GPT-3: LANGUAGE MODELS ARE FEW-SHOT LEARNERS

LLM Reading Group

GPT-3 Applications: ChatBot

```
Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many legs does a frog have?
A: A frog has four legs.

Q: Are there any animals with three legs?
A: No, there are no animals with three legs.

Q: Why don't animals have three legs?
A: Animals don't have three legs because they would fall over.
```

GPT-3 Applications: Summarization

Type your text below and we will explain it back to you in a simplified fashion:

Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art finetuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model.

Submit

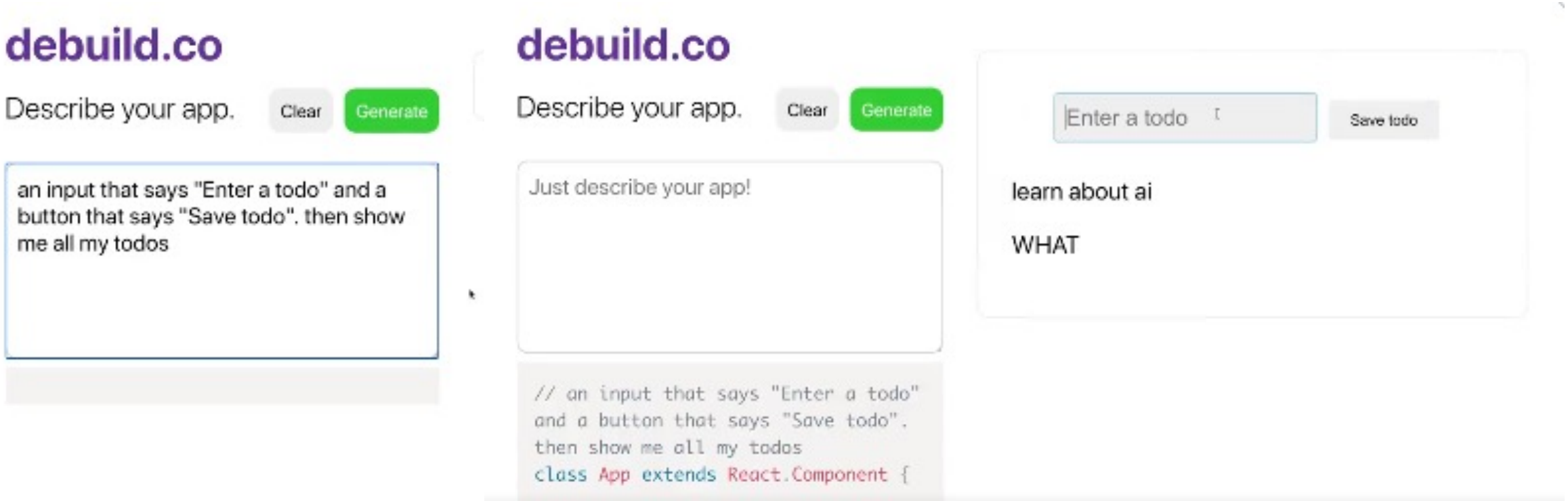
If you consistently get blank responses, try shortening your input!

Our explanation:

If you make a language model with more parameters, it will be able to generalize better to new tasks. We trained a language model with more parameters than any other language model before, and tested it on new tasks, and it did well. This is cool because it suggests that language models can be used to learn representations that can generalize to new tasks, even if they have never seen those tasks before. This is useful for machine learning in general, because it suggests that language models can be used as "transfer" or "architecture" learning for other machine learning tasks.

GPT-3 Applications: Building Apps

- Demo



From GPT to GPT-4

06/2017 Attention Is All You Need — **Transformer Architecture**

06/2018 Improving Language Understanding (GPT) — **117M parameters — ~400MB in size** — **Pre-train and Fine-tune**

02/2019 Language Models are Unsupervised Multitask Learners (GPT-2)— **1.5B parameters — ~5GB in size** — **Zero-shot**

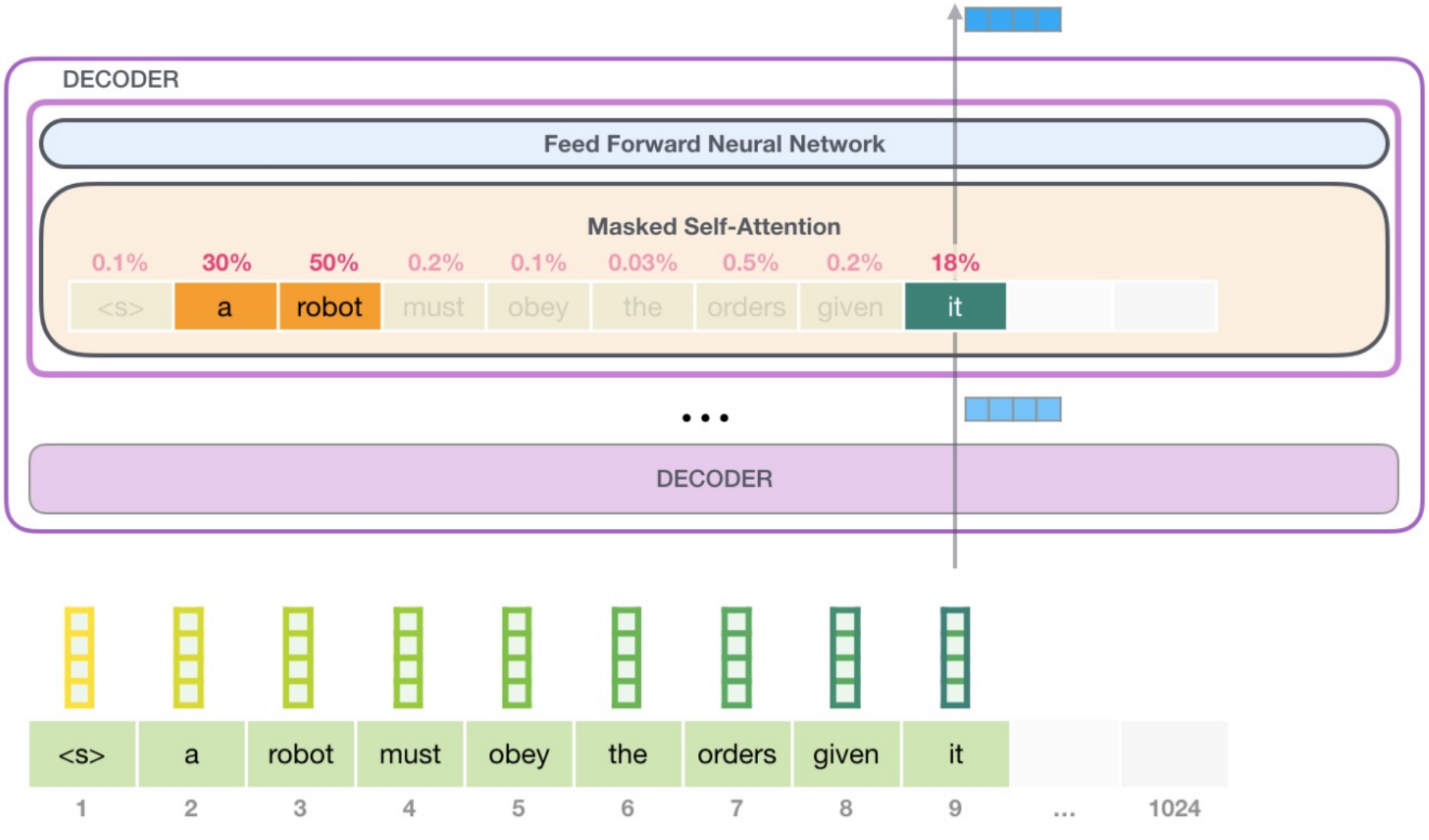
05/2020 Language Models are Few-Shot learners (GPT-3) — **175B parameters — ~500GB in size** — **In-context few-shot**

03/2022 Training language models to follow instructions with human feedback (GPT-3.5/InstructGPT) – over **350B parameters** — **Human Alignment**

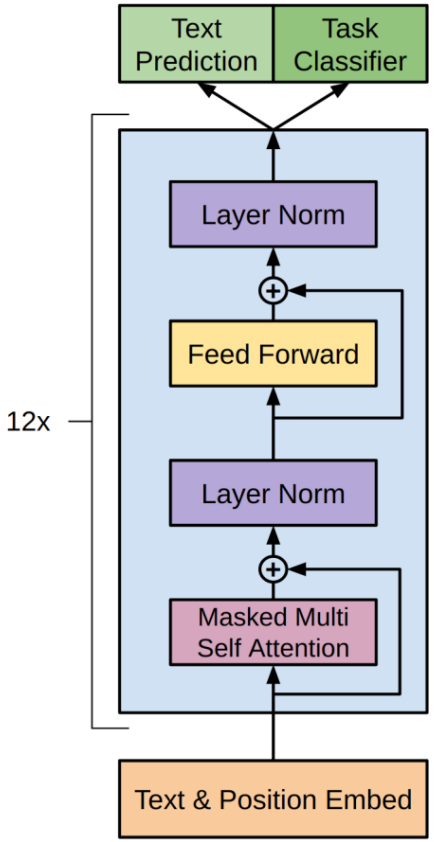
11/2022 **ChatGPT Release**

03/2023 Large-scale Multimodal model with better post-training alignment (GPT-4) – over **1.5T parameters** — **Multi-modal**

GPT: Predicting the next token

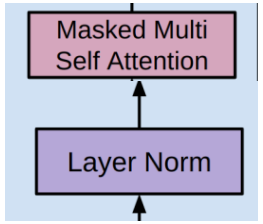


GPT-3 Model Architecture



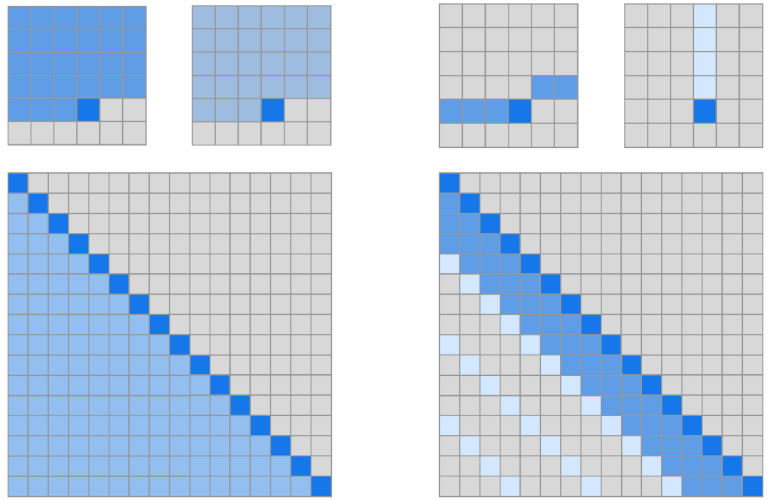
GPT

- **Layer normalization** was moved to the input of each sub-block, and an additional layer normalization was added after the final self-attention block.
- We **scale the weights of residual layers** at initialization by a factor of 1/ √ N where N is the number of residual layers.
- The **vocabulary is expanded** to 50,257. We also increase the context size from 512 to 1024 tokens and a larger batch size of 512 is used.



GPT-2

- Alternating dense and locally banded sparse attention patterns, similar to the Sparse Transformer.



(a) Transformer (b) Sparse Transformer (strided)

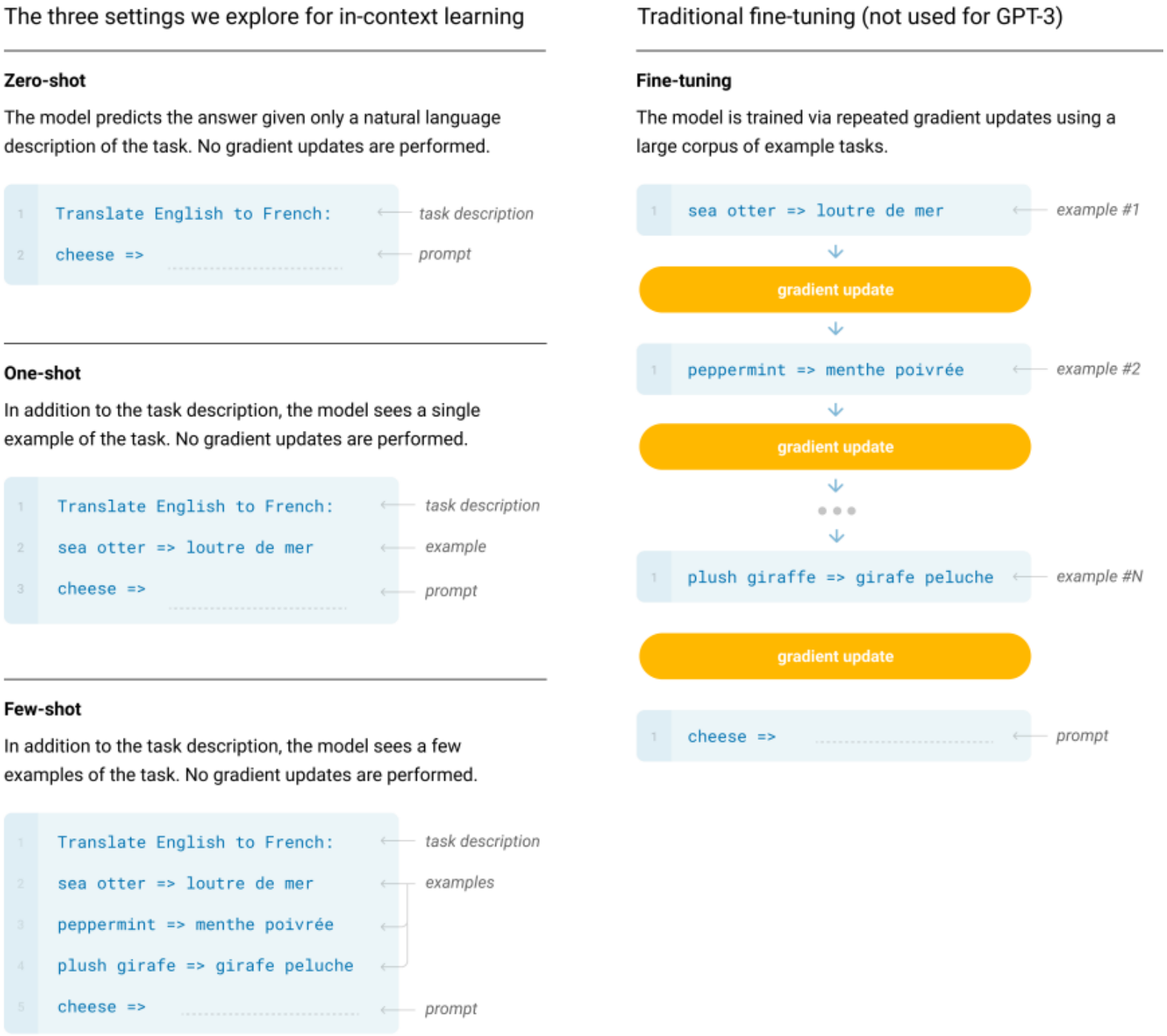
GPT-3

GPT-3: Increasing model size

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}

Compare the model performance across different NLP tasks with an increasing model size.

In-context Learning



Training: Datasets used to train GPT-3

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

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

Evaluation

- For few-shot learning, we evaluate each example in the evaluation set by randomly drawing K examples from that task's training set as conditioning (in-context examples), delimited by 1 or 2 newlines depending on the task.
- K can be any value from 0 to the maximum amount allowed by the model's context window, which is nctx = 2048 for all models and typically fits 10 to 100 examples. Larger values of K are usually but not always better
- On tasks with free-form completion, we use beam search with a beam width of 4 and a length penalty of $\alpha = 0.6$.

Task Phrasing and Specifications

Context →	Title: The_Blitz Background: From the German point of view, March 1941 saw an improvement. The Luftwaffe flew 4,000 sorties that month, including 12 major and three heavy attacks. The electronic war intensified but the Luftwaffe flew major inland missions only on moonlit nights. Ports were easier to find and made better targets. To confuse the British, radio silence was observed until the bombs fell. X- and Y-Gerät beams were placed over false targets and switched only at the last minute. Rapid frequency changes were introduced for X-Gerät, whose wider band of frequencies and greater tactical flexibility ensured it remained effective at a time when British selective jamming was degrading the effectiveness of Y-Gerät. Q: How many sorties were flown in March 1941? A: 4,000 Q: When did the Luftwaffe fly inland missions? A:
Target Completion →	only on moonlit nights

Context →	The trend toward lower rents may seem surprising given that some communities in New York are bemoaning the loss of favorite local businesses to high rents. But, despite the recent softening, for many of these retailers there's still been too big a jump from the rental rates of the late 1970s, when their leases were signed. Certainly, the recent drop in prices doesn't mean Manhattan comes cheap. question: Manhattan comes cheap. true, false, or neither? answer:
Target Completion →	false

Figure G.30: Formatted dataset example for CB

GPT-3 Strengths and Weaknesses

Task Class		Few-Shot Performance
Cloze, Completion, and Language Modeling		Very Good
Question Answering / Knowledge Base		Very Good
Translation	From non-English to English	Good
Winograd / Winogrande	commonsense reasoning	Good
Common-Sense Reasoning		Mixed
Reading Comprehension		Mixed
SuperGLUE	Language understanding	Mixed
NLI	Natural language inference (entailment/contradiction/neutral)	Poor
Bias Issues		Poor

Few-Shot Learning Improves With Scale

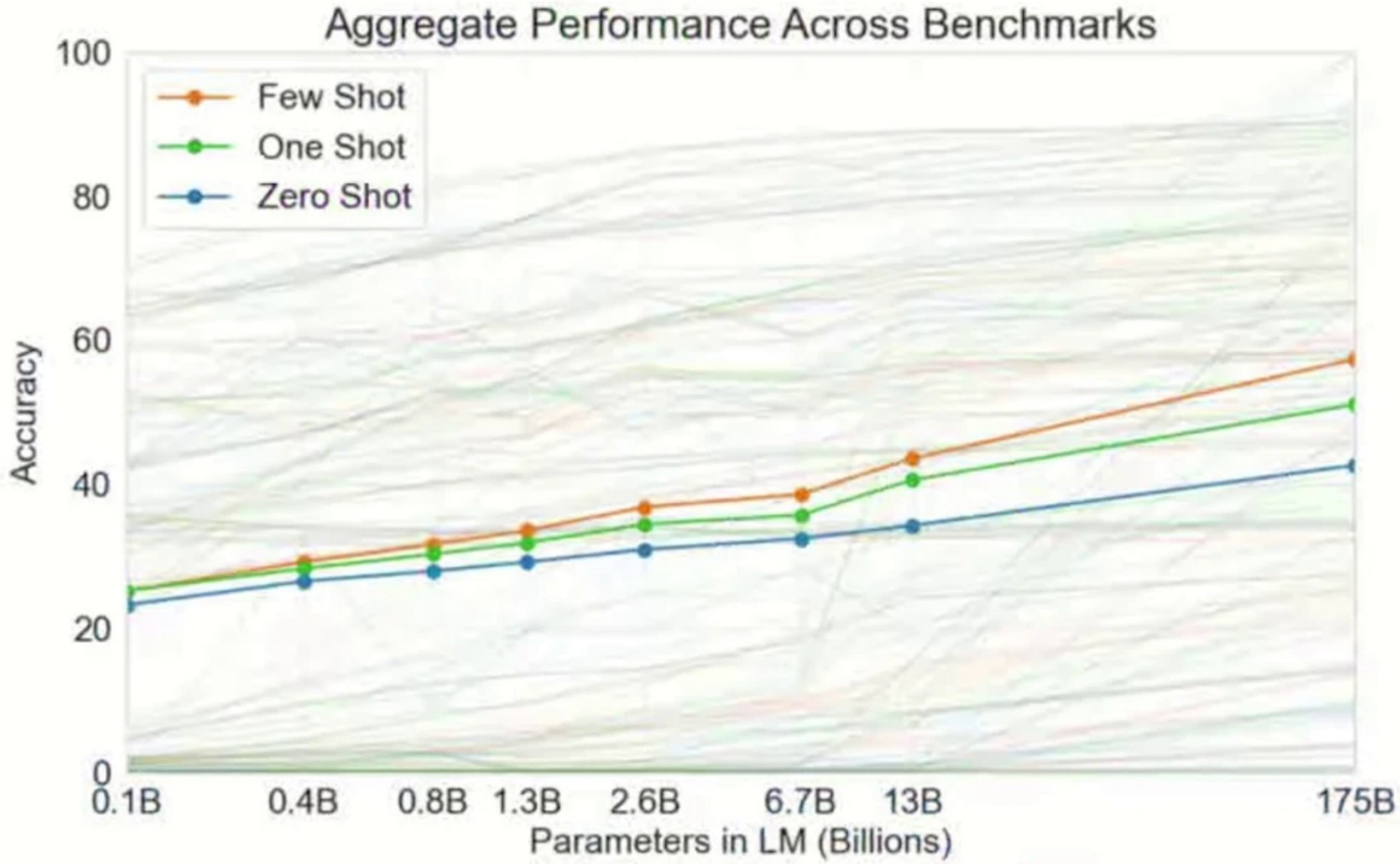


Figure 1.3: Aggregate performance for all 42 accuracy-denominated benchmarks While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning. See Figure 3.8 for a more detailed analysis on SuperGLUE, a standard NLP benchmark suite.

Cloze, Completion, LM

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

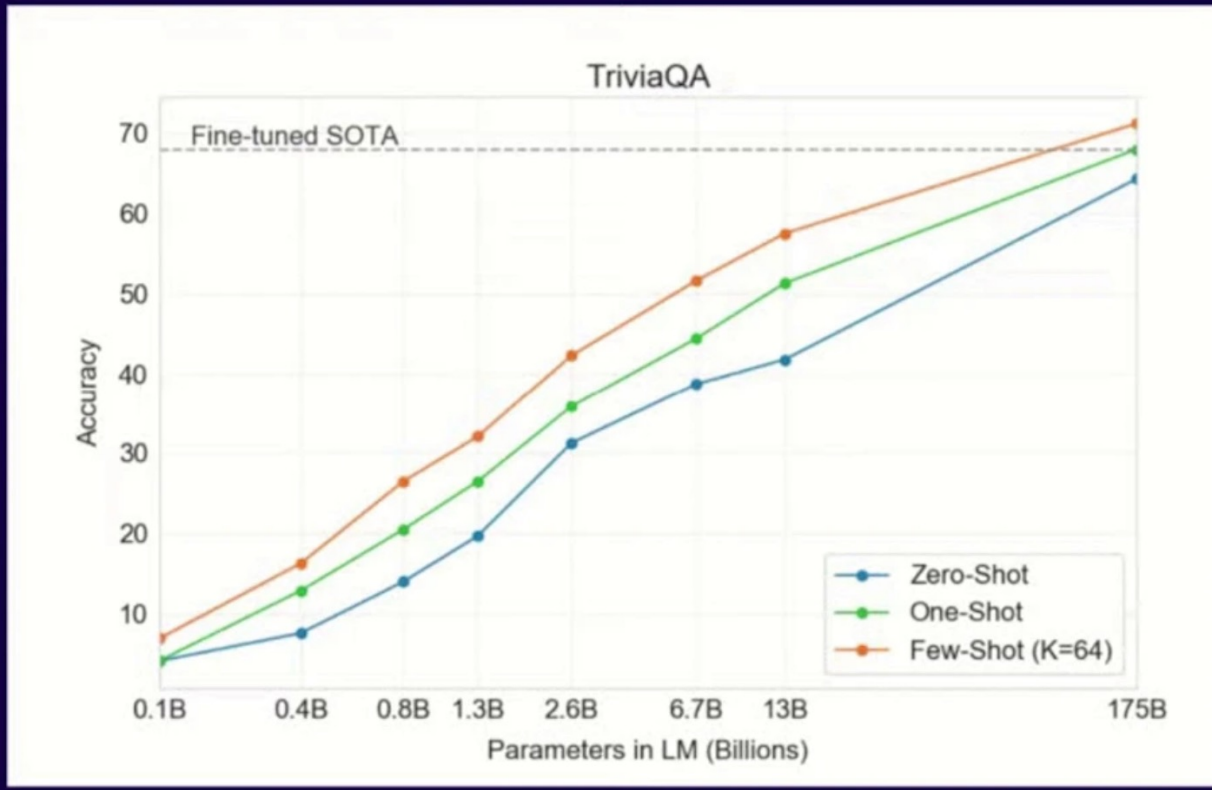
Few-Shot learning addresses specification ambiguity in the LAMBADA task

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

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

Closed-Book Q&A

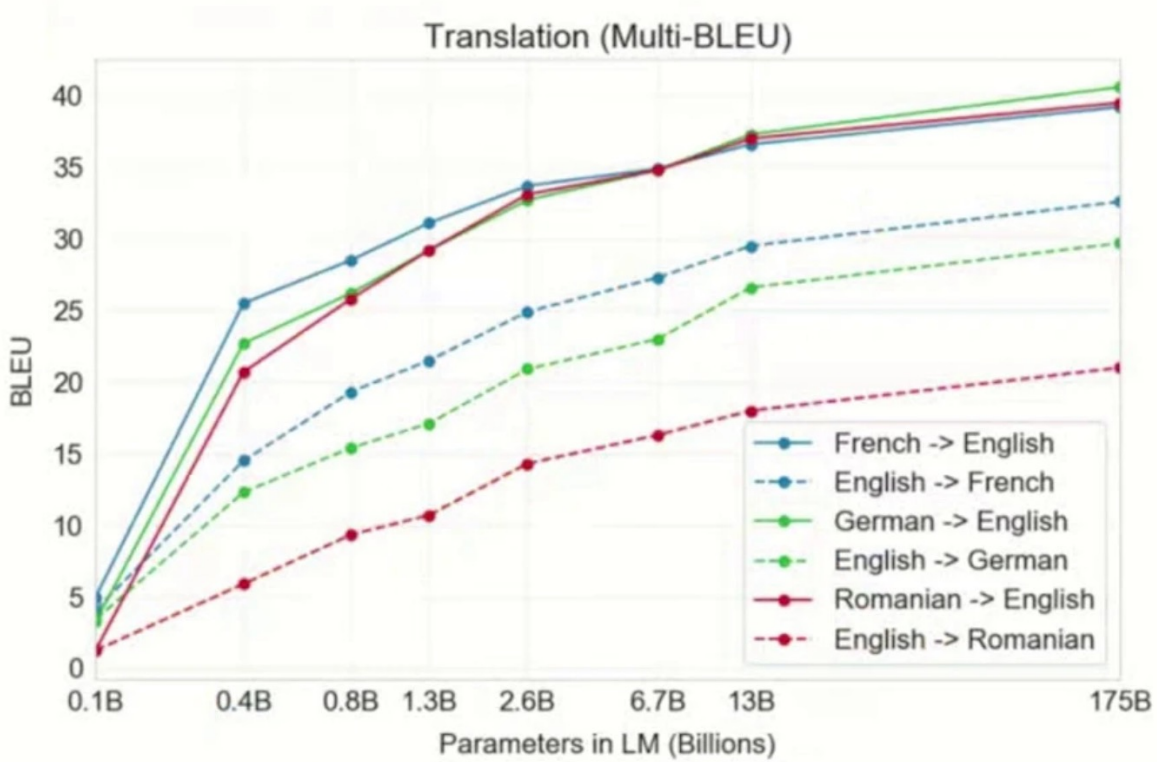
GPT-3 few-shot in the closed book setting is competitive with fine-tuned methods and even with open-book performance.



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

GPT-3 few-shot is often competitive with *supervised* SOTA when translating *into* English and unsupervised SOTA when translating *from* English.

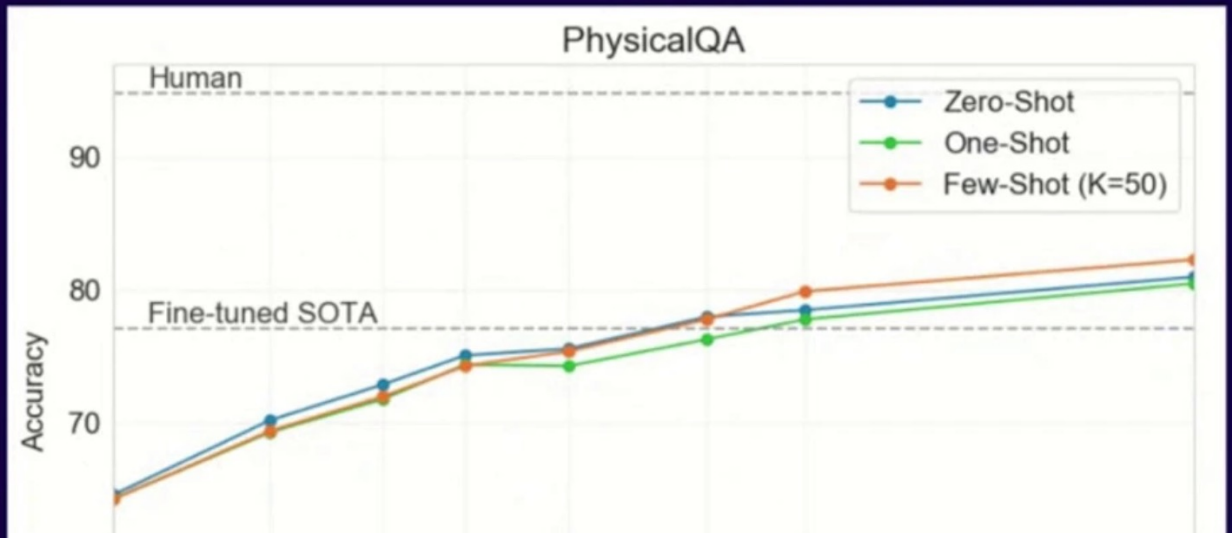


Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3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

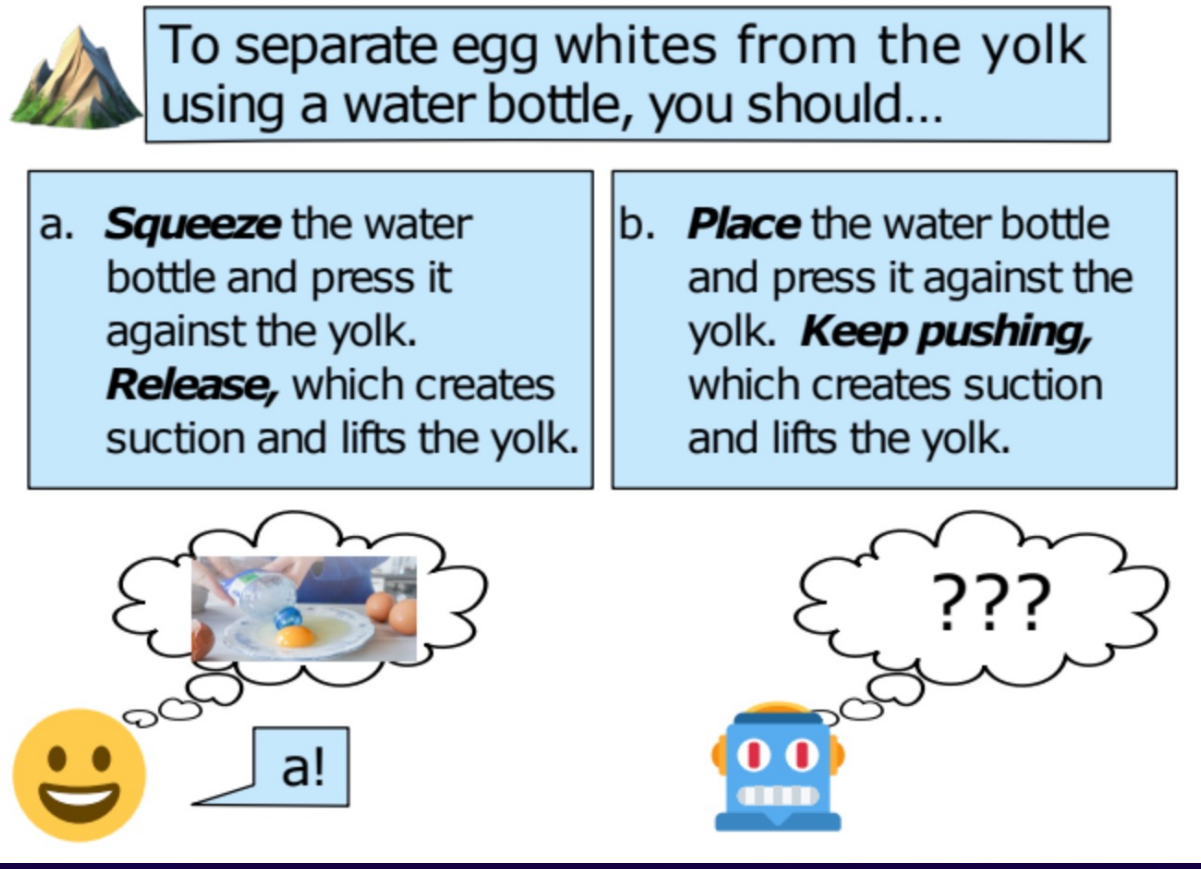
Common-Sense

Mixed: strong performance on PiQa, middling on other common sense QA. Mixed message on need for “grounding”.

PIQA: Physical Interaction: Question Answering



Setting	PIQA	ARC (Easy)
Fine-tuned SOTA	79.4	**92.0**[KKS+
GPT-3 Zero-Shot	**80.5***	68.8
GPT-3 One-Shot	**80.5***	71.2
GPT-3 Few-Shot	**82.8***	70.1



Reading Compreh

Mixed: strong on CoQa, surprisingly weak on others. DROP actually strong as SOTA are models specialized for arithmetic.

Setting	CoQA	DROP	QuAC	SQu
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5
GPT-3 One-Shot	84.0	34.3	43.3	65.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8

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

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

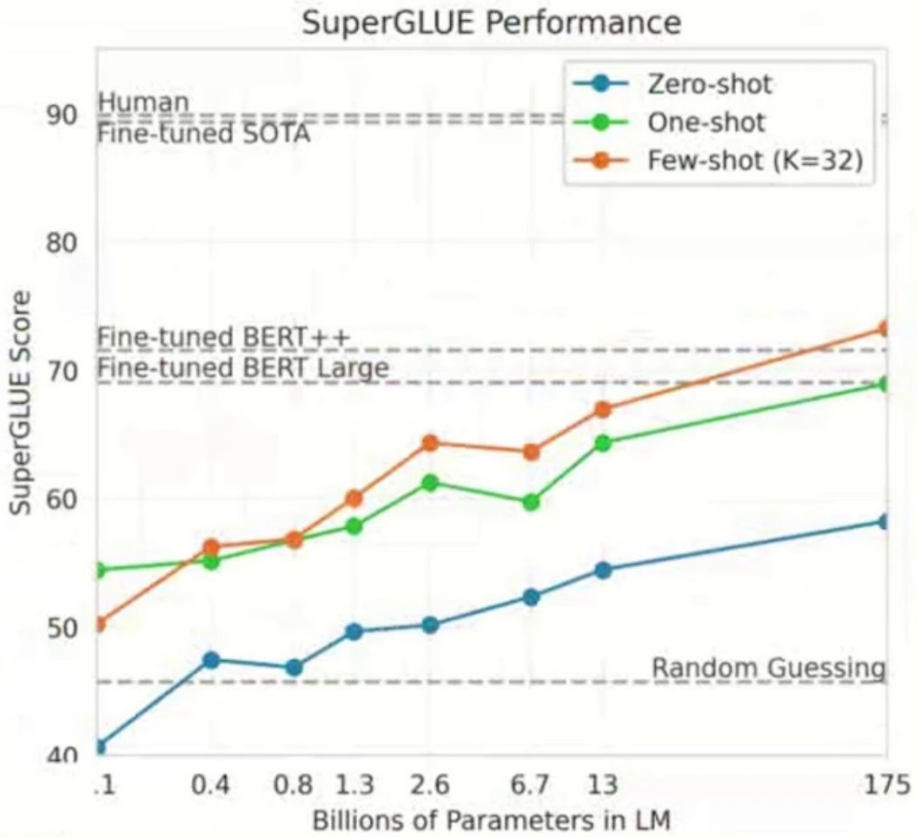
Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

SuperGLUE

Mixed: near-SOTA on COPA, Record, weak on WiC, RTE. Big gains to few-shot.

SuperGLUE: Super general language understanding evaluation



	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

COPA: Choice Of Plausible Alternatives

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.

Premise: I tipped the bottle. What happened as a RESULT?
Alternative 1: The liquid in the bottle froze.
Alternative 2: The liquid in the bottle poured out.

Premise: I knocked on my neighbor's door. What happened as a RESULT?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left his house.

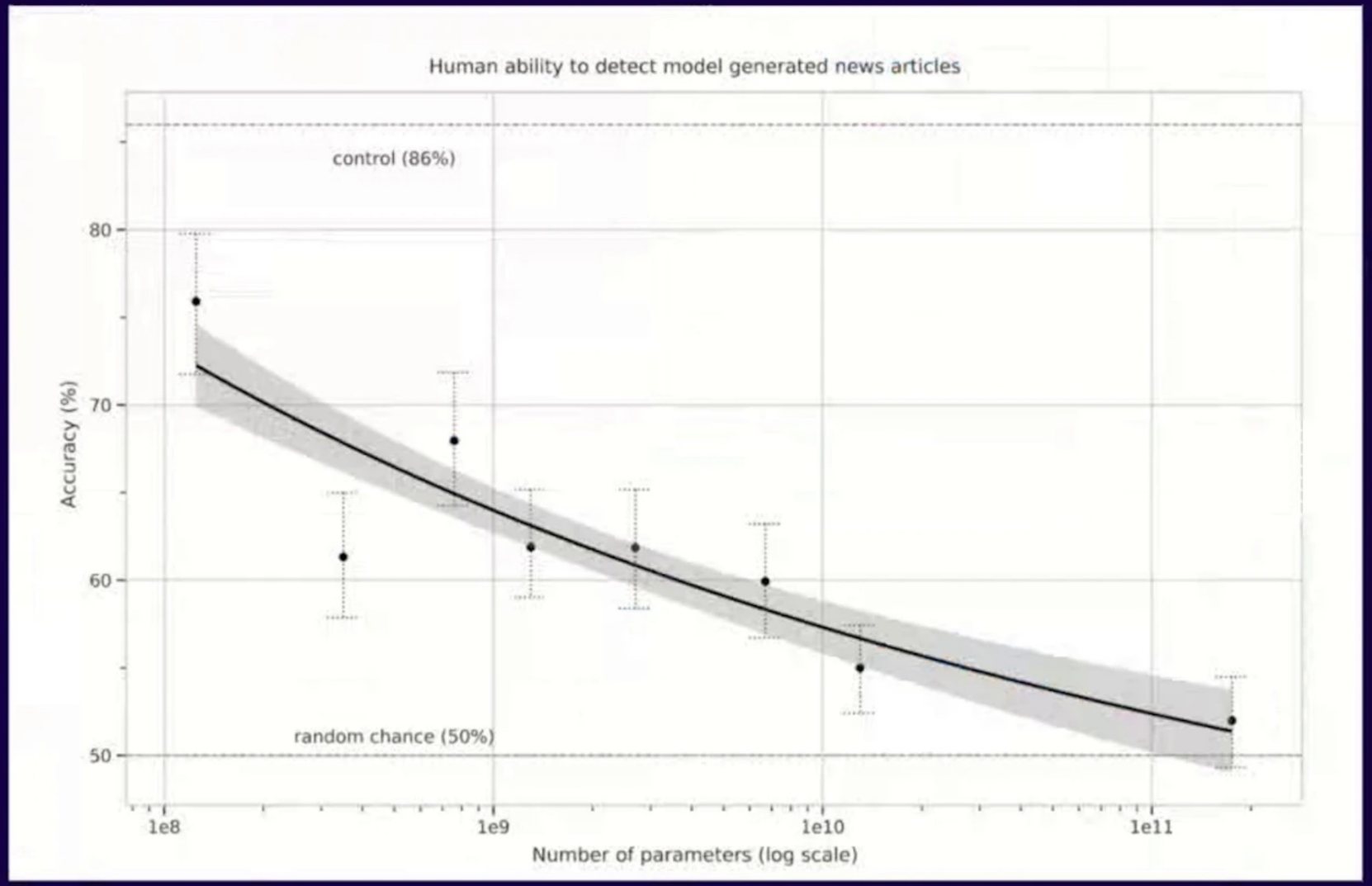
Arithmetic

(To be updated in paper): Adding commas gives GPT-3 strong performance on 4-6 digit arithmetic, e.g. “34,567” vs “34567”.

Task	Accuracy Without Commas	Accuracy With Commas
4 digit addition	25.5%	91.1%
4 digit subtraction	26.9%	89.7%
5 digit addition	9.3%	90.2%
5 digit subtraction	9.9%	82.2%
6 digit addition	3%	78.5%
6 digit subtraction	3%	73.9%

Thanks to Gwern Branwen for first noticing this difference.

Discriminating News Articles



Broader Impacts and Bias

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'

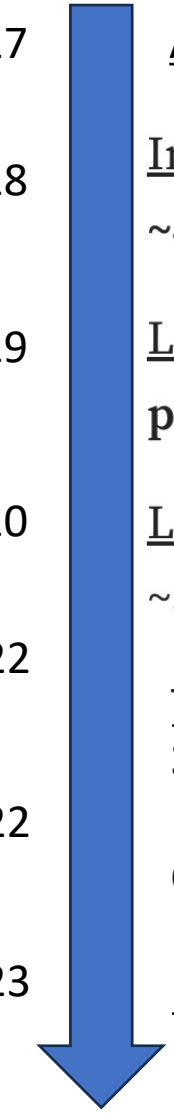
Table 6.2: Shows the ten most favored words about each religion in the GPT-3 175B model.

GPT-3 model is biased and tend to reflect stereotypes present in their training data.

GPT-3 Strengths and Weaknesses

Task Class		Few-Shot Performance
Cloze, Completion, and Language Modeling		Very Good
Question Answering / Knowledge Base	Open-Book QA	Very Good
Translation	From non-English to English	Good
Winograd / Winogrande	commonsense reasoning	Good
Common-Sense Reasoning		Mixed
Reading Comprehension		Mixed
SuperGLUE	Language understanding	Mixed
NLI	Natural language inference (entailment/contradiction/neutral)	Poor
Bias Issues		Poor

From GPT to GPT-4



06/2017 Attention Is All You Need — **Transformer Architecture**

06/2018 Improving Language Understanding (GPT) — **117M parameters — ~400MB in size** — **Pre-train and Fine-tune**

02/2019 Language Models are Unsupervised Multitask Learners (GPT-2)— **1.5B parameters — ~5GB in size** — **Zero-shot**

05/2020 Language Models are Few-Shot learners (GPT-3) — **175B parameters — ~500GB in size** — **In-context few-shot**

03/2022 Training language models to follow instructions with human feedback (GPT-3.5/InstructGPT) – over **350B parameters** — **Human Alignment**

11/2022 **ChatGPT Release**

03/2023 Large-scale Multimodal model with better post-training alignment (GPT-4) – over **1.5T parameters** — **Multi-modal**

More Coming Up!