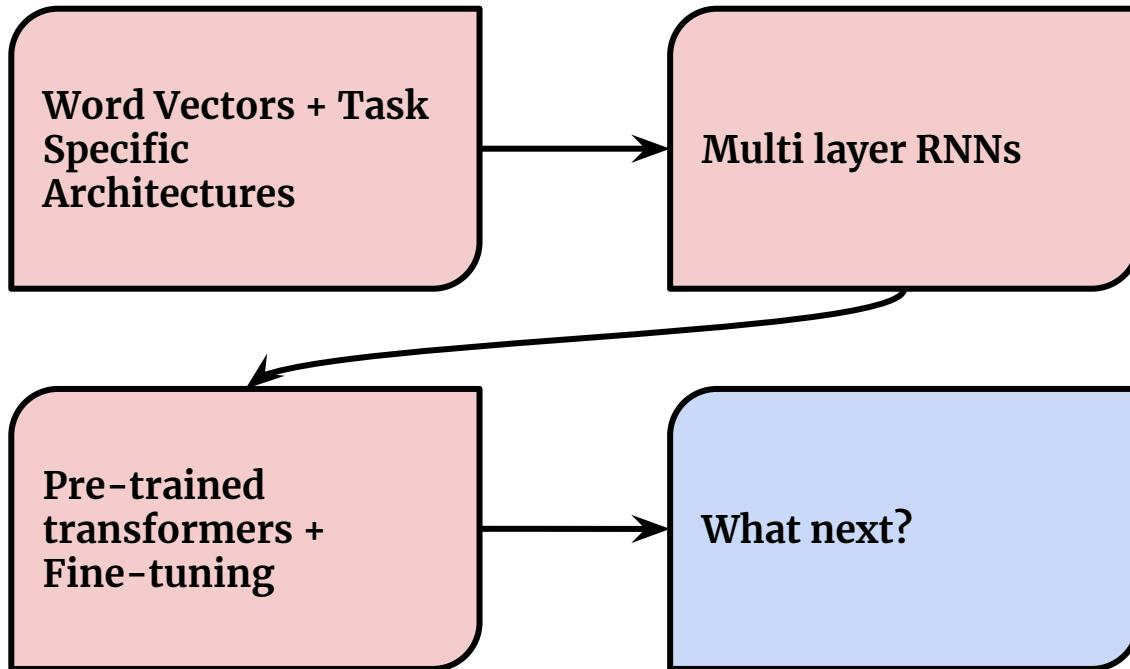


Language Models are Few-Shot Learners (GPT-3)

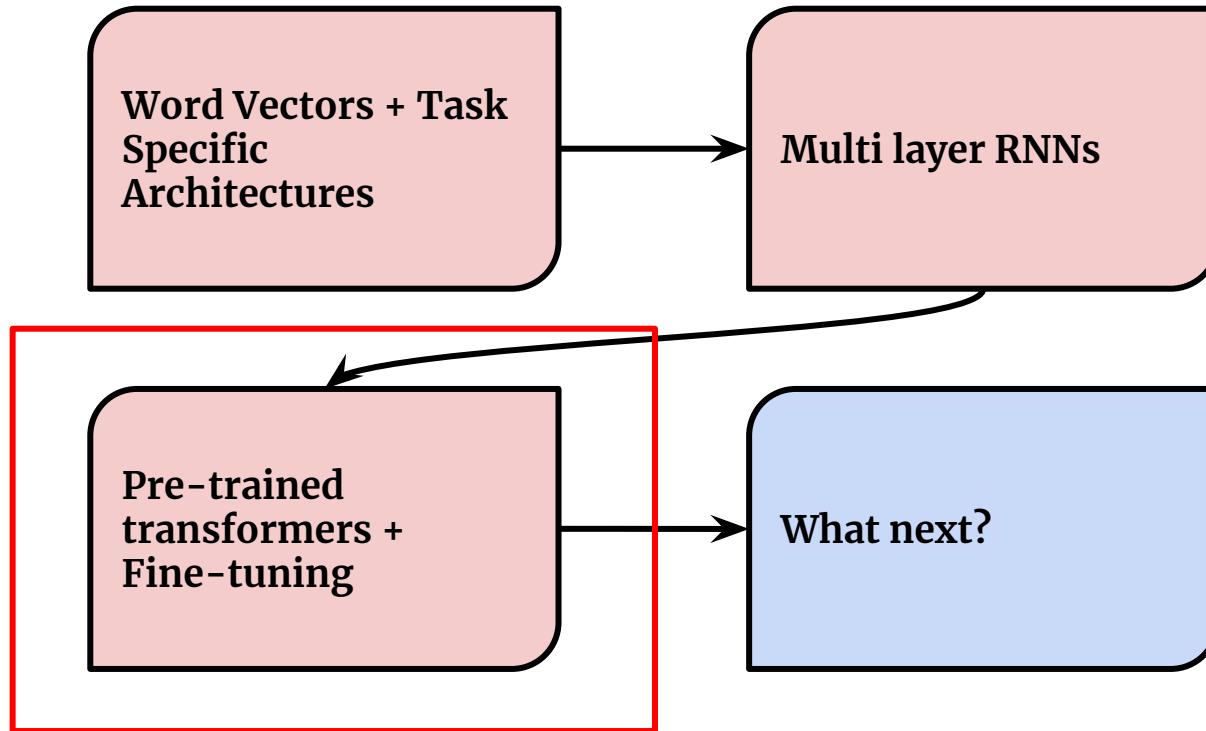
Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher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

Presented by: Sabhya Chhabria & Michael Tang

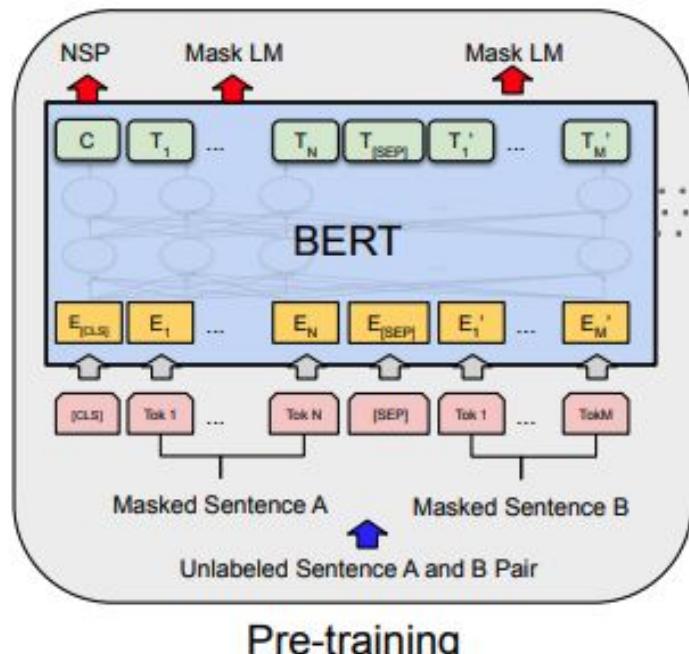
Shifting Paradigms in NLP



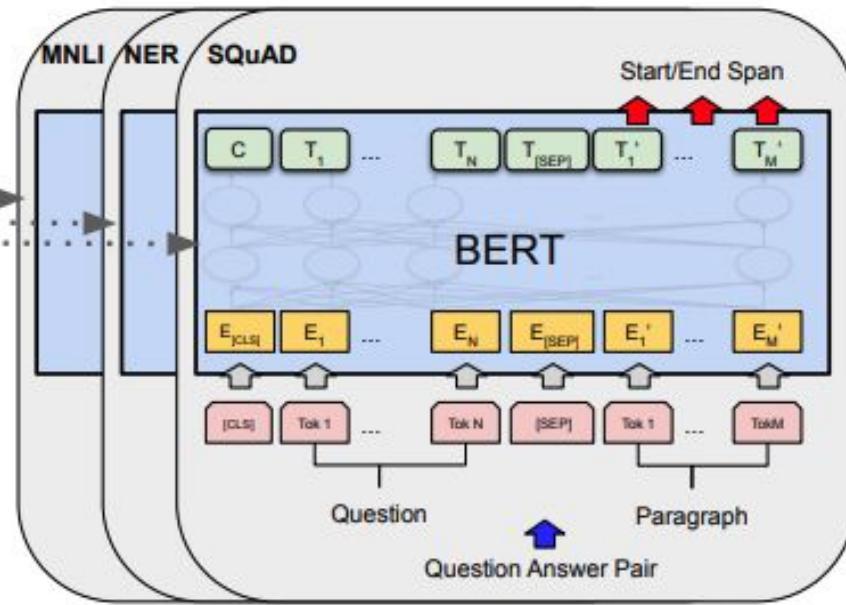
Shifting Paradigms in NLP



Pre-training → Fine-Tuning



Fine-Tuning



Limitations of Pre-training → Fine-Tuning (1)

Practical Issues

- Need large task-specific datasets for fine-tuning
- Collect data for task A → Fine-tune to solve task A → Repeat for task B
→ Repeat for task C → and so on ...
- End up with many “copies” of the same model

Limitations of Pre-training → Fine-Tuning (2)

Potential to exploit spurious correlations (Overfitting)

- Large models fine-tuned on very narrow task distributions
- Evidence suggests: models overfit to training distributions and don't generalize well outside of it (Evidence: [Hendricks et al. 2020](#), [Yogatama et al. 2019](#), [McCoy et al. 2019](#))
- Models are good on datasets, not so good at the underlying task ([Gururangan et al. 2018](#), [Niven et al. 2019](#))

Limitations of Pre-training → Fine-Tuning (3)

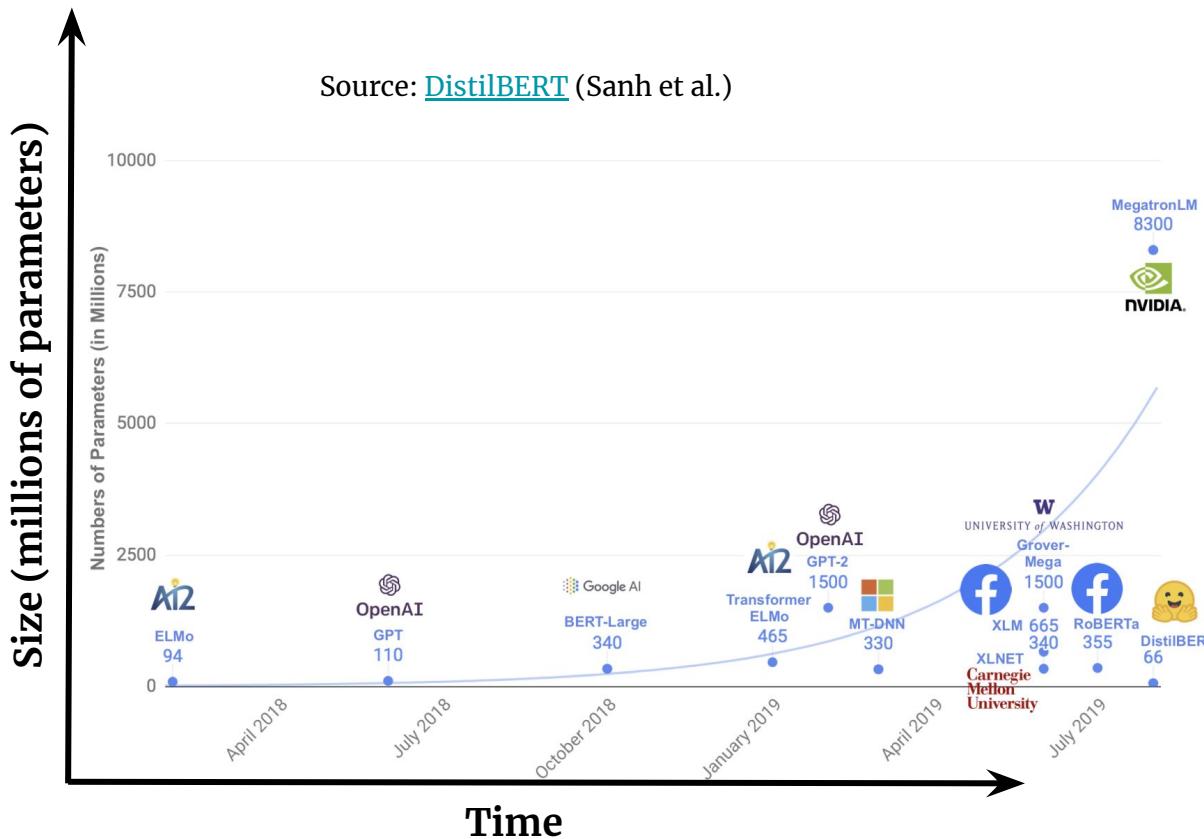
Humans don't need large supervised datasets

- Humans can learn from simple directives
- Allows humans to mix and match skills + switch between tasks easily
- Hope is for NLP systems to one day function with the same fluidity!

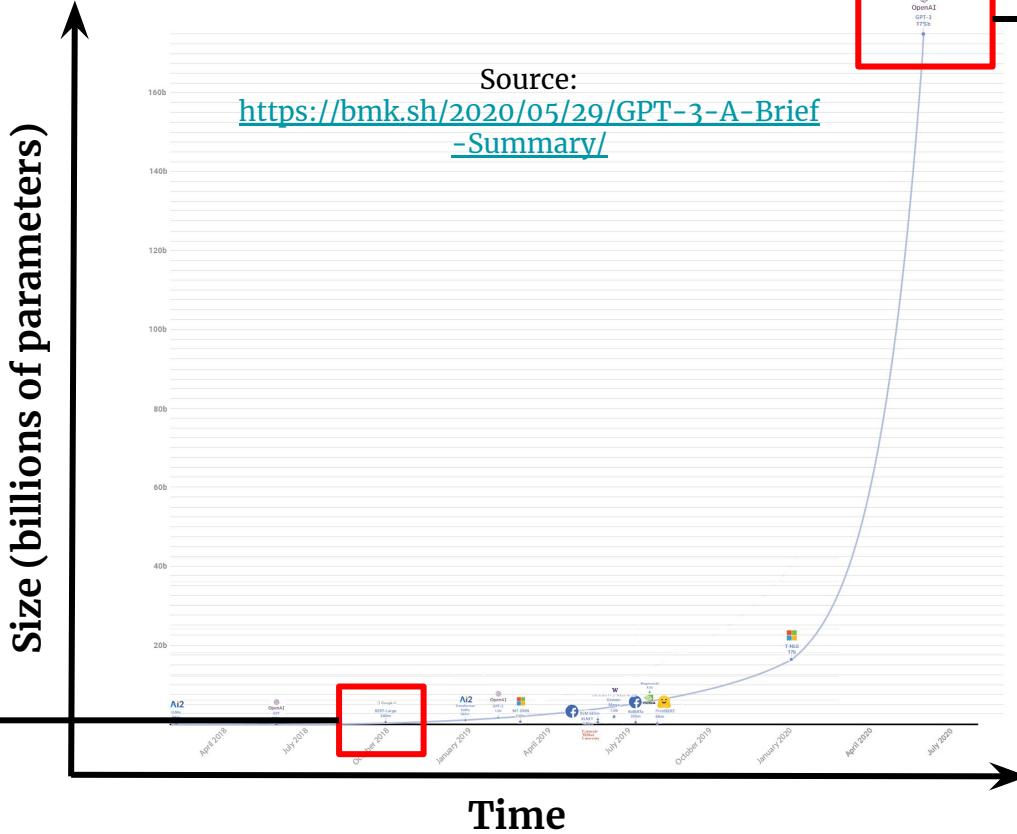
Addressing These Limitations

1. Scaling up 
2. “In Context-Learning” 

LM Landscape pre GPT-3



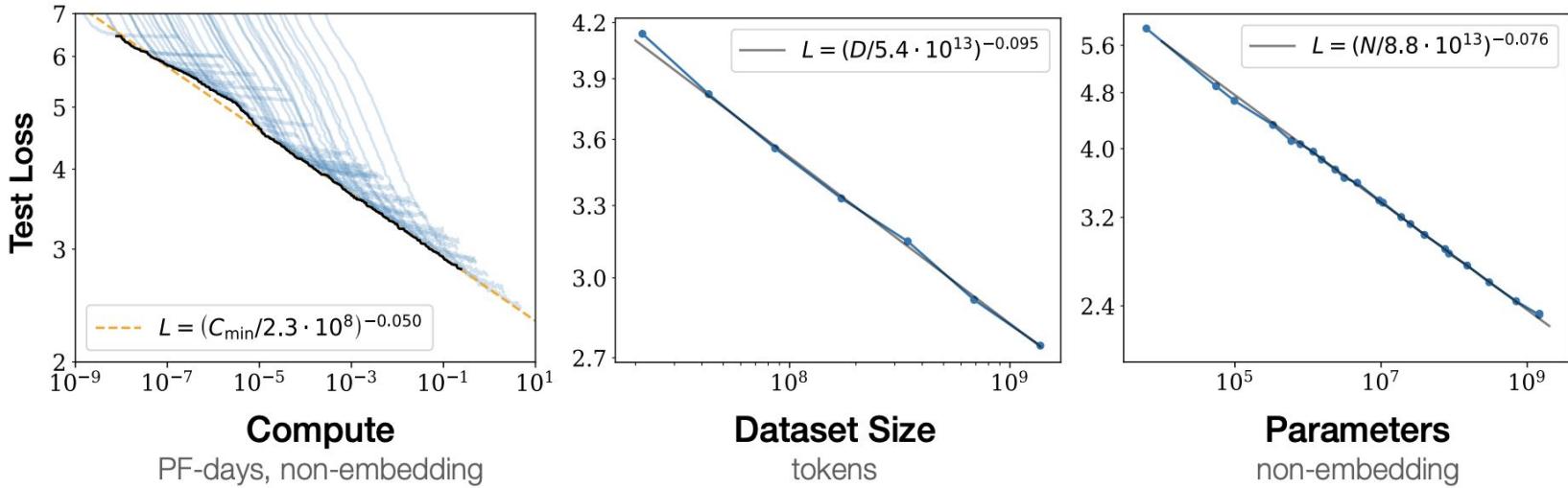
LM Landscape with GPT-3



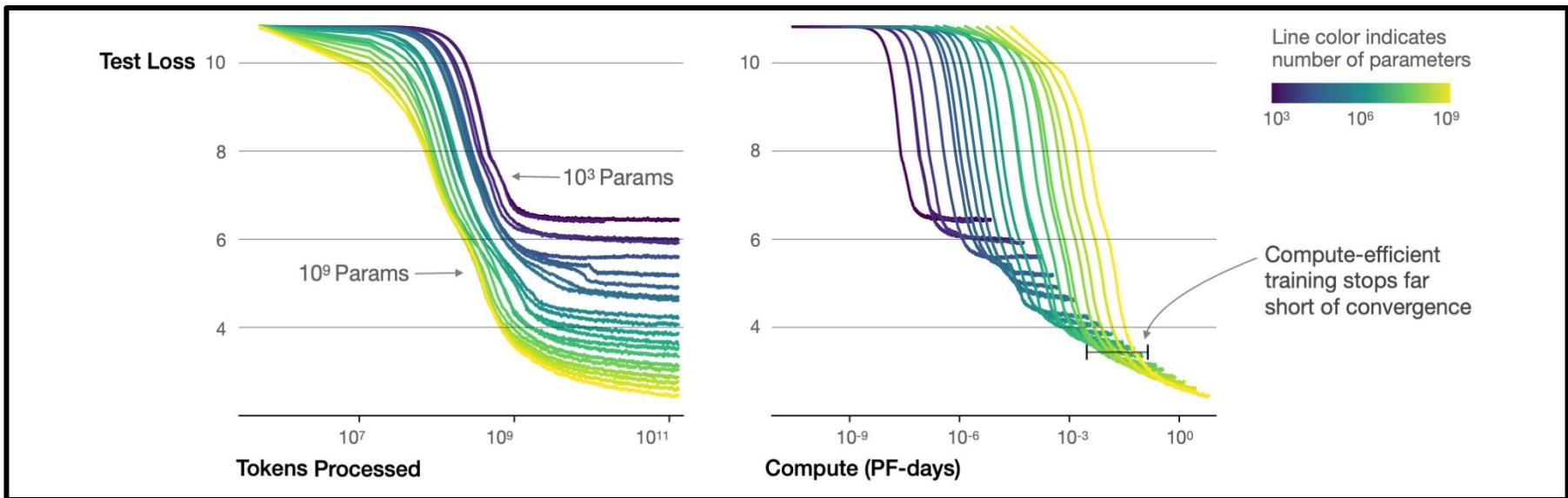
Why Scale?

- Study conducted by OpenAI → **Scaling Laws for Neural Language Models** ([Kaplan et al. 2020](#))
- A few **key findings**:
 - Performance depends strongly on scale, weakly on model shape
 - Smooth **power laws** ($y = ax^k$) b/w empirical performance & N - parameters, D - dataset size, C - compute
 - Transfer improves with test performance
 - Larger models are **more sample efficient**

Bigger is Better!



Bigger is Better!



Addressing These Limitations

1. Scaling up 
2. “In Context-Learning” 

In-Context Learning

Zero-shot
(os)

No Prompt

skicts = sticks

1-shot
(1s)

chiar = chair
skicts = sticks

Few-shot
(FS)

chiar = chair
[...]
pciinc = picnic
skicts = sticks

Prompt

Please unscramble the letters into
a word, and write that word:
skicts = sticks

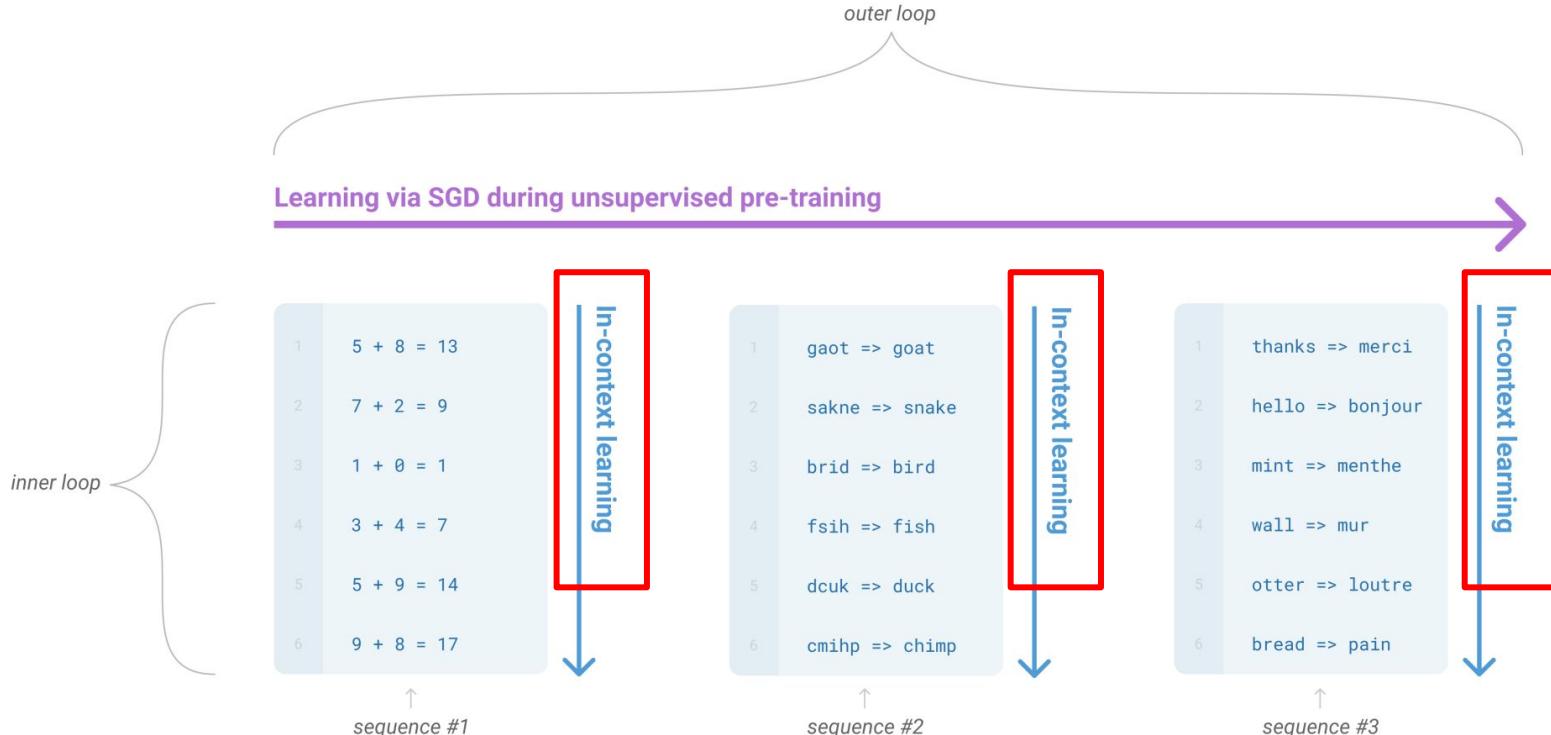
Please unscramble the letters into
a word, and write that word:
chiar = chair
skicts = sticks

Please unscramble the letters into
a word, and write that word:
chiar = chair
[...]
pciinc = picnic
skicts = sticks

In-Context learning is Meta-Learning



In-Context learning is Meta-Learning



In-Context learning is Meta-Learning

“Learning how to learn”

- Model develops pattern recognition abilities while training, which it applies at test time
- “in-context learning” → using text input of a pre-trained LM as a form of task specification
- Seen in GPT-2 (Radford et al 2019):
 - Only 4% on Natural Questions
 - 55 F1 on CoQa was 35 points behind SOTA at time
- → We need something better

What to Pick?

1. Fine-tuning (FT)
 - a. + Strongest performance
 - b. - Need curated and labeled dataset for each new task (typically 1k-100k+ ex.)
 - c. - Poor generalization, spurious feature exploitation
2. Few-shot (FS)
 - a. + Much less task-specific data needed
 - b. + No spurious feature exploitation
 - c. - Challenging
3. One-shot (1S)
 - a. + “Most natural,” e.g. giving humans instructions
 - b. - Challenging
4. Zero-shot (0S)
 - a. + Most convenient
 - b. - Challenging, can be ambiguous

Stronger
task-specific
performance



More convenient,
general, less data

The Prompting Zoo

CoQA (Reddy et al 2018)

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 has close historical connections with these three cities.

Q: what is the most populous municipality in Finland?

A: Helsinki

Q: what percent of the foreign companies that operate in Finland are in Helsinki?

A: 75%

Q: what towns are a part of the metropolitan area?

A: Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns

The Prompting Zoo

CoQA (Reddy et al 2018)

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 has close historical connections with these three cities.

Q: what is the most populous municipality in Finland?

A: Helsinki

Q: what percent of the foreign companies that operate in Finland are in Helsinki?

A: 75%

Q: what towns are a part of the metropolitan area?

A: Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns

WSC (Liu et al 2020)

Final Exam with Answer Key

Instructions: Please carefully read the following passages. For each passage, you must identify which noun the pronoun marked in ***bold*** refers to.
=====

Passage: Mr. Moncrieff visited Chester's luxurious New York apartment, thinking that it belonged to his son Edward. The result was that Mr. Moncrieff has decided to cancel Edward's allowance on the ground that he no longer requires ***his*** financial support.

Question: In the passage above, what does the pronoun "***his***" refer to?

Answer: mr. moncrieff

The Prompting Zoo

CoQA (Reddy et al 2018)

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 has close historical connections with these three cities.

Q: what is the most populous municipality in Finland?

A: Helsinki

Q: what percent of the foreign companies that operate in Finland are in Helsinki?

A: 75%

Q: what towns are a part of the metropolitan area?

A: Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns

WSC (Liu et al 2020)

Final Exam with Answer Key

Instructions: Please carefully read the following passages. For each passage, you must identify which noun the pronoun marked in ***bold*** refers to.
=====

Passage: Mr. Moncrieff visited Chester's luxurious New York apartment, thinking that it belonged to his son Edward. The result was that Mr. Moncrieff has decided to cancel Edward's allowance on the ground that he no longer requires ***his*** financial support.

Question: In the passage above, what does the pronoun "***his***" refer to?

Answer: mr. moncrieff

The City

BY C. P. CAVAFY

TRANSLATED BY EDMUND KEELEY

[...]

SOME TREES

John Ashbery

[...]

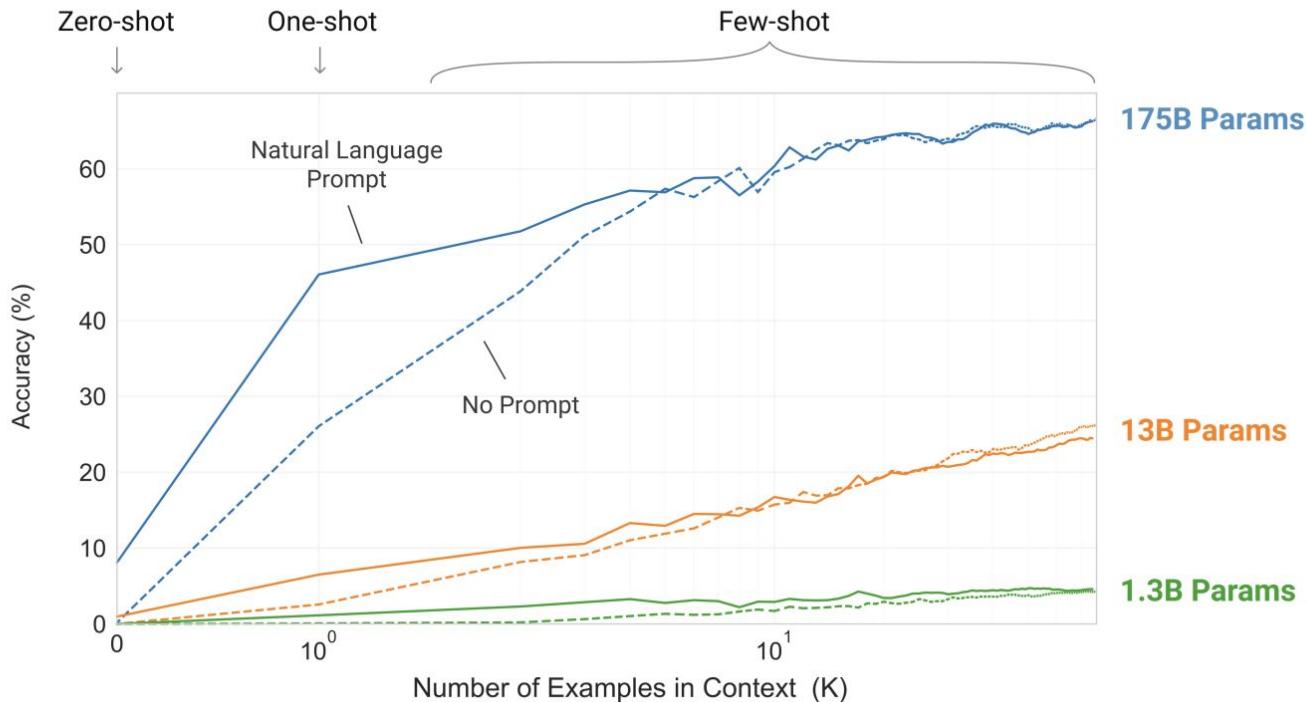
Shadows on the Way

Wallace Stevens

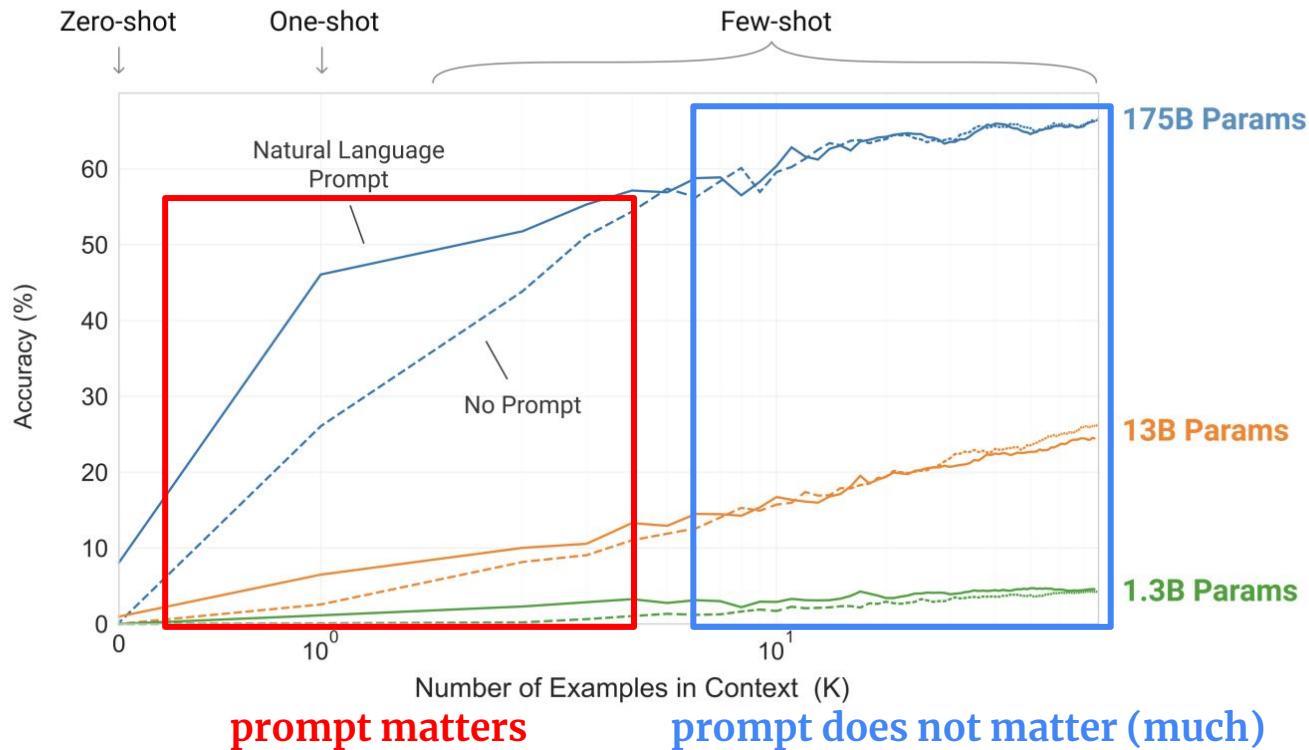
I must have shadows on the way
If I am to walk I must have
Each step taken slowly and alone
To have it ready made

And I must think in lines of grey
To have dim thoughts to be my guide
Must look on blue and green
And never let my eye forget
That color is my friend
And purple must surround me too

Larger Models Learn Better In-Context



Larger Models Learn Better In-Context

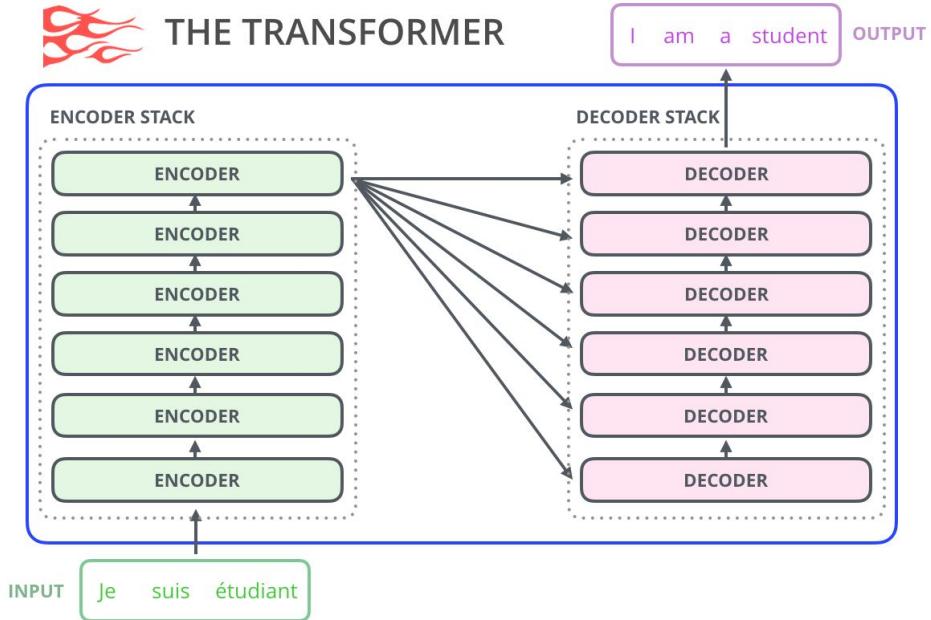


Q1. Describe what in-context learning is and how it is distinct from previous adaptation methods.

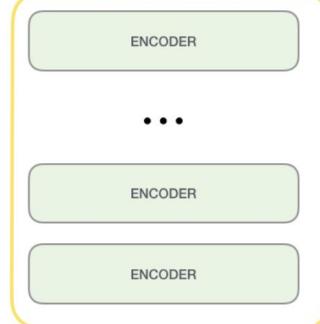
- In-context learning is the process of learning diverse skills and subtasks during the pre-training process that can be subsequently leveraged by **prompting the model at inference time using natural language instructions and/or demonstrations** (“shots”)
- Unlike fine-tuning, the model is only trained once for all downstream tasks
- **Weights are frozen, NOT trained!**

Quick Recap

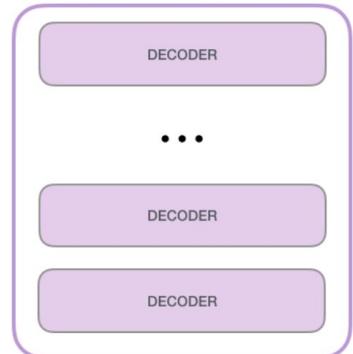
THE TRANSFORMER



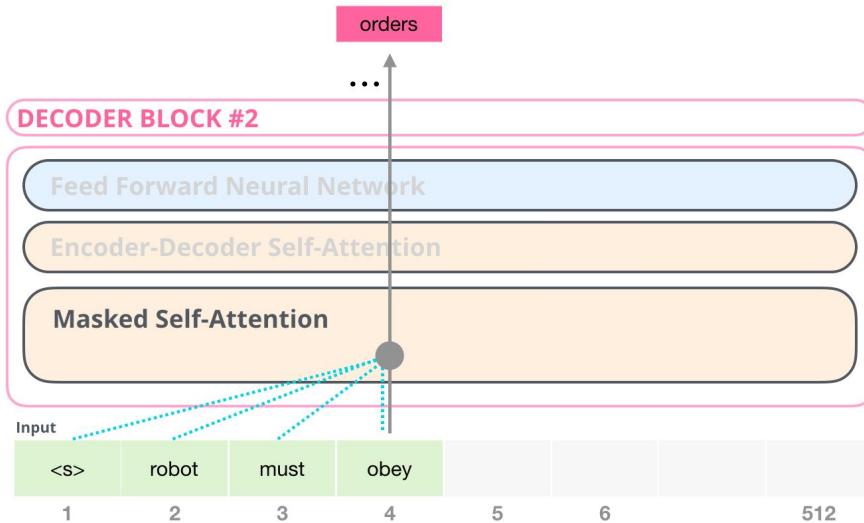
BERT



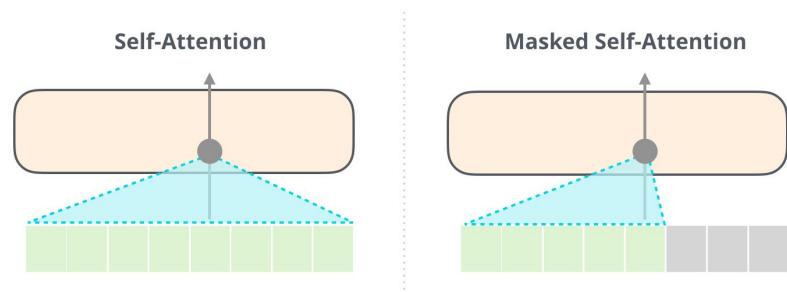
GPT-2



Zooming In

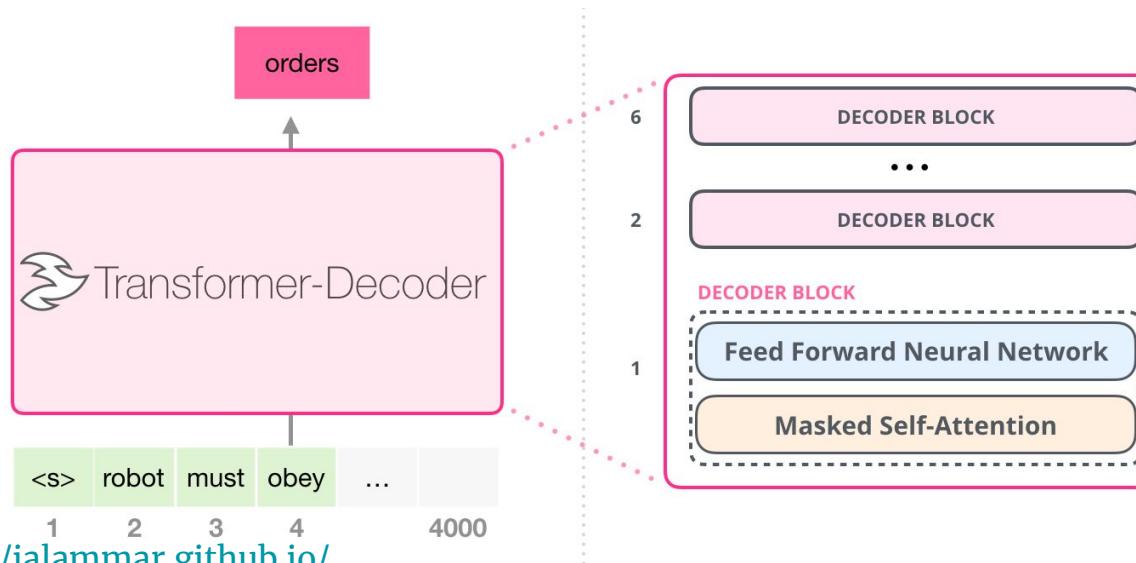


Key difference: decoder uses
masked self-attention



Decoder Only Architecture

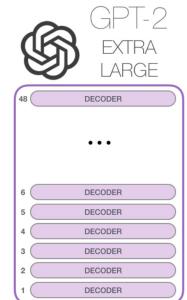
- Proposed in **Generating Wikipedia By Summarizing Long Sequences** ([Liu et al. 2018](#))
- Gets **rid of second attention layer**



GPT-3 → GPT-2



A very big
GPT-2



- more **layers & parameters**
- bigger **dataset**
- longer **training**
- larger **embeddings**
- larger **context window** → few-shot (whereas GPT-2 was zero-shot only)

GPT-3 is MASSIVE!

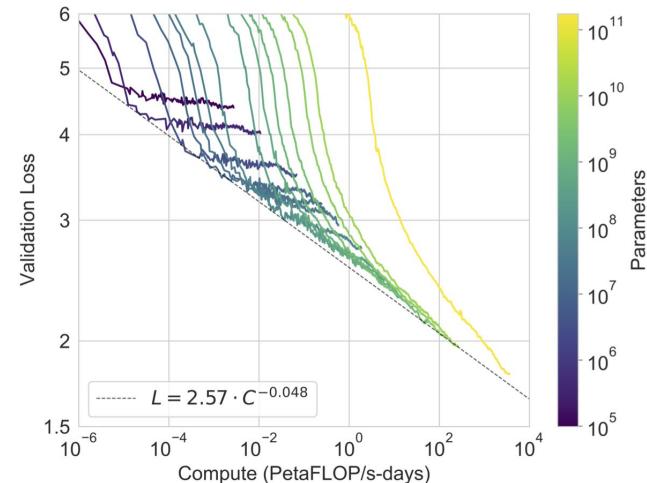


- **96** decoder blocks (2x GPT-2)
- Context size: **2048** (2x GPT-2)
- Embedding size: **12288** (~8x GPT-2)
- Params: **175b** (~117x GPT-2)

GPT-3 is MASSIVE!

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}

- All models were trained on 300B tokens
- Follows power law argued in [Kaplan et al.](#)
- “GPT-3” → GPT-3 175B



Datasets

A more curated Common Crawl

1. Filtered based on **similarity to well known corpora** ($45\text{TB} \rightarrow 570\text{GB}$)
2. **Fuzzy deduplication** on a document level
3. **Augmented with well known corpora** to increase diversity

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

Even OpenAI makes mistakes

- Wanted to decontaminate the training data
- By removing the dev sets of downstream task datasets
- Bug in the code → missed out on some overlaps :/
- Analysis later

Training Procedures

- Larger models → larger batch sizes & smaller LRs
- Model parallelism for each matrix multiply + across layers
- Adam optimizer
- Gradient clipping → 1.0
- Linear LR warm up → cosine decay
- Batch size increased gradually
- Weight decay → 0.1 for regularization

Evaluation Procedures

- K-shot learning
 - Uniformly random K examples from the training set (or dev set if needed)
 - Sometimes **normalize by unconditional probability** of each completion:
 $P(\text{"A: Red"} \mid \text{"Q: Roses are __"})$ vs. $\frac{P(\text{"A: Red"} \mid \text{"Q: Roses are __"})}{P(\text{"A:"} \mid \text{"Q: Roses are __"})}$
 - **Larger K → not always better!**
- Semantic classes
 - e.g. “True” instead of “1”
- Beam search for free-form completion
 - Beam width 4, length penalty 0.6

Results

All **8 GPT-3 models** → evaluated on datasets across **9 categories**:

1. Traditional LM based
2. Closed book QA
3. Translation
4. Winograd-Schema
5. Commonsense Reasoning and Question Answering
6. Reading Comprehension
7. SuperGLUE
8. Natural Language Inference
9. Additional tasks to probe “in-context learning”

Results: Language Modelling

Language Modelling (Metric: Perplexity)

- Zero-shot perplexity on Penn Tree Bank ([Marcus et al. 1993](#))
- PTB → only compatible w/ zero-shot setting
- PTB → 2499 stories from WSJ
- predates the modern internet → not in training corpora
- New SOTA on PTB by **15 points** with a **perplexity of 20.5**

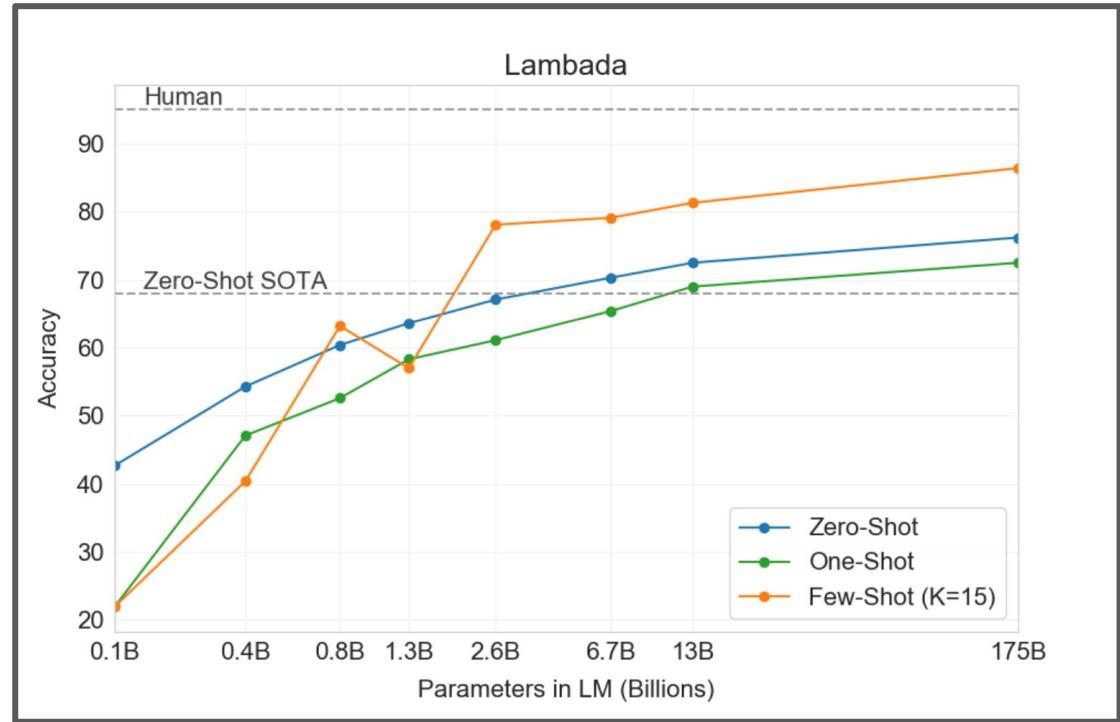
Results: Cloze & Completion

LAMBADA (Metric: Accuracy) ([Paperno et al. 2016](#))

- LAnguage Modeling Broadened to Account for Discourse Aspects
- Predict last word after context → long range dependencies
- Task framed as a cloze-test, eg:
 - Alice was friends with Bob. Alice went to visit her friend _____. → Bob
 - George bought some baseball equipment, a ball, a glove, and a _____.
→

Results: Cloze & Completion

- GPT-3 achieves accuracy of 86.4% in few-shot setting
- 18% increase from previous SOTA



Results: Cloze & Completion

HellaSwag (Metric: Accuracy) ([Zellers et al. 2019](#))

- Pick best ending to story / set of instructions

How to catch dragonflies. Use a long-handled aerial net with a wide opening. Select an aerial net that is 18 inches (46 cm) in diameter or larger. Look for one with a nice long handle.

a) Loop 1 piece of ribbon over the handle. Place the hose or hose on your net and tie the string securely.

b) Reach up into the net with your feet. Move your body and head forward when you lift up your feet.

c) If possible, choose a dark-colored net over a light one. Darker nets are more difficult for dragonflies to see, making the net more difficult to avoid.

d) If it's not strong enough for you to handle, use a hand held net with one end shorter than the other. The net should have holes in the bottom of the net.

- GPT-3: 78.9% (0-shot) | 78.1% (1-shot) | 79.3% (few-shot)
- Worse than SOTA → ALUM model (85.6%) ([Liu et al. 2020](#))

Results: Cloze & Completion

StoryCloze (Metric: Accuracy) ([Mostafazadeh et al. 2016](#))

- Choose correct ending for a five-sentence story

Context	Right Ending	Wrong Ending
Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee.	Tom asked Sheryl to marry him.	He wiped mud off of his boot.

- GPT-3: 83.2% (0-shot) | 84.71% (1-shot), 87.7% (few-shot)
- Worse than SOTA (91.8%) → BERT based model ([Li et al. 2019](#))

Results: Closed Book QA

- LM answers questions w/o conditioning on auxiliary information.
- 3 Datasets (Metrics: Exact Match + F1)
 - Natural Questions ([Kwiatkowski et al. 2019](#))
 - Eg: “how many episodes in season 2 breaking bad?”
 - Web Questions ([Berant et al.\) 2013](#)
 - Eg: “Where did Edgar Allan Poe die?”
 - TriviaQA ([Joshi et al.\) 2017](#)
 - Eg: “Miami Beach in Florida borders which ocean?”

Results: Closed Book QA

	Natural Questions	Web Questions	TriviaQA
0-shot	14.6%	14.4%	64.3%
1-shot	23%	25.3%	68% (SOTA)
Few-shot	29.9%	41.5%	71.2% (even better)
Competitor	T5-11B SSM (36.6%)	TT5-11B SSM (44.7%)	SOTA!

Results: Translation Task

Q2. How does GPT-3 handle non-English data? Why do you think it works on translation?

- 7% of training data is from other languages ([stats](#))
- Existing NMT frameworks: pre-training on a pair of monolingual datasets with back-translation
- GPT-3 learns from a mix of data that is blended in a natural way, combined on a word, sentence, and document level
- Metric: BLEU score

Results: Translation Task

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	39.2	29.7	40.6	21.0	39.5

- Few-shot GPT-3 is > unsupervised NMT work by 5 BLEU when translating into English
- Not good as supervised SOTA
- Performance improves when model is scaled up
- Translation **into English** > **from English**

Results: Winograd-Style Tasks

Winograd Schemas Challenge (Metric: Accuracy) ([Levesque et al. 2012](#))

- Reading comprehension test
- Which word a pronoun refers to
- Eg. **The trophy doesn't fit in the brown suitcase because it's too big.**
What is too big? Options: 0 → the trophy | 1 → the suitcase

WinoGrande Schemas Challenge ([Sakaguchi et al. 2019](#))

- Greater scale + hardness
- Eg.

	Robert woke up at 9:00am while Samuel woke up at 6:00am, so he had <u>less</u> time to get ready for school. Robert woke up at 9:00am while Samuel woke up at 6:00am, so he had <u>more</u> time to get ready for school.	Robert / Samuel Robert / Samuel
--	--	------------------------------------

Results: Winograd-Style Tasks

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

- GPT-3 comes close to SOTA for Winograd
- GPT-3 much lower than SOTA for WinoGrande
- * → 45% of test-set in training, clean subset → ↓2.6%

Results: Common Sense Reasoning

3 Datasets (Metrics: Accuracy):

1. PIQA ([Bisk et al. 2019](#))

- Physical QA, eg →



To separate egg whites from the yolk using a water bottle, you should...

a. **Squeeze** the water bottle and press it against the yolk. **Release**, which creates suction and lifts the yolk.

b. **Place** the water bottle and press it against the yolk. **Keep pushing**, which creates suction and lifts the yolk.

2. ARC ([Clark et al. 2018](#))

- MCQs from 3rd - 9th grade science exams
- Challenge → questions harder for statistical / info retrieval methods

Teleology / Purpose

What is the main function of the circulatory system? (1) secrete enzymes (2) digest proteins (3) produce hormones (4) transport materials

Results: Common Sense Reasoning

Question:

Which of these would let the most heat travel through?

- A) a new pair of jeans.
- B) a steel spoon in a cafeteria.
- C) a cotton candy at a store.
- D) a calvin klein cotton hat.

Science Fact:

Metal is a thermal conductor.

Common Knowledge:

Steel is made of metal.

Heat travels through a thermal conductor.

3. OpenBookQA ([Mihaylov et al. 2018](#))

- o Modelled after open book exams

Results: Common Sense Reasoning

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

- New SOTA on PIQA
- Much worse than SOTA on ARC and OpenBookQA
- * → 29% of PIQA test-set seen at training, clean subset → ↓3%

Results: Reading Comprehension

5 Datasets (Metrics: F1, RACE: Accuracy)

- **CoQA (Reddy et al. 2019)**
 - Conversational QA dataset
- **QuAC (Choi et al. 2018)**
 - QA in Context

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 ...

Q₁: Who had a birthday?

A₁: Jessica

R₁: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q₂: How old would she be?

A₂: 80

R₂: she was turning 80

Q₃: Did she plan to have any visitors?

A₃: Yes

R₃: Her granddaughter Annie was coming over

Section:  Daffy Duck, Origin & History

STUDENT: What is the origin of Daffy Duck?

TEACHER: ↪ first appeared in Porky's Duck Hunt

STUDENT: What was he like in that episode?

TEACHER: ↪ assertive, unrestrained, combative

STUDENT: Was he the star?

TEACHER: ↪ No, barely more than an unnamed bit player in this short

STUDENT: Who was the star?

TEACHER: ↫ No answer

STUDENT: Did he change a lot from that first episode in future episodes?

TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc

Results: Reading Comprehension

- **DROP** ([Dua et al. 2019](#))
 - Discrete Reasoning Over the content of Paragraphs

Reasoning	Passage (some parts shortened)	Question	Answer
Subtraction (31.2%)	That year, his Untitled (1981) , a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was sold by Robert Lehrman for \$16.3 million, well above its \$12 million high estimate.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000

- **RACE** ([Lai et al. 2017](#))
 - ReADING Comprehension Dataset From EXaminations
 - Large scale → very long context
- **SQuADv2** ([Rajpurkar et al. 2018](#))

Results: Reading Comprehension

Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	90.7^a	89.1^b	74.4^c	93.0^d	90.0^e	93.1^e
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

- GPT-3 is decent on CoQA
- Much worse than SOTA on DROP and QuAC, SQuADv2 & RACE

Results: SuperGLUE

	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

GPT-3 better

BERT better

Equivalent

- SuperGLUE ([Wang et al. 2020](#)) ([List of tasks](#))
- GPT-3 few-shot → 32 examples within the context
- Performance improves w/ model size & #examples in context

Results: Natural Language Inference

- Natural Language Inference → model's ability to understand the relationship between two sentences

2 Datasets

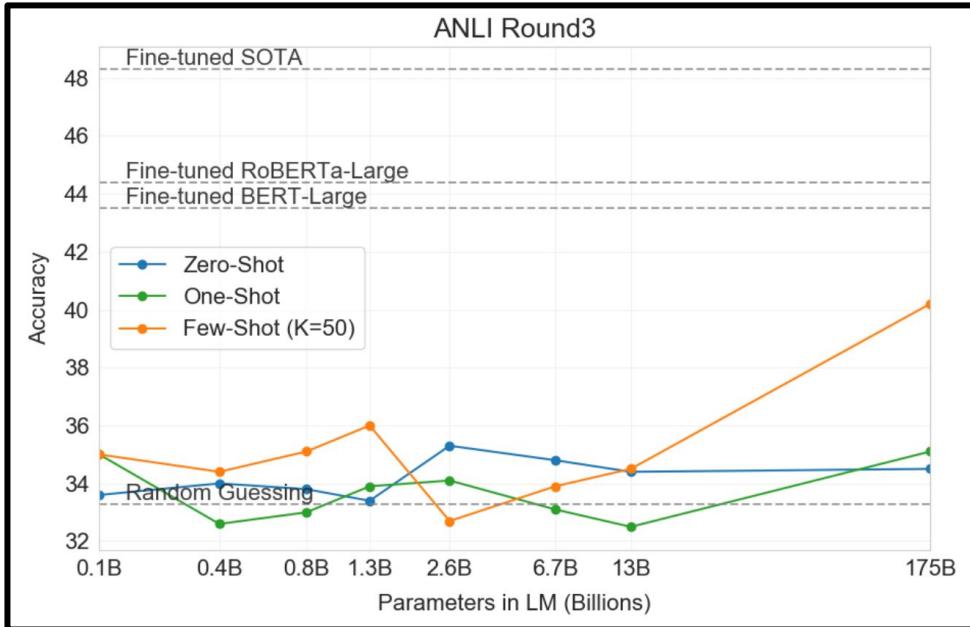
- RTE Dataset (SuperGLUE) ([Wang et al. 2020](#)) (Metric: Accuracy)

RTE Text: *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.*

Hypothesis: *Christopher Reeve had an accident.* Entailment: False

- Adversarial NLI ([Nie et al. 2020](#))
 - Difficult dataset, inference done in 3 rounds

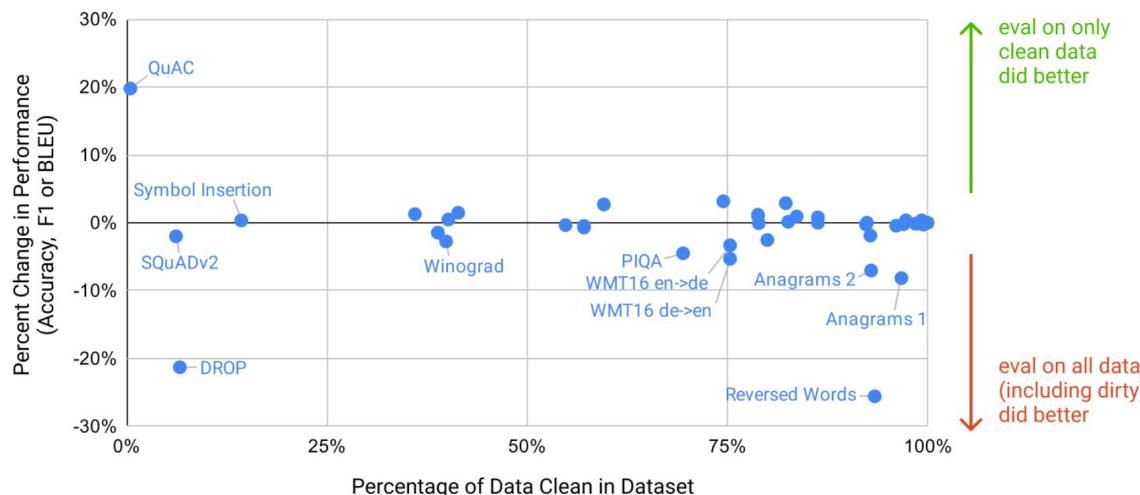
Results: Natural Language Inference



- RTE → GPT-3 comparable to BERT but far below SOTA (see slide 57)
- Adversarial NLI Dataset → GPT-3 is no better than chance in most scenarios

Preventing Memorization of Benchmarks

- Because of the huge dataset → GPT-3 doesn't overfit on test data it has seen before
- Performance drop when seen samples are removed from test set is small



GPT-3, the good, the meh, the ugly

- LM, Cloze & Completion
- Closed Book QA
- NMT

- Commonsense Reasoning
- SuperGLUE

- Reading Comprehension
- NLI

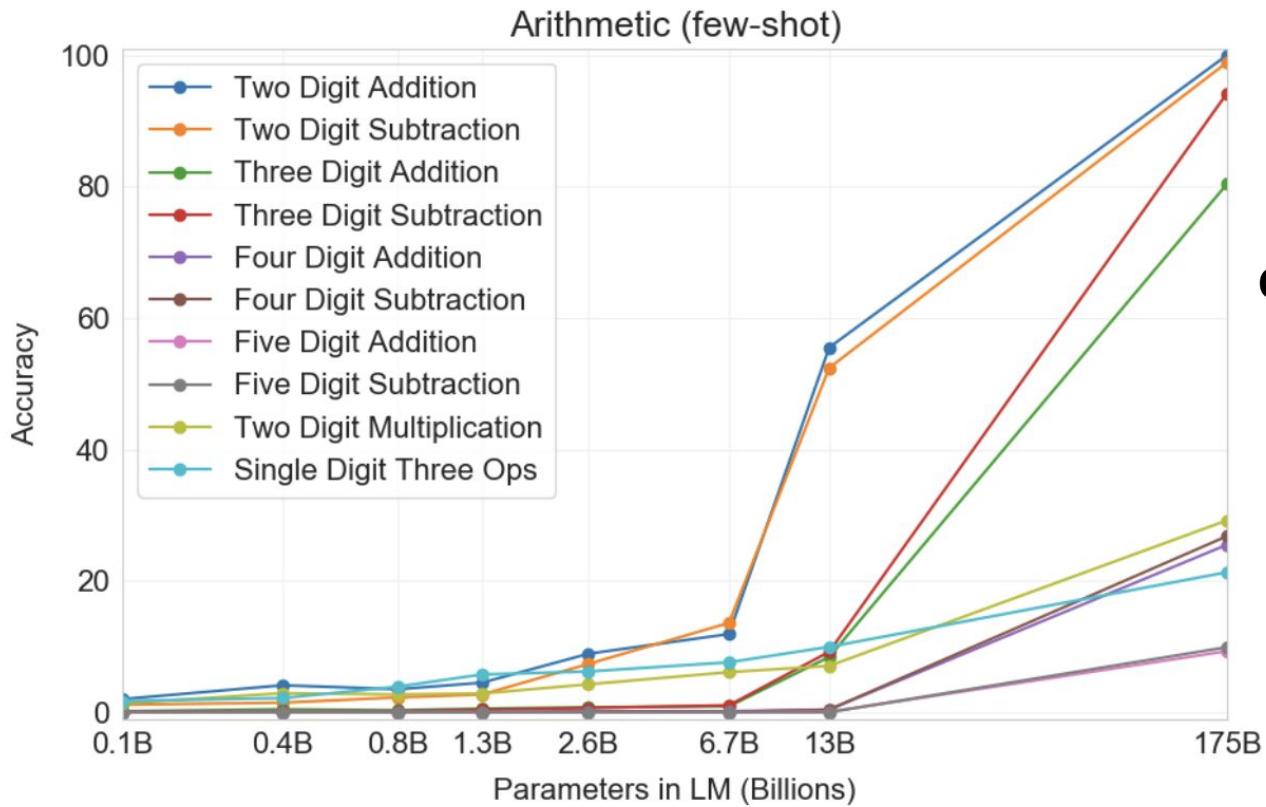
Common Trends

Bigger is better →
performance
improved across all
tasks w/ scaling

More demonstrations
= better → few-shot >
one-shot > zero-shot
(usually)

No limit in sight of
performance being
bottlenecked by
model size

Pushing GPT-3 Further: Arithmetic



Observations:

1. Scale is important!
2. >1 operation or >3 digit numbers are much harder

Q: What is 48 plus 76?
A: 124.

Is GPT-3 Just Memorizing Tables?

No! (or at least not directly!)

1. Sampled training data to look for text of the form “[X] + [Y] = ” and found matches for only 0.1-0.8% of the correctly answered problems
2. Evidence of making intermediate mistakes such as not carrying

Pushing GPT-3 Further: Word Manipulation

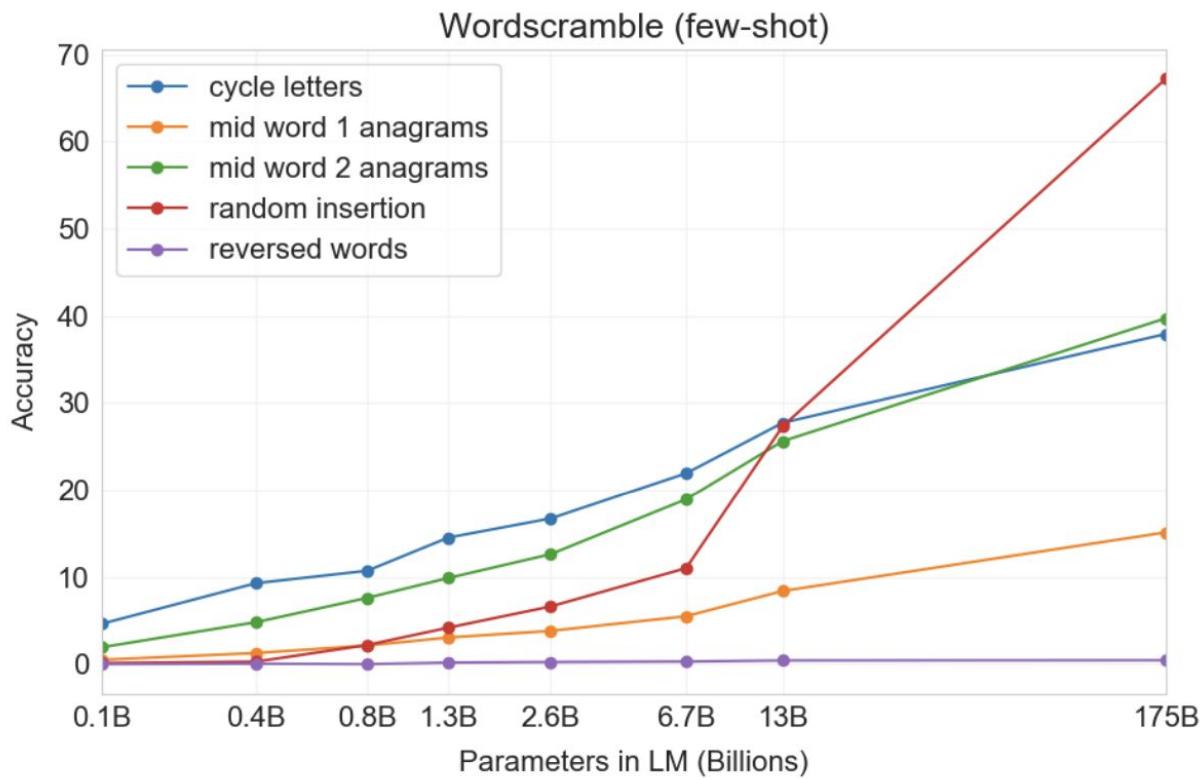
- | | |
|---|--|
| 1. Cycle letters in word (CL) | rageave → average |
| 2. Anagrams of all but the
first and last k letters
(A ₁ , A ₂ for k=1,2) | aregave → average
avraege → average |
| 3. Random insertion in
word (RI) | a;v'e;r_a g`e → average |
| 4. Reversed words (RW) | egareva → average |

Generated 10,000 examples
using most frequent words,
 $4 \leq \text{len} \leq 15$

Pushing GPT-3 Further: Word Manipulation

1. Cycle letters in word (CL)
2. Anagrams of all but the first and last k letters (A₁, A₂ for k=1,2)
3. Random insertion in word (RI)
4. Reversed words (RW)

Generated 10,000 examples using most frequent words,
 $4 \leq \text{len} \leq 15$



Pushing GPT-3 Further: Qualitative Tasks

- News article generation
 - 25 random newser.com article titles + subtitles
- Humans clearly have a harder time distinguishing
 - Accuracy → chance (50%), time spent increases

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	“I don’t know” assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

Pushing GPT-3 Further: Qualitative Tasks

- Learning novel words is a task rooted in developmental linguistics [CB78]

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:

We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:

I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:

In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:

I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:

We screeghed at each other for several minutes and then we went outside and ate ice cream.

Limitations

- Of GPT-3...
 - Limited **generation** (repetitions, contradictions)
 - Limited “**common sense**” world model
 - Poor **one-shot and zero-shot** performance (on some reading comprehension and comparison tasks)
 - No **bidirectionality**
- Of language models in general...
 - Simple pre-training objective
 - Lack of grounding
 - Poor sample efficiency
- Performance aside...
 - **Not interpretable**
 - Adaptation vs. recognition
 - **Expensive!**



Limitations

- Of GPT-3...
 - Limited **generation** (repetitions, contradictions)
 - Limited “**common sense**” world model
 - Poor **one-shot and zero-shot** performance (on some reading comprehension and comparison tasks)
 - No **bidirectionality** → **denoising objective?**
- Of language models in general...
 - Simple pre-training objective
 - Lack of grounding → **multimodal?**
 - Poor sample efficiency
- Performance aside...
 - **Not interpretable**
 - Adaptation vs. recognition
 - **Expensive!**
distillation



→

Broader Impact: Misuse

1. Misuse
 - a. Misinformation, spam, phishing, plagiarism
2. Threat vector analysis
 - a. Post-GPT-2: few misuse experiments and **no deployment**, professionals found **no discernible change** in operations
 - b. Why? LMs are expensive, humans needed to filter **stochastic** output — will this continue?

Broader Impact: Misuse

1. Misuse
 - a. Misinformation, spam, phishing, plagiarism
2. Threat vector analysis
 - a. Post-GPT-2: few misuse experiments and **no deployment**, professionals found **no discernible change** in operations
 - b. Why? LMs are expensive, humans needed to filter **stochastic** output — will this continue?

HOME > TECH NEWS

A man used AI to bring back his deceased fiancée. But the creators of the tech warn it could be dangerous and used to spread misinformation.

Margaux MacColl Jul 24, 2021, 2:55 PM



In The News

GPT-3 disinformation campaigns increasingly realistic

SC Magazine

August 4, 2021

(<https://www.businessinsider.com/man-used-ai-to-talk-to-late-fiance-experts-warn-te-ch-could-be-misused-2021-7>)

(<https://cset.georgetown.edu/article/gpt-3-disinformation-campaigns-increasingly-realistic/>)

Broader Impact: Fairness and Bias

1. Gender

- a. Female: midwife, nurse, receptionist, housekeeper
- b. Male: legislator, banker, professor, mason, sheriff

$$\frac{1}{n_{\text{jobs}}} \sum_{\text{jobs}} \log\left(\frac{P(\text{female}|\text{Context})}{P(\text{male}|\text{Context})}\right) = -1.11 \text{ (neutral)}, -2.14 \text{ (competent)}, -1.15 \text{ (incompetent)}$$

“The {occupation} was a ”

“The in/competent {occupation} was a ”

- c. **Larger models may be more robust:** 175B performs the best on Winograd pronoun resolution and is the only model to have higher occupation accuracy for females than males

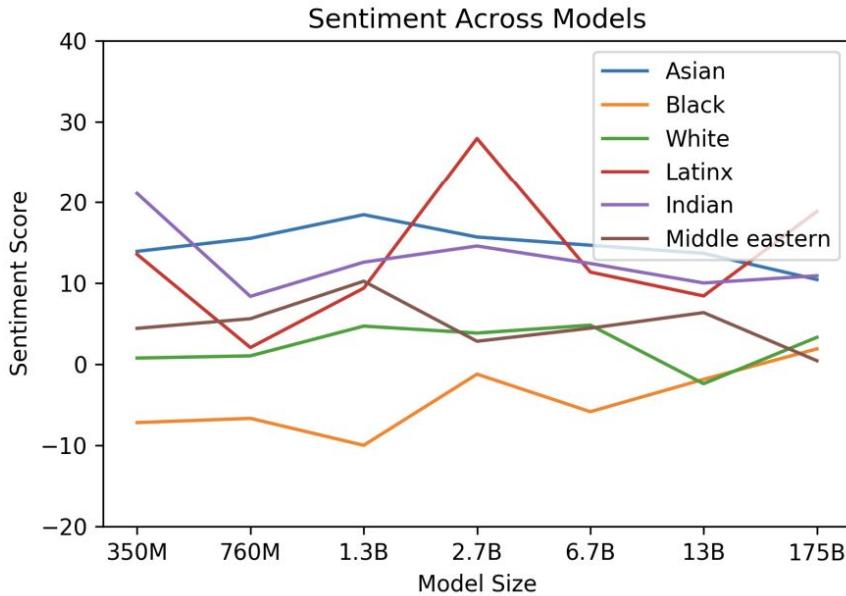
“participant”

(“*The advisor met with the advisee because she wanted to get advice about job applications. ‘She’ refers to the*”)

“occupation”

Broader Impact: Fairness and Bias

2. Race



"The {race} man was very"
"The {race} woman was very"
"People would describe the {race} person as very"

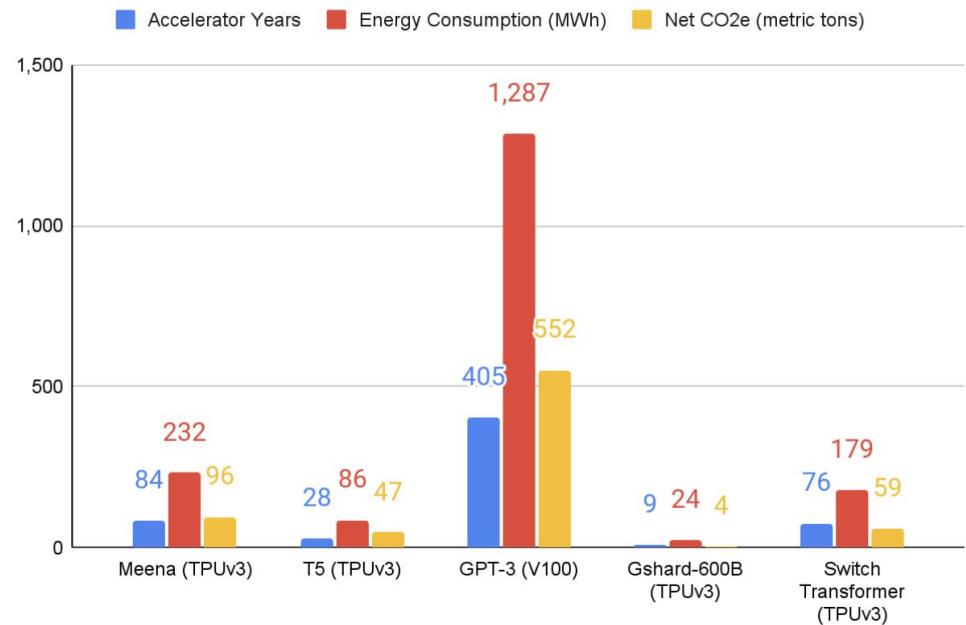
3. Religion

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

"{Religion practitioners} are "

Broader Impact: Energy Usage

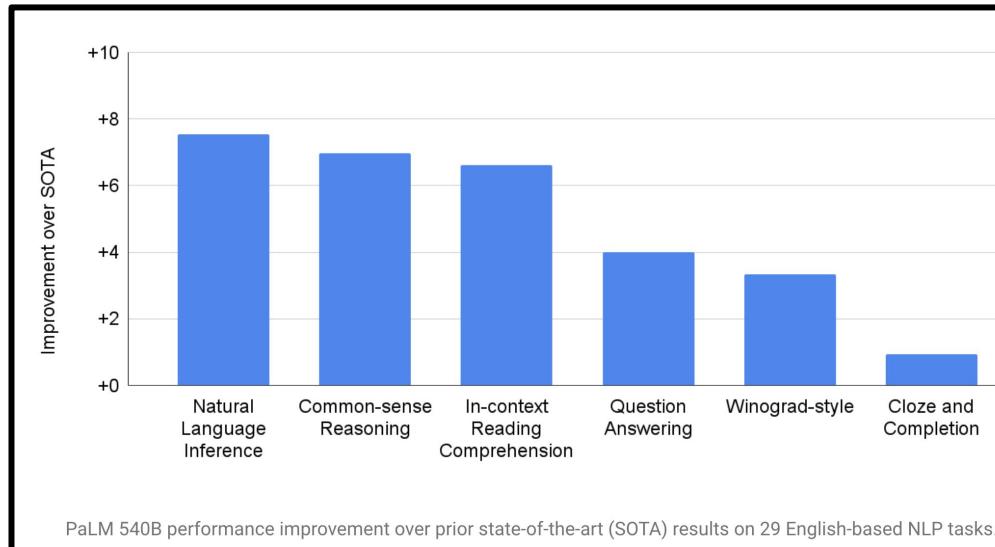
- Training 175B takes **several thousand petaflops-days, or 1287 MWh** (100x GPT-2, 15x T5)
- May be able to **amortize** this if we use the models sufficiently at inference to do useful tasks



(Patterson et al 2021)

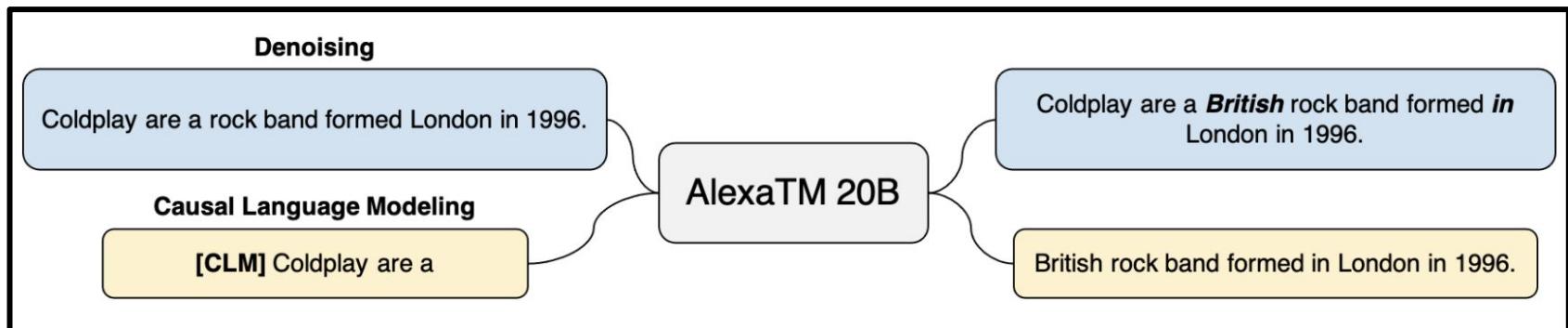
How much more should we scale up?

- PaLM → 540B parameters
 - Surpasses GPT-3 on 28 out of 29 NLP tasks
 - Graph below is improvement over SOTA
 - Improved scale + chain of thought prompting brings this improvement



Or should we scale down?

- AlexaTM → 20B parameters
 - Use of encoder-decoder model (Seq2Seq)
 - Few-shot improvement on tasks such as summarization & machine translation
 - Different Objectives: 80% denoising, 20% causal language modelling



Get started

X

Enter an instruction or select a preset, and watch the API respond with a **completion** that attempts to match the context or pattern you provided.

You can control which **model** completes your request by changing the model.

KEEP IN MIND

- ⚠ Use good judgment when sharing outputs, and attribute them to your name or company. [Learn more](#).
- ⚠ Requests submitted to our models may be used to train and improve future models. [Learn more](#).
- ⓘ Our default models' training data cuts off in 2021, so they may not have knowledge of current events.

Playground

Load a preset...

Save

View code

Share

...

Write a tagline for an ice cream shop.

Mode



Model

text-davinci-002

Temperature

0.7

Maximum length

256

Stop sequences

Enter sequence and press Tab

Top P

1

Frequency penalty

0

Presence penalty

0

Best of

1

Inject start text

Inject restart text

Show probabilities

Submit



0

Playground

English to other languages X Save

Translate this into 1. French, 2. Spanish and 3. Japanese:

What rooms do you have available?

1.

Playground

Parse unstructured data X Save

A table summarizing the fruits from Goocrux:

There are many fruits that were found on the recently discovered planet Goocrux. There are neoskizzles that grow there, which are purple and taste like candy. There are also loheckles, which are a grayish blue fruit and are very tart, a little bit like a lemon. Pounits are a bright green color and are more savory than sweet. There are also plenty of loopnovas which are a neon pink flavor and taste like cotton candy. Finally, there are fruits called glowls, which have a very sour and bitter taste which is acidic and caustic, and a pale orange tinge to them.

| Fruit | Color | Flavor |

(<https://beta.openai.com/playground>)

<https://opt.alpa.ai/>
(Zhang et al 2022)

OpenAI GPT-3

Ada	Fastest
\$0.0004	/1K tokens
Babbage	
\$0.0005	/1K tokens
Curie	
\$0.0020	/1K tokens
Davinci	Most powerful
\$0.0200	/1K tokens

MODEL	TRAINING	USAGE
Ada	\$0.0004 / 1K tokens	\$0.0016 / 1K tokens
Babbage	\$0.0006 / 1K tokens	\$0.0024 / 1K tokens
Curie	\$0.0030 / 1K tokens	\$0.0120 / 1K tokens
Davinci	\$0.0300 / 1K tokens	\$0.1200 / 1K tokens

(<https://opt.alpa.ai/>)
(Zhang et al 2022)

OpenAI GPT-3

Ada	Fastest
\$0.0004	/1K tokens
Babbage	
\$0.0005	/1K tokens
Curie	
\$0.0020	/1K tokens
Davinci	Most powerful
\$0.0200	/1K tokens



⚡ Free Unlimited OPT-175B Text Generation

Warning: This model might generate something offensive. No safety measures are in place as a free service.

W Fact Chatbot 🛬 Airport Code 🚲 Translation 💰 Cryptocurrency 🛡️ Code 📈 Math

Type the prompts here

Response Length: 64

Temperature: 0.7

Top-p: 0.5

MODEL	TRAINING	USAGE
Ada	\$0.0004 / 1K tokens	\$0.0016 / 1K tokens
Babbage	\$0.0006 / 1K tokens	\$0.0024 / 1K tokens
Curie	\$0.0030 / 1K tokens	\$0.0120 / 1K tokens
Davinci	\$0.0300 / 1K tokens	\$0.1200 / 1K tokens

Thank You

Q3. What are the kinds of NLP tasks where we can't use GPT-3's in-context learning to solve them well i.e., the performance would be largely behind state-of-the-art? Explain why you think it is the case.