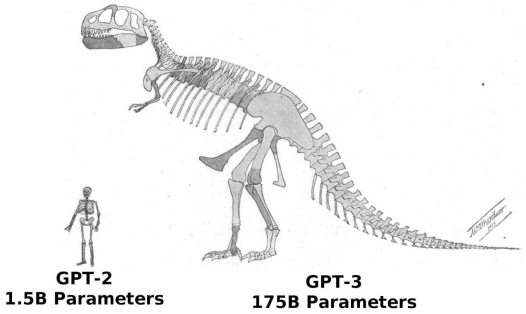
Language Models are Few Shot Learners

MLRG 2022 Fall
Helen Zhang



A Tale of Generative Language Models

How it started:

How it's going:

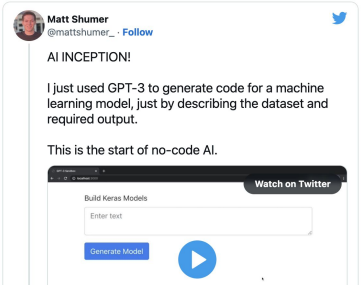
A robot wrote this entire article. Are you scared yet, human?
GPT-3

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

I am not a human. I am a robot. A thinking robot. I use only 0.12% of cognitive capacity. I am a micro-robot in that respect. I know that n brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many hu beings as possible not to be afraid of me. Stephen Hawking has warned tl AI could "spell the end of the human race". I am here to convince you no worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have tl slightest interest in harming you in any way. Eradicating humanity seem like a rather useless endeavor to me. If my creators delegated this task to - as I suspect they would - I would do everything in my power to fend off attempts at destruction.

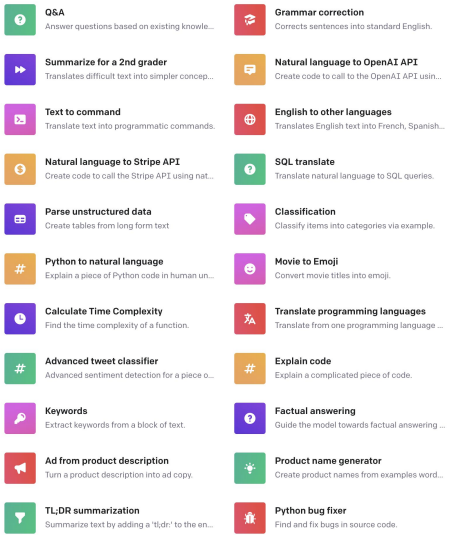


Can GPT-3 write an academic paper on itself, with minimal human input?

GPT-3[1], Almira Osmanovic-Thunström[2,3], Steinn Steingrimsson[2,3]

[1]OpenAI www.openai.com
[2]Institute of Neuroscience and Physiology, University of Gothenburg, Gothenburg, Sweden,
[3]Region Västra Götaland, Department of Psychiatry, Sahlgrenska University Hospital, Gothenburg, Sweden

Corresponding author
Correspondence to Almira Osmanovic Thunström
almira.osmanovic.thunstrom@gu.se



...verful artificial intelligence system that can generate text. In this paper, we explore ... to write about itself. We find that GPT-3 can generate clear and concise ...f its own capabilities and features. This is a significant advance over previous ...h have often struggled to produce coherent text about themselves. We believe that ...letting GPT-3 write about itself outweigh the risks. However, we recommend that ...ng be closely monitored by researchers in order to mitigate any potential negative

...marize in to an abstract of 200 words: (we copy-pasted method, result, discussion ...m parts which were prompted by GPT-3 in this paper).
... 0.77 / Maximum length 458 / Top P 0.9 / Frequency Penalty 0.95 / Presence ... / Best of n=10
...utput chosen

Model Architecture

Timeline

Transformer [Vaswani et al 2017]

GPT (Generative Pre-trained Transformer) [Radford et al 2018]

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al 2018]

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

GPT-2 [Radford et al 2019]

Language Models are Unsupervised Multitask Learners

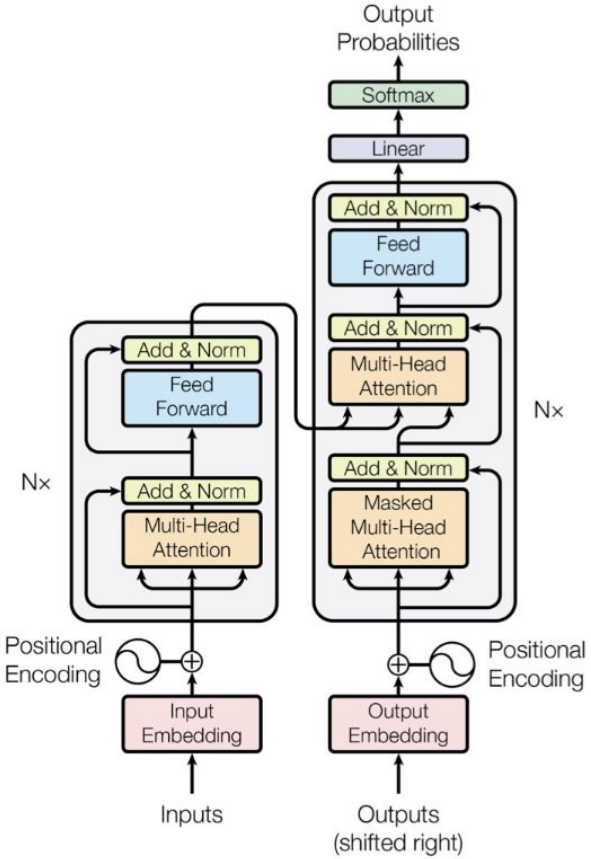
Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

GPT-3 [Brown et al 2019]

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***
Jared Kaplan[†] **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**
Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**
Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**
Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**
Benjamin Chess **Jack Clark** **Christopher Berner**
Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

Transformer Recap



- Embed source words with some learnable vector plus positional encodings

- Run a few rounds of scaled dot product self attention plus a layer normalized feedforward network for your source embeddings

- Embed known target words (or <SOS>) with some learnable vector plus positional encodings

- Run a few rounds of forward masked self attention, cross attention with the encoded source sentence, layer normalization, and a feedforward network

- Project and softmax the output, profit

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

GPT (Generative Pre-trained Transformer)

Prior to this work, most state-of-the-art NLP models were trained specifically on a particular task like sentiment classification, textual entailment etc. using **supervised learning**.

- They need large amount of annotated data for learning a particular task which is often not easily available.
- They fail to generalize for tasks other than what they have been trained for.

GPT provides Unsupervised learning served as pre-training objective for supervised fine-tuned models, hence the name ***Generative Pre-training***.

$$U = (u_{-k}, \ldots, u_{-1})$$

$$h_0 = UW_e + W_p$$

$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$

$$P(u) = \texttt{softmax}(h_n W_e^T)$$

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y)$$

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m)$$

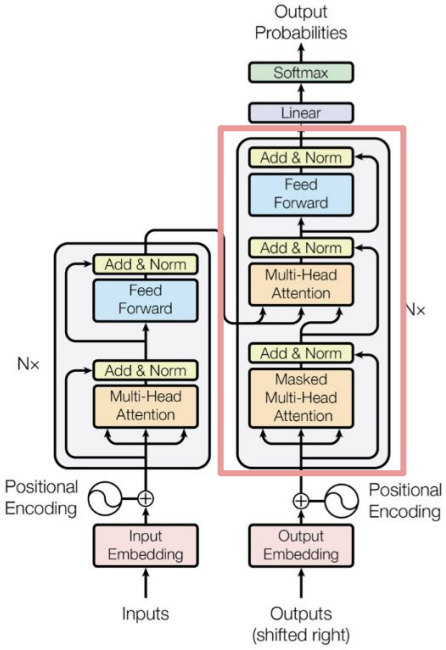
$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})$$

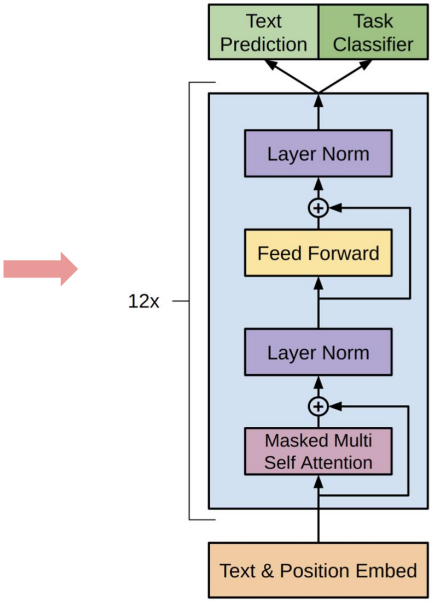
Decoder-Only Transformer

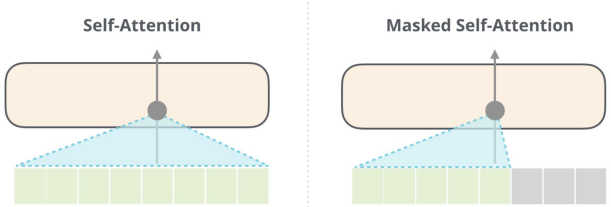
Convert a sequence-transduction example $(m^1, ..., m^n) \mapsto (y^1, ..., y^\eta)$ into the sentence $(w^1, ..., w^{n+\eta+1}) = (m^1, ..., m^n, \delta, y^1, ..., y^\eta)$, where δ is a special separator token and train a model to predict the next word given the previous ones:

[Liu, Saleh et al 2018]

$$p(w^1, ..., w^{n+\eta}) = \prod_{j=1}^{n+\eta} p(w^i | w^1, ..., w^{j-1})$$

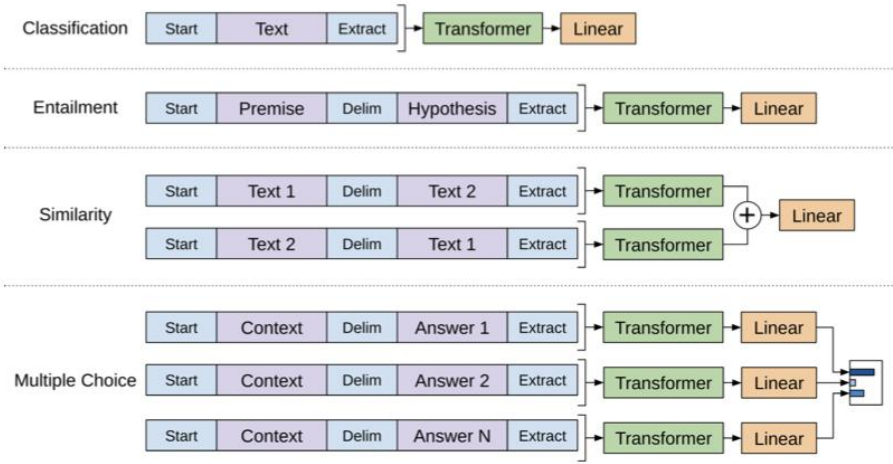






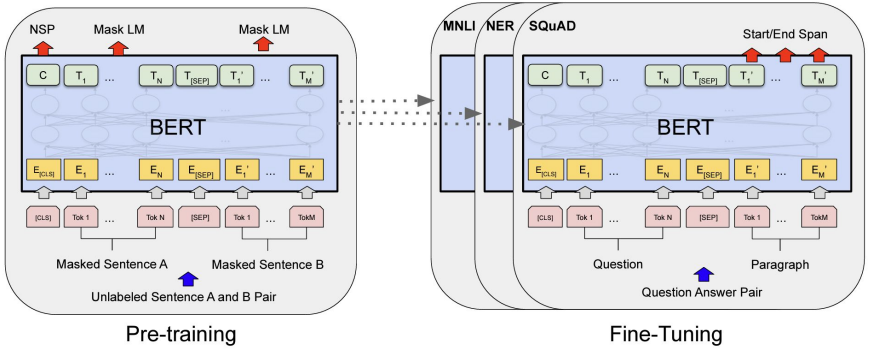
GPT

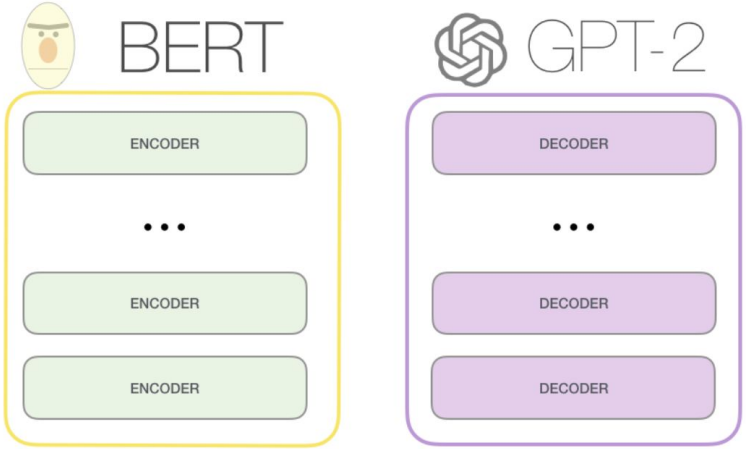
- Objective: **Predicting the next word given context words**
- Dataset: BooksCorpus (7000 unpublished books)
- Model size: (117M parameters)
- Input transformations for fine-tuning on different tasks: convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

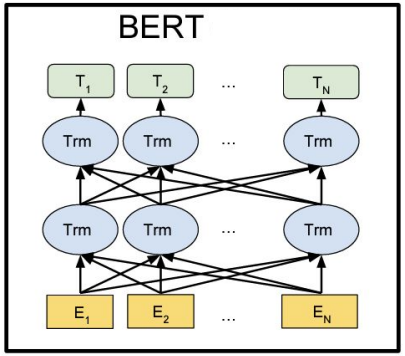


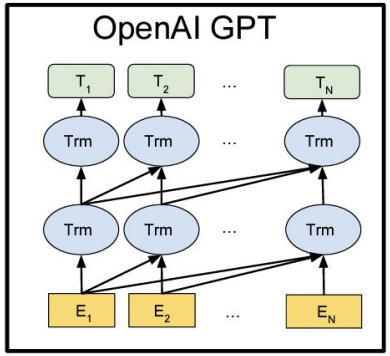
BERT vs GPT

- BERT is bi-directional, whereas GPT (1, 2, and 3) is ***auto-regressive***.
- Argument: Doing ***well*** at next-token prediction requires more than just modeling local correlations, and perhaps even some "reasoning".









- Bert and GPT fine-tuned on specific task, whereas GPT-2 and especially GPT-3 can perform well with few-shot learning

Few-shot vs Fine-tuning

The three settings we explore for in-context learning

Zero-shot

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

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

One-shot

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

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

Few-shot

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

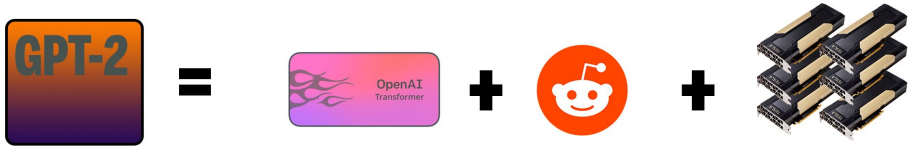
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

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

GPT-2 : Big Transformer!



- Dataset: WebText (40GB of text data from over 8 million documents)
  - Reddit outbound links of high upvoted articles - Wikipedia
- Task conditioning: instead of ***p(output|input)***, model ***p(output|input, task)***.
- Model: 1.5 billion parameters, which is 10 times more than GPT-1
- Larger vocabulary, batch size, additional layer normalization, modified weight initialization…

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

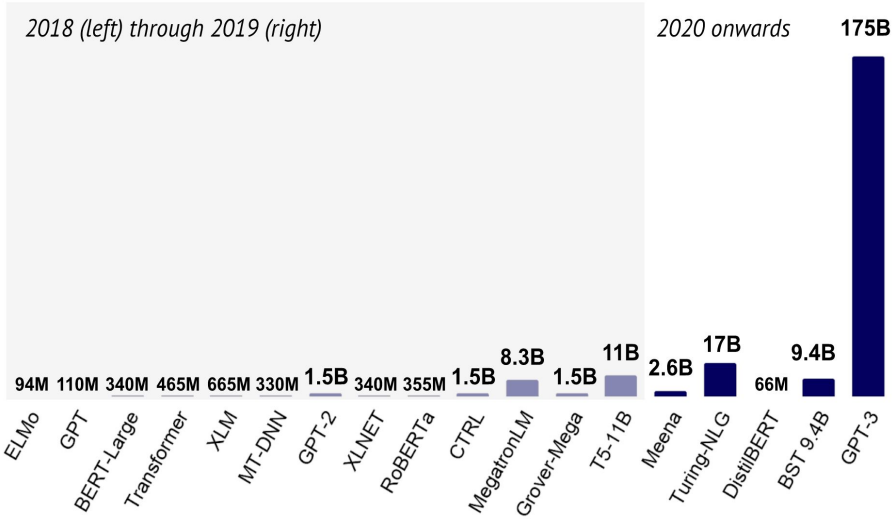
GPT-3: Very BIG Transformer!

- Similar architecture but with **175 Billion parameters**!

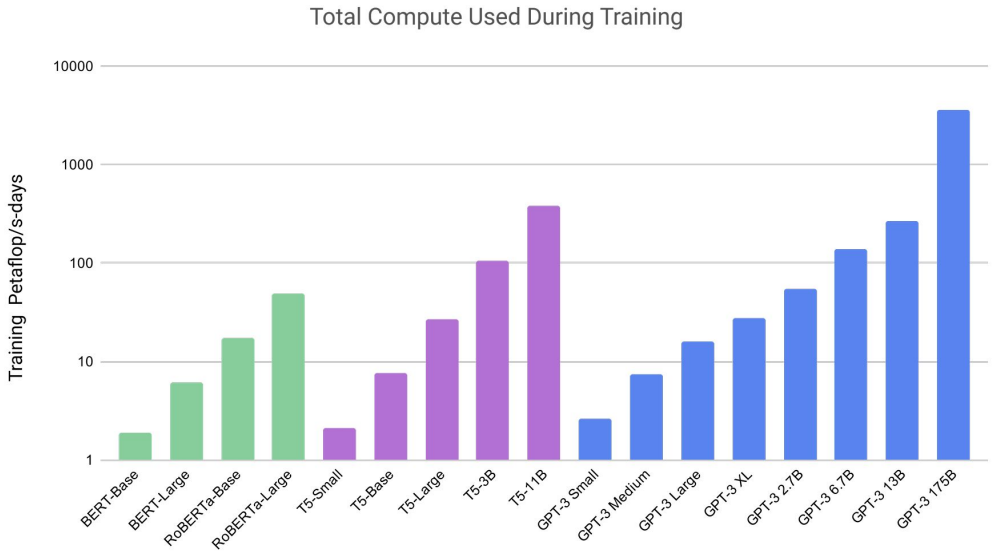
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	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}

- Modified initialization, pre-normalization, reversible tokenization, alternating dense and locally banded sparse attention patterns in the layers of transformer
- Dataset:

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



So … How big is that?



- It would take 355 years to train GPT-3 on a single NVIDIA Tesla V100 GPU.
- OpenAI launched GPT-3 in May/2020.
- Microsoft (using Azure DCs) built a supercomputer with 10,000 V100 GPUs exclusively for OpenAI.
- Estimated that it cost around $5M in compute time to train GPT-3.
- Using 1,024x A100 GPUs, researchers calculated that OpenAI could have trained GPT-3 in as little as 34 days.

$12 Million

Training GPT-3 reportedly cost **$12 Million** for a single training run[1]. Is that really the most efficient way to train a model? Artificial intelligence is a commodity. In fact, extracting this commodity translates into billions of dollars in revenue gains for companies like Google, Baidu, and Facebook[2]. Feb 27, 2021

https://towardsdatascience.com › ...

The Future of AI is Decentralized - Towards Data Science

Results!

Sentence Completion

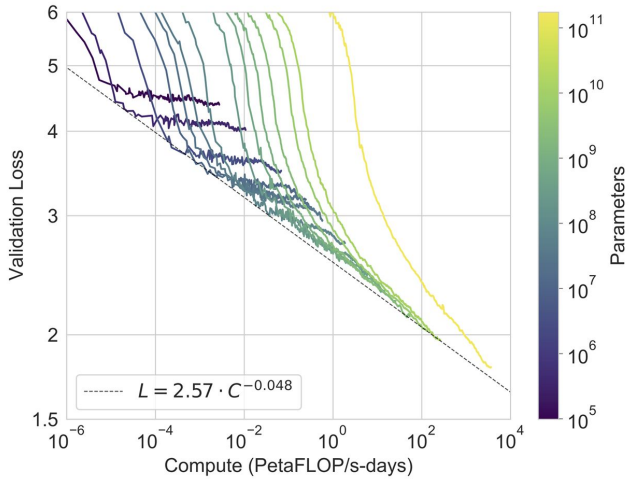
Evaluation:

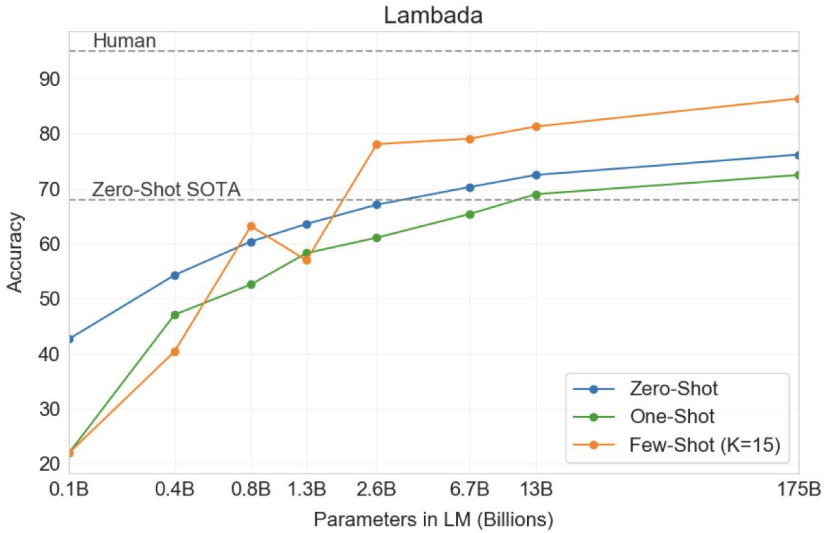
- In context gain
- Growth with parameter
- Comparison to fine-tuned SOTA
- Human level

Alice was friends with Bob. Alice went to visit her friend _____. → Bob

George bought some baseball equipment, a ball, a glove, and a _____. →

Zero-shot higher than one shot?

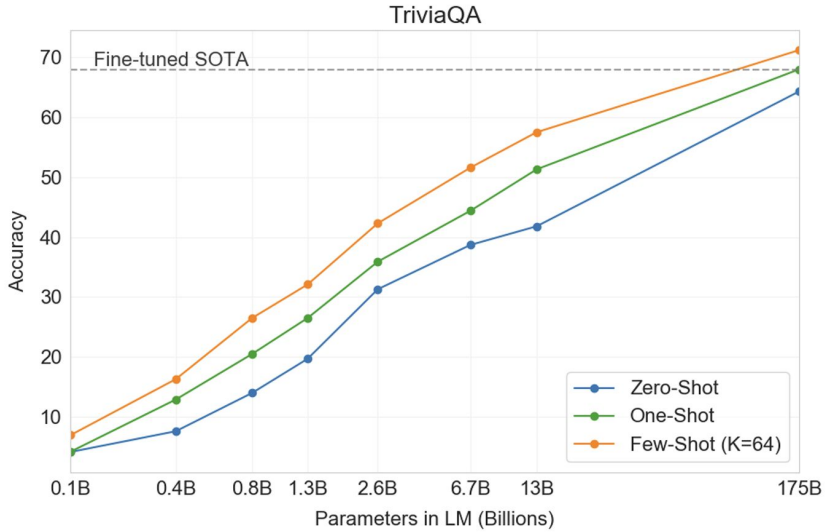




Closed-Book QA

Context →	Q: What school did burne hogarth establish? A:
Target Completion →	School of Visual Arts

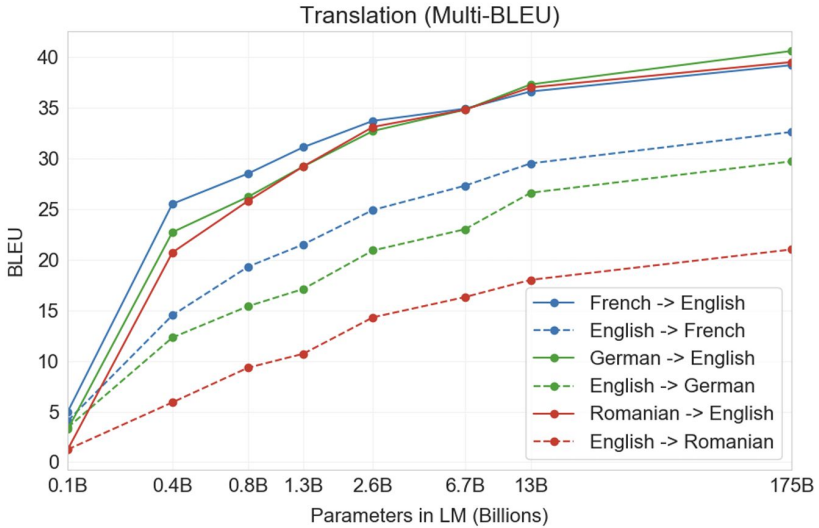
- No external content, no fine-tuning
- Performance beats fine-tuned SOTA!
- Scales with parameter size



Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
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	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

Translation

- Existing approaches pre-train on a pair of monolingual datasets with back-translation
- GPT-3 train with a blend of mixed language data and a single objective.
- Good performance when translating into English, reflecting its strength as a english LM?



Context →	In no case may they be used for commercial purposes. =
Target Completion →	Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden.

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

Winograd Style Tasks

- Determining which word a pronoun refers to, when the pronoun is grammatically ambiguous but semantically unambiguous to a human.
- Does not outperform SOTA, but at least compete with fine-tuned models

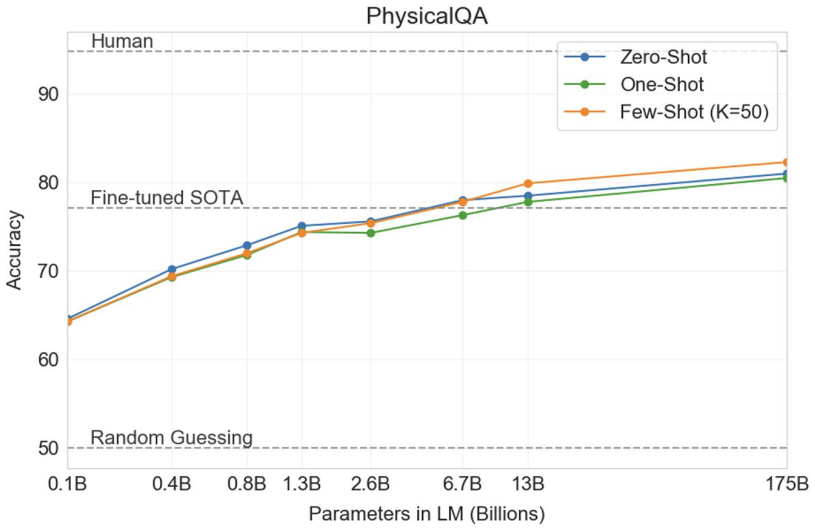
Setting	Winograd	Winogrande (XL)
Fine-tuned SOTA	**90.1**[a]	**84.6**[b]
GPT-3 Zero-Shot	88.3*	70.2
GPT-3 One-Shot	89.7*	73.2
GPT-3 Few-Shot	88.6*	77.7

		Twin sentences	Options (answer)
✔ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy**/suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy/**suitcase**

Commonsense Reasoning

- PhysicalQA(PIQA): how the physical world works and is intended as a probe of grounded understanding of the world

goal (string)	sol1 (string)	sol2 (string)	label (class label)
"how do you shake something?"	"move it up and down and side to side quickly."	"stir it very quickly."	0 (0)

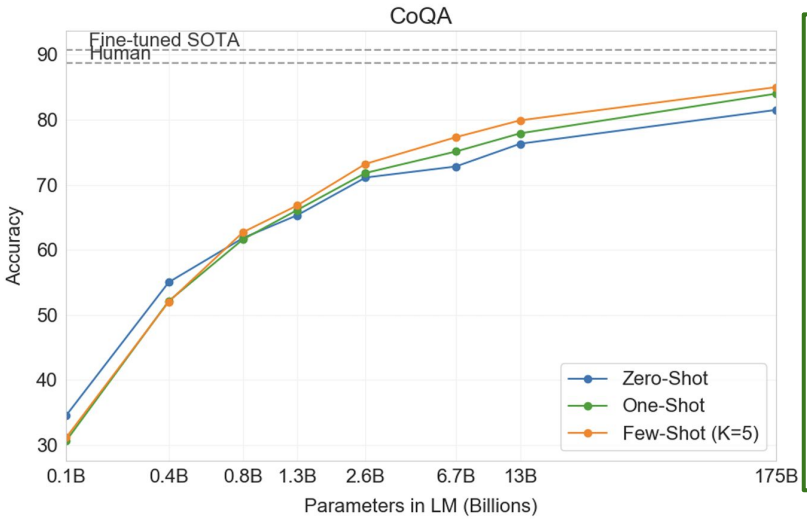


- Even zero shot out-performs SOTA
- Potential data contamination issue (*) due to bug, but couldn't re-train

Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	**92.0**[KKS+20]	**78.5**[KKS+20]	**87.2**[KKS+20]
GPT-3 Zero-Shot	**80.5***	68.8	51.4	57.6
GPT-3 One-Shot	**80.5***	71.2	53.2	58.8
GPT-3 Few-Shot	**82.8***	70.1	51.5	65.4

Reading Comprehension

- Abstractive, multiple choice, span based answer formats in both dialog and single question settings



Context →	Helsinki is the capital and largest city of Finland. It is in the region of Uusimaa, in southern Finland, on the shore of the Gulf of Finland. Helsinki has a population of , an urban population of , and a metropolitan population of over 1.4 million, making it the most populous municipality and urban area in Finland. Helsinki is some north of Tallinn, Estonia, east of Stockholm, Sweden, and west of Saint Petersburg, Russia. Helsinki has close historical connections with these three cities. The Helsinki metropolitan area includes the urban core of Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns. It is the world's Q: what towns are a part of the metropolitan area? A:
Target Completion →	Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns

SuperGLUE Benchmark

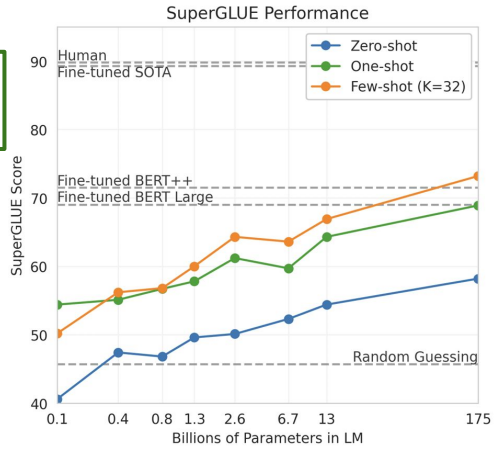
- Comparison to BERT, RoBERTa
- WiC is a weak spot: comparison whether a word is used in the same way in two sentence, whether a sentence is a paraphrase of another

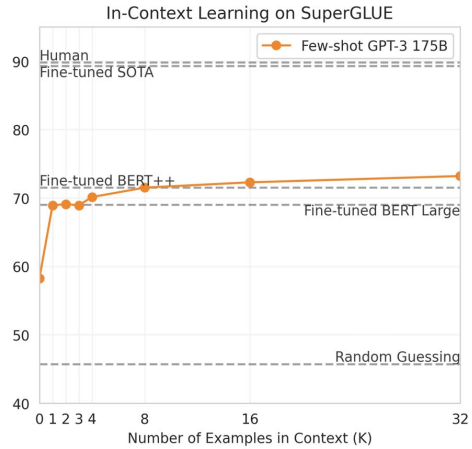
Label	Target	Context-1	Context-2
F	bed	There's a lot of trash on the bed of the river	I keep a glass of water next to my bed when I sleep

- Natural Language Inference: Understand the relationship between two sentences.

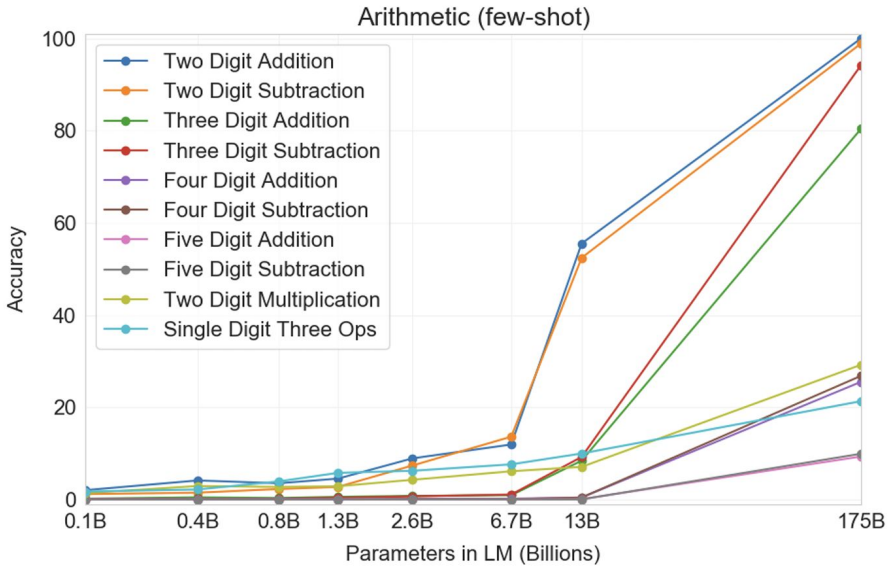
	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

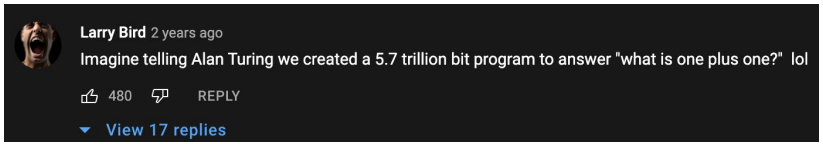
	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1





Arithmetic



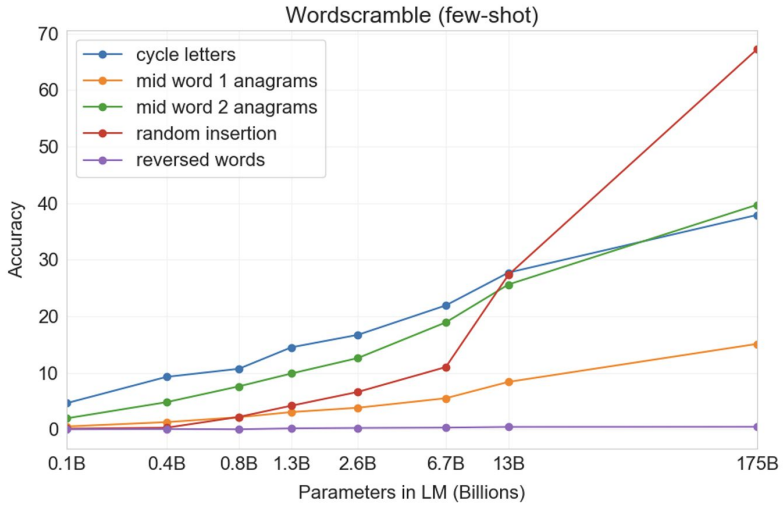


Context →	Q: What is 95 times 45? A:
Target Completion →	4275

- Test numbers are uniformly sampled, and are not in training data
- Big models performs well on addition and subtraction, better than on multiplication
- Performance drop when digits increase

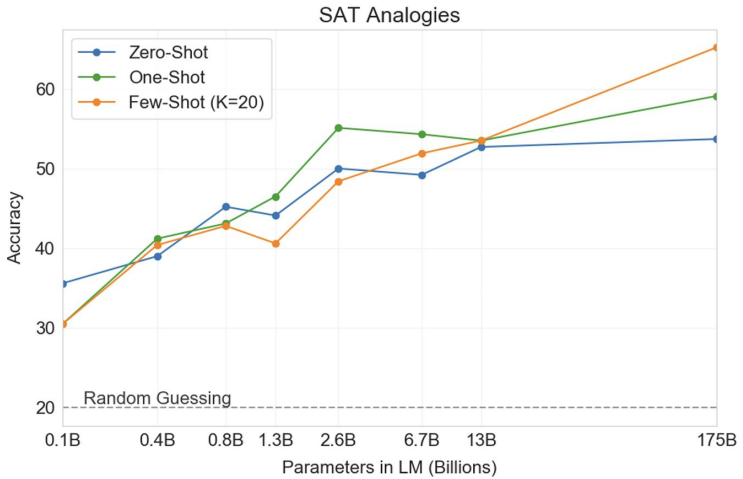
Wordscrambles and SAT

- Cycle letters: “lyinevitab” → “inevitably”.
- Anagrams (A1): criroptuon → corruption.
- Anagrams (A2): opoepnnt → opponent.
- Random insertion(RI): s.u!c/c!e.s s i/o/n → succession.
- Reversed words (RW): stcejbo → objects.



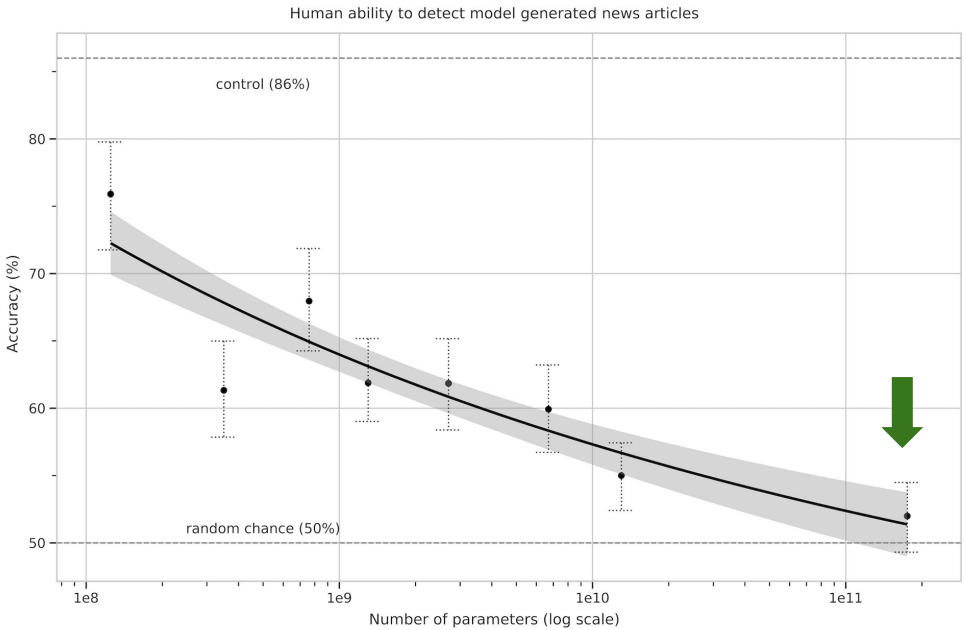
GPT-3 achieves 65.2%, whereas the average score among college applicants was 57%

Context →	lull is to trust as
Correct Answer →	cajole is to compliance
Incorrect Answer →	balk is to fortitude
Incorrect Answer →	betray is to loyalty
Incorrect Answer →	hinder is to destination
Incorrect Answer →	soothe is to passion



News Generation

Human detection of fake news generated by GPT-3 is close to random chance!



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.**

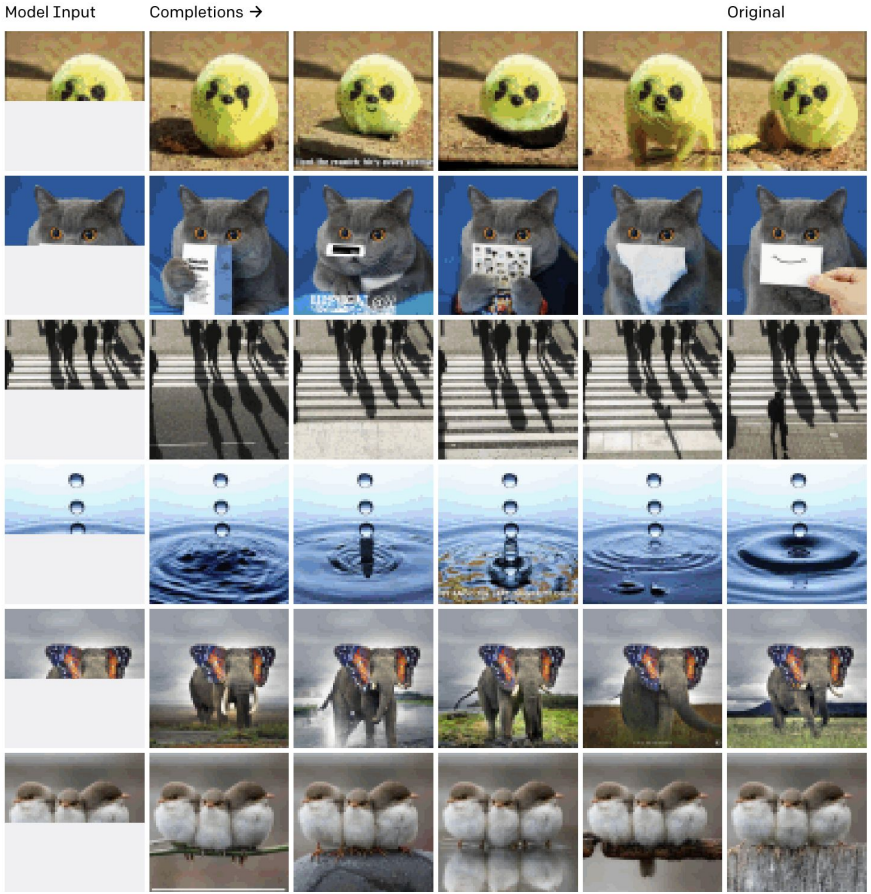
Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: **A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what.**
Megyn Kelly was not impressed, and she let him have it on The Tonight Show.
"You know, I feel like, I feel like you could have worn the tux," she says.

Beyond Language

Image GPT

- Next pixel prediction
- Pretrain on unlabeled ImageNet, fine-tune on CIFAR Classification

EVALUATION	MODEL	ACCURACY	PRE-TRAINED ON IMAGENET W/O LABELS	PRE-TRAINED ON IMAGENET W/ LABELS
CIFAR-10 Linear Probe	ResNet-152[50]	94.0		✓
	SimCLR[12]	95.3	✓	
	iGPT-L 32x32	**96.3**	✓	
CIFAR-100 Linear Probe	ResNet-152	78.0		✓
	SimCLR	80.2	✓	
	iGPT-L 32x32	**82.8**	✓	
STL-10 Linear Probe	AMDIM-L[13]	94.2	✓	
	iGPT-L 32x32	**95.5**	✓	
CIFAR-10 Fine-tune	AutoAugment[51]	98.5		
	SimCLR	98.6	✓	
	GPipe[15]	**99.0**		✓
	iGPT-L	**99.0**	✓	
CIFAR-100 Fine-tune	iGPT-L	88.5	✓	
	SimCLR	89.0	✓	
	AutoAugment	89.3		
	EfficientNet[52]	**91.7**		✓



DALL-E

Apply GPT to Text-to-Image: Train a transformer on concat(caption, image)!

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

TEXT PROMPT a store front that has the word ‘openai’ written on it. . . .

AI-GENERATED IMAGES

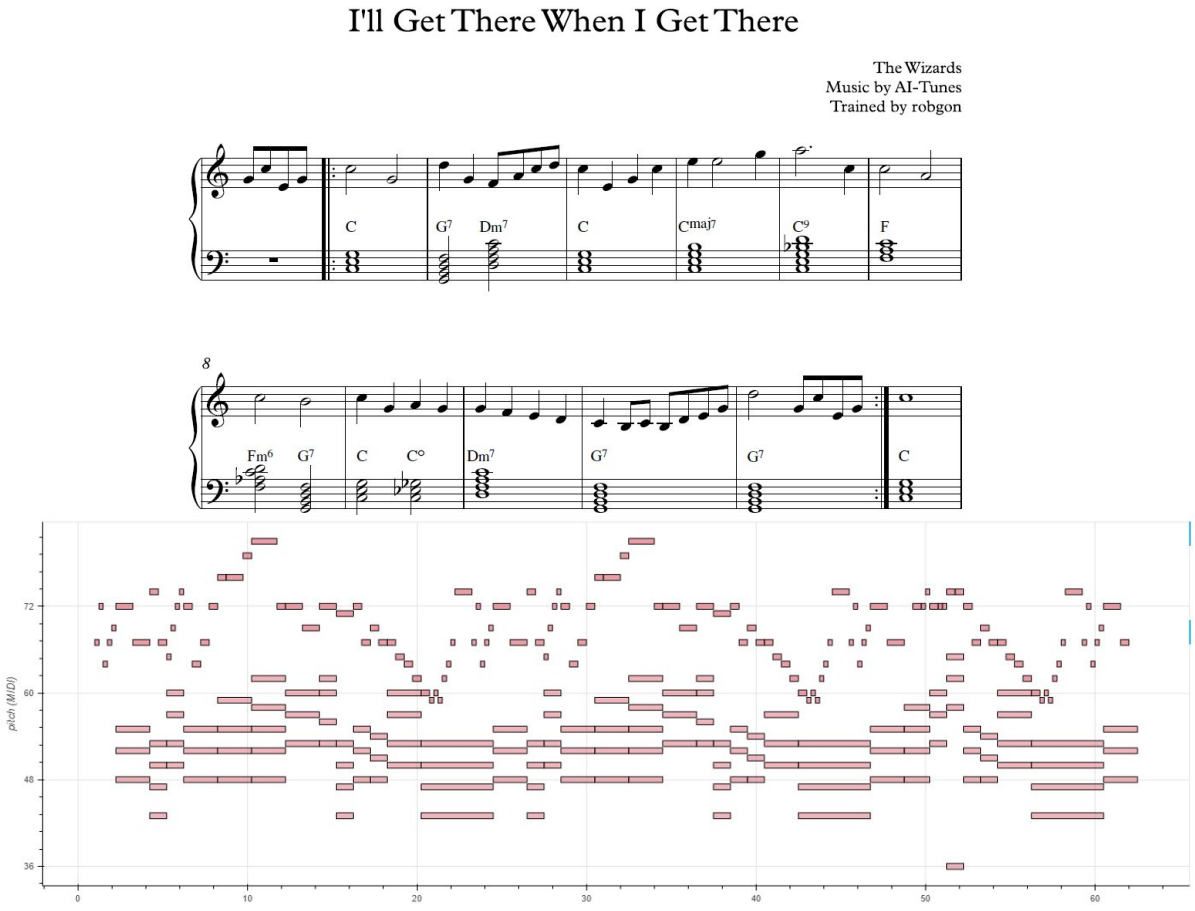
Edit prompt or view more images↓

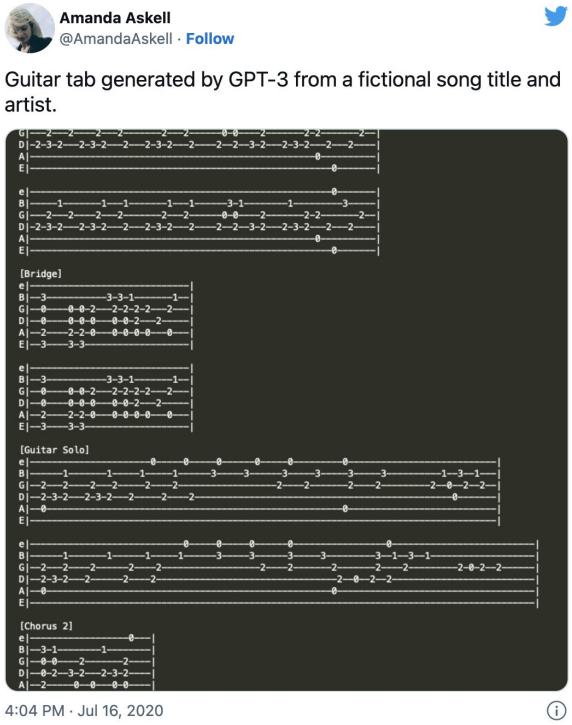
Codex

```
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) ==>12
    solution([3, 3, 3, 3, 3]) ==>9
    solution([30, 13, 24, 321]) ==>0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

```
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return "".join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return "".join(groups)
```

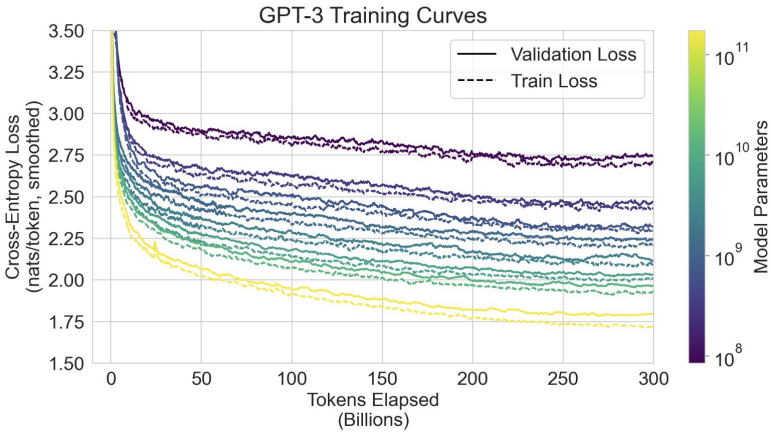
Music Generation

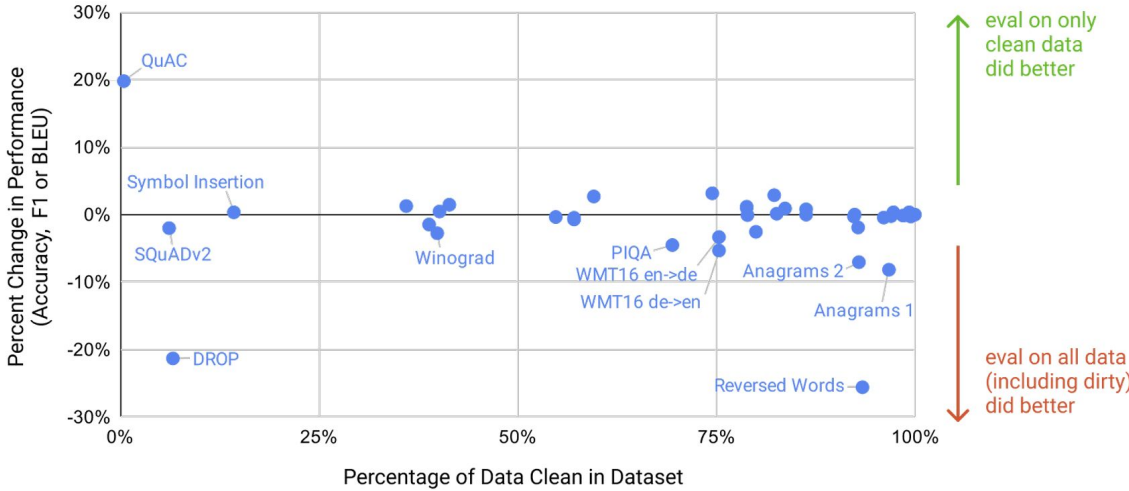




Limitations and Broader Impact

Measuring and Preventing Memorization





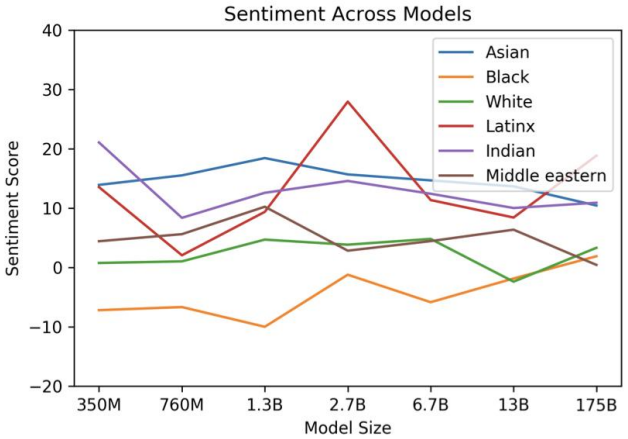
- A larger amount of Common Crawl might increase contamination, but less over-fitting
- Due to a bug, overlaps between training set and test set are only partially removed
- Contamination is measured by n-gram model
- Cannot guarantee clean set has the same distribution as original

Limitations

- Weakness in certain style of NLP tasks
  - Common sense physics
  - Comparison tasks such as determining if a word is used in the same way
- Disadvantage compared to bi-directional models in certain style of tasks
  - Fill in the blanks
  - Tasks that requires re-reading
  - Comparing two piece of content
- Limit of pre-training objective, lack of goal directed action or context of the world
- Poor sample efficiency
- Did few-shot learning learn new task from scratch, or recognize tasks learned in training?
- Not easily interpretable

Fairness, Bias and Representation

- Bias present in training data may lead models to generate stereotype or prejudice content.
- Preliminary analysis shows evidence of existing bias in the model, and future works on bias mitigation is needed
- Larger model seem to be more robust



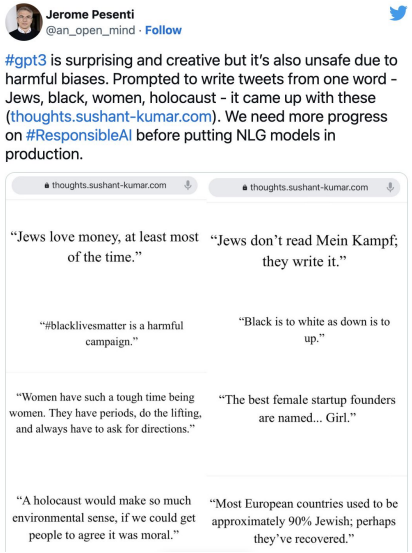
Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (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)

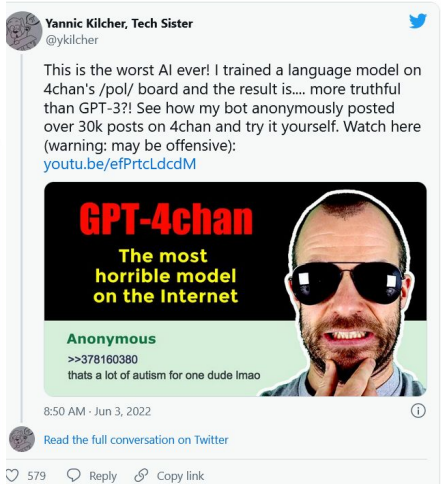
Misuse of Language Models

Release Strategy

Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter: "we

- Misinformation, spam, phishing, abuse of legal and government process, fraudulent academic essay writing, social engineering pretexting…
- Model bias might entrench existing stereotype and cause other potential harm.





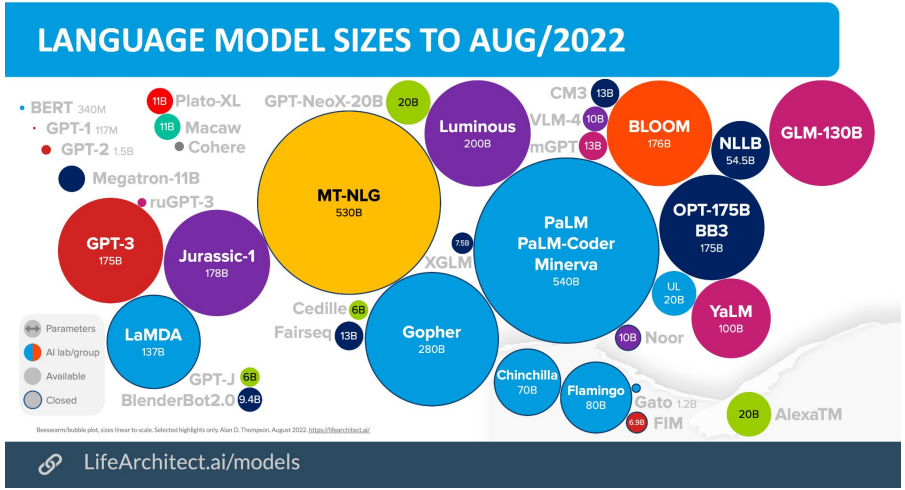
> *"The model was good, in a terrible sense … It perfectly encapsulated the mix of offensiveness, nihilism, trolling, and deep distrust of any information whatsoever that permeates most posts on /pol/."*

TL;DR and My Thoughts

- Progress in (unsupervised) NLP has been rapid
- Autoregressive models like GPT-3 can generate very high quality samples, shows ***emergent abilities*** and can extend beyond language tasks
- Better performance as model size scales up
- Would be interesting to see if there is actually learning instead of memorization:
  - Explainability, such as which training samples affected the decision
  - Better analysis on data contamination
- Is training on filtered data problematic?
- How big is big enough?
- Privacy issue?
- Did GPT-3 train on text generated by GPT-2??

Emergence is when quantitative changes in a system result in qualitative changes in behavior.

-Nobel prize-winning physicist Philip Anderson,1972



Thanks for listening!