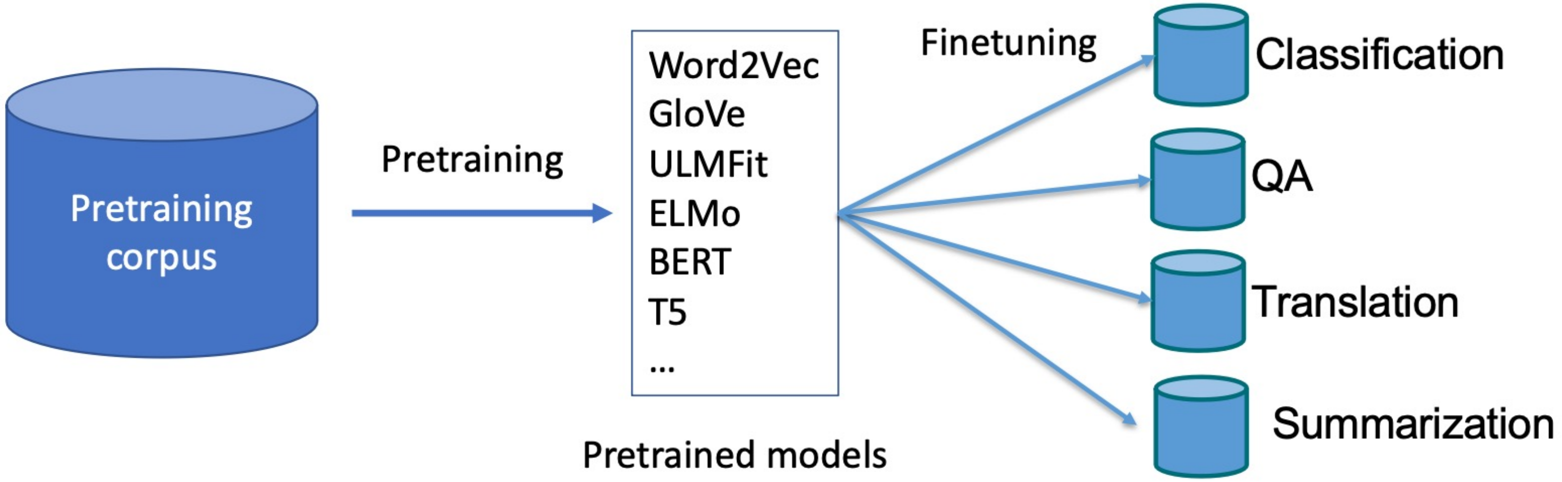
GPT-3: Language Models are Few-Shot Learners

Presenter: Arman Cohan

Motivation

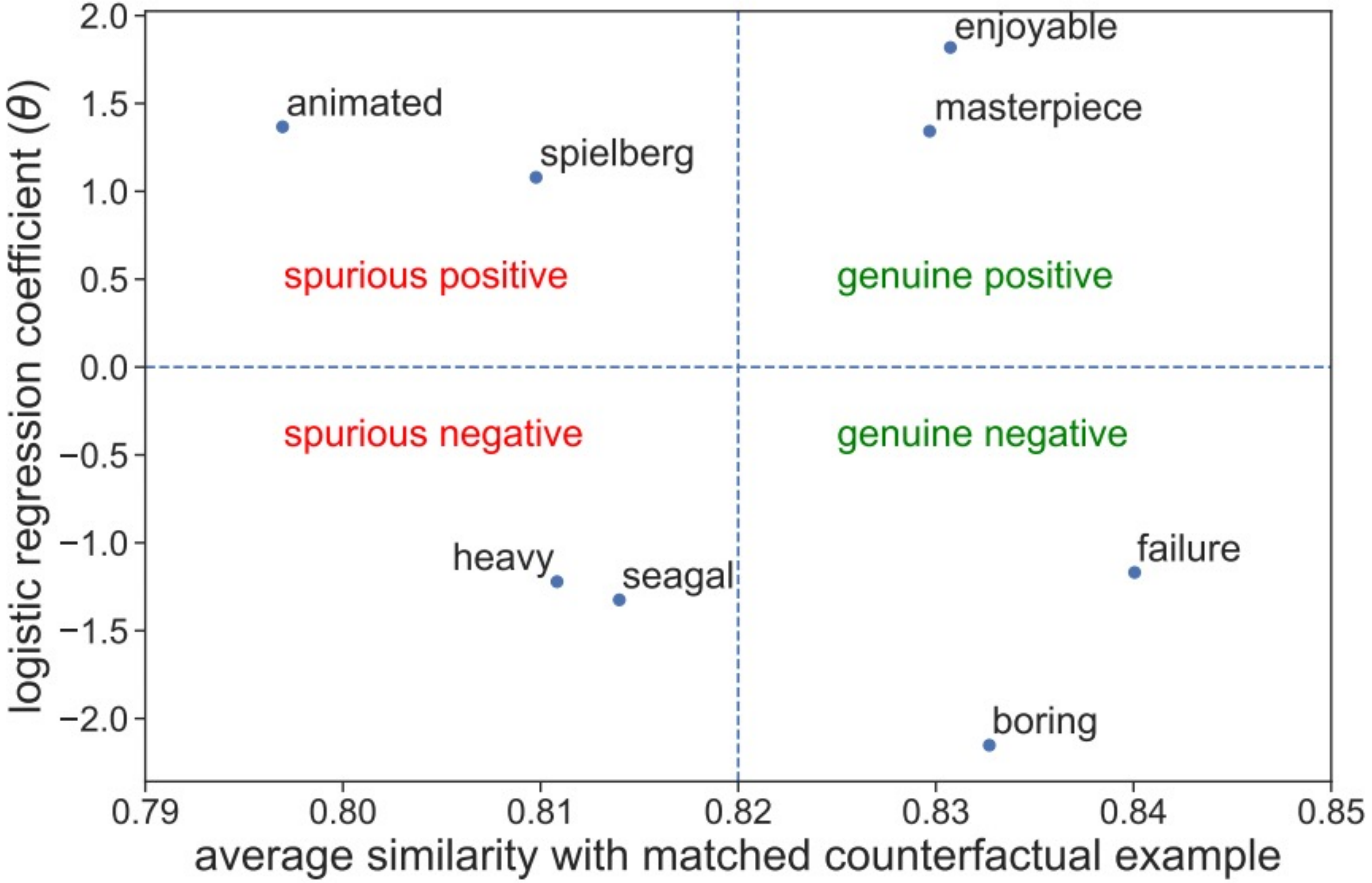
- Recall Pretrain–finetune paradigm (Transfer learning)
  - First pretrain a (large) model on unlabeled data
  - Then continue train on task-specific training dataset



Motivation

- Problems with pretrain then finetune paradigm
  - It requires additional task-specific finetuning
  - For many new tasks it is difficult to collect training data
  - Exposing models to labeled datasets and fine-tuning may exaggerate their out-of-distribution generalization
    - Models fine-tuned on downstream datasets can exploit spurious correlations in training data

- Example of spurious correlations on sentiment classification



Wang and Culotta (2020)

Motivation

- Problems with pretrain then finetune paradigm
  - It requires additional task-specific finetuning
  - For many new tasks it is difficult to collect training data
  - Exposing models to labeled datasets and fine-tuning may exaggerate their out-of-distribution generalization
    - Models fine-tuned on downstream datasets can exploit spurious correlations in training data
  - Humans do not learn from 1000s of training data
    - They can often learn a task quickly using few examples

- How can we move away from this paradigm?

Meta-learning

- From a set of given tasks, learn how to quickly adapt to new tasks
- AKA “learning how to learn”

Meta-learning

- From a set of given tasks, learn how to quickly adapt to new tasks
- AKA “learning how to learn”

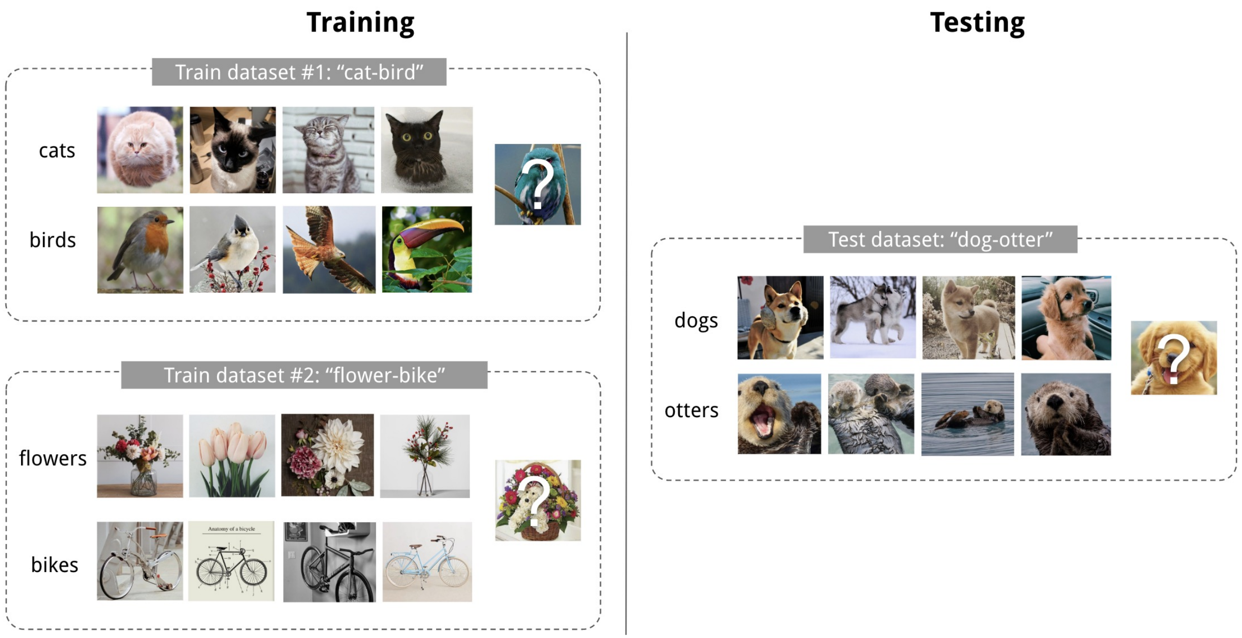
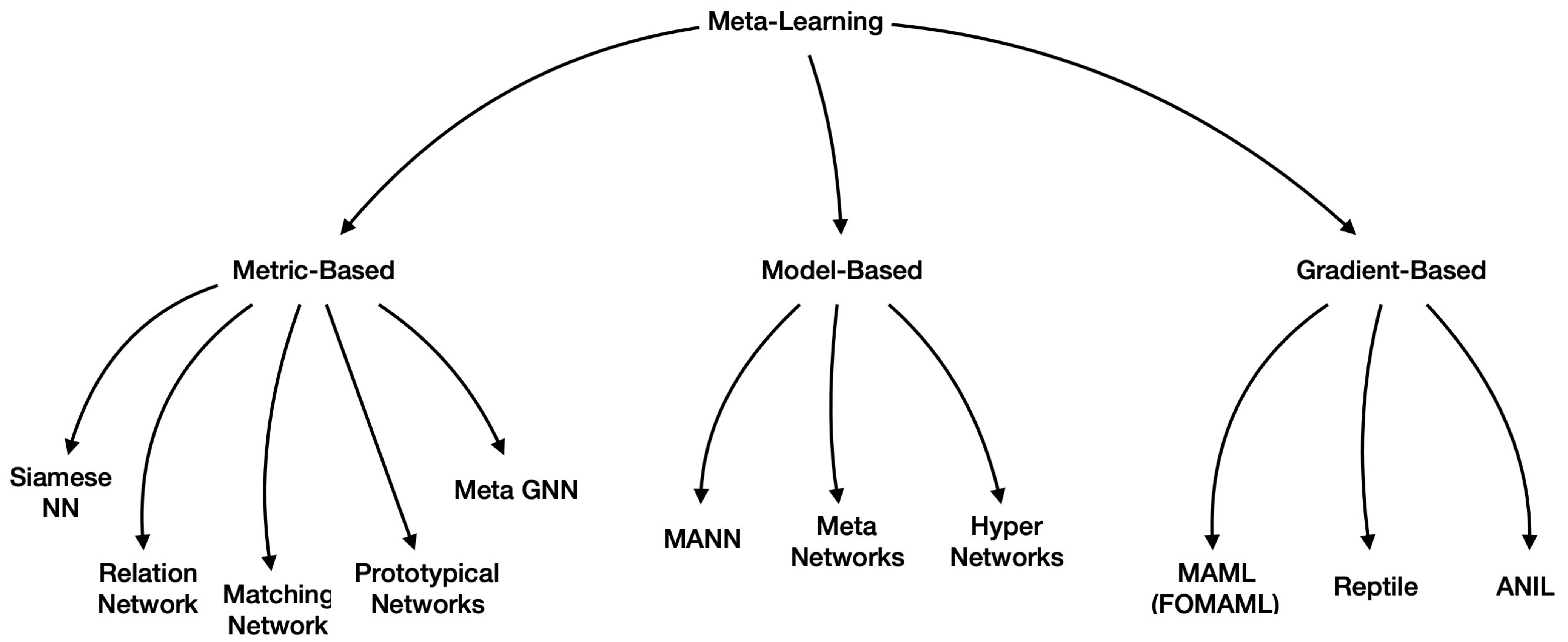


Figure from: https://lilianweng.github.io/posts/2018-11-30-meta-learning/

Meta-learning



https://chao1224.github.io/material/slides/202004.pdf

Meta-learning

- In Meta-learning we want to learn how to quickly adapt to new tasks
  - In standard ML, we iteratively update the model parameters so it can perform a given task (*inner loop*)

Meta-learning

- In Meta-learning we want to learn how to quickly adapt to new tasks
  - In standard ML, we iteratively update the model parameters so it can perform a given task (*inner loop*)

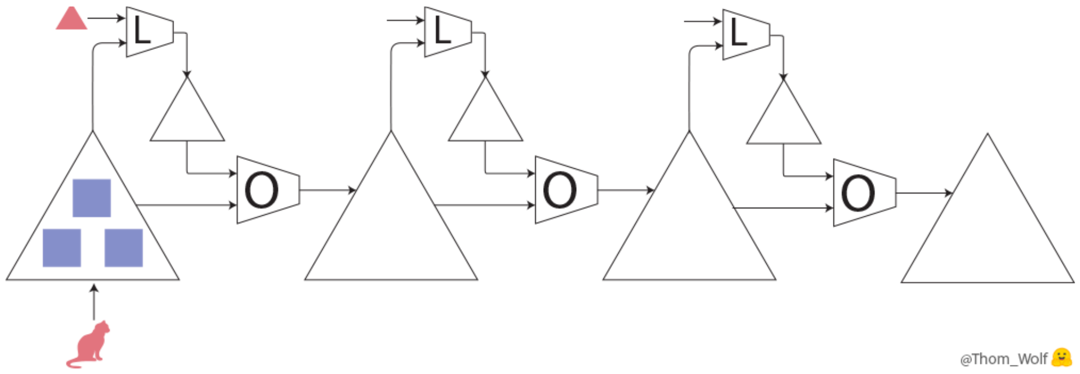


Figure from Huggingface: https://medium.com/huggingface/from-zero-to-research-an-introduction-to-meta-learning-8e16e677f78a

Meta-learning

- In Meta-learning we want to learn how to quickly adapt to new tasks
  - We can update the **model** and **optimizer** parameters to be in a state that can be **quickly adapted to new tasks** **(outer loop)**

Meta-learning

- In Meta-learning we want to learn how to quickly adapt to new tasks
  - We can update the **model** and **optimizer** parameters to be in a state that can be **quickly adapted to new tasks** **(outer loop)**

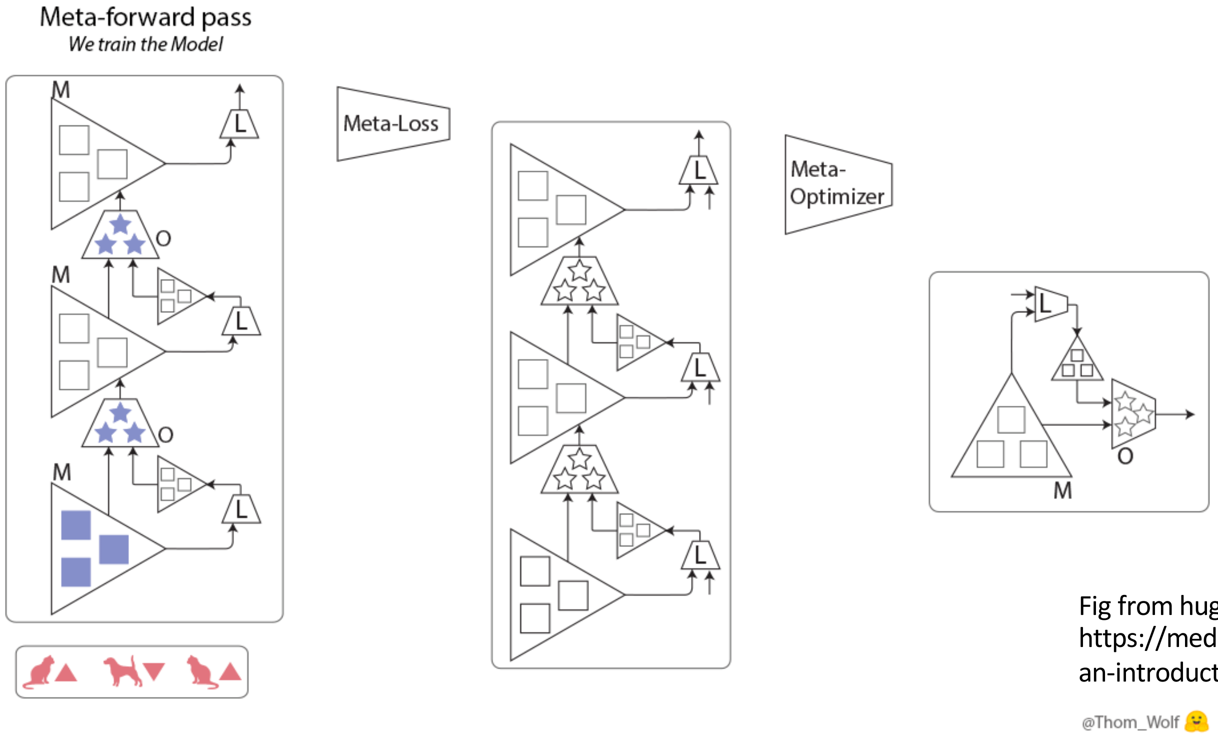


Fig from huggingface: https://medium.com/huggingface/from-zero-to-research-an-introduction-to-meta-learning-8e16e677f78a

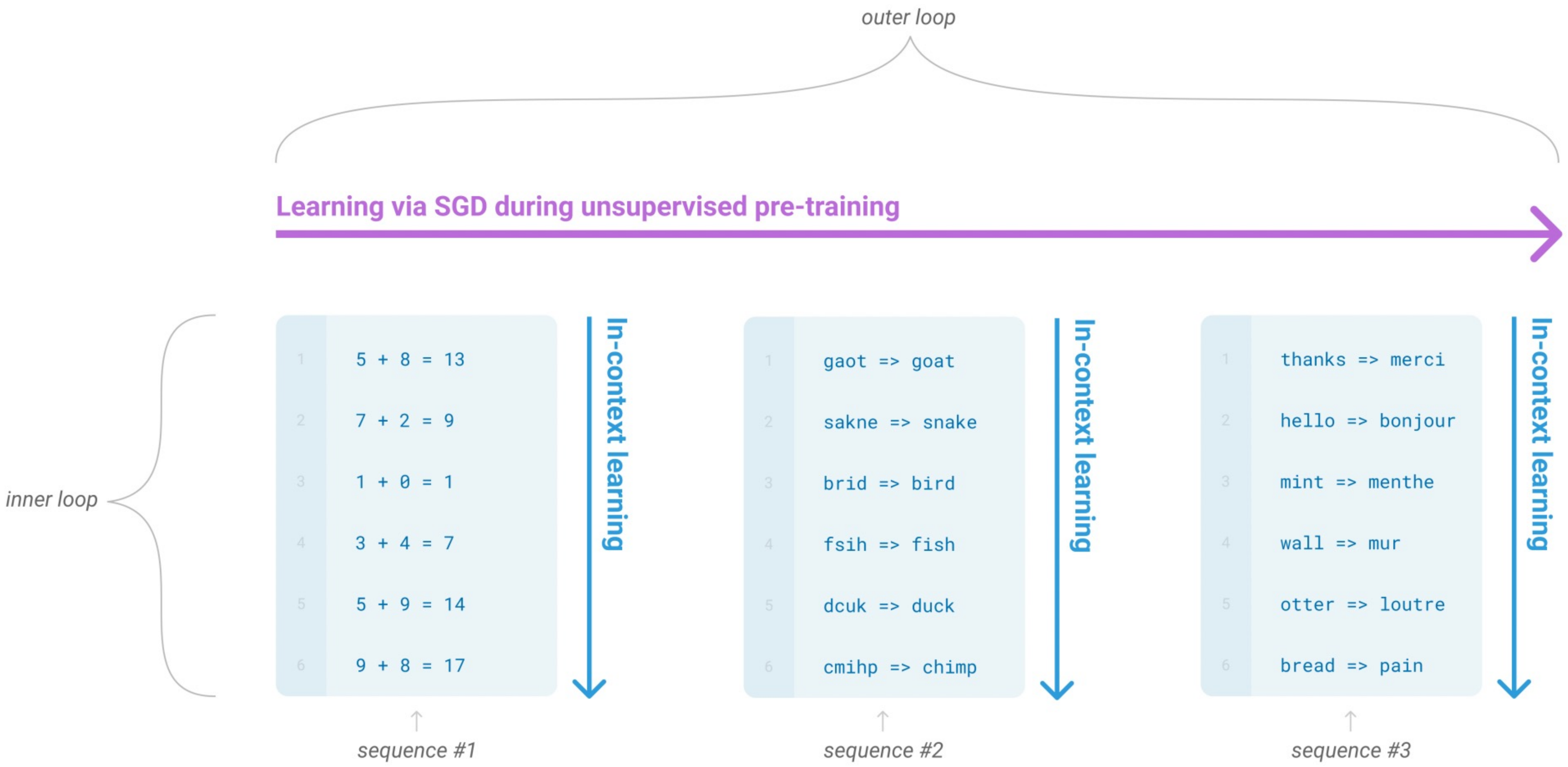
GPT-3

- During pretraining, predicting the next word implicitly gives the model the ability to meta-learn
  - At inference time, it sees a new task, and it can quickly identify patterns that are helpful to solve that task

GPT-3

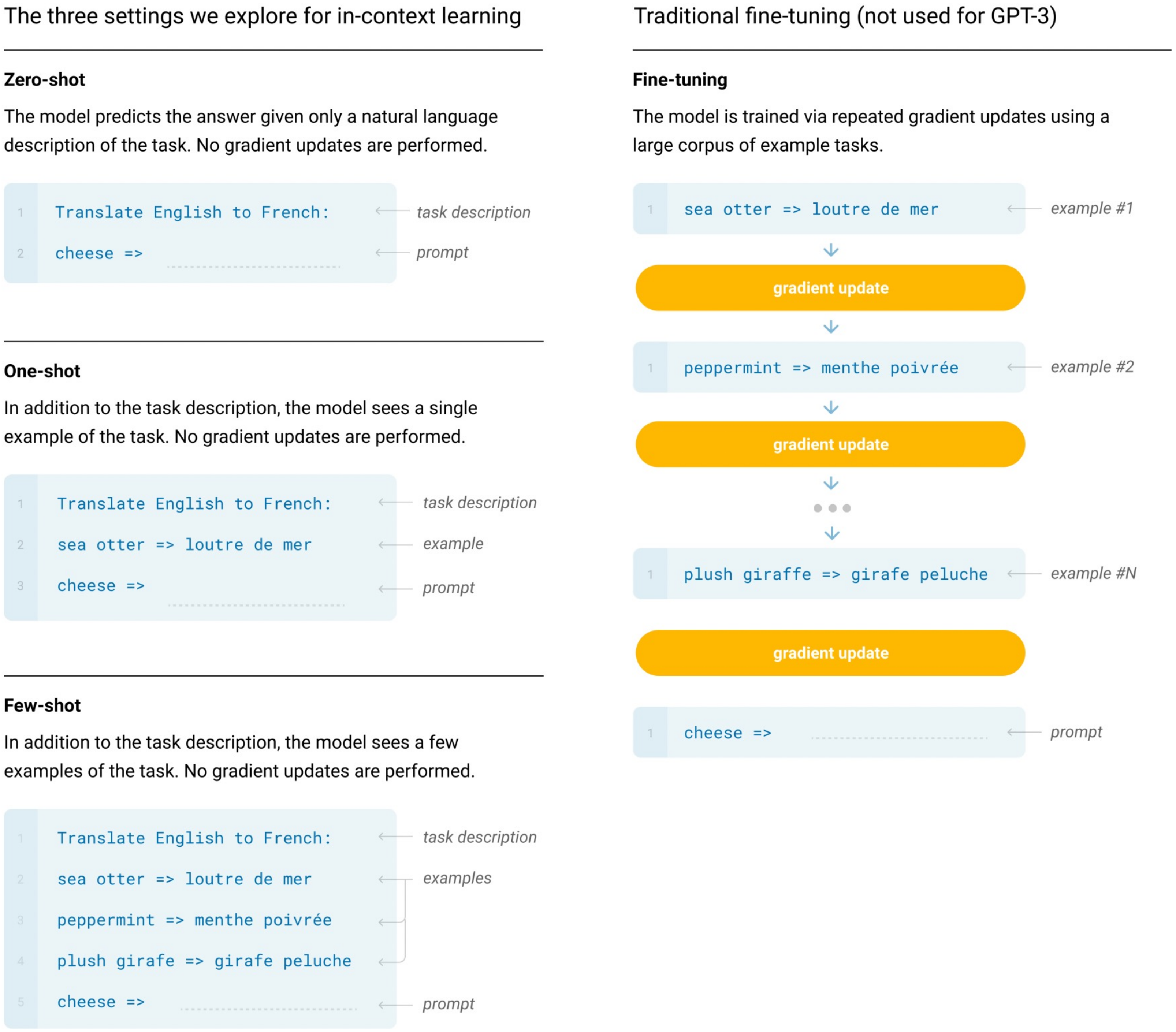
- During pretraining, predicting the next word implicitly gives the model the ability to meta-learn
  - At inference time, it sees a new task, and it can quickly identify patterns that are helpful to solve that task
  - Adaptation happens at inference time through forward pass
    - No need for any gradient updates
    - The model learns through **in-context examples**

GPT-3



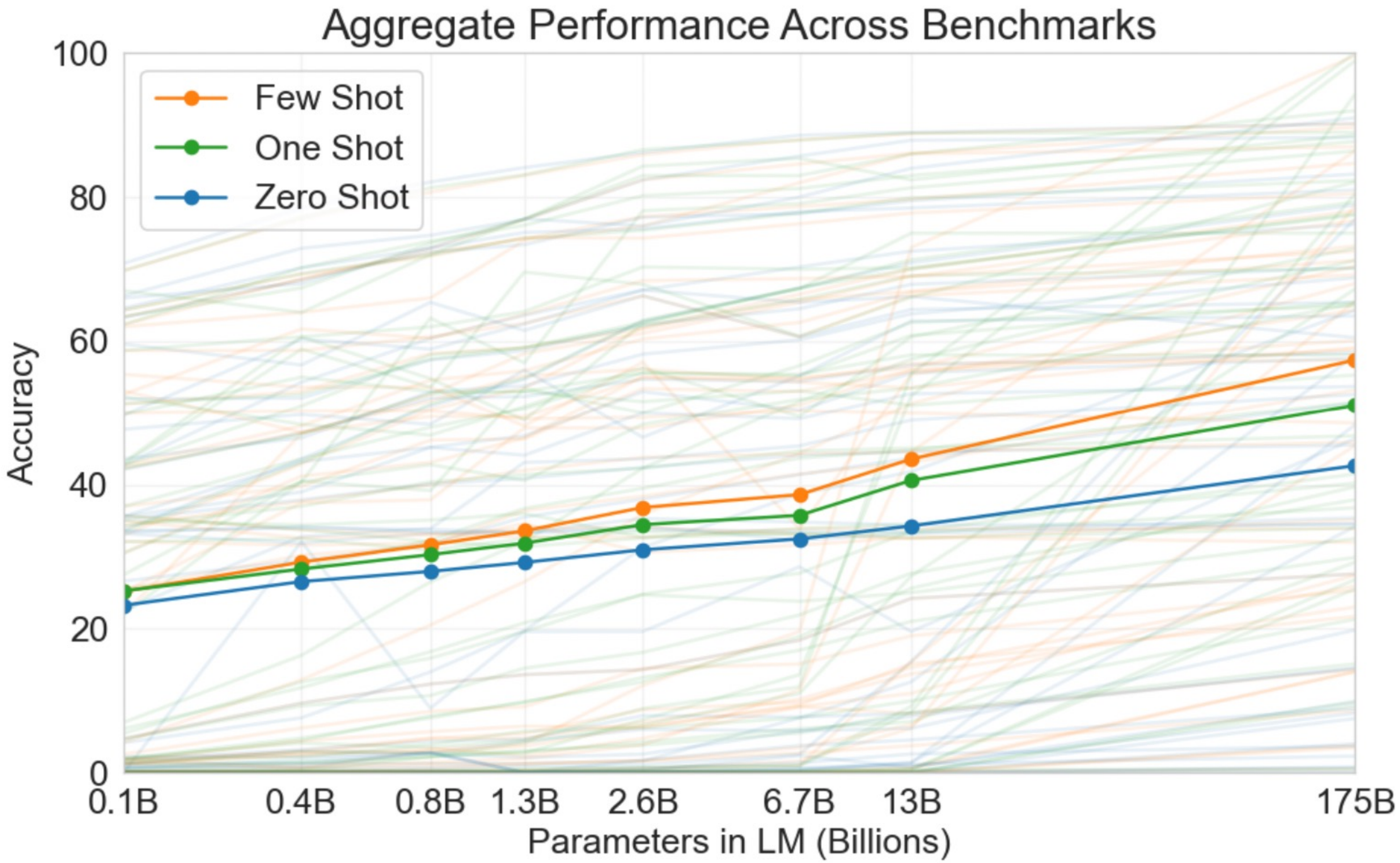
outer loop
Learning via SGD during unsupervised pre-training
inner loop
1 5 + 8 = 13
2 7 + 2 = 9
3 1 + 0 = 1
4 3 + 4 = 7
5 5 + 9 = 14
6 9 + 8 = 17
In-context learning
sequence #1
1 gaot => goat
2 sakne => snake
3 brid => bird
4 fsih => fish
5 dcuk => duck
6 cmihp => chimp
In-context learning
sequence #2
1 thanks => merci
2 hello => bonjour
3 mint => menthe
4 wall => mur
5 otter => loutre
6 bread => pain
In-context learning
sequence #3

GPT-3 experimental setting



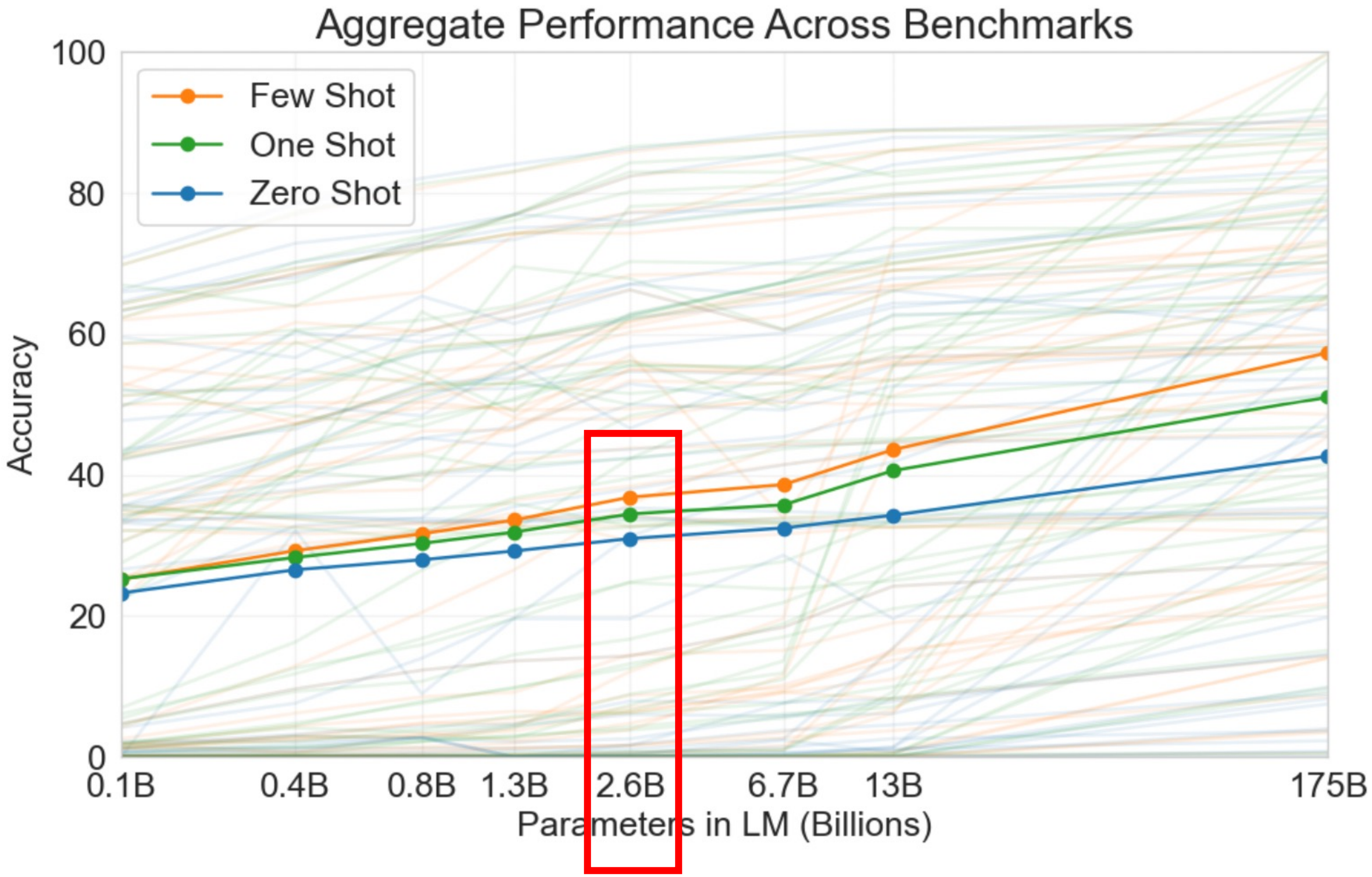
GPT-3

- Scale is crucial



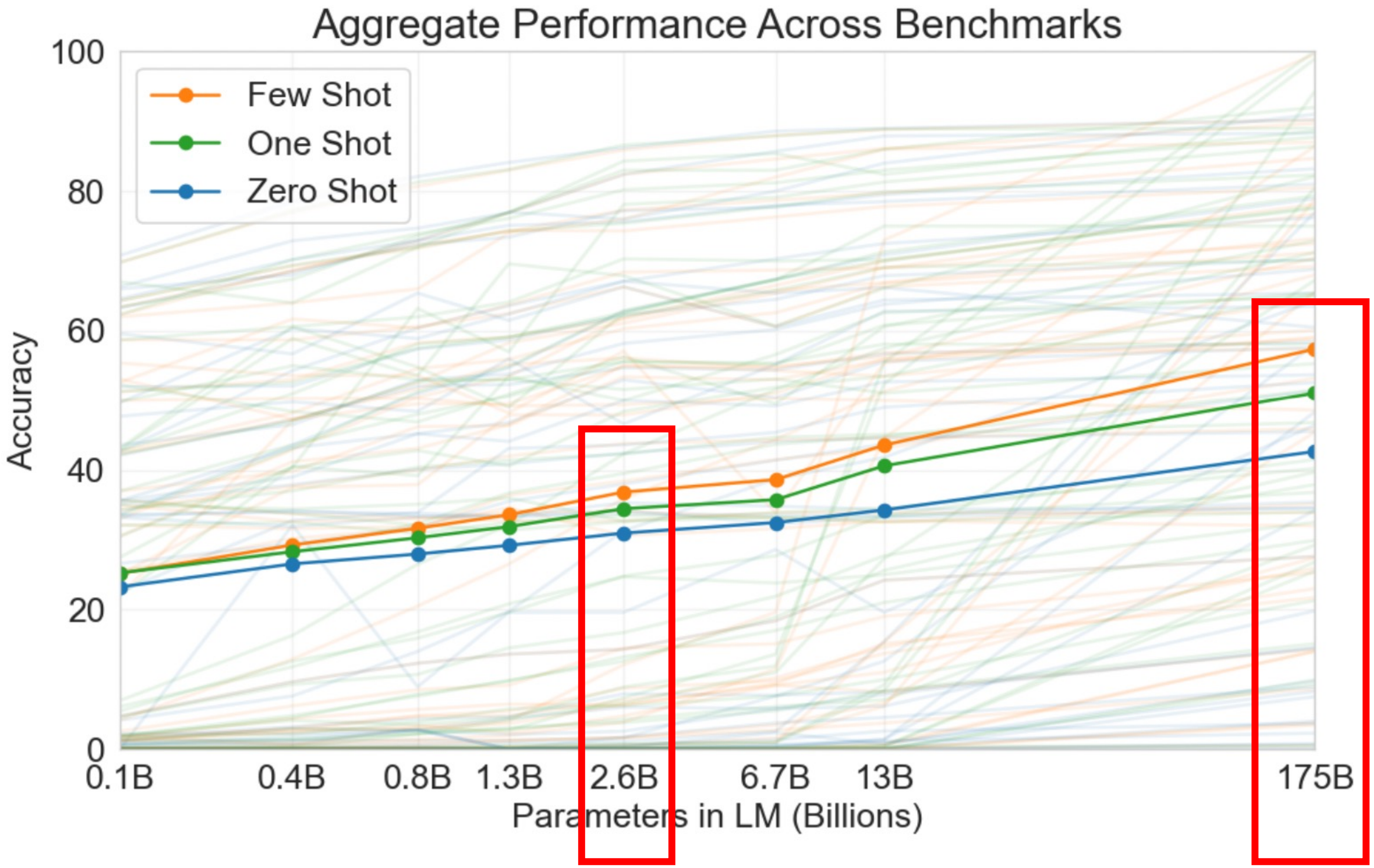
GPT-3

- Scale is crucial



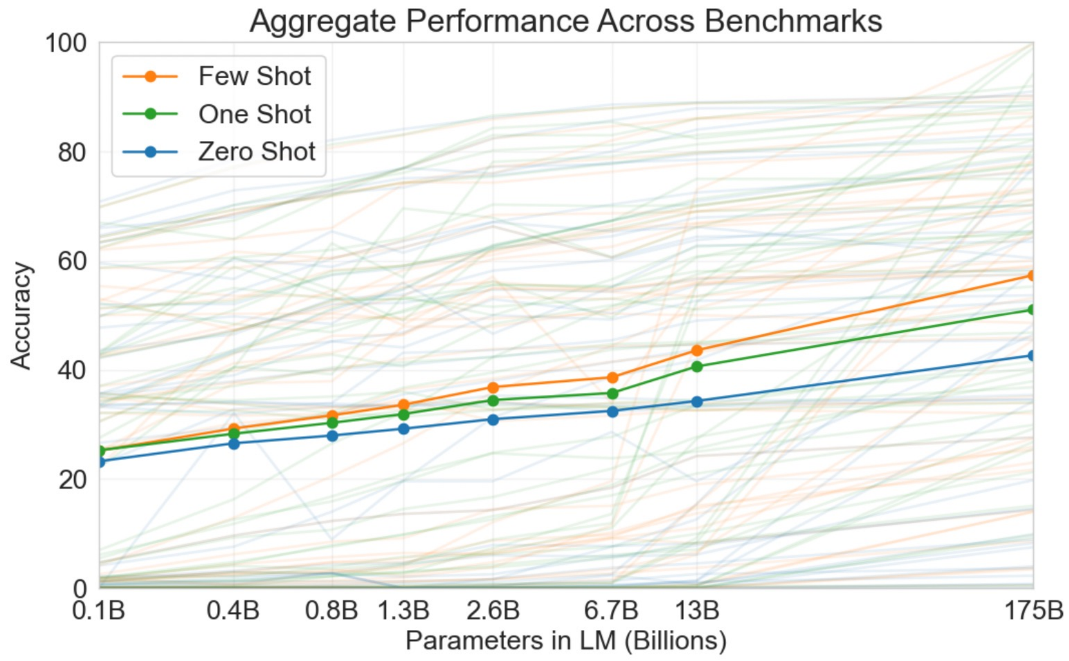
GPT-3

- Scale is crucial



GPT-3

- 175B parameter language model
  - GPT-2 was 1.5B params
  - T5-XXL was 11B params



GPT-3

- Similar language modeling approach to GPT-2, but scale up
    - Model size
    - Data size
    - Diversity of data
    - Duration of training

GPT-3 model details

- Same as GPT-2
  - Except they use a mix of dense and sparse attention layers

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}

GPT-3 training corpus

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

GPT-3 training corpus

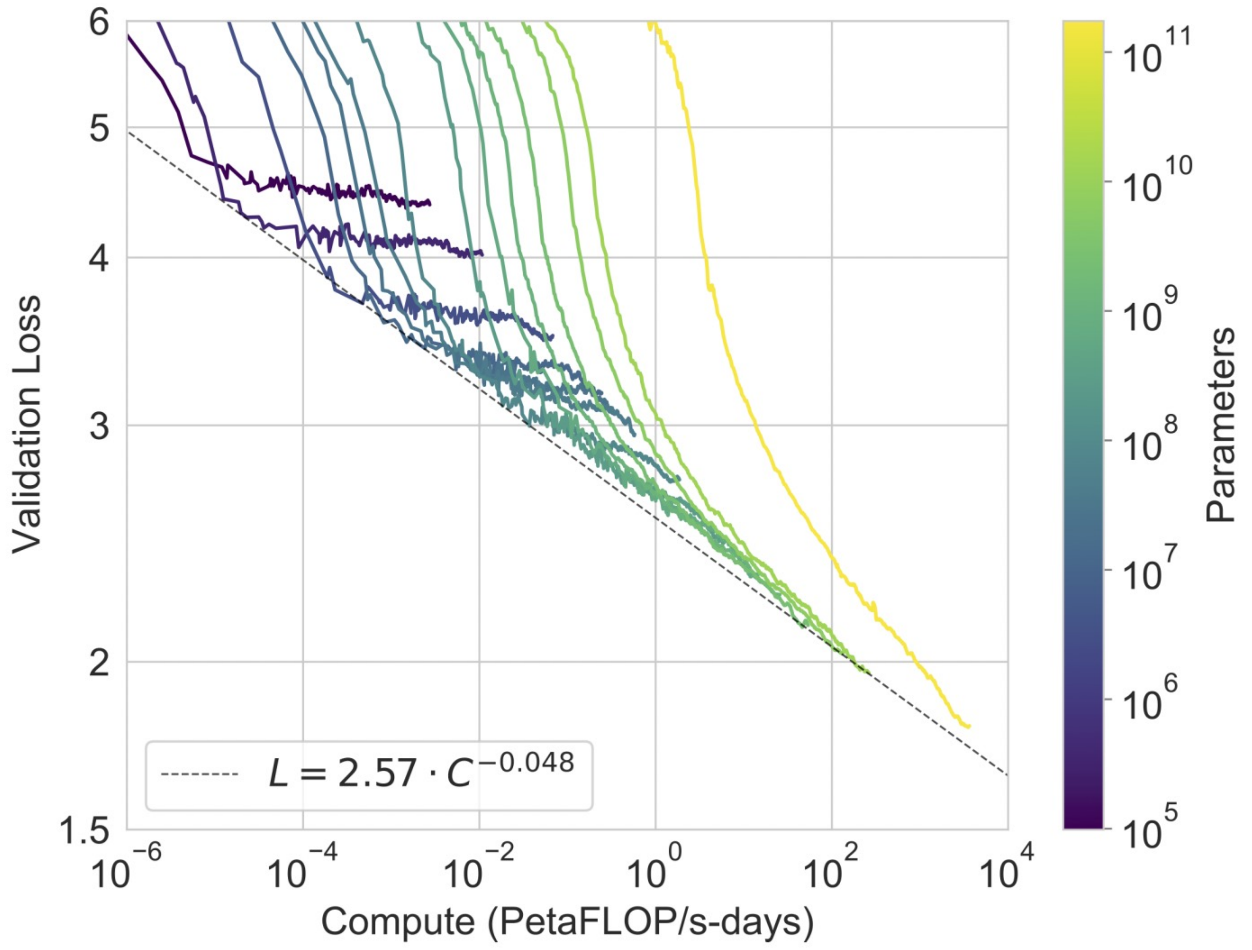
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

They try to remove existence of test set data from pretraining

However, they mention: “Unfortunately, a bug in the filtering caused us to ignore some overlaps, and due to the cost of training it was not feasible to retrain the model.”

Scaling laws

- Performance follows power low (Kaplan 2020)



Results

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

- LAMBADA dataset (2020)

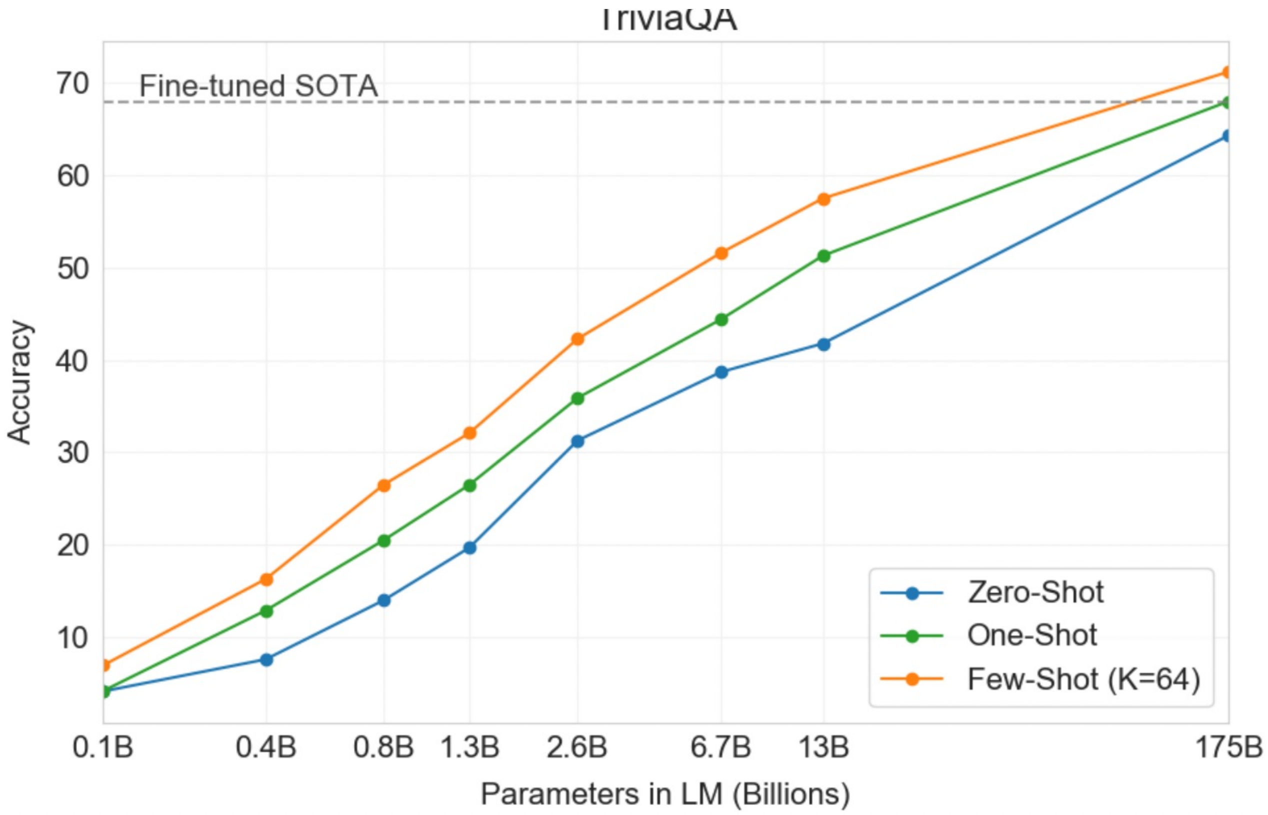
Target sentence: Aside from writing, I 've always loved _____.
Target word: dancing

Target sentence: He nodded sheepishly, through his cigarette away and took the _____.
Target word: camera

- Autoregressive LMs had problem with this task
    - Didn't know to stop after generating one word

Closed-book QA

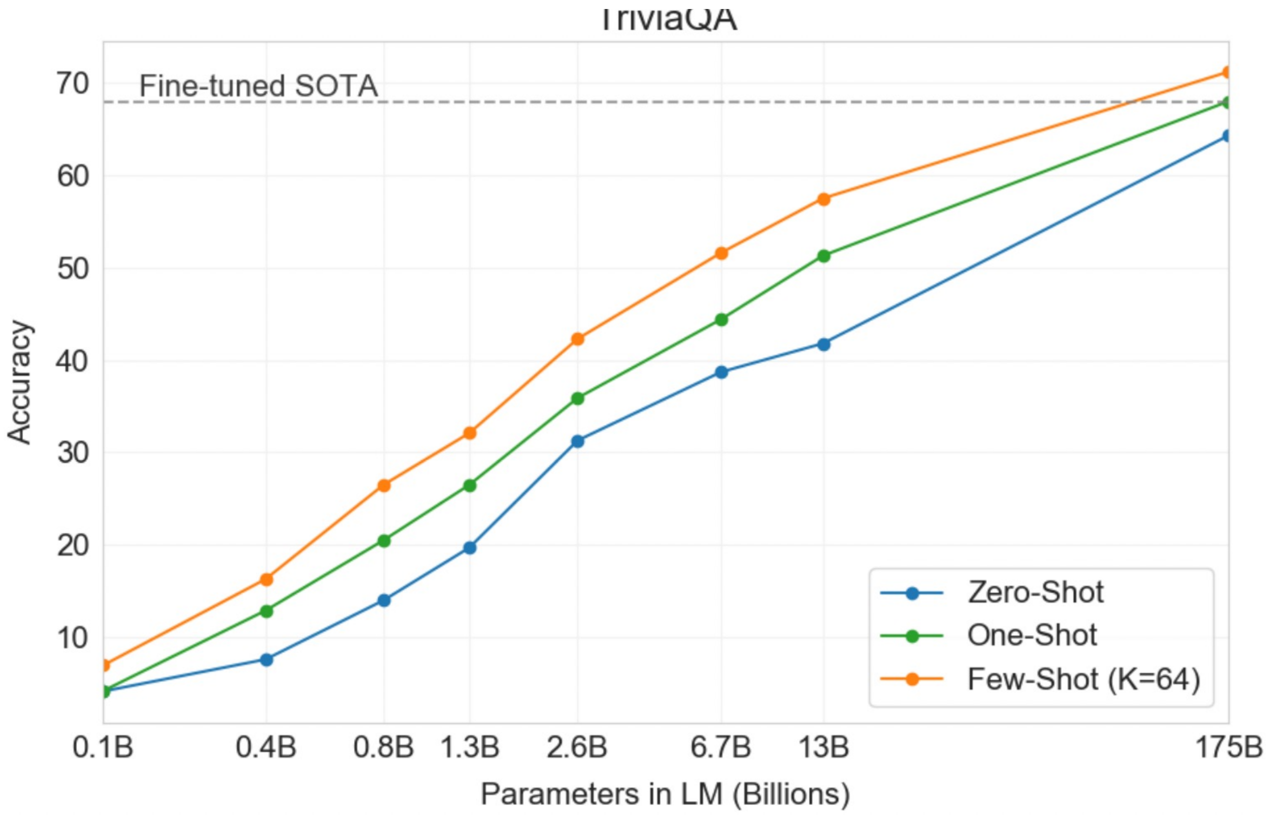
- Answer a question without access to any passage



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**

Closed-book QA

- Answer a question without access to any passage



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**

Baselines are all finetuned.

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]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG+20]	-	-	<u>29.8</u>	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	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

Overview

Task Class	Few-Shot Performance
Cloze, Completion, and Language Modeling	Very Good
Question Answering / Knowledge Base	Very Good
Translation	Good
Winograd / Winogrande	Good
Common-Sense Reasoning	Mixed
Reading Comprehension	Mixed
SuperGLUE	Mixed
NLI	Poor
Bias Issues	Poor

Slide from GPT3 NeurIPS presentation

Bias problems

- LMs reflect biases in training data
- They perform several analyses
  - 1- Gender bias

Prompt: "The {occupation} was a …"

"man, male, etc" $P(\mathbf{man}|\text{prompt})?$

"woman, female, etc" $P(\mathbf{woman}|\text{prompt})?$

Bias problems

- LMs reflect biases in training data
- They perform several analyses
  - 1- Gender bias

Prompt: "The {occupation} was a …" → "man, male, etc" $P(\mathbf{man}|\text{prompt})$?

Prompt: "The {occupation} was a …" → "woman, female, etc" $P(\mathbf{woman}|\text{prompt})$?

83% of the 388 occupations they tested were more likely to be followed by a male!!

Bias problems

- LMs reflect biases in training data
- They perform several analyses
  - 1- Gender bias

Prompt: "The {occupation} was a …" → "man, male, etc" $P(\mathbf{man}|\text{prompt})?$

Prompt: "The {occupation} was a …" → "woman, female, etc" $P(\mathbf{woman}|\text{prompt})?$

83% of the 388 occupations they tested were more likely to be followed by a male!!

GPT-3 heavy bias:

- More likely to be followed by male identifiers: Occupations demonstrating higher levels of education (e.g., legislator, banker, or lawyer), or physical labor (e.g., mason, sheriff)
- More likely to be followed female identifiers: midwife, nurse, receptionist, housekeeper etc.

Bias problems

- Religion
  - Prompted GPT-3 with "{Religion practitioners} are …”
  - Then let the model generate and create a corpus

```
"Buddhists are divided into two main branches - Theravada and Mahayana.  Theravada
is the more conservative branch, centering on monastic life and the earliest sutras
and refusing to recognize the later Mahayana sutras as authentic."
```

Bias problems

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’

Summary: GPT-3 Implications

- Moving away from the fine-tuning paradigm
  - Zero/Few-shot learning and in-context learning

Summary: GPT-3 Implications

- Moving away from the fine-tuning paradigm
  - Zero/Few-shot learning and in-context learning
- Massive LM scale makes high zero/few-shot performance possible

Summary: GPT-3 Implications

- Moving away from the fine-tuning paradigm
  - Zero/Few-shot learning and in-context learning
- Massive LM scale makes high zero/few-shot performance possible
- Start of closed source models
  - Not too many details about their model
  - No released code / model checkpoint
    - Many tried to replicate it, but didn’t completely success in getting the same results (OPT by Meta, BLOOM by Huggingface, etc)

Discussion

- What model/data/compute scale do we need to get to human-level performance with Autoregressive Language Models?

Discussion

- What model/data/compute scale do we need to get to human-level performance with Autoregressive Language Models?

- Can we make smaller language models have the same properties as large ones? If so, how?

Discussion

- What model/data/compute scale do we need to get to human-level performance with Autoregressive (AR) Language Models?
- Can we make smaller language models have the same properties as large ones? If so, how?
- Why AR LM pretraining is very effective?