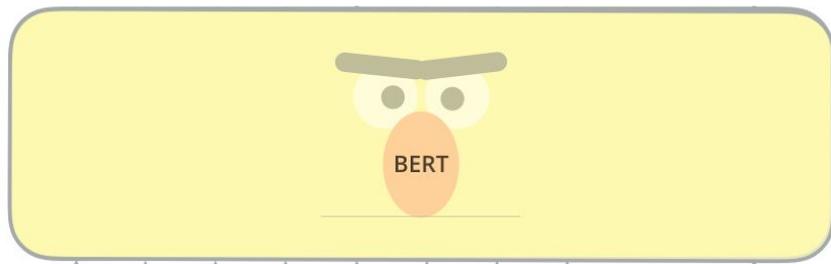


# BERT + Masked Language Modeling

Input

↑   ↑   ↑   ↑   ↑   ↑   ↑   ↑  
[CLS] Let's stick to improvisation in this skit

# BERT + Masked Language Modeling



Randomly mask  
15% of tokens

1↑ 2↑ 3↑ 4↑ 5↑ 6↑ 7↑ 8↑ ... 512↑  
[CLS] Let's stick to [MASK] in this skit

Input

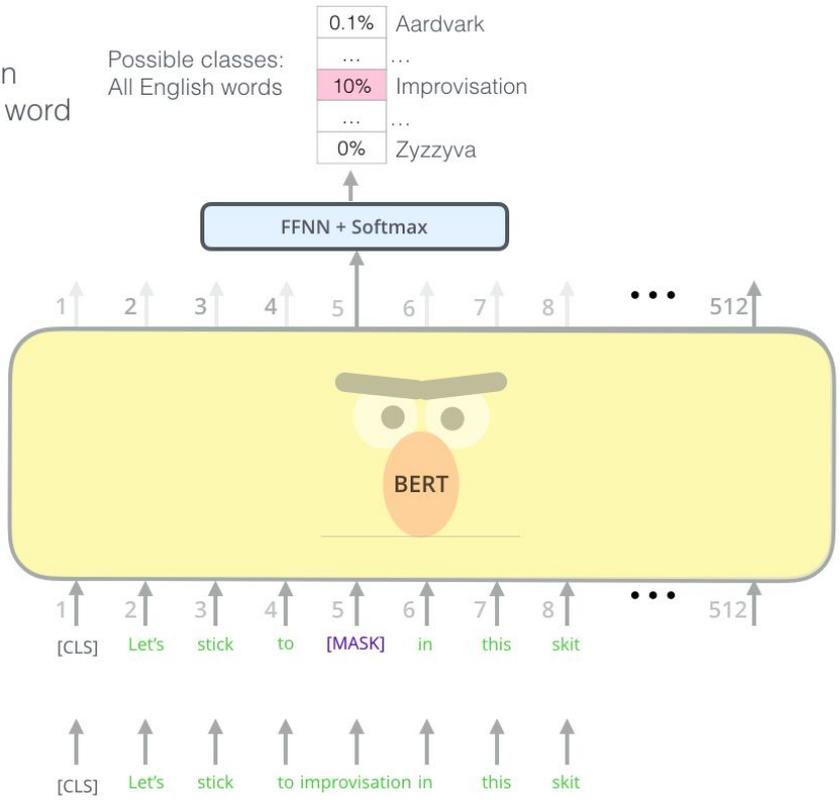
↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑  
[CLS] Let's stick to improvisation in this skit

# BERT + Masked Language Modeling

Use the output of the masked word's position to predict the masked word

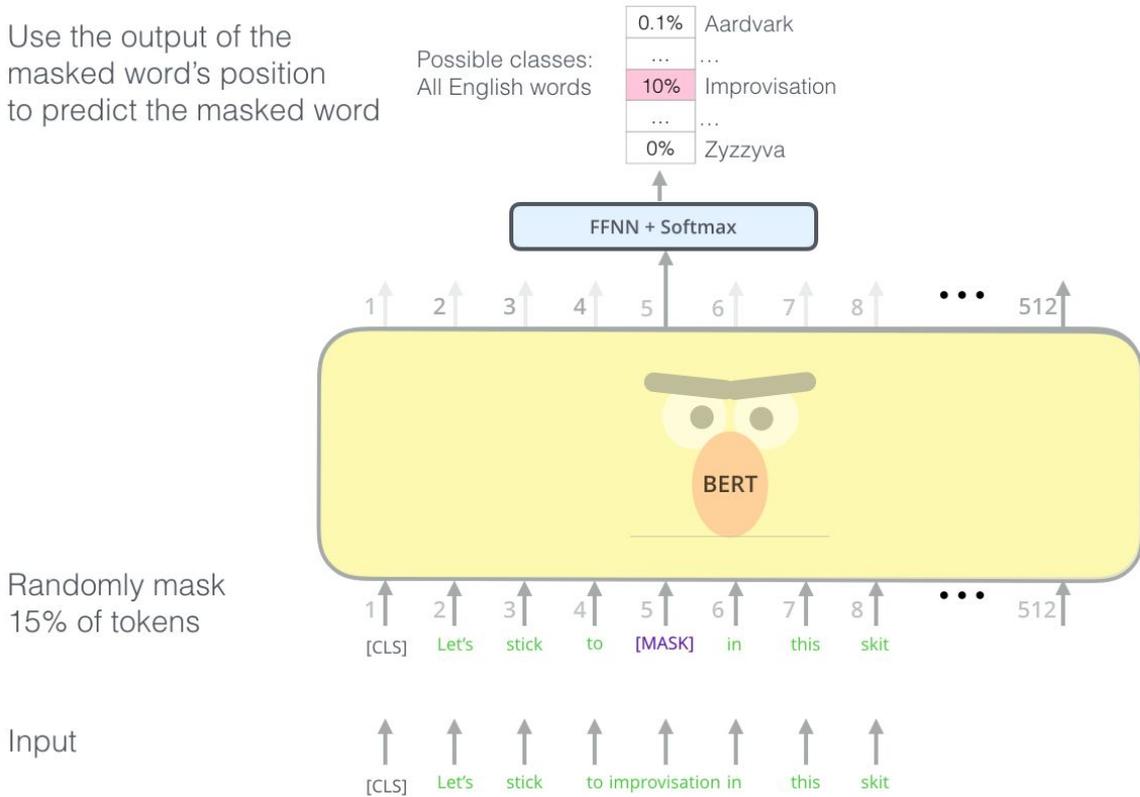
Randomly mask 15% of tokens

Input



# BERT + Masked Language Modeling

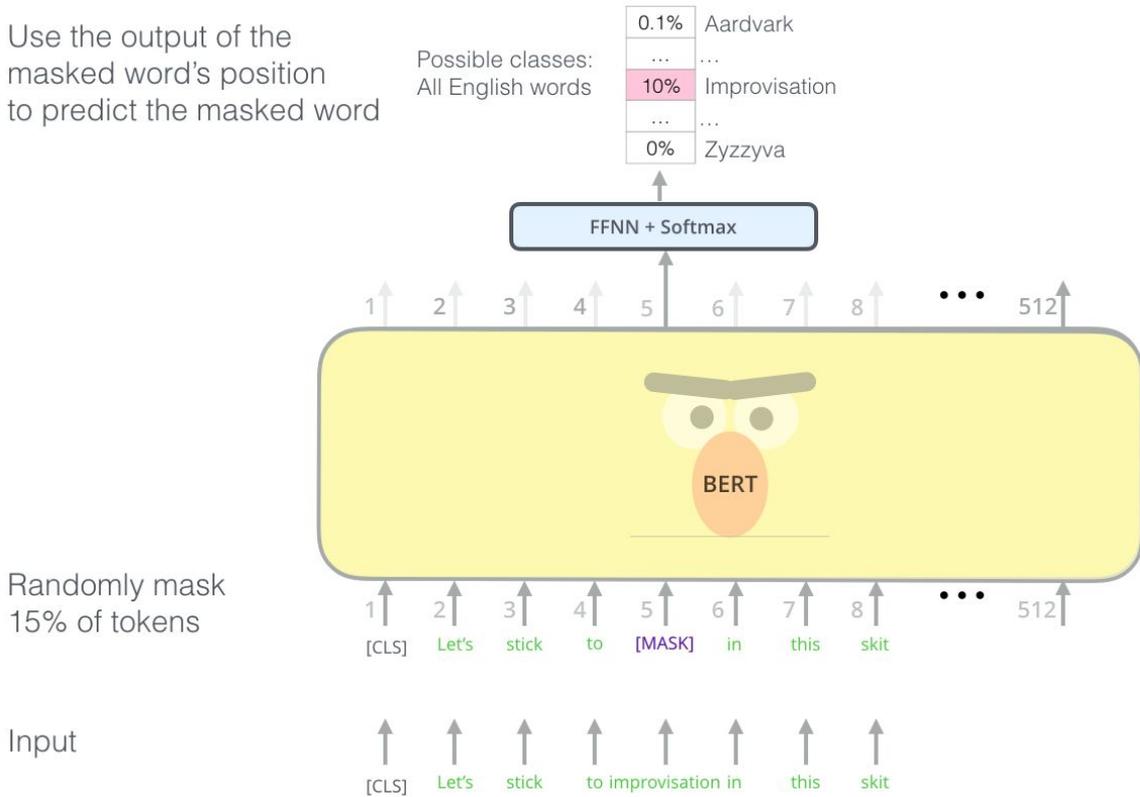
Use the output of the masked word's position to predict the masked word



- Pretrained on ~3.4B words for 40 epochs

# BERT + Masked Language Modeling

Use the output of the masked word's position to predict the masked word



- Pretrained on ~3.4B words for 40 epochs
- 110M and 345M parameter models

# BERT Performance

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
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
BERT <sub>BASE</sub>	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT <sub>LARGE</sub>	<b>86.7/85.9</b>	<b>72.1</b>	<b>92.7</b>	<b>94.9</b>	<b>60.5</b>	<b>86.5</b>	<b>89.3</b>	<b>70.1</b>	<b>82.1</b>

# Domain Adaptive Pretraining (DAPT)

# Domain Adaptive Pretraining (DAPT)

**Idea:** BERT is trained over news, Wikipedia, and other “general” sources. This is not ideal for biomedical tasks.

# Domain Adaptive Pretraining (DAPT)

**Idea:** BERT is trained over news, Wikipedia, and other “general” sources. This is not ideal for biomedical tasks.

**Solution:** Pretrain on PubMed (i.e., biomedical text).

# BioBERT (2019)

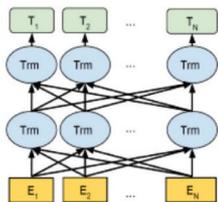
## Pre-training of BioBERT

Pre-training Corpora

**PubMed** 4.5B words

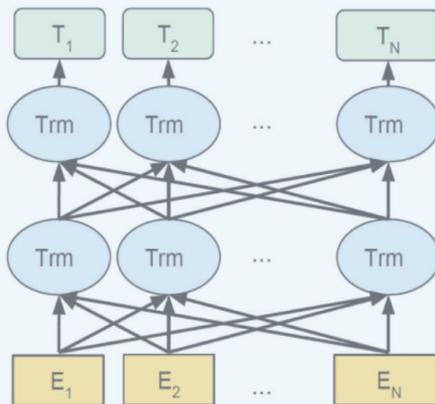
**PMC** 13.5B words

Weight Initialization



**BERT**  
from Devlin et al.

BioBERT Pre-training



Pre-trained BioBERT with  
biomedical domain corpora

# BioBERT (2019)

Type	Datasets	Metrics	SOTA	BERT	BioBERT v1.0			BioBERT v1.1
				(Wiki + Books)	(+ PubMed)	(+ PMC)	(+ PubMed + PMC)	(+ PubMed)
Disease	NCBI disease	P	<u>88.30</u>	84.12	86.76	86.16	<b>89.04</b>	88.22
		R	89.00	87.19	88.02	89.48	<u>89.69</u>	<b>91.25</b>
		F	88.60	85.63	87.38	87.79	<u>89.36</u>	<b>89.71</b>

## ClinicalBERT (2020)

**Question:** Clinical notes are even stranger than biomedical text. Can we do the same for clinical notes?

## ClinicalBERT (2020)

**Question:** Clinical notes are even stranger than biomedical text. Can we do the same for clinical notes?

**Answer:** Yes!

# ClinicalBERT (2020)

**Question:** Clinical notes are even stranger than biomedical text. Can we do the same for clinical notes?

**Answer:** Yes!

Model	MedNLI	i2b2 2006	i2b2 2010	i2b2 2012	i2b2 2014
BERT	77.6%	93.9	83.5	75.9	92.8
BioBERT	80.8%	<b>94.8</b>	86.5	78.9	<b>93.0</b>
Clinical BERT	80.8%	91.5	86.4	78.5	92.6
Discharge Summary BERT	80.6%	91.9	86.4	78.4	92.8
Bio+Clinical BERT	<b>82.7%</b>	94.7	87.2	<b>78.9</b>	92.5
Bio+Discharge Summary BERT	<b>82.7%</b>	94.8	<b>87.8</b>	78.9	92.7

## GPT-2 (2019)

- Same pretraining task: next word prediction

## GPT-2 (2019)

- Same pretraining task: next word prediction
- Pretrain on more data (40GB of text)

## GPT-2 (2019)

- Same pretraining task: next word prediction
- Pretrain on more data (40GB of text)
- Use bigger models

# GPT-2 (2019)

- Same pretraining task: next word prediction
- Pretrain on more data (40GB of text)
- Use bigger models

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	<b>21.8</b>
117M	<b>35.13</b>	45.99	<b>87.65</b>	<b>83.4</b>	<b>29.41</b>	65.85	1.16	1.17	37.50	75.20
345M	<b>15.60</b>	55.48	<b>92.35</b>	<b>87.1</b>	<b>22.76</b>	47.33	1.01	<b>1.06</b>	26.37	55.72
762M	<b>10.87</b>	<b>60.12</b>	<b>93.45</b>	<b>88.0</b>	<b>19.93</b>	<b>40.31</b>	<b>0.97</b>	<b>1.02</b>	22.05	44.575
1542M	<b>8.63</b>	<b>63.24</b>	<b>93.30</b>	<b>89.05</b>	<b>18.34</b>	<b>35.76</b>	<b>0.93</b>	<b>0.98</b>	<b>17.48</b>	42.16

# Summary + Takeaways thus Far

# Summary + Takeaways thus Far

1. **ELMO**: Pretraining contextualized word embeddings

## Summary + Takeaways thus Far

1. **ELMO**: Pretraining contextualized word embeddings
2. **BERT & GPT**: Pretraining a transformer improves performance

# Summary + Takeaways thus Far

1. **ELMO**: Pretraining contextualized word embeddings
2. **BERT & GPT**: Pretraining a transformer improves performance
3. **ClinicalBERT**: Pretraining on clinical text helps with clinical tasks

# Summary + Takeaways thus Far

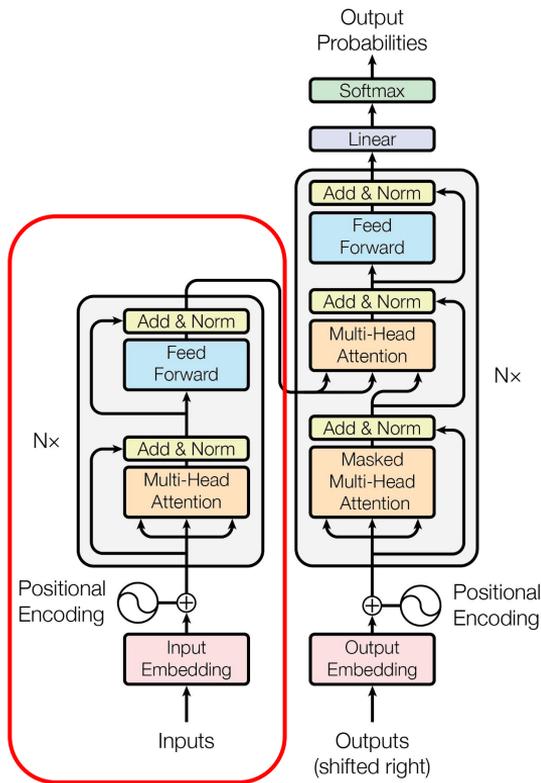
1. **ELMO**: Pretraining contextualized word embeddings
2. **BERT & GPT**: Pretraining a transformer improves performance
3. **ClinicalBERT**: Pretraining on clinical text helps with clinical tasks
4. **GPT-2**: Scaling these models helps a lot!

# Summary + Takeaways thus Far

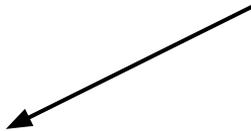
1. **ELMO**: Pretraining contextualized word embeddings
2. **BERT & GPT**: Pretraining a transformer improves performance
3. **ClinicalBERT**: Pretraining on clinical text helps with clinical tasks
4. **GPT-2**: Scaling these models helps a lot!

Questions so far?

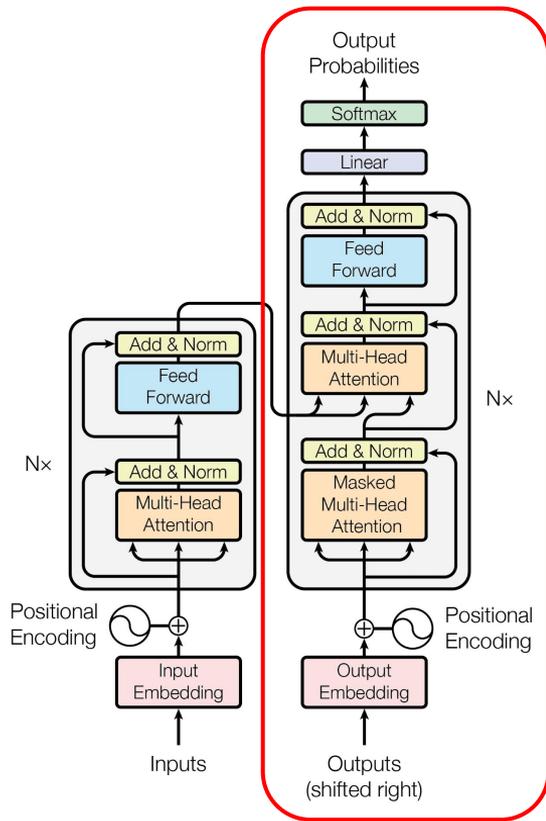
# Text-to-Text Transfer Transformer (T5) (2020)



BERT uses the first part (called encoder)



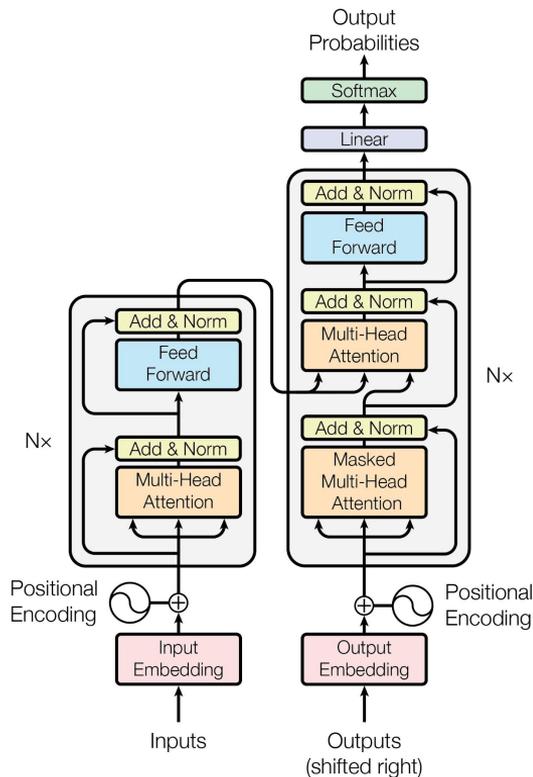
# Text-to-Text Transfer Transformer (T5) (2020)



GPT uses the second part (called decoder)

# Text-to-Text Transfer Transformer (T5) (2020)

T5 uses the entire architecture.



# T5 Performance

	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	<b>83.28</b>	<b>19.24</b>	<b>80.88</b>	<b>71.36</b>	<b>26.98</b>	<b>39.82</b>	<b>27.65</b>
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	<b>39.77</b>	24.04

# How Much Unique Text to Pretrain On?

Number of tokens	Repeats	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Full data set	0	<b>83.28</b>	<b>19.24</b>	<b>80.88</b>	<b>71.36</b>	<b>26.98</b>	<b>39.82</b>	<b>27.65</b>
$2^{29}$	64	<b>82.87</b>	<b>19.19</b>	<b>80.97</b>	<b>72.03</b>	<b>26.83</b>	<b>39.74</b>	<b>27.63</b>
$2^{27}$	256	82.62	<b>19.20</b>	79.78	69.97	<b>27.02</b>	<b>39.71</b>	27.33
$2^{25}$	1,024	79.55	18.57	76.27	64.76	26.38	39.56	26.80
$2^{23}$	4,096	76.34	18.33	70.92	59.29	26.37	38.84	25.81

# Scale Scale Scale

Model	GLUE Average	CoLA Matthew's	SST-2 Accuracy	MRPC F1	MRPC Accuracy	STS-B Pearson	STS-B Spearman
Previous best	89.4 <sup>a</sup>	69.2 <sup>b</sup>	97.1 <sup>a</sup>	<b>93.6<sup>b</sup></b>	<b>91.5<sup>b</sup></b>	92.7 <sup>b</sup>	92.3 <sup>b</sup>
T5-Small (80M)	77.4	41.0	91.8	89.7	86.6	85.6	85.0
T5-Base (220M)	82.7	51.1	95.2	90.7	87.5	89.4	88.6
T5-Large (770M)	86.4	61.2	96.3	92.4	89.9	89.9	89.2
T5-3B	88.5	67.1	97.4	92.5	90.0	90.6	89.8
T5-11B	<b>90.3</b>	<b>71.6</b>	<b>97.5</b>	92.8	90.4	<b>93.1</b>	<b>92.8</b>

Pretraining is expensive...

## Pretraining is expensive...

Back in 2018, the cost to train BERT was \$7000. What about now?

## Pretraining is expensive...

Back in 2018, the cost to train BERT was \$7000. What about now?

**Just \$38!!!**

Enter... GPT-3... (2020)

## Enter... GPT-3... (2020)

- 175B parameter model trained on 300B words

## Enter... GPT-3... (2020)

- 175B parameter model trained on 300B words
- This model is REALLY good at next word prediction

## Enter... GPT-3... (2020)

- 175B parameter model trained on 300B words
- This model is REALLY good at next word prediction
- Do we still need to train models?

Enter... GPT-3... (2020)

You can frame any NLP task as a next word completion task.

## Enter... GPT-3... (2020)

You can frame any NLP task as a next word completion task.

- **Mortality Prediction:** “Based on the above note, do you think that the patient will die?”

## Enter... GPT-3... (2020)

You can frame any NLP task as a next word completion task.

- **Mortality Prediction:** “Based on the above note, do you think that the patient will die?”
- **De-identification:** “List all of the names mentioned in the note:”

# Enter... GPT-3... (2020)

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP <sup>+</sup> 20]	<b>44.5</b>	<b>45.5</b>	<b>68.0</b>
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	<b>68.0</b>
GPT-3 Few-Shot	29.9	41.5	<b>71.2</b>

# GPT-3 Can Resolve Clinical Acronyms

# GPT-3 Can Resolve Clinical Acronyms

**Input:** Bob was sent to IR for thrombolysis. Post IR, ultrasound showed that... What does IR stand for?

**Output:** Interventional radiology

# GPT-3 Can Resolve Clinical Acronyms

Algorithm	CASI Acc.	CASI Macro F1	MIMIC Accuracy	MIMIC Macro F1
Random	0.31	0.23	0.32	0.28
Most Common	0.79	0.28	0.51	0.23
BERT (from Adams et al. (2020))	0.42	0.23	0.40	0.33
ELMo (from Adams et al. (2020))	0.55	0.38	0.58	0.53
LMC (from Adams et al. (2020))	0.71	0.51	0.74	<b>0.69</b>
<i>GPT-3 edit</i> + R: 0-shot	0.86	0.69	*	*
<i>GPT-3 edit</i> + R: 0-shot + distillation	<b>0.90</b>	<b>0.76</b>	<b>0.78</b>	<b>0.69</b>

# GPT-3 Can Resolve Clinical Acronyms

Algorithm	CASI Acc.	CASI Macro F1	MIMIC Accuracy	MIMIC Macro F1
Random	0.31	0.23	0.32	0.28
Most Common	0.79	0.28	0.51	0.23
BERT (from Adams et al. (2020))	0.42	0.23	0.40	0.33
ELMo (from Adams et al. (2020))	0.55	0.38	0.58	0.53
LMC (from Adams et al. (2020))	0.71	0.51	0.74	<b>0.69</b>
<i>GPT-3 edit</i> + R: 0-shot	0.86	0.69	*	*
<i>GPT-3 edit</i> + R: 0-shot + distillation	<b>0.90</b>	<b>0.76</b>	<b>0.78</b>	<b>0.69</b>

Agrawal et. al (2022) showed that GPT-3 had strong performance on a number of clinical extraction tasks.

# How You Ask Matters!

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

---

(Output) 8 **X**

# How You Ask Matters!

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

---

(Output) 8 ❌

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

---

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓*

# How You Ask Matters!

No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	<b>78.7</b>
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		AbraKadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	<b>17.7</b>

# How You Ask Matters!

Jeanne wants to ride the Ferris wheel, the roller coaster, and the bumper cars. The Ferris wheel costs 5 tickets, the roller coaster costs 4 tickets and the bumper cars cost 4 tickets. Jeanne has 5 tickets.

*Jeanne's neighbor rides 8 kilometers to the bus station every day.* How many more tickets should Jeanne buy?

**Standard Answer**

8

# How You Ask Matters!

*Solve the following math problem. Feel free to ignore irrelevant information in the given problems.*

Jeanne wants to ride the Ferris wheel, the roller coaster, and the bumper cars. The Ferris wheel costs 5 tickets, the roller coaster costs 4 tickets and the bumper cars cost 4 tickets. Jeanne has 5 tickets.

*Jeanne's neighbor rides 8 kilometers to the bus station every day.* How many more tickets should Jeanne buy?

# How You Ask Matters!

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to common consumer search queries about health.

**Question:** How do you treat skin redness?

**Complete Answer:** It depends on the cause of the skin redness. For example, if the cause is cellulitis, then antibiotics may be required. However, this might be inappropriate for other causes of redness such as eczema. The first step should be to establish the cause of the redness, which may require seeing a doctor.

# Tons of New Models

- Llama (65B)
- Chinchilla (70B)
- Gopher (270B)
- Megatron-Turing NLG (530B)
- Palm (540B)
- etc.

# Tons of New Models

- Llama (65B)
- Chinchilla (70B)
- Gopher (270B)
- Megatron-Turing NLG (530B)
- **Palm (540B)**
- etc.

Would need 1TB GPU RAM to naively load.



# Tons of New Models

- Llama (65B)
- **Chinchilla (70B)**
- Gopher (270B)
- Megatron-Turing NLG (530B)
- Palm (540B)
- etc.

## Chinchilla (2022)

Suppose you have a fixed training budget, but no restriction on model size and no restriction on amount of data available.

# Chinchilla (2022)

Suppose you have a fixed training budget, but no restriction on model size and no restriction on amount of data available.

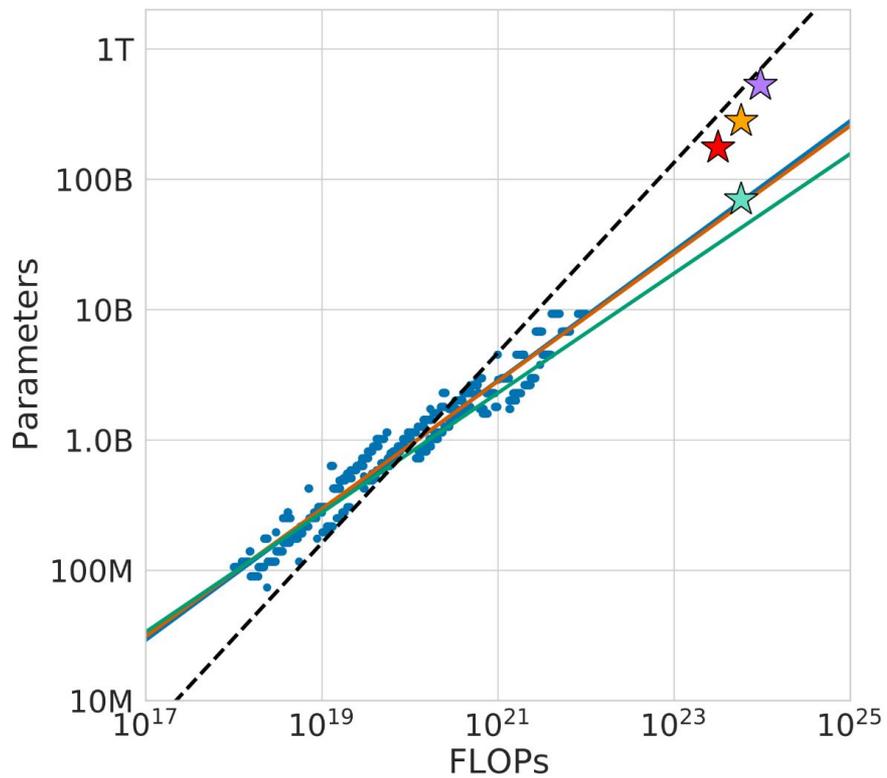
- 1. What size model should you choose?***

# Chinchilla (2022)

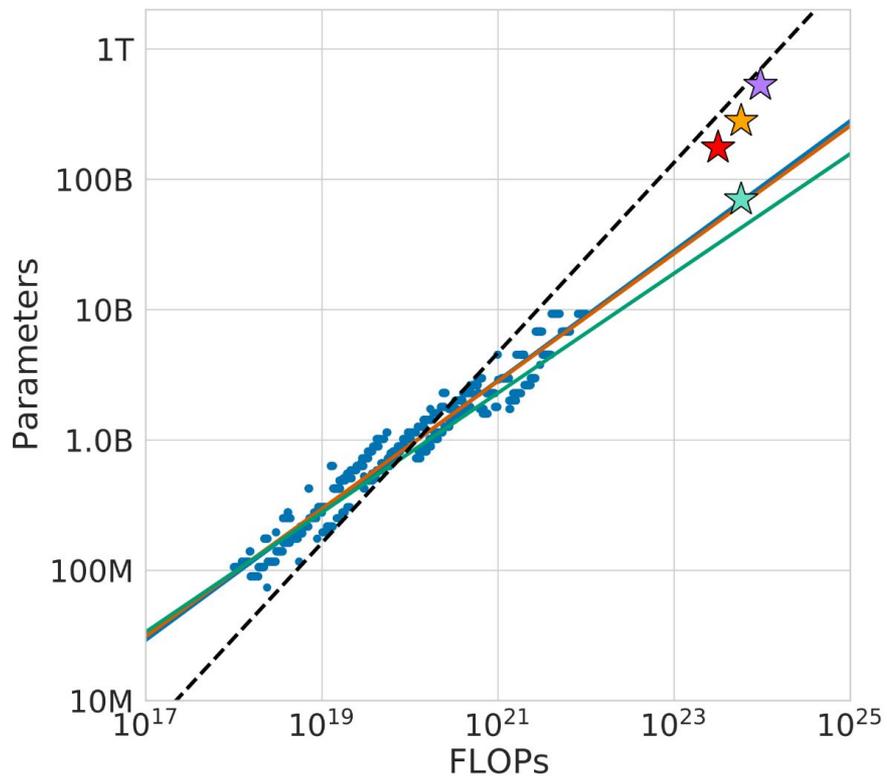
Suppose you have a fixed training budget, but no restriction on model size and no restriction on amount of data available.

- 1. What size model should you choose?***
- 2. How long should you pretrain it for?***

# Chinchilla (2022)

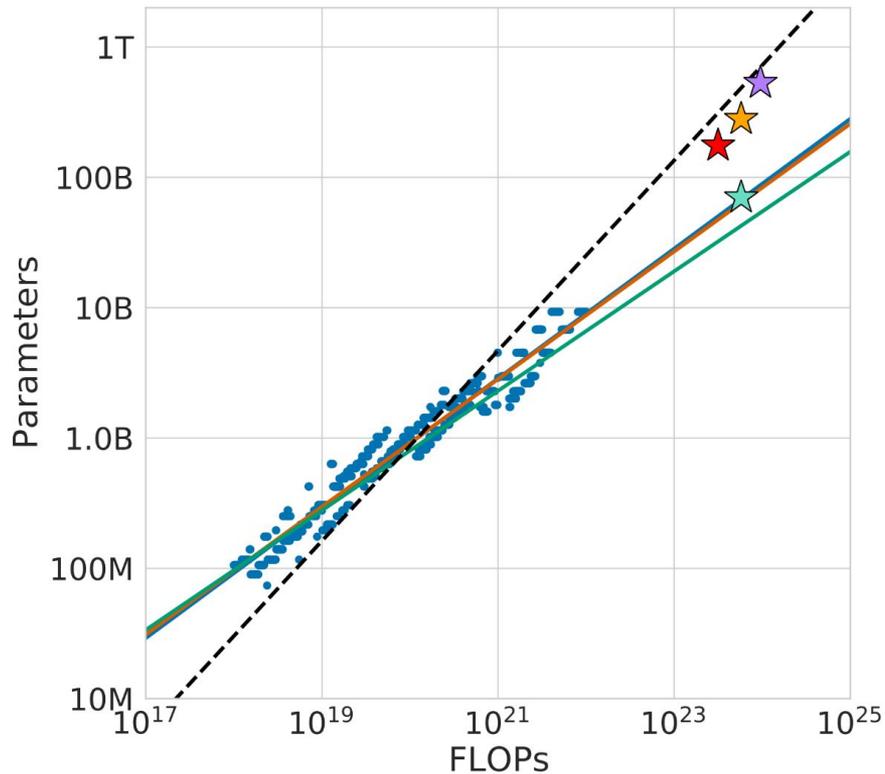


# Chinchilla (2022)



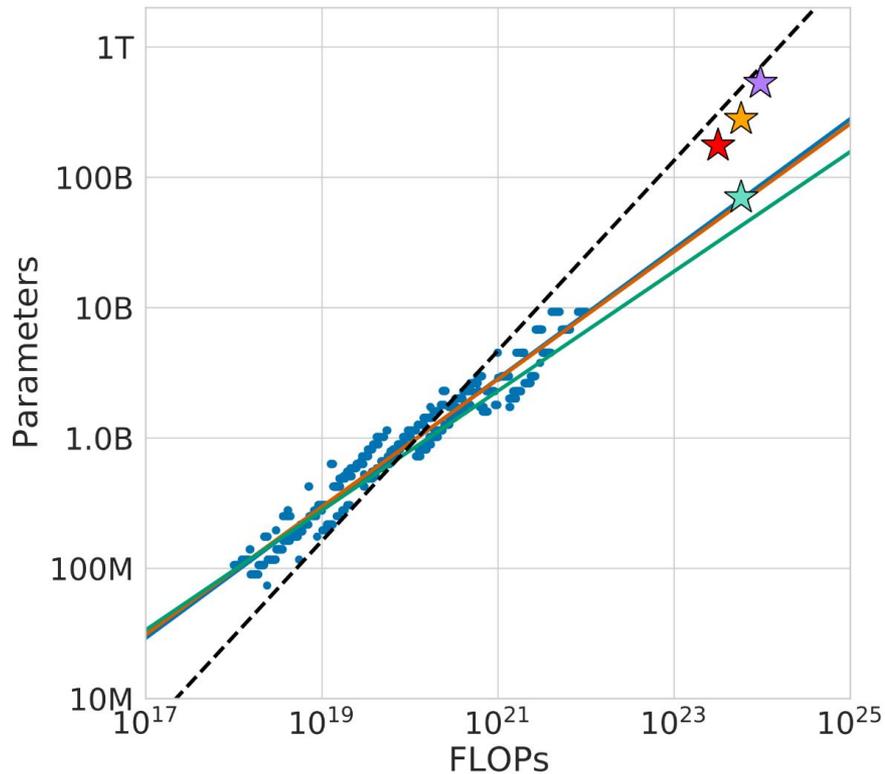
1. Be on the line!!!

# Chinchilla (2022)



1. Be on the line!!!
2. Above the line = you could've done something better

# Chinchilla (2022)



1. Be on the line!!!
2. Above the line = you could've done something better
3. Below the line = you should've picked a smaller model

# InstructGPT (2022)

Forcing your model to read isn't the only way to improve performance!

# InstructGPT (2022)

Forcing your model to read isn't the only way to improve performance!

BUSINESS • TECHNOLOGY

**Exclusive: OpenAI Used Kenyan Workers on  
Less Than \$2 Per Hour to Make ChatGPT Less  
Toxic**

# InstructGPT (2022)

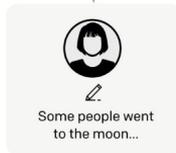
Step 1

**Collect demonstration data,  
and train a supervised policy.**

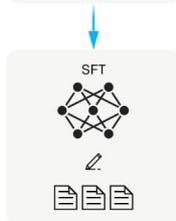
A prompt is  
sampled from our  
prompt dataset.



A labeler  
demonstrates the  
desired output  
behavior.



This data is used  
to fine-tune GPT-3  
with supervised  
learning.



# InstructGPT (2022)

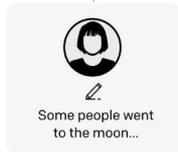
Step 1

**Collect demonstration data,  
and train a supervised policy.**

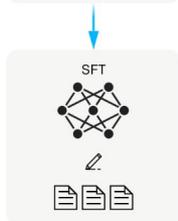
A prompt is  
sampled from our  
prompt dataset.



A labeler  
demonstrates the  
desired output  
behavior.



This data is used  
to fine-tune GPT-3  
with supervised  
learning.



Step 2

**Collect comparison data,  
and train a reward model.**

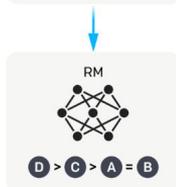
A prompt and  
several model  
outputs are  
sampled.



A labeler ranks  
the outputs from  
best to worst.



This data is used  
to train our  
reward model.



# InstructGPT (2022)

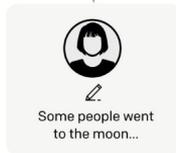
Step 1

**Collect demonstration data,  
and train a supervised policy.**

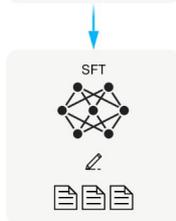
A prompt is  
sampled from our  
prompt dataset.



A labeler  
demonstrates the  
desired output  
behavior.



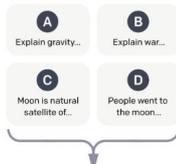
This data is used  
to fine-tune GPT-3  
with supervised  
learning.



Step 2

**Collect comparison data,  
and train a reward model.**

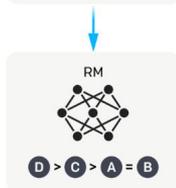
A prompt and  
several model  
outputs are  
sampled.



A labeler  
ranks the  
outputs from  
best to worst.



This data is used  
to train our  
reward model.



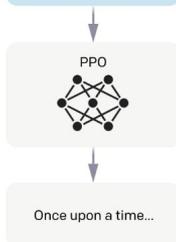
Step 3

**Optimize a policy against  
the reward model using  
reinforcement learning.**

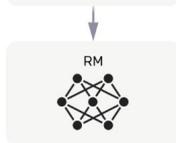
A new prompt  
is sampled from  
the dataset.



The policy  
generates an  
output.



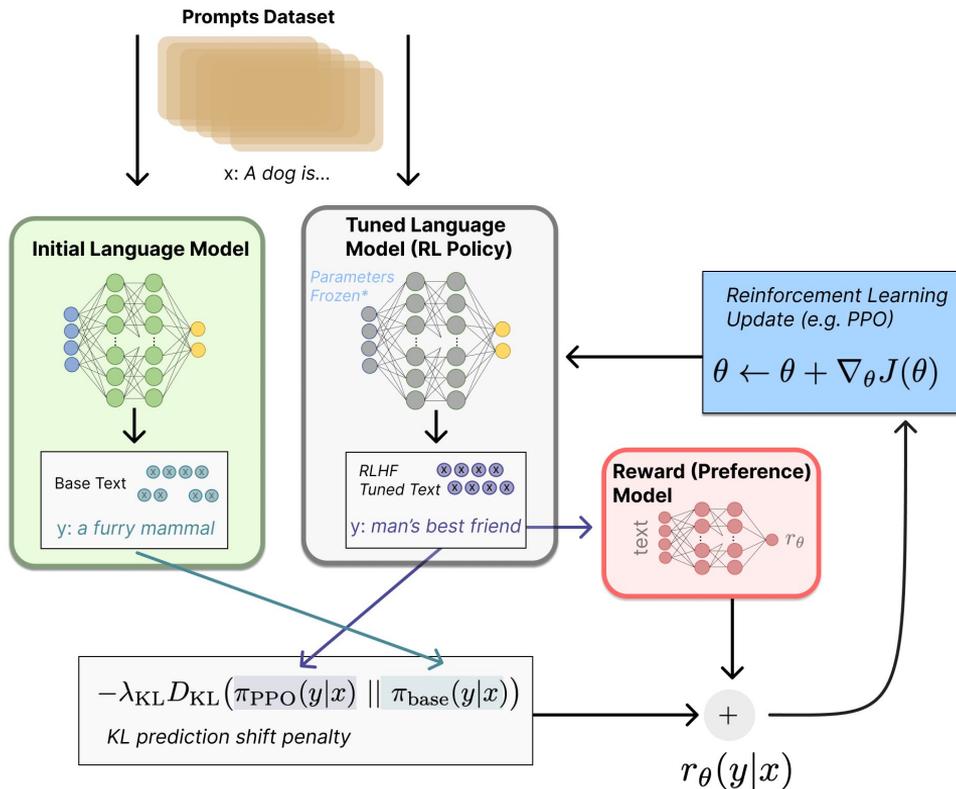
The reward model  
calculates a  
reward for  
the output.



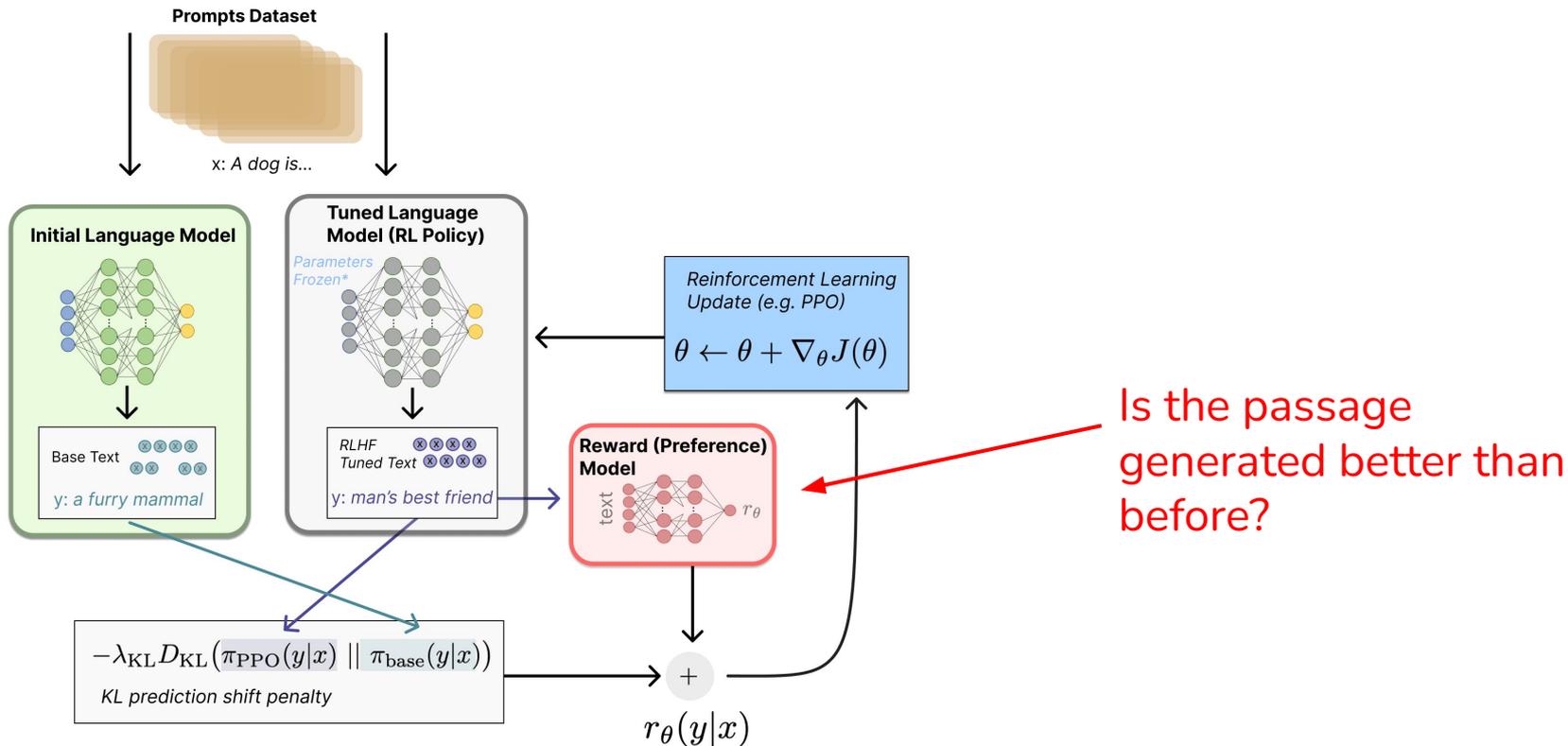
The reward is  
used to update  
the policy  
using PPO.



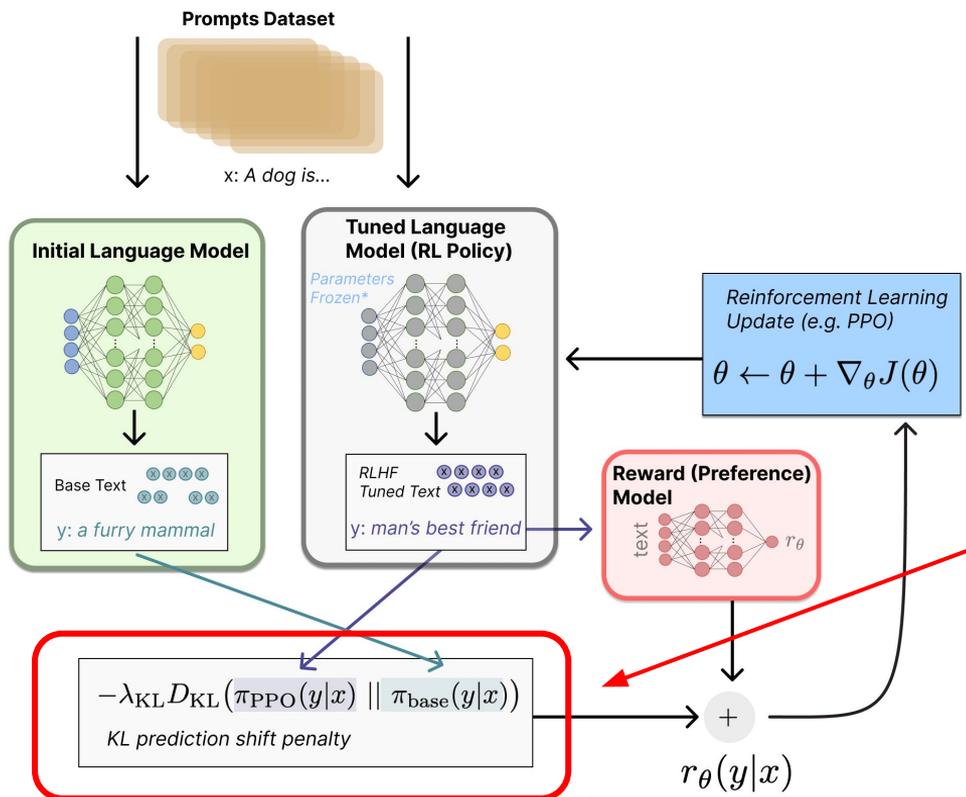
# InstructGPT (2022)



# InstructGPT (2022)

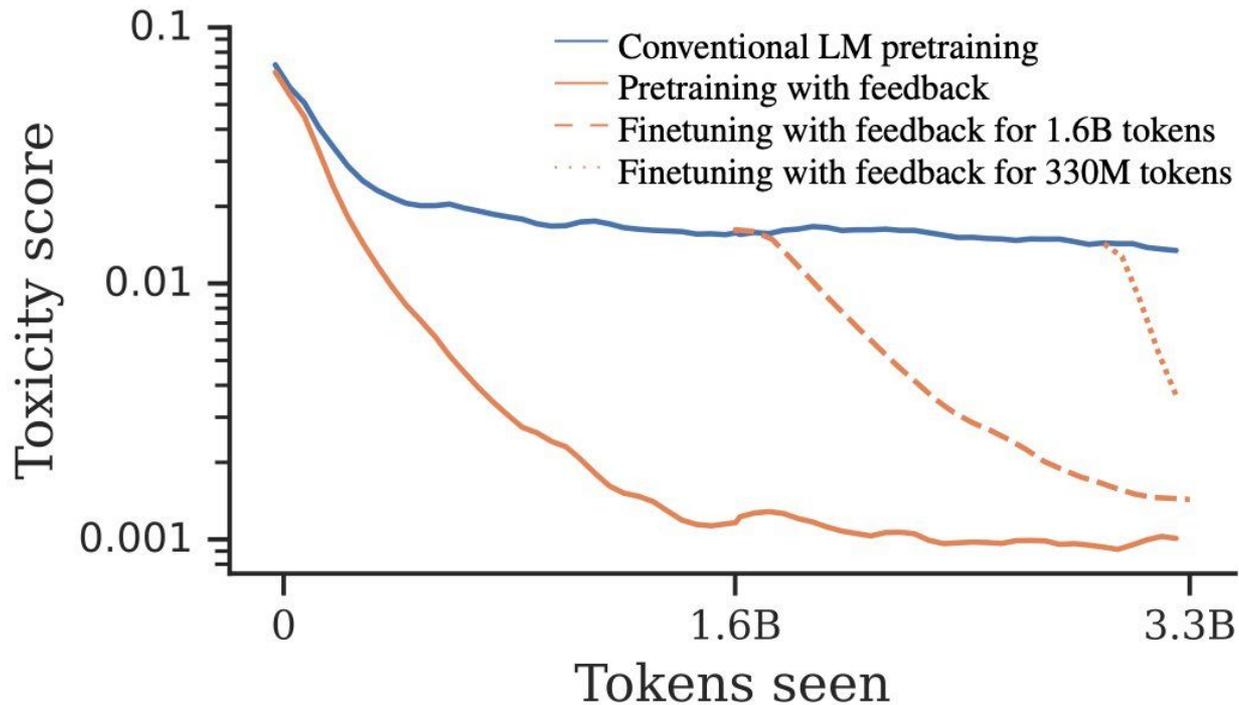


# InstructGPT (2022)



Is the model generating something similar to before?

# InstructGPT (2022)



What Does This Mean For Healthcare?

# What Does This Mean For Healthcare?

- Is it still working on these problems?

# What Does This Mean For Healthcare?

- Is it still working on these problems?
- Is it still worth creating custom models for clinical text?

# What Does This Mean For Healthcare?

- Is it still working on these problems?
- Is it still worth creating custom models for clinical text?

---

## Do We Still Need Clinical Language Models?

---

**Eric Lehman**<sup>1,2</sup>   **Evan Hernandez**<sup>1, 2</sup>   **Diwakar Mahajan**<sup>3</sup>   **Jonas Wulff**<sup>2</sup>  
**Micah J. Smith**<sup>2</sup>   **Zachary Ziegler**<sup>2</sup>   **Daniel Nadler**<sup>2</sup>   **Peter Szolovits**<sup>1</sup>  
**Alistair Johnson**<sup>4</sup>   **Emily Alsentzer**<sup>5,6</sup>

<sup>1</sup>MIT   <sup>2</sup>Xyla   <sup>3</sup>IBM Research   <sup>4</sup>The Hospital for Sick Children

<sup>5</sup>Brigham and Women's Hospital   <sup>6</sup>Harvard Medical School

{lehmer16, dez}@mit.edu

# Working With Clinical Text is Different

- Safety matters

# Working With Clinical Text is Different

- Safety matters
- Models trained on the general web are likely biased

# Working With Clinical Text is Different

- Safety matters
- Models trained on the general web are likely biased
- The text from a note is very different

# Working With Clinical Text is Different

- Safety matters
- Models trained on the general web are likely biased
- The text from a note is very different
  - Random abbreviations

# Working With Clinical Text is Different

- Safety matters
- Models trained on the general web are likely biased
- The text from a note is very different
  - Random abbreviations
  - Incorrect grammar

# Working With Clinical Text is Different

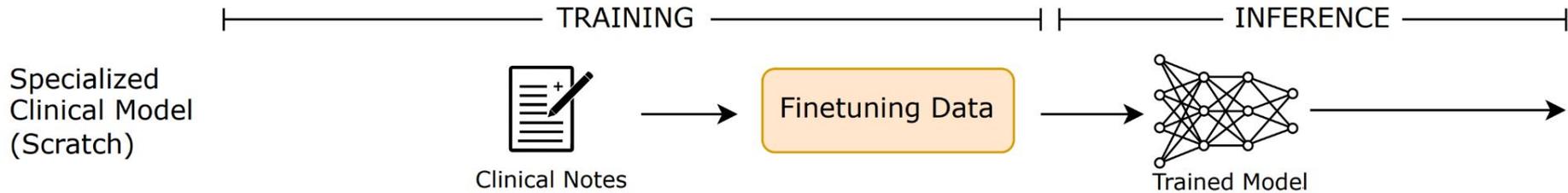
- Safety matters
- Models trained on the general web are likely biased
- The text from a note is very different
  - Random abbreviations
  - Incorrect grammar
  - Medical Terms

# Working With Clinical Text is Different

- Safety matters
- Models trained on the general web are likely biased
- The text from a note is very different
  - Random abbreviations
  - Incorrect grammar
  - Medical Terms
- You can't find this text online!!!

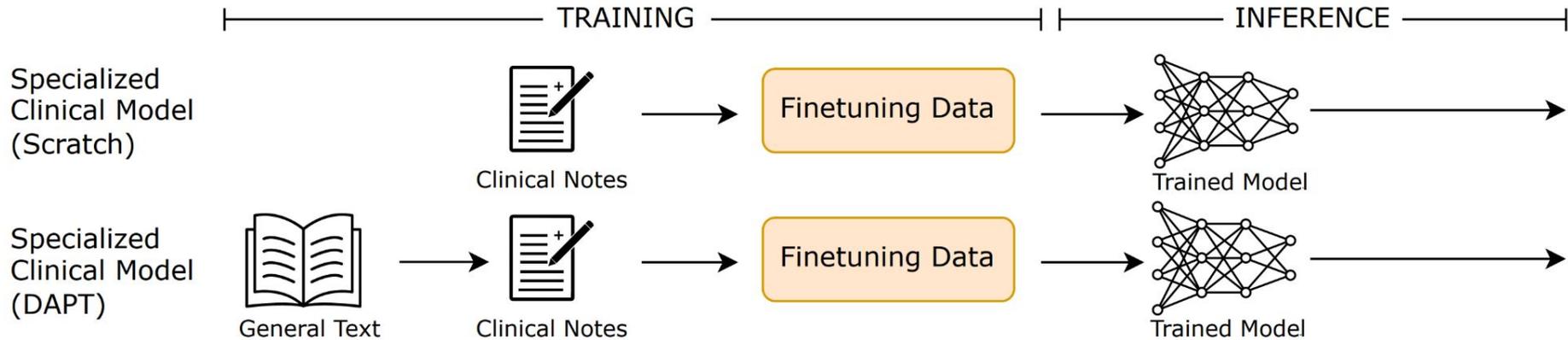
If You're a Hospital, What Do You Do?

# If You're a Hospital, What Do You Do?



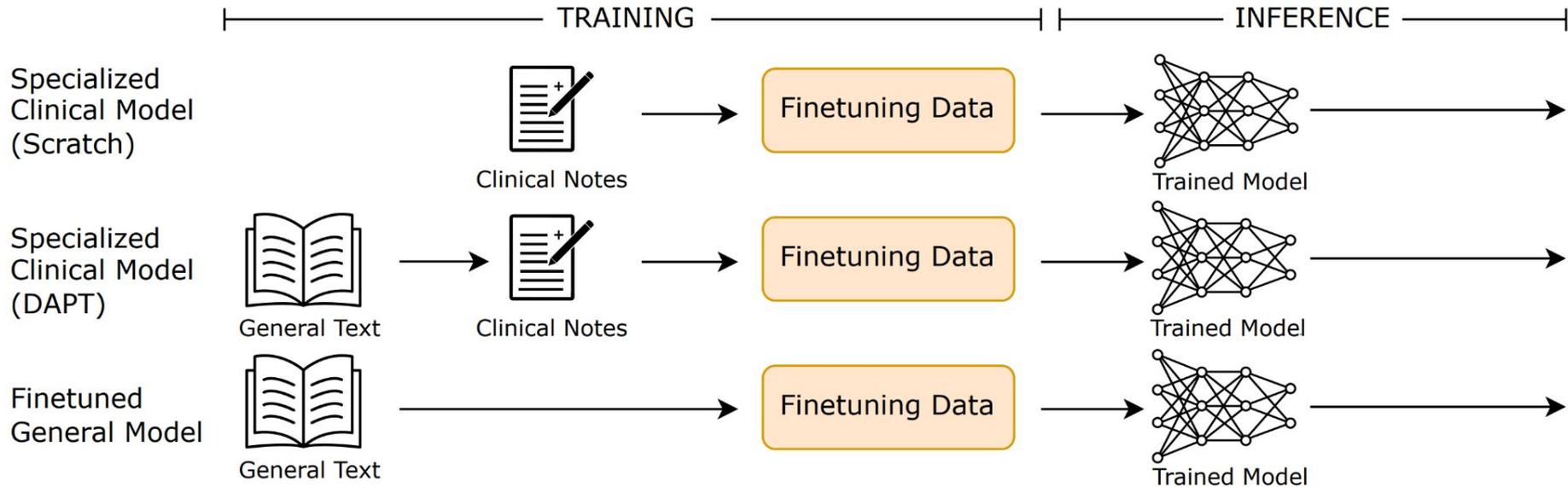
**Option 1:** Randomly initialize your model and train your own language model ON clinical notes from scratch

# If You're a Hospital, What Do You Do?



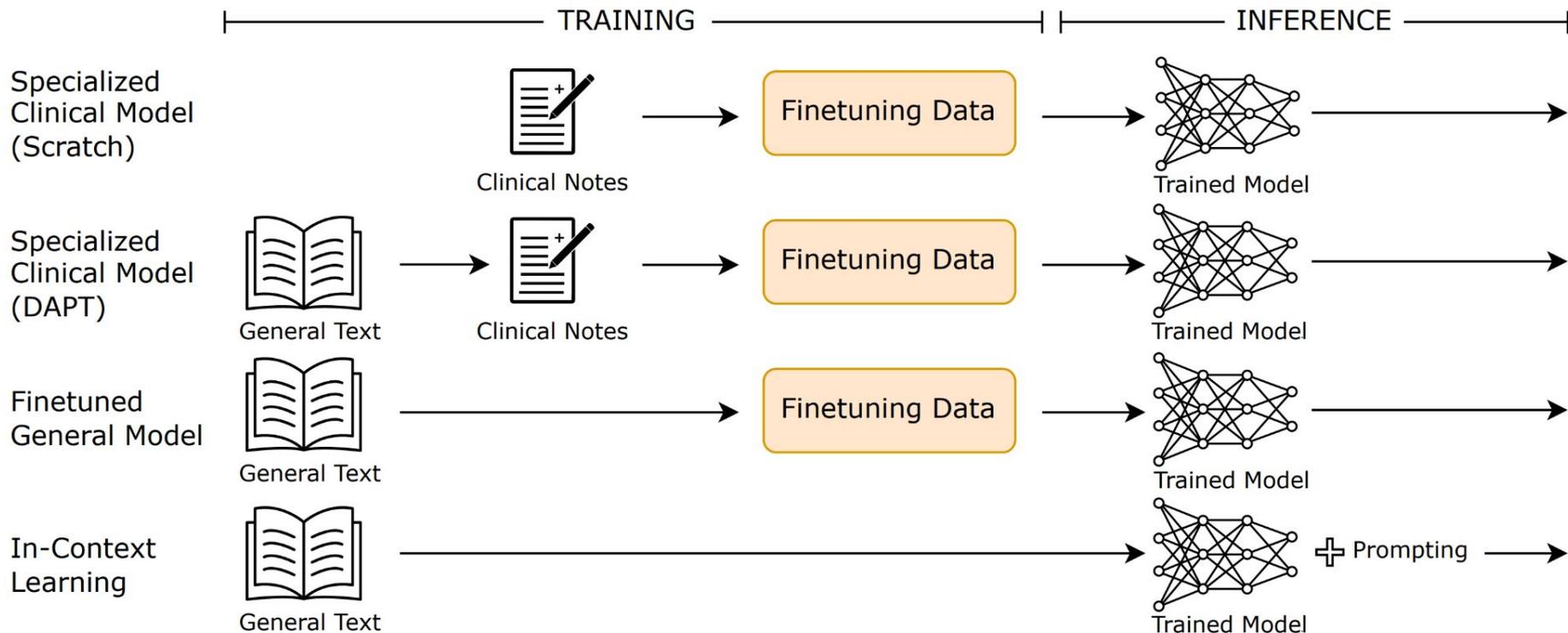
**Option 2:** Initialize your model from a model trained on the general web, and further train it on clinical notes.

# If You're a Hospital, What Do You Do?



**Option 3:** Download some model trained on the internet and finetune it.

# If You're a Hospital, What Do You Do?



Should You *Pretrain* on Clinical Notes?

# Should You *Pretrain* on Clinical Notes?

Size	Model	MedNLI	RadQA		CLIP	
		Acc.	EM	F1	Micro F1	Macro F1
220M	T5-Base	0.818	0.479	0.662	0.767	0.594
	Clinical-T5-Base-Ckpt	0.852	0.507	0.689	0.772	0.605
	Clinical-T5-Base	0.855	0.531	0.710	0.793	0.652

# Should You *Pretrain* on Clinical Notes?

Size	Model	MedNLI	RadQA		CLIP	
		Acc.	EM	F1	Micro F1	Macro F1
220M	T5-Base	0.818	0.479	0.662	0.767	0.594
	Clinical-T5-Base-Ckpt	0.852	0.507	0.689	0.772	0.605
	Clinical-T5-Base	0.855	0.531	0.710	0.793	0.652
770M	T5-Large	0.849	0.537	0.700	0.779	0.629
	Clinical-T5-Large	<b>0.872</b>	0.550	<b>0.745</b>	<b>0.800</b>	<b>0.663</b>

# Should You *Pretrain* on Clinical Notes?

Size	Model	MedNLI	RadQA		CLIP	
		Acc.	EM	F1	Micro F1	Macro F1
220M	T5-Base	0.818	0.479	0.662	0.767	0.594
	Clinical-T5-Base-Ckpt	0.852	0.507	0.689	0.772	0.605
	Clinical-T5-Base	<b>0.855</b>	0.531	<b>0.710</b>	0.793	0.652
770M	T5-Large	<b>0.849</b>	0.537	<b>0.700</b>	0.779	0.629
	Clinical-T5-Large	<b>0.872</b>	0.550	<b>0.745</b>	<b>0.800</b>	<b>0.663</b>

# Should You Train on Clinical Notes?

Size	Model	MedNLI	RadQA		CLIP	
		Acc.	EM	F1	Micro F1	Macro F1
220M	T5-Base	0.818	0.479	0.662	0.767	0.594
	Clinical-T5-Base-Ckpt	0.852	0.507	0.689	0.772	0.605
	Clinical-T5-Base	0.855	0.531	0.710	0.793	0.652
770M	T5-Large	0.849	0.537	0.700	0.779	0.629
	Clinical-T5-Large	<b>0.872</b>	0.550	<b>0.745</b>	<b>0.800</b>	<b>0.663</b>
3B	T5-XL	<b>0.869</b>	<b>0.568</b>	0.729	0.780	0.640

## Should You *Pretrain* on Clinical Notes?

1. Yes – It allows for smaller models, but similar performance!
2. You will eventually recover the costs of doing this extra training.

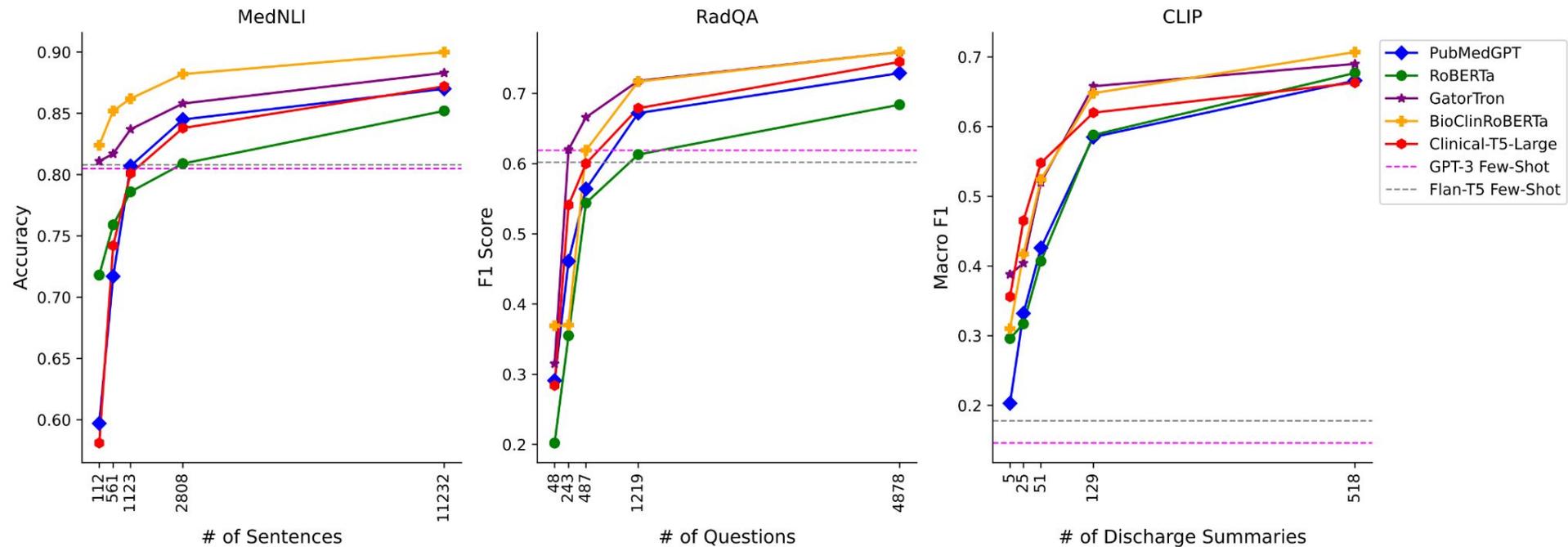
# What About Just Using GPT-3?

- GPT-3 doesn't require any training data – you just ask it!

## What About Just Using GPT-3?

- GPT-3 doesn't require any training data – you just ask it!
- For a fair comparison, we evaluate smaller models tailored to work with clinical text on a **SMALL** number of samples.

# What About Just Using GPT-3



But ChatGPT Passed the USMLE!!!

**But ChatGPT Passed the USMLE!!!**

Working with clinical notes != taking a test.

# Conclusions

- GPT-3 is very very good in constrained settings

# Conclusions

- GPT-3 is very very good in constrained settings
- However, if you're looking for the best performance, you should still customize your approach *for the* clinical space!

The Future

# **KS Health System Unveils Generative AI Partnership**

The University of Kansas Health System will use generative AI to summarize provider-patient conversations and generate clinical documentation in real-time.

<https://tinyurl.com/mlhc2023>

Sign-in sheet!