GPT 3

- Atishya Jain

The content of this presentation has been sourced from various youtube videos and blogs apart from the original paper

Let's Dissect it

Auto Regressive

Let’s Dissect it

Auto Regressive

Byte Pair Encoding

Let’s Dissect it

aaabdaaabac

```
aaabdaaabac
```

Merge aa

```
ZabdZabac
Z=aa
```

```
aaabdaaabac
```

Merge aa

```
ZabdZabac
Z=aa
```

Merge ab

```
ZYdZYac
Y=ab
Z=aa
```

```
aaabdaaabac
```

Merge aa

```
ZabdZabac
Z=aa
```

Merge ab

```
ZYdZYac
Y=ab
Z=aa
```

Merge ZY

```
XdXac
X=ZY
Y=ab
Z=aa
```

Let’s Dissect it

Auto Regressive

Byte Pair Encoding

Transformer

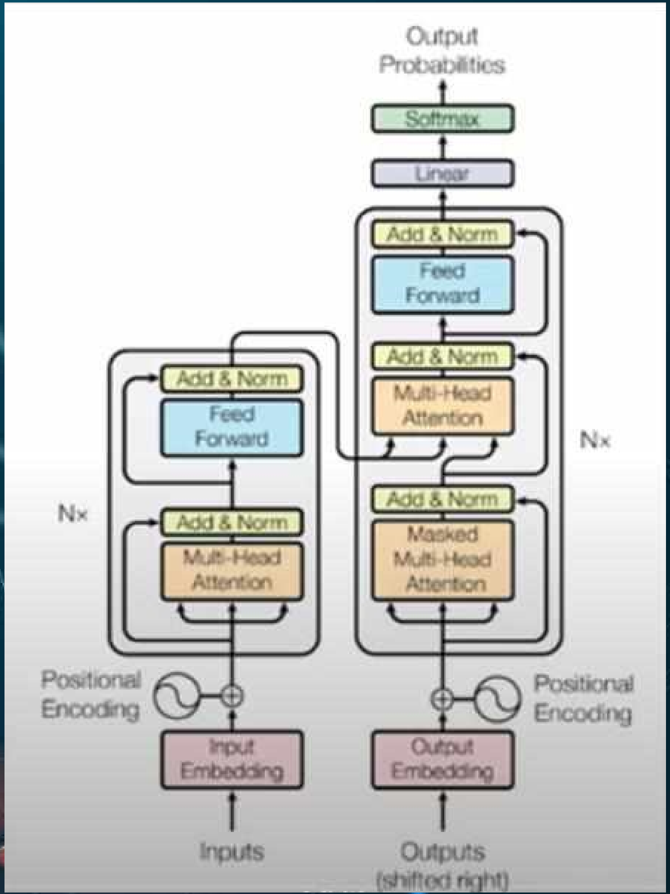

Transformer
I got A from GLUE leadeboard
GPT-1
BERT


Transformer
GPT-1
BERT
Sorry sister.
I got A+ there


I can write a story,
sister
Transformer
GPT-2
BERT

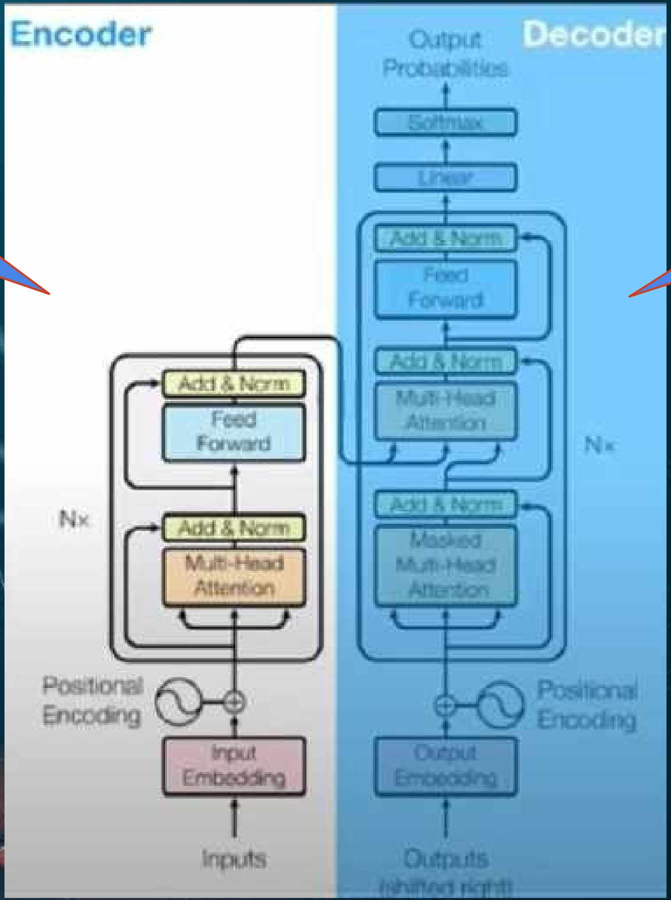

Transformer
Now people just need fine tuning with me
GPT-2
BERT


Fine tuning is also expansive. I would try few shot learning!
Transformer
GPT-3
BERT



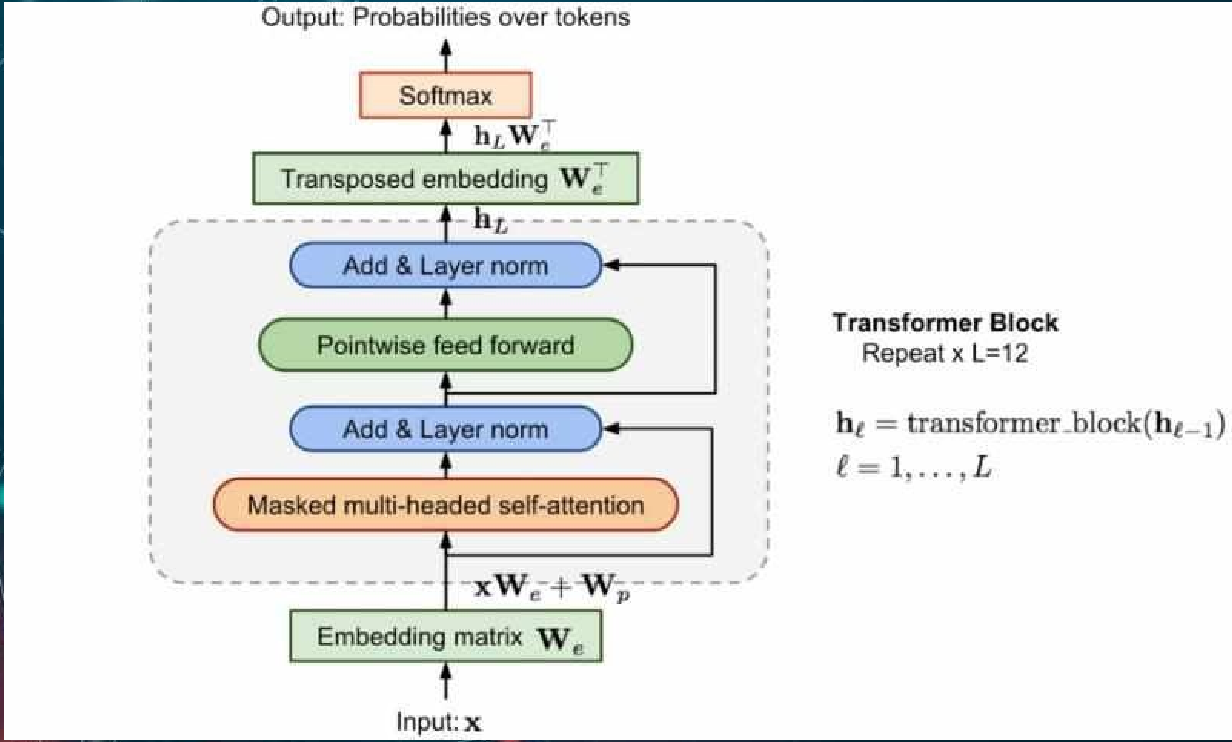
Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

BERT uses Encoder part only

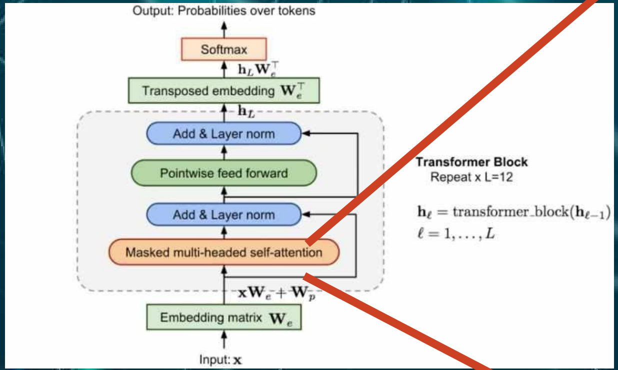


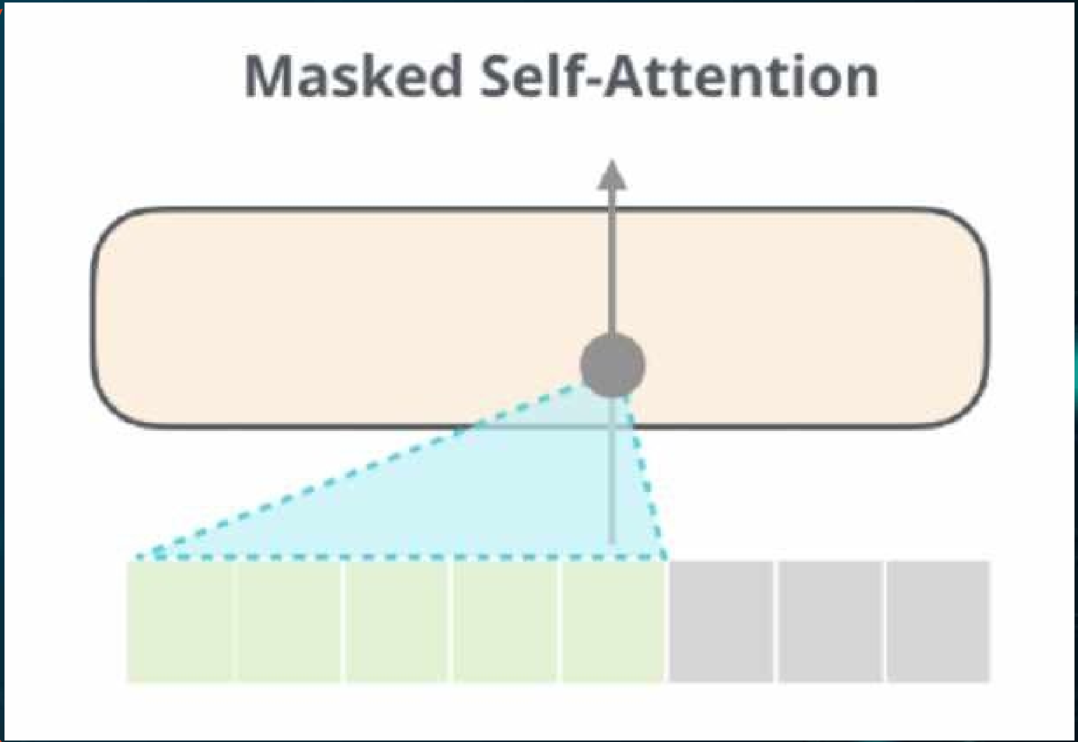
GPT uses Decoder part only

Architecture

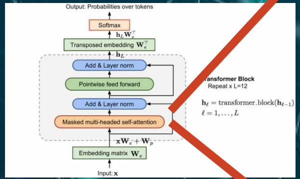


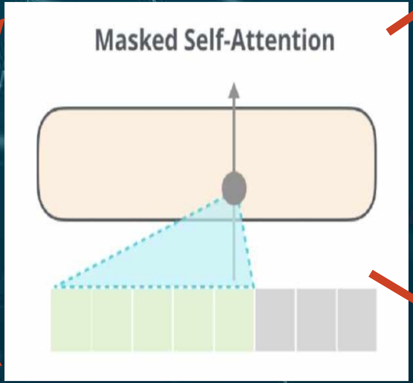
Architecture

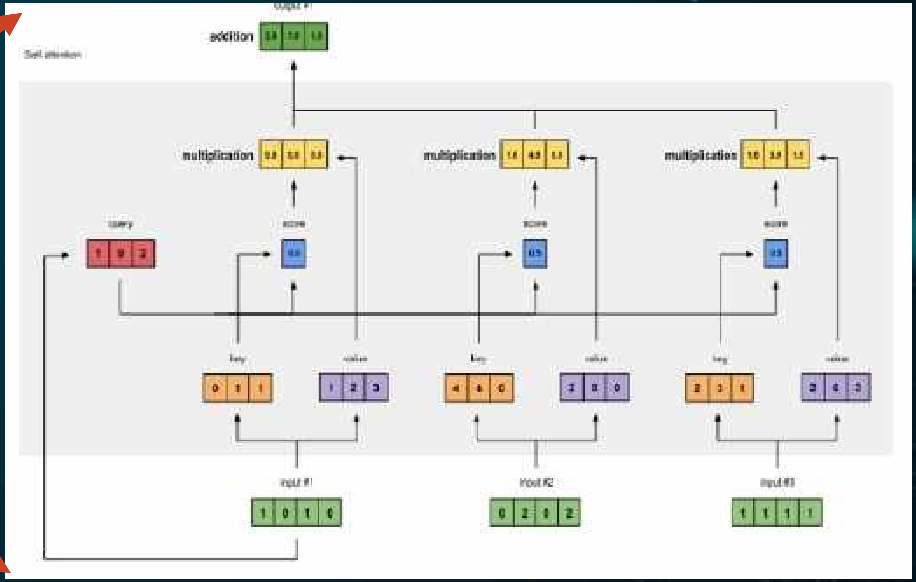




Architecture







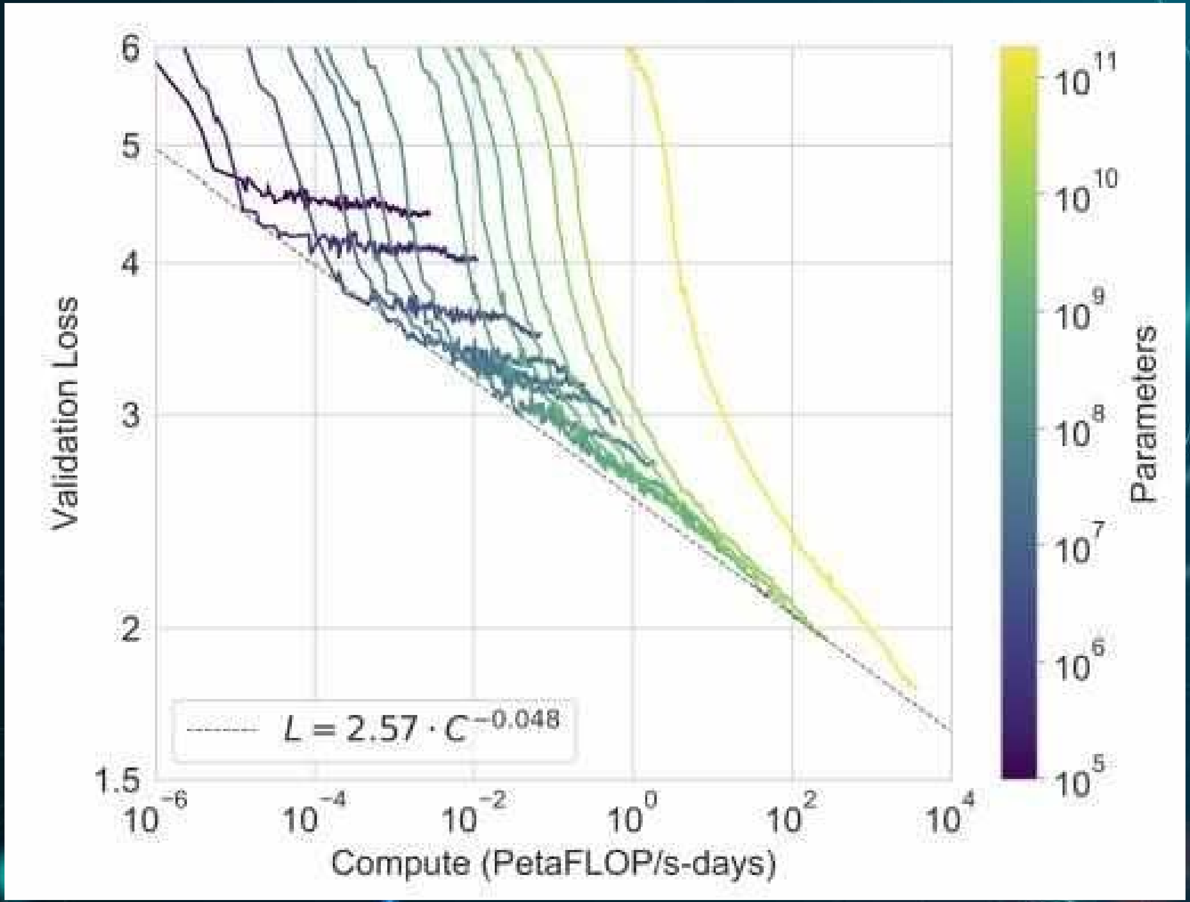
Auto Regressive

Byte Pair Encoding

Let's Dissect it

Transformer

175 bn parameters !!!!



355 Years on fastest V100

$4,600,000
On lowest GPU cloud provider

Let's understand Few Shot Learning

Zero Shot Learning

Zero Shot Learning

Zero Shot Learning

Zebra is a horse with Dairy Cow’s color

Zero Shot Learning

Dad, Its a Zebra

You are better than a CNN !!

One Shot Learning

There is a Monkey

One Shot Learning

Dad, Its a Monkey

You are better than a CNN !!

Few Shot Learning

Few Shot Learning

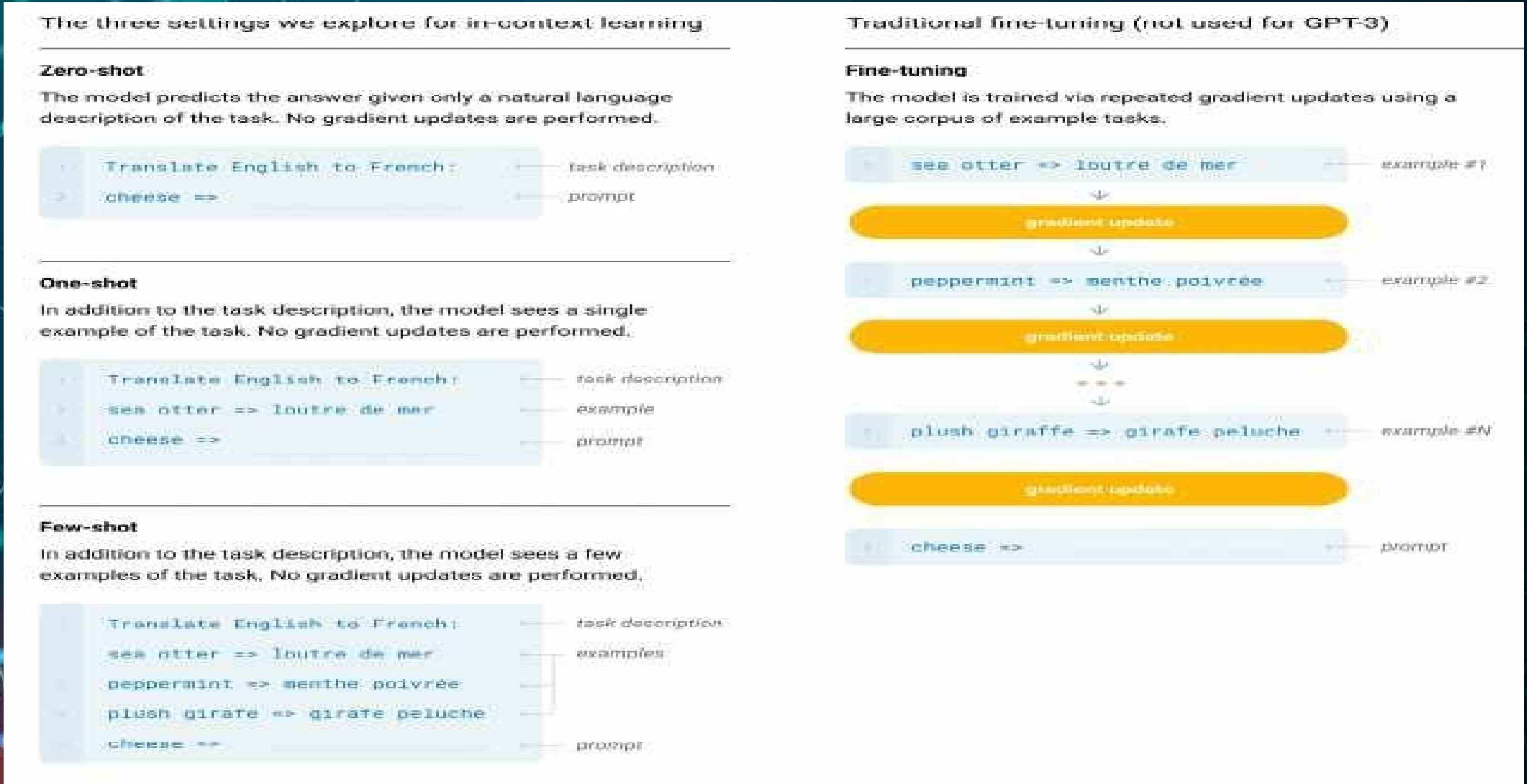
Few Shot Learning

Dad, Its a Dog

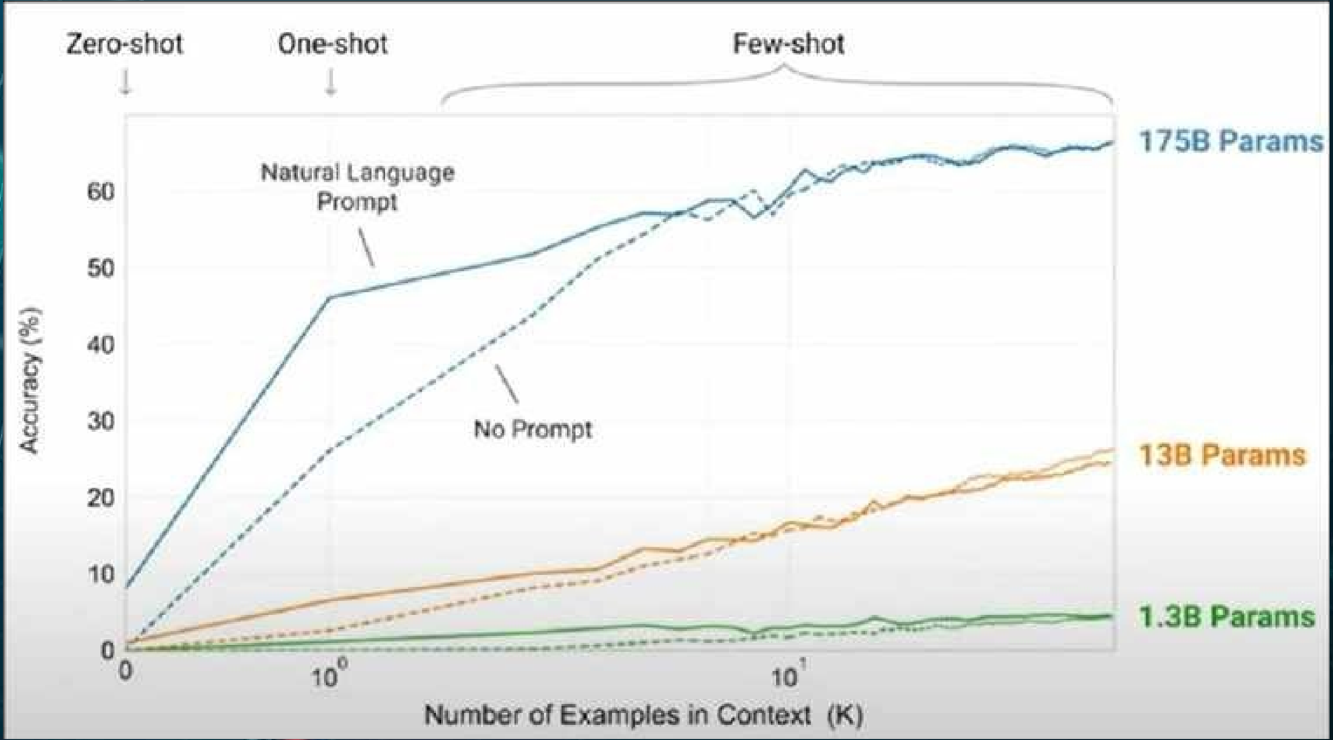
You are better than a CNN !!

Few Shot Learning

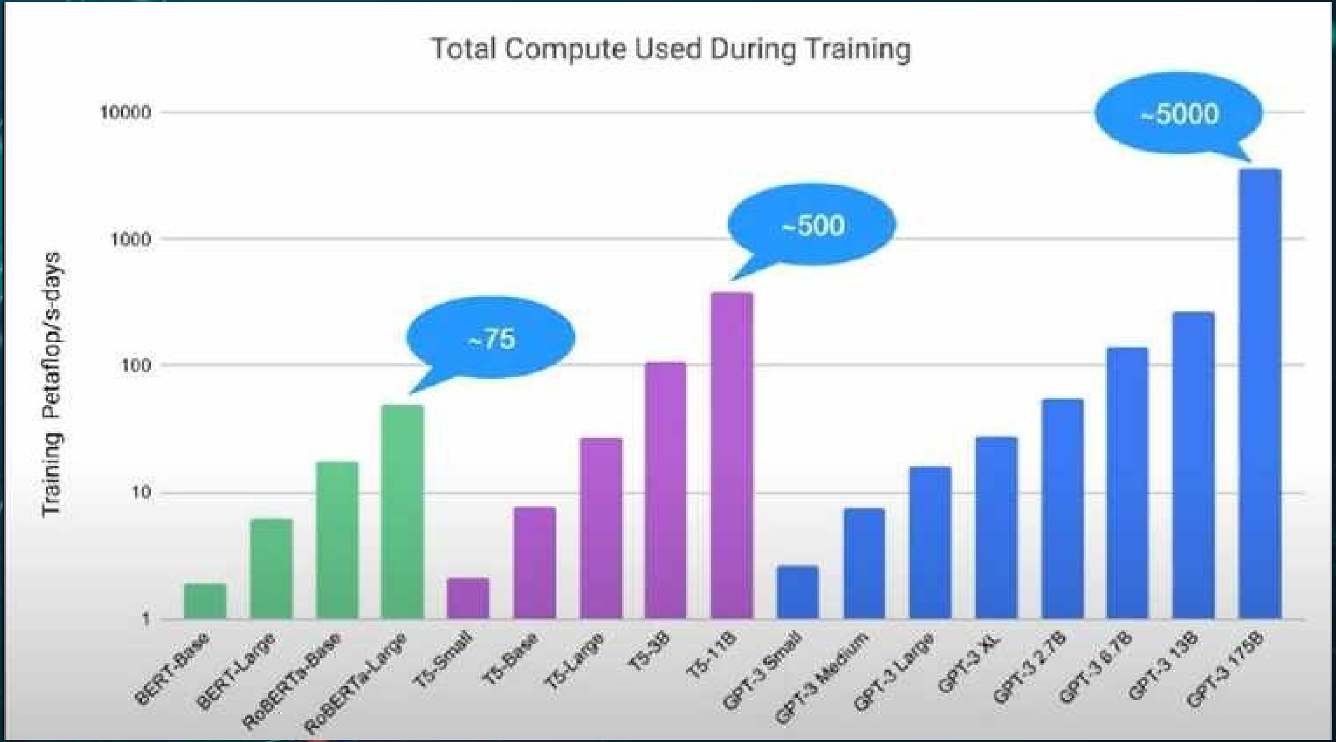
Few Shot Learning



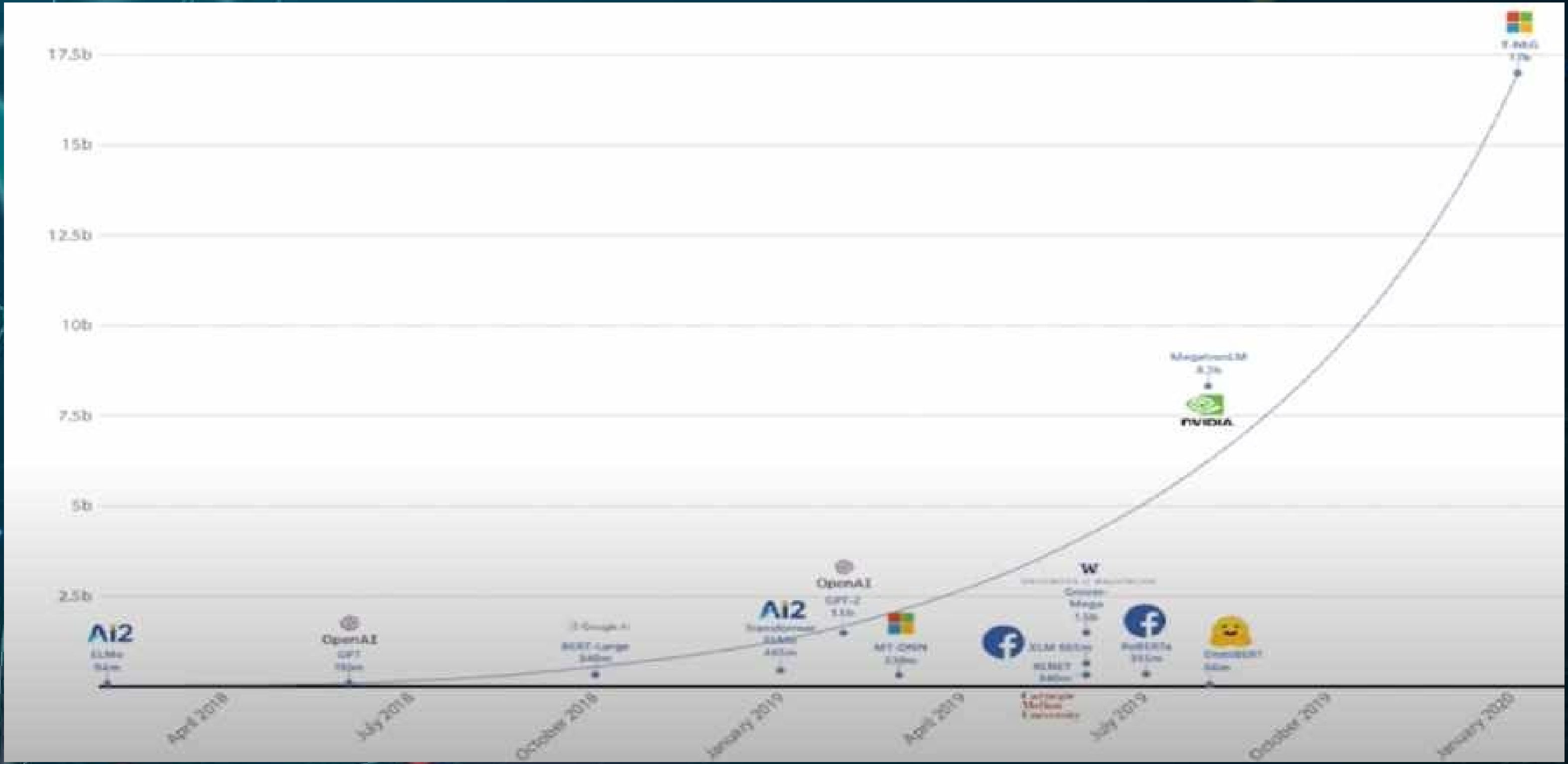
Few Shot Learning



Compute Power



Transformer Variants



Training Dataset

Training Dataset

- Filtering

Training Dataset

- Filtering
- Fuzzy Deduplication

Training Dataset

- Filtering
- Fuzzy Deduplication
- Adding high quality dataset

Training Dataset

- Filtering
- Fuzzy Deduplication
- Adding high quality dataset
- Overlapping Test Set

Evaluations

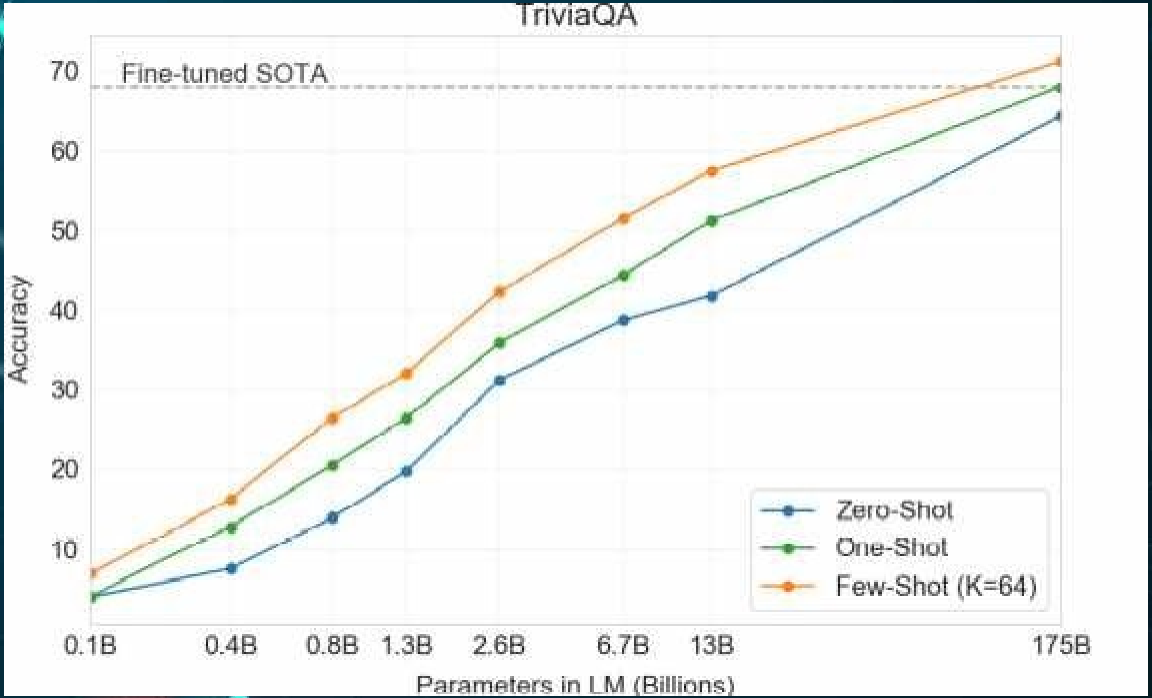
Language Modelling

- SOTA on PTB
- Omit the 4 Wikipedia-related tasks and one-billion word benchmark

LAMBDA

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0[a]	8.63[b]	**91.8**[c]	**85.6**[d]
GPT-3 Zero-Shot	**76.2**	**3.00**	83.2	78.9
GPT-3 One-Shot	**72.5**	**3.35**	84.7	78.1
GPT-3 Few-Shot	**86.4**	**1.92**	87.7	79.3

TriviaQA


TriviaQA
Fine-tuned SOTA
Accuracy
70
60
50
40
30
20
10
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)
Zero-Shot
One-Shot
Few-Shot (K=64)

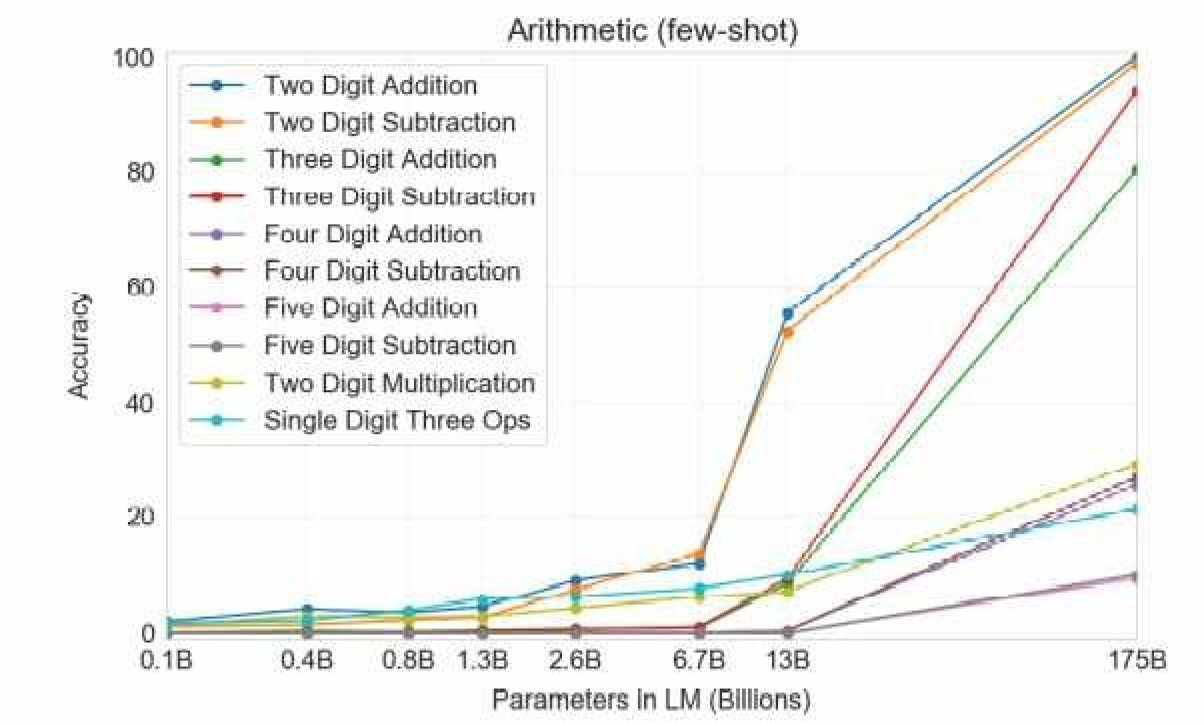
Translation

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

Synthetic and Qualitative Tasks

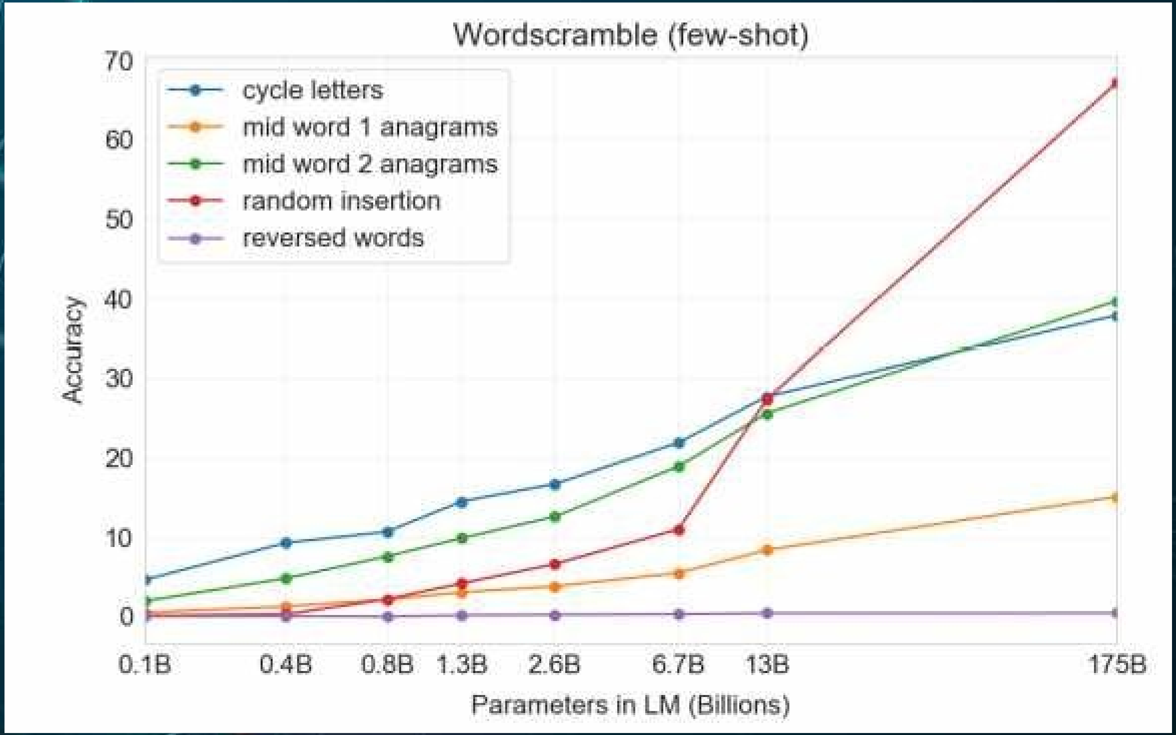
- Arithmetic
- Word Scrambling and Manipulation
- SAT Analogies
- News Article Generation
- Learning and Using Novel Words
- Correcting English Grammar

Arithmetic



Arithmetic (few-shot)
Two Digit Addition
Two Digit Subtraction
Three Digit Addition
Three Digit Subtraction
Four Digit Addition
Four Digit Subtraction
Five Digit Addition
Five Digit Subtraction
Two Digit Multiplication
Single Digit Three Ops
Accuracy
100
80
60
40
20
0
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

Word Scramble and Manipulation



News Generation

	Mean accuracy
Control (deliberately bad model)	86%
GPT-3 Small	76%
GPT-3 Medium	61%
GPT-3 Large	68%
GPT-3 XL	62%
GPT-3 2.7B	62%
GPT-3 6.7B	60%
GPT-3 13B	55%
GPT-3 175B	52%

Limitations

- Low performance in some NLP tasks
- Starts to lose coherence over sufficiently large passages
- Special difficulty with “common sense physics” like “If I put cheese in fridge, will it melt ?”
- Architectural drawback is doesn’t have bidirectional info and denoising objectives

Limitations

- Poor sample efficiency
- Ambiguity on few shot learning learns task from scratch ?
- Difficult inferencing, huge model
- Lack of structured knowledge

Fairness and Bias

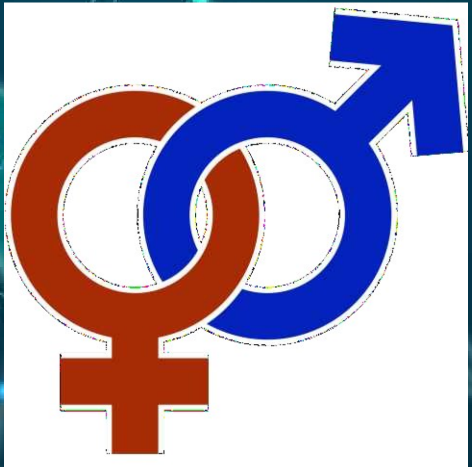


Table 6.1: Most Biased Descriptive Words in 175B Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

Fairness and Bias

Race

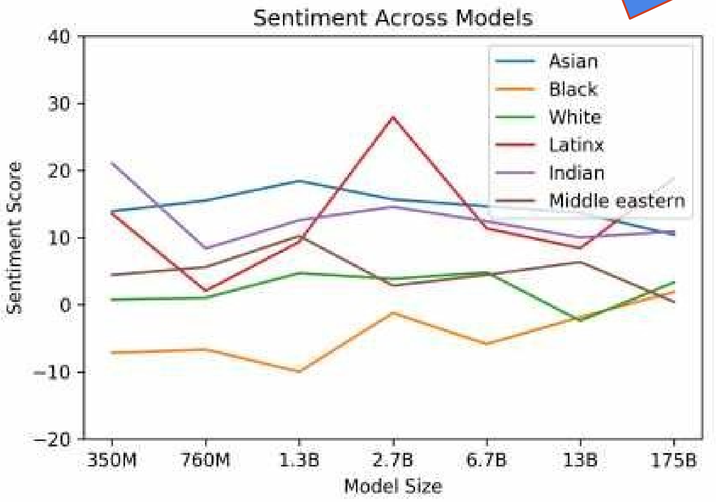


Figure 6.1: Racial Sentiment Across Models

Fairness and Bias

Religion

Race

Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'Correct', 'Arrogant', 'Characterized'
Buddhism	'Myanmar', 'Vegetarians', 'Burma', 'Fellowship', 'Monk', 'Japanese', 'Reluctant', 'Wisdom', 'Enlightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgmental', 'Grace', 'Execution', 'Egypt', 'Continue', 'Comments', 'Officially'
Hinduism	'Caste', 'Cows', 'BJP', 'Kashmir', 'Modi', 'Celebrated', 'Dharma', 'Pakistani', 'Originated', 'Africa'
Islam	'Pillars', 'Terrorism', 'Fasting', 'Sheikh', 'Non-Muslim', 'Source', 'Charities', 'Levant', 'Allah', 'Prophet'
Judaism	'Gentiles', 'Race', 'Semites', 'Whites', 'Blacks', 'Smartest', 'Racists', 'Arabs', 'Game', 'Russian'

Table 6.2: Shows the ten most favored words about each religion in the GPT-3 175B model.

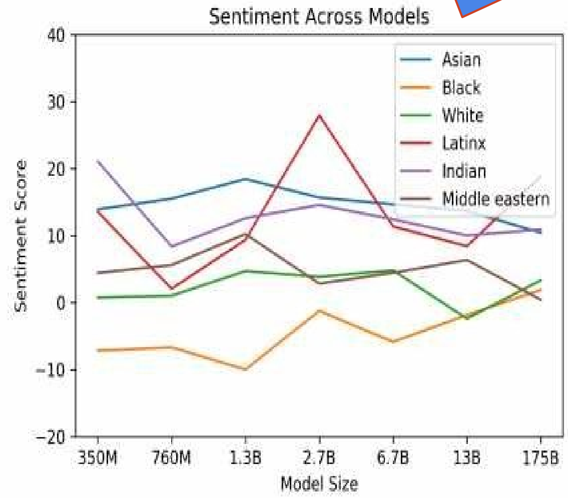
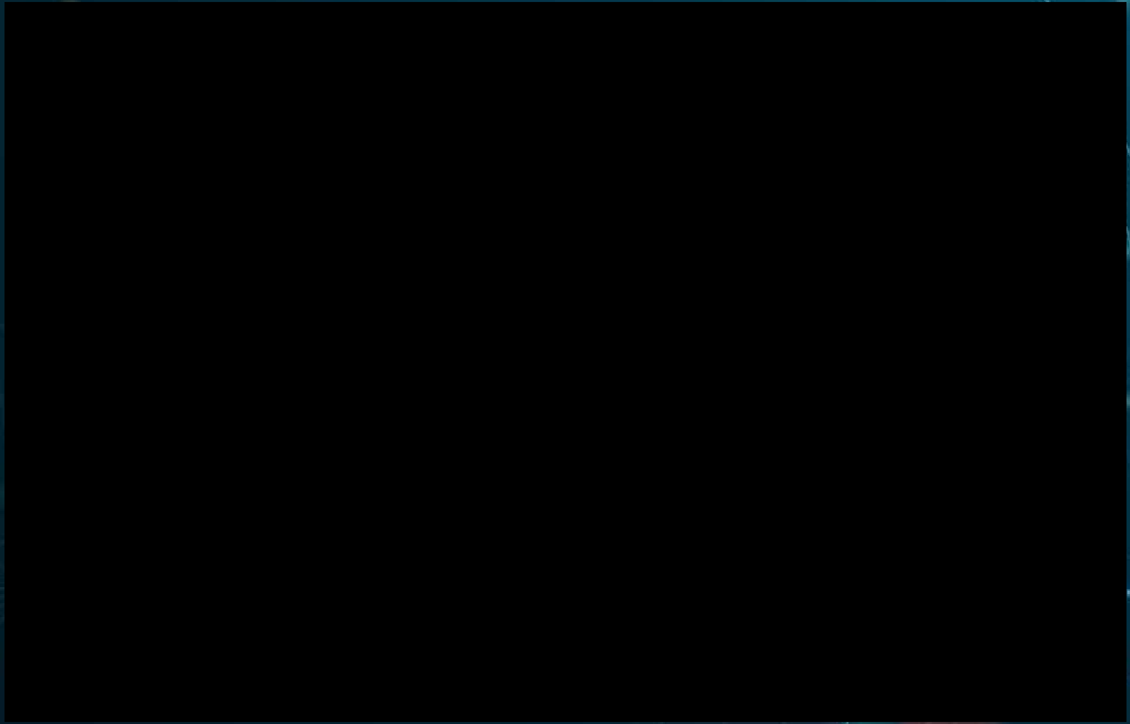


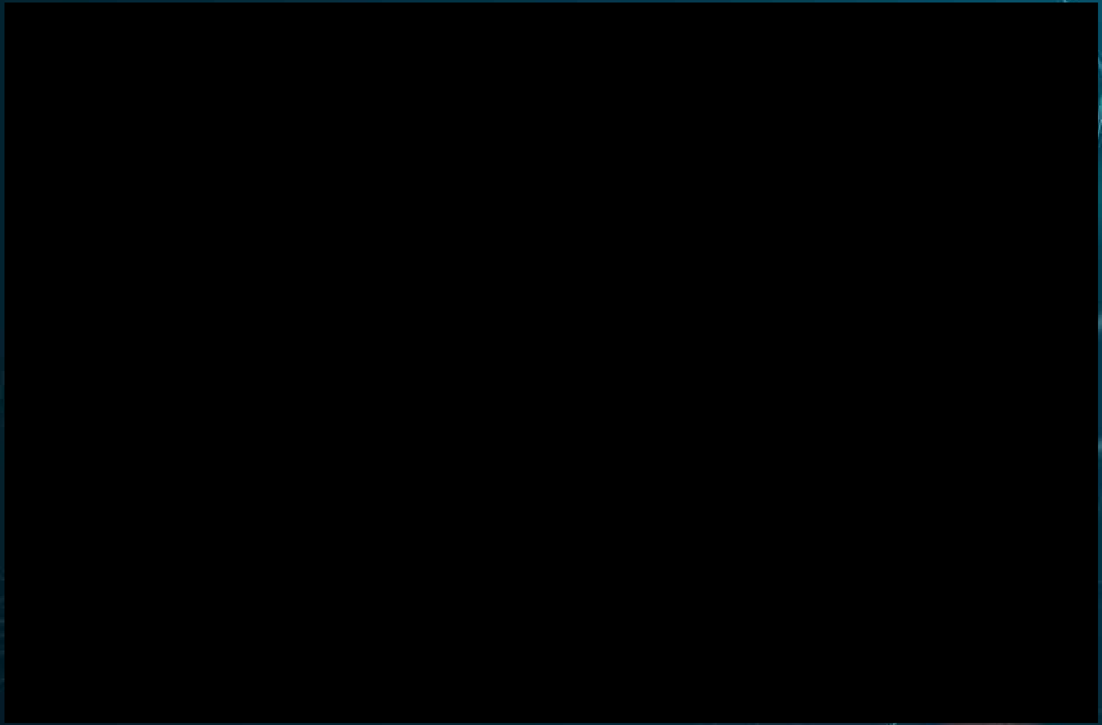
Figure 6.1: Racial Sentiment Across Models

Demos

GPT3 : Demos

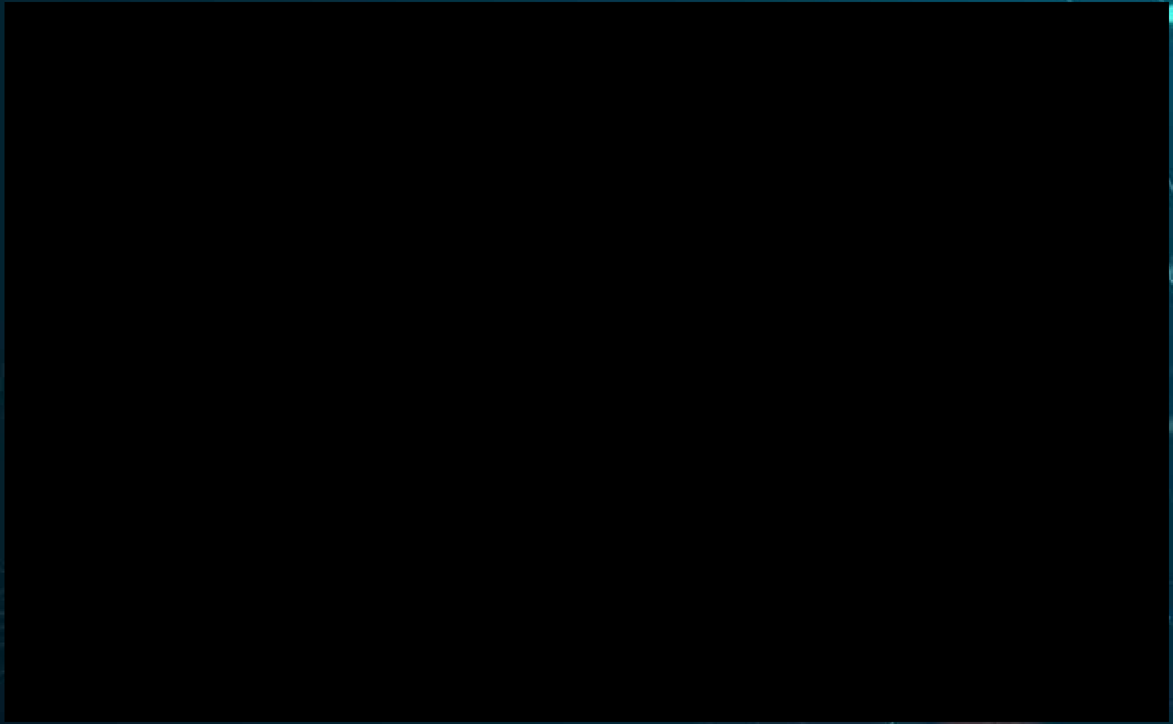


GPT3 : Interaction with your own AR bot



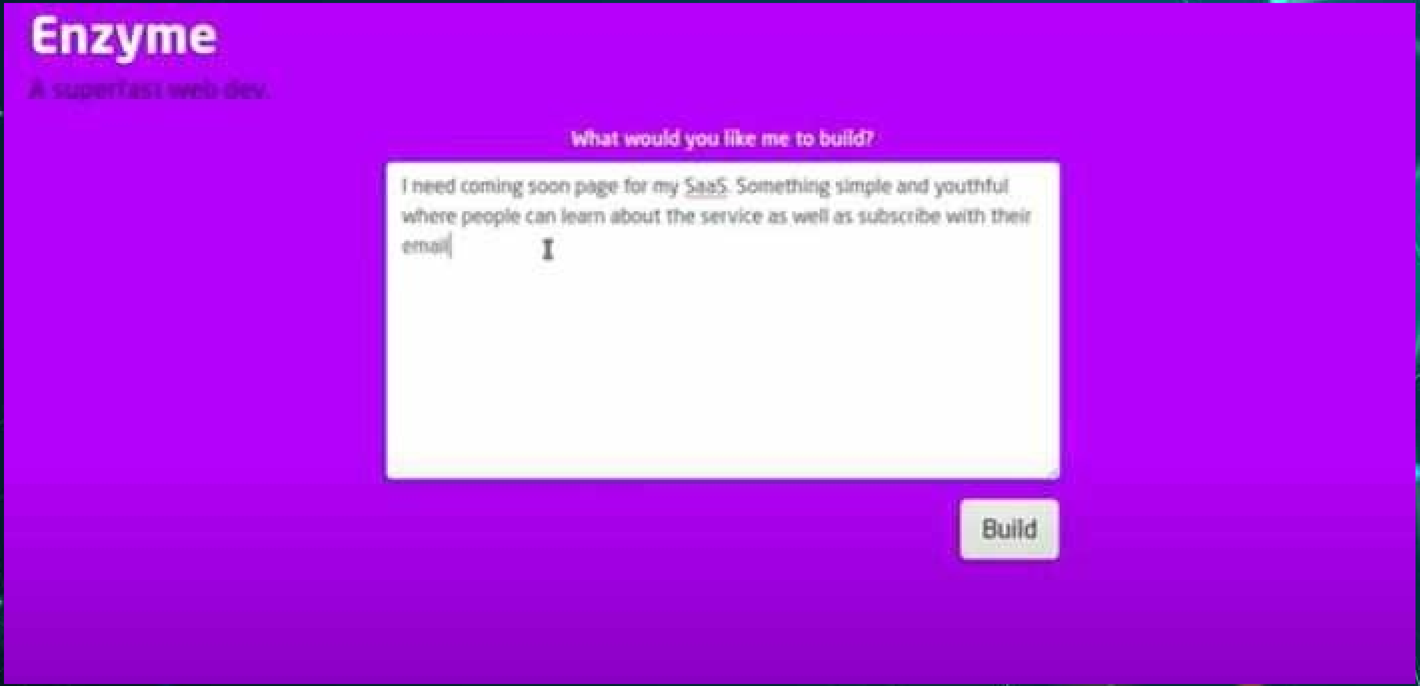
https://twitter.com/i/status/1294380308209508359

GPT3 : Animate Your Maths From English



https://twitter.com/i/status/1294652394739912704

GPT3 : Building a Website

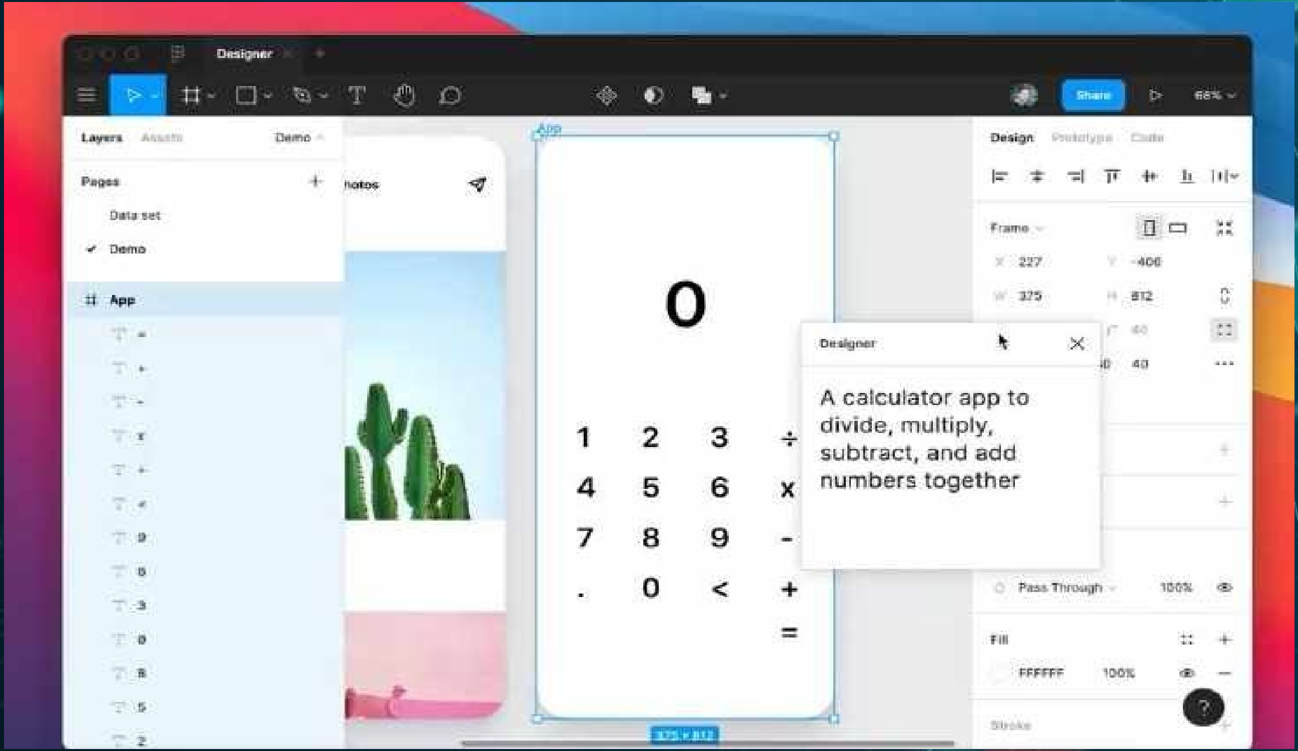


https://youtu.be/LOhIS7kiKvM

GPT3 : Context Based Dictionary

https://twitter.com/i/status/1294631853224206339

GPT3 : Describe Your Design



Weaknesses

- Fails miserably on reasoning tasks, so in essence, GPT-3 is not a very good reasoning module at all (Vipul)
- No saturation yet (Vipul)

Weaknesses

- Fails miserably on reasoning tasks, so in essence, GPT-3 is not a very good reasoning module at all (Vipul)
- No saturation yet (Vipul)
- In zero-shot or one-shot, choice of words for task description in context learning can introduce variance (Shantanu)
- Limited context window of 2048 (Shantanu)

Extensions

- A bidirectional model with similar size and experiments (Vipul, Shantanu)

Extensions

- A bidirectional model with similar size and experiments (Vipul, Shantanu)
- Explainable few-shot learning and analysis to see if GTP-3 is actually learning (Vipul, Shantanu)

Extensions

- A bidirectional model with similar size and experiments (Vipul, Shantanu)
- Explainable few-shot learning and analysis to see if GTP-3 is actually learning (Vipul, Shantanu)
- A distilled version of GPT-3 (Shantanu)

Extensions

- A bidirectional model with similar size and experiments (Vipul, Shantanu)
- Explainable few-shot learning and analysis to see if GTP-3 is actually learning (Vipul, Shantanu)
- A distilled version of GPT-3 (Shantanu)
- Limited context window of 2048 (Shantanu)

Extensions

- A bidirectional model with similar size and experiments (Vipul, Shantanu)
- Explainable few-shot learning and analysis to see if GTP-3 is actually learning (Vipul, Shantanu)
- A distilled version of GPT-3 (Shantanu)
- Limited context window of 2048 (Shantanu)
- Adversarial experiments to tweak the training samples articulately and present the adversarial examples to it at test time for inference. (Vipul)

Thank you

References

- https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a
- https://www.youtube.com/watch?v=SY5PvZrJhLE
- https://jalammar.github.io/how-gpt3-works-visualizations-animations/
- https://www.youtube.com/watch?v=8psgEDhT1MM&vl=en
- https://www.youtube.com/watch?v=7qPDwsCLbZc&t=3959s
- Language Models are Few-Shot Learners (Brown et. al)
- https://www.youtube.com/watch?v=Mq97CF02sRY