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

Tom B. Brown, Benjamin Mann, Nick Ryder,
Melanie Subbiah et al.

Presented by:

Alex Li
Bernal Jimenez

Overview

- Zero, One, and Few-shot Learning
- GPT-3 Architecture
- Training
- Results
- Measuring and Preventing Memorization
- Limitations
- Broader Impacts
- Demos

Trend over Time

Task-specific architectures using shared word vector representations

RNNs with multiple layers of shared representations

Pretraining and Fine tuning

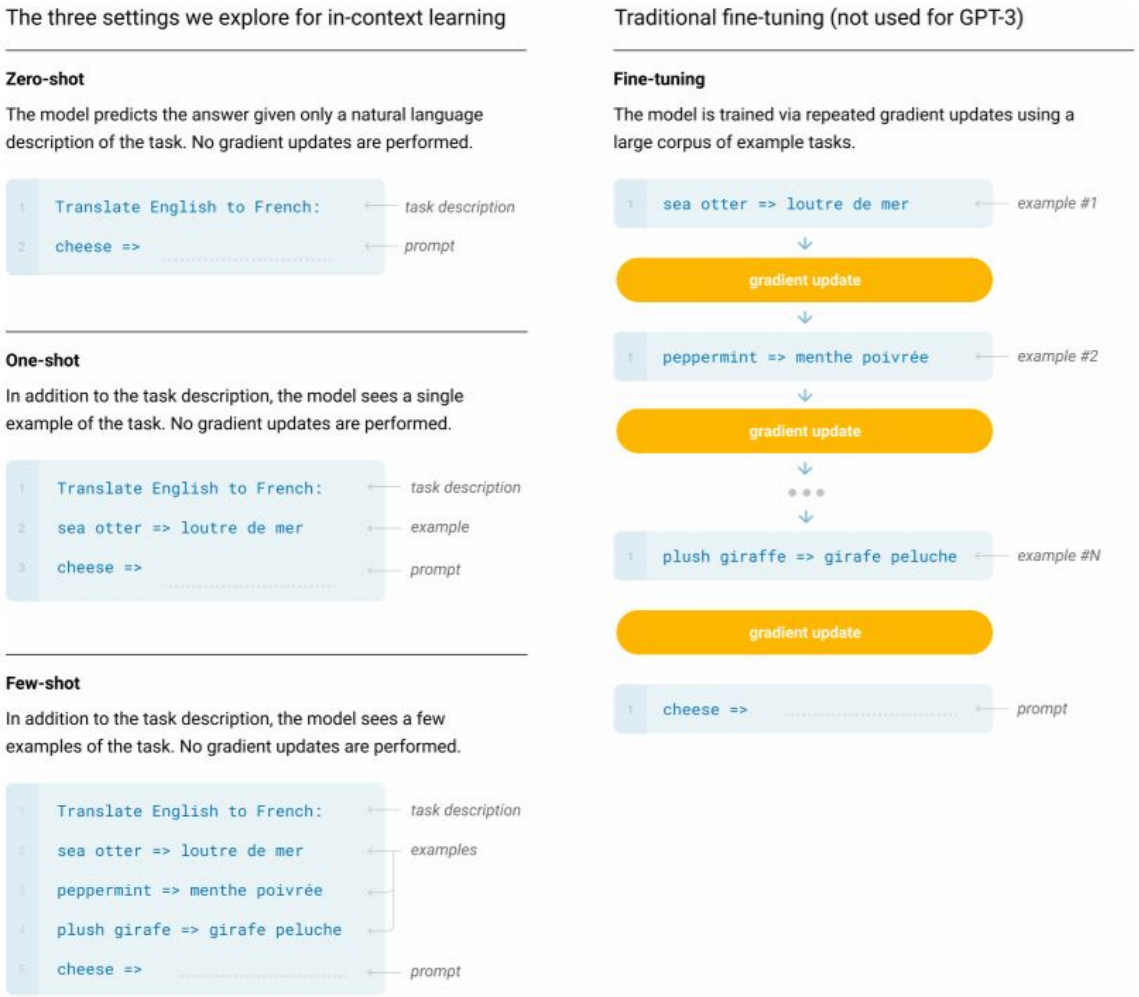
The next step: Few-shot learning

Humans don’t need fine-tuning

Models may only work well on training-like data

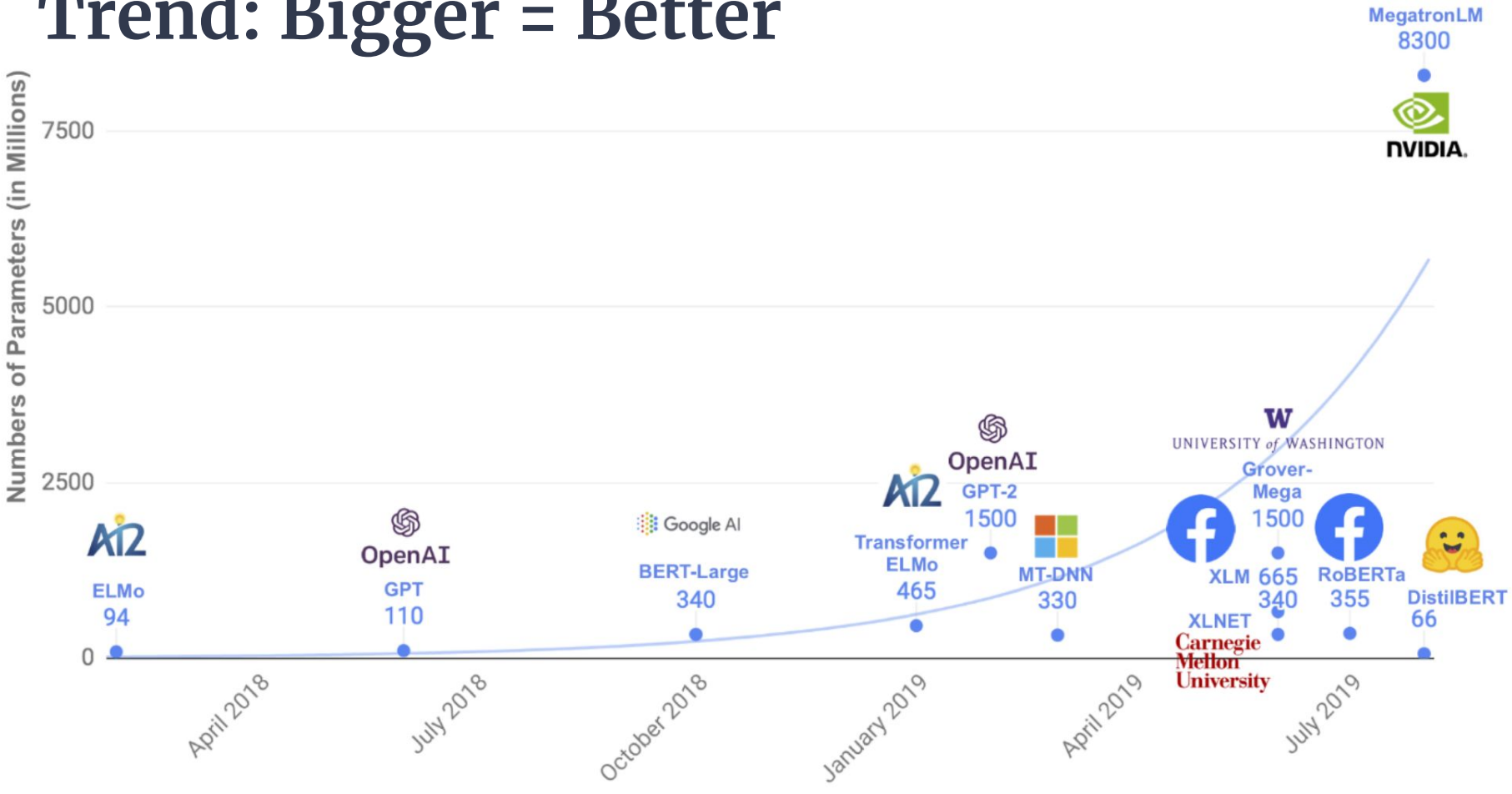
Large sets of training data are hard to find

Approach

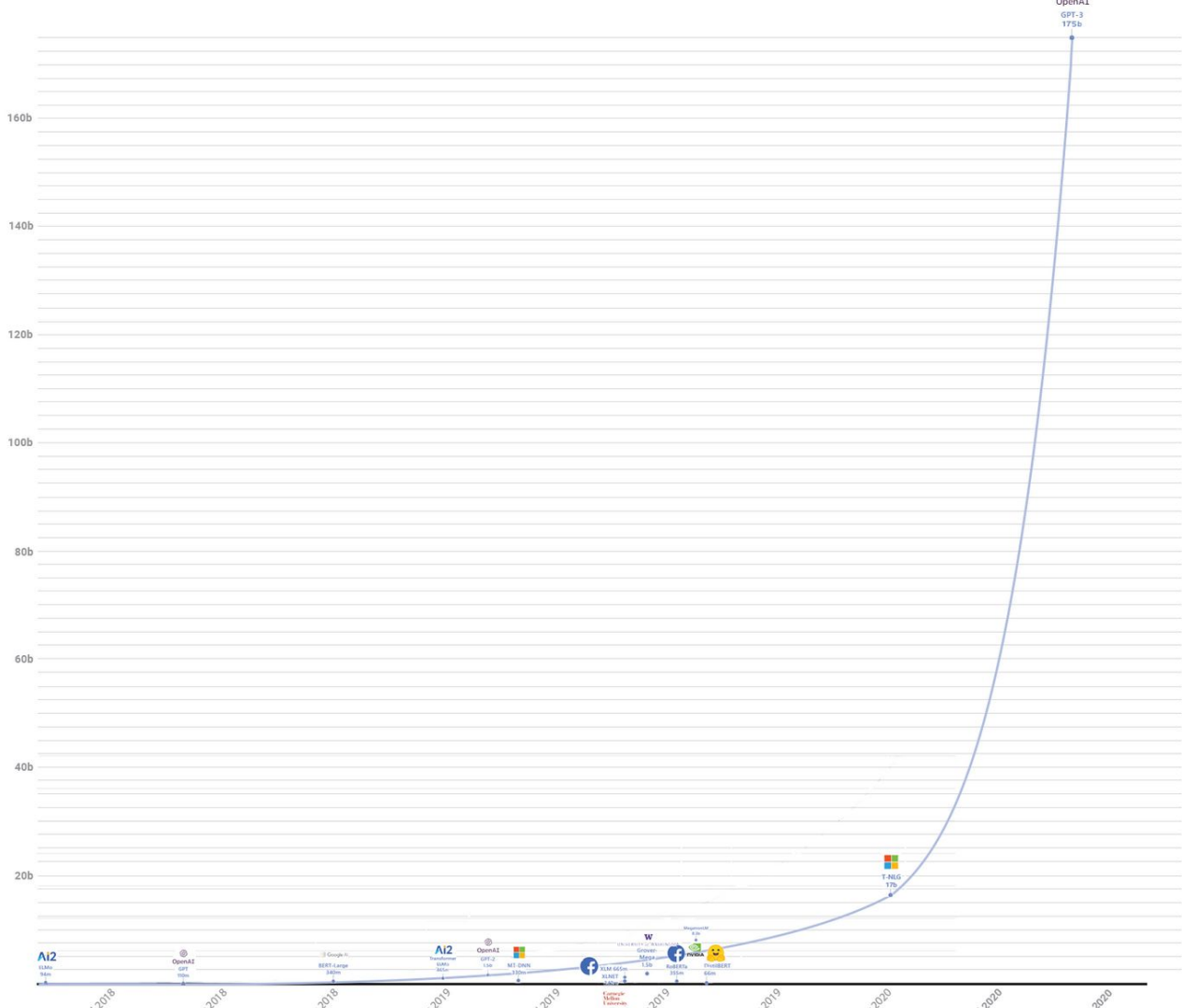

The three settings we explore for in-context learning
Zero-shot
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.
Translate English to French:
task description
cheese =>
prompt
One-shot
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.
Translate English to French:
task description
sea otter => loutre de mer
example
cheese =>
prompt
Few-shot
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.
Translate English to French:
task description
sea otter => loutre de mer
examples
peppermint => menthe poivrée
plush girafe => girafe peluche
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.
sea otter => loutre de mer
example #1
gradient update
peppermint => menthe poivrée
example #2
gradient update
plush giraffe => girafe peluche
example #N
gradient update
cheese =>
prompt

“Since in-context learning involves absorbing many skills and tasks within the parameters of the model, it is plausible that in-context learning abilities might show similarly strong gains with scale.”

Trend: Bigger = Better



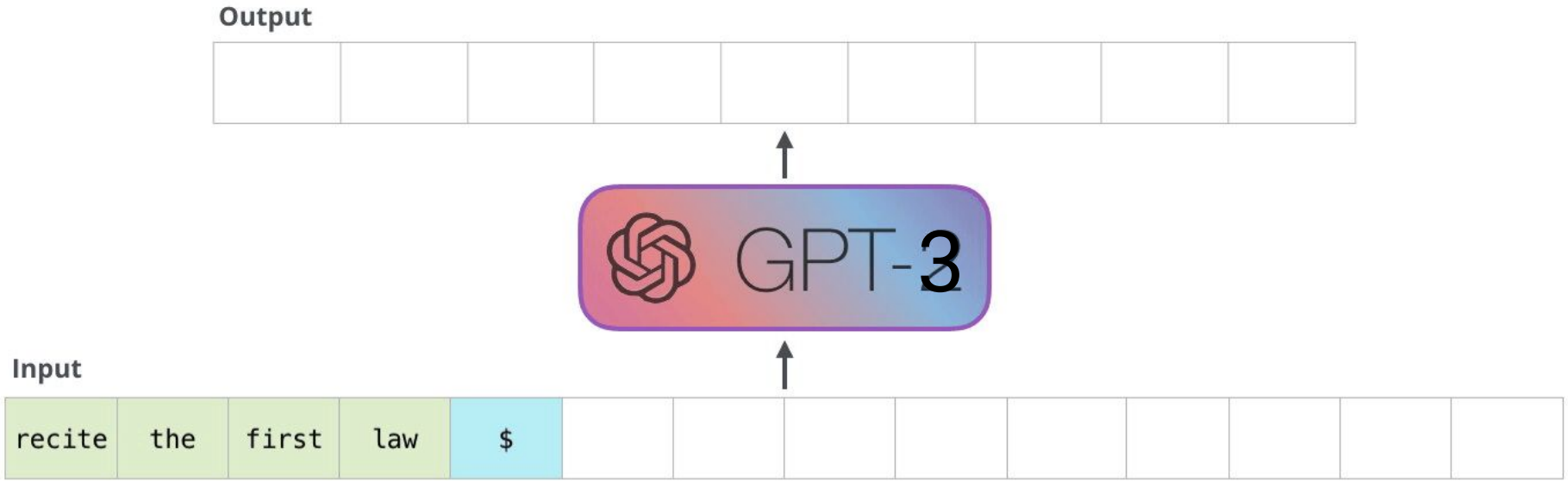
175 Billion Parameters!



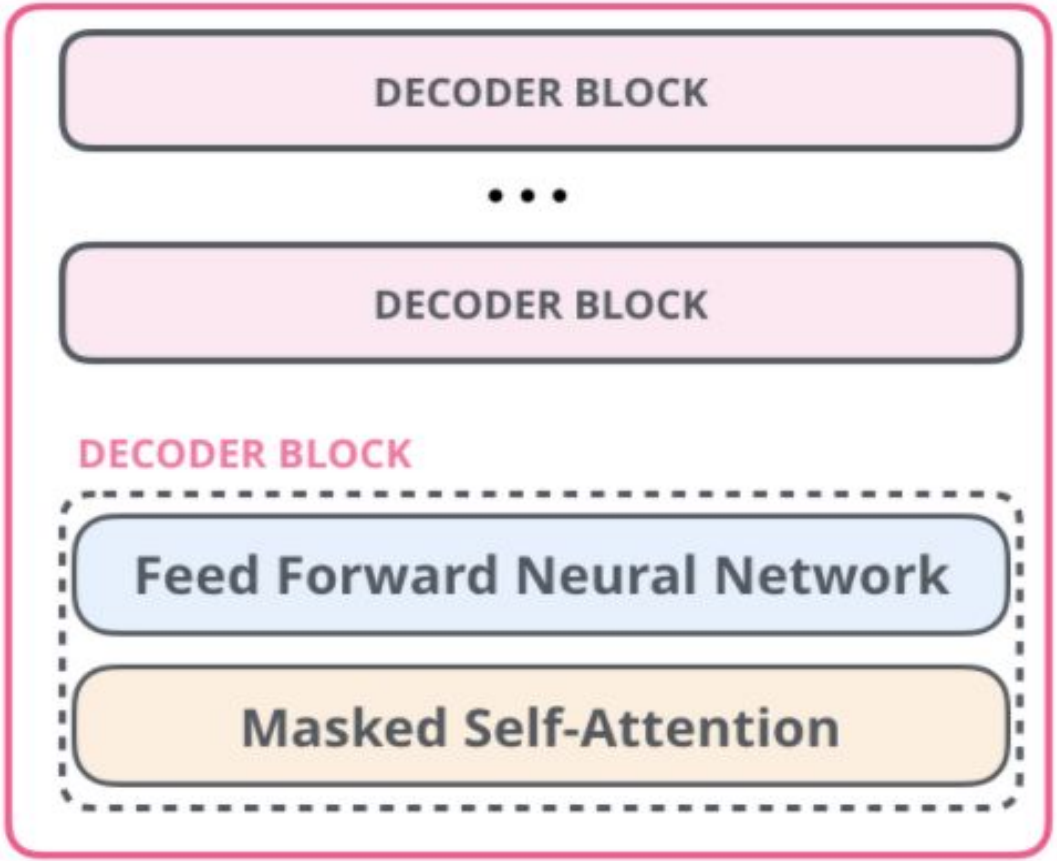
Architecture

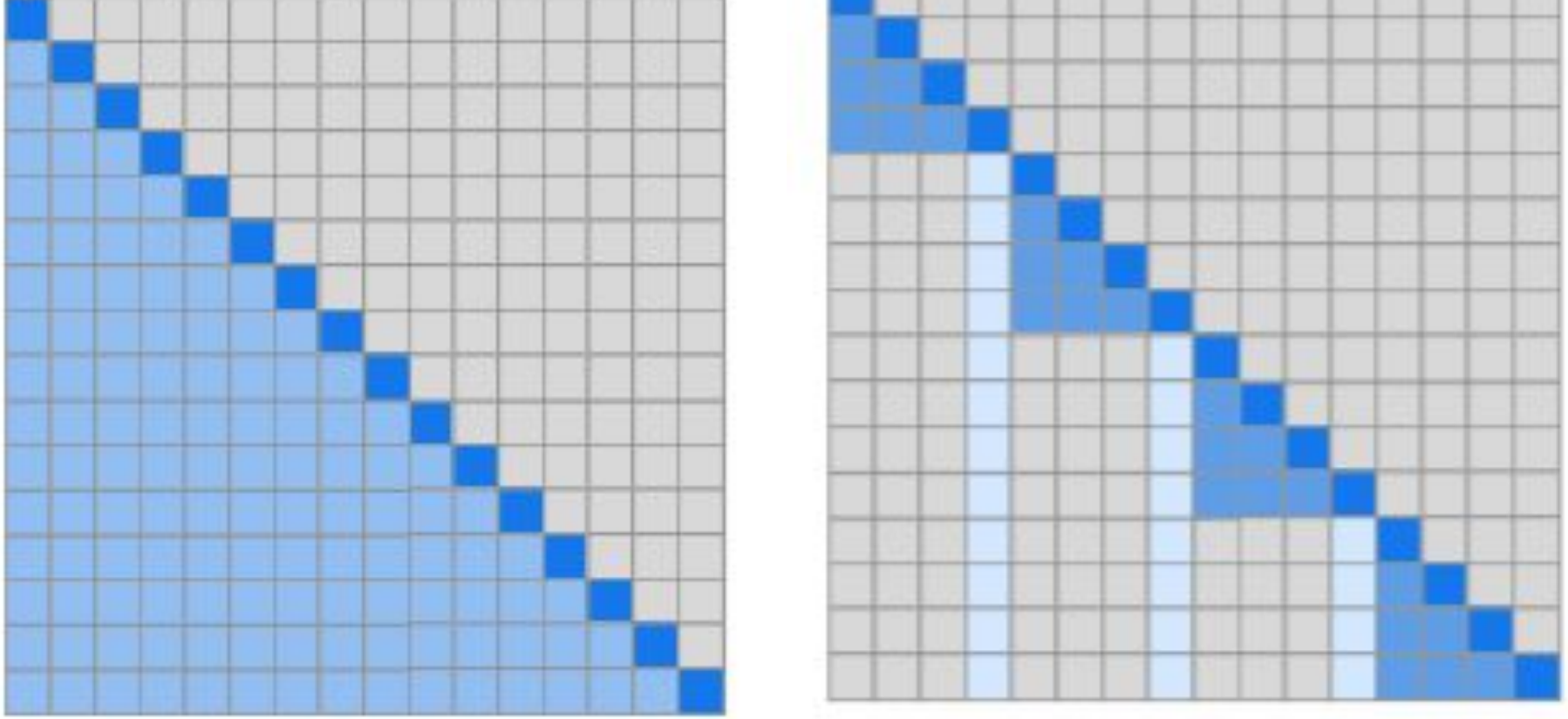
Architecture

A larger GPT-2



Context size: 2048


96 layers
DECODER BLOCK
DECODER BLOCK
DECODER BLOCK
Feed Forward Neural Network
Masked Self-Attention
96 attention heads
Input Dimensionality: 12288



Masked Alternating Sparse Self-Attention

Training

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}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Models Trained

Training Data

More training time spent on higher quality data

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

Filtering Low Quality Documents

High quality - Wikipedia, WebText

Low Quality - Unfiltered Common Crawl

Classify Common Crawl using logistic regression on term frequency vectors

Add if

np.random.pareto(9) > 1 - score

Removing Duplicate Documents

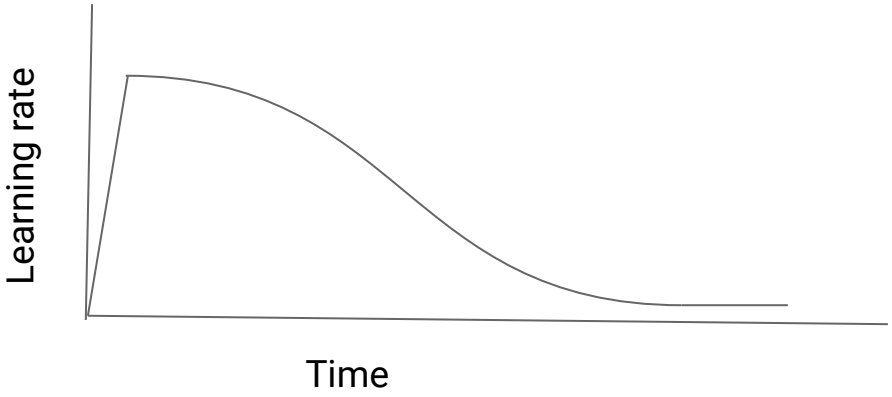
Deduplicate documents within each dataset to prevent redundancy and preserve the validation dataset.

Removing Contaminated data

- Some datasets contain examples of test data.
- Remove colliding 13-grams and a window around them.

Training

- Adam optimizer
- Gradient norm clipped at 1.0
- Weight decay of .1
- Uses full window with a ‘end of text token’



Learning rate: Cosine decay with warmup, LR minimizes at 10% the max after 260B tokens

Evaluation

Multiple choice: Get average likelihood of each token in the answer (besides ‘Answer:")

Binary classification: Multiple choice with the answers “True" and “False”

Free form completion: Beam search with width 4 and length penalty .6

Scoring dependent on the standard for the dataset (F1 similarity, BLEU, or exact match)

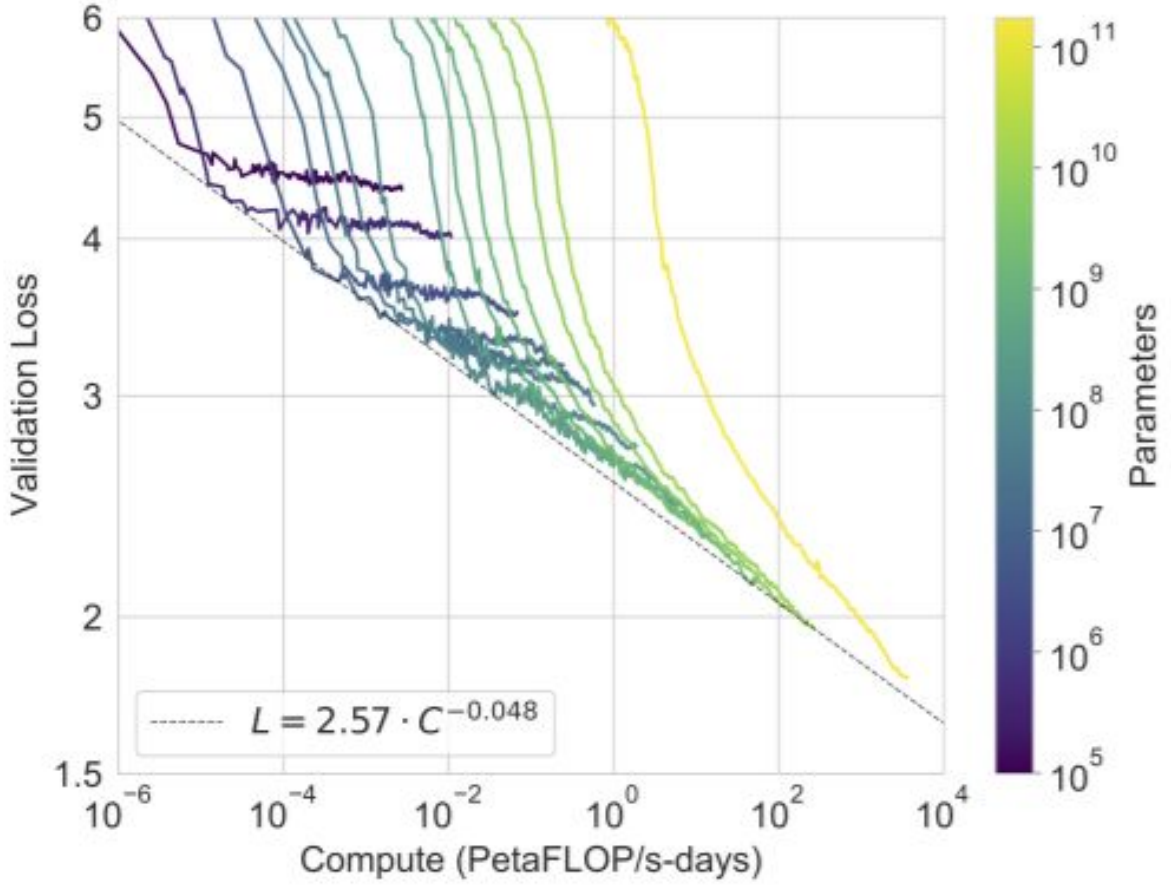
Evaluation

Examples for few-shot drawn from the training set

Use development set instead of private test set

Results

Loss on Validation Set



Language Modeling

Task	GPT-3	State of the art
Penn Tree Bank (Perplexity)	**20.5 (0-shot)**	35.8
LAMBADA (Predict last word)	**84.4%** (Few-shot)	68.4%
HellaSwag (Finish story)	78.1% (Few-shot)	**85.6%**
StoryCloze (Finish story)	87.7% (Few-shot)	**91.1%**

Context →	Fill in blank: She held the torch in front of her. She caught her breath. "Chris? There's a step." "What?" "A step. Cut in the rock. About fifty feet ahead." She moved faster. They both moved faster. "In fact," she said, raising the torch higher, "there's more than a ____. ->
Target Completion →	step

Figure G.21: Formatted dataset example for LAMBADA

LAMBADA formatting - Works well with few-shot, poorly with one-shot.

Context →	Making a cake: Several cake pops are shown on a display. A woman and girl are shown making the cake pops in a kitchen. They
Correct Answer →	bake them, then frost and decorate.
Incorrect Answer →	taste them as they place them on plates.
Incorrect Answer →	put the frosting on the cake as they pan it.
Incorrect Answer →	come out and begin decorating the cake as well.

Figure G.9: Formatted dataset example for HellaSwag

Context →	Bob went to the gas station to fill up his car. His tank was completely empty and so was his wallet. The cashier offered to pay for his gas if he came back later to pay. Bob felt grateful as he drove home.
Correct Answer →	Bob believed that there were good people in the world.
Incorrect Answer →	Bob contemplated how unfriendly the world was.

Figure G.17: Formatted dataset example for StoryCloze

HellaSwag and StoryCloze

Closed-Book Question Answering

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**

Context →	Q: 'Nude Descending A Staircase' is perhaps the most famous painting by which 20th century artist? A:
Target Completion →	MARCEL DUCHAMP
Target Completion →	r mutt
Target Completion →	duchamp
Target Completion →	marcel duchamp
Target Completion →	R.Mutt
Target Completion →	Marcel duChamp
Target Completion →	Henri-Robert-Marcel Duchamp
Target Completion →	Marcel du Champ
Target Completion →	henri robert marcel duchamp
Target Completion →	Duchampian
Target Completion →	Duchamp
Target Completion →	duchampian
Target Completion →	marcel du champ
Target Completion →	Marcel Duchamp
Target Completion →	MARCEL DUCHAMP

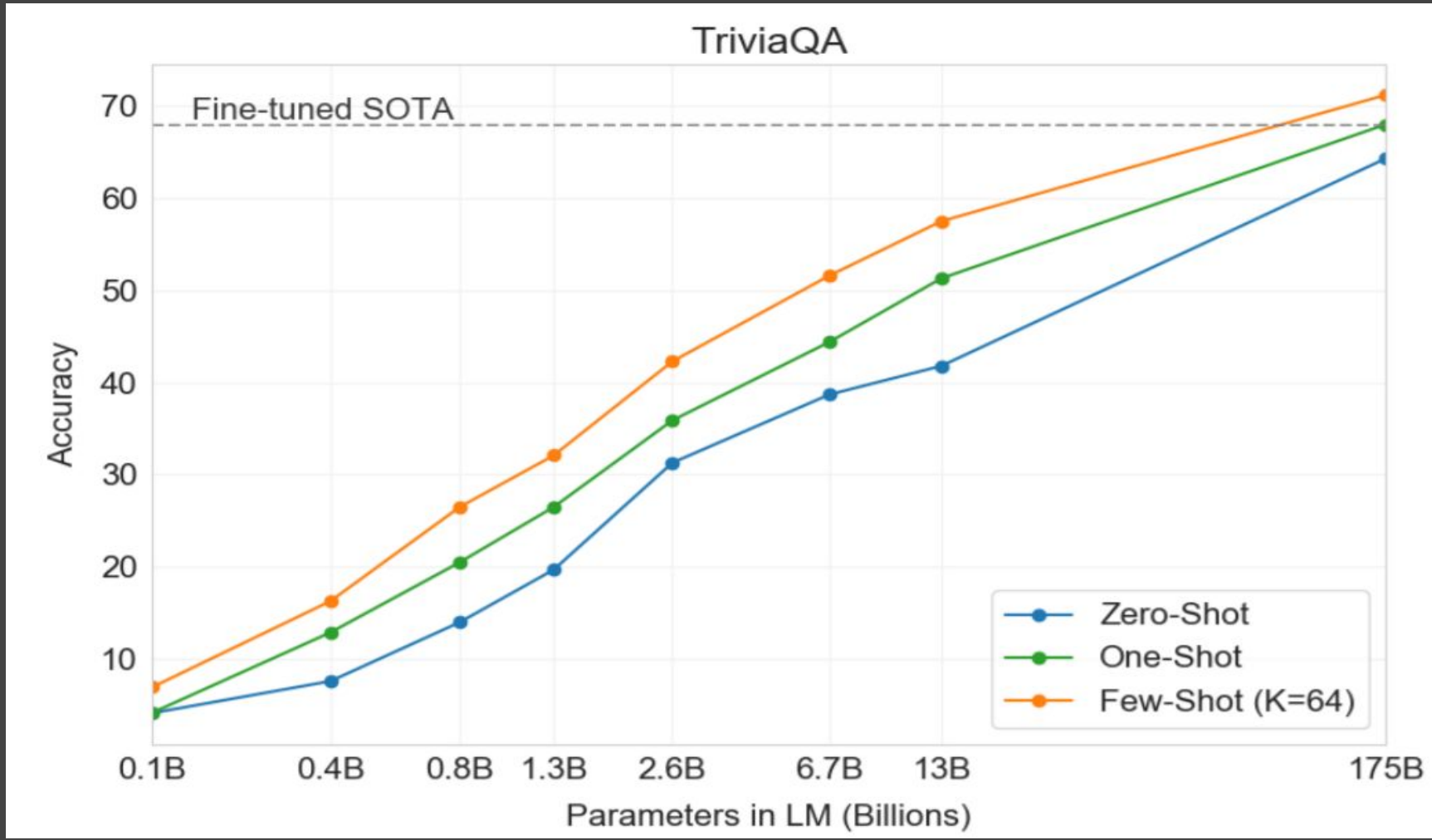
Figure G.34: Formatted dataset example for TriviaQA. TriviaQA allows for multiple valid completions.

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

Figure G.35: Formatted dataset example for WebQA

Context →	Q: Who played tess on touched by an angel? A:
Completion →	Delloreese Patricia Early (July 6, 1931 { November 19, 2017), known professionally as Della Reese

Figure G.24: Formatted dataset example for Natural Questions

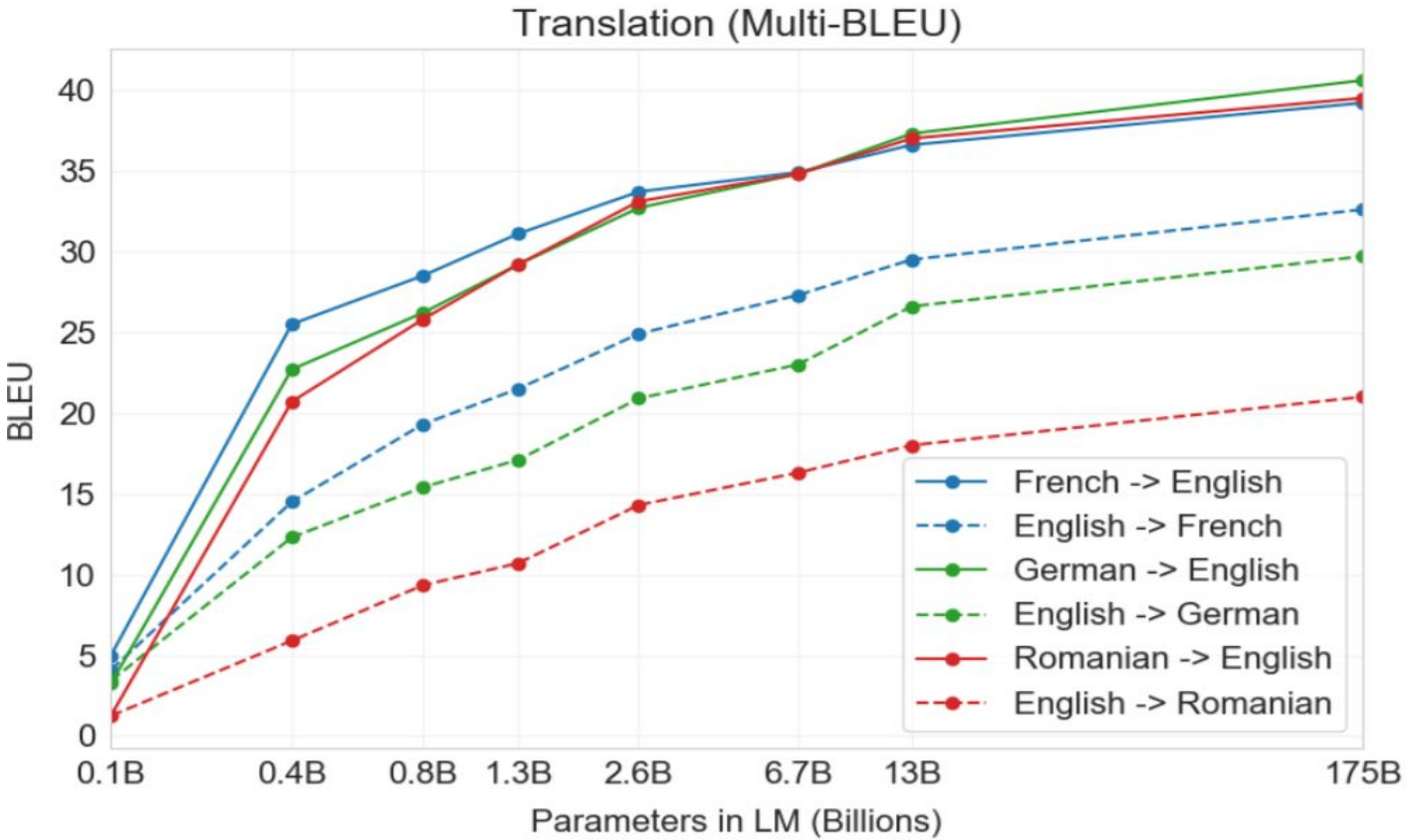

TriviaQA
Fine-tuned SOTA
70
60
50
40
30
20
10
Accuracy
Zero-Shot
One-Shot
Few-Shot (K=64)
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

Translation

Automatic translation capabilities since training has 7% non-english corpus.

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

De: German, Ro= Romanian



Score change with size

Winograd(e)

Correct Context →	Grace was happy to trade me her sweater for my jacket. She thinks the sweater
Incorrect Context →	Grace was happy to trade me her sweater for my jacket. She thinks the jacket
Target Completion →	looks dowdy on her.

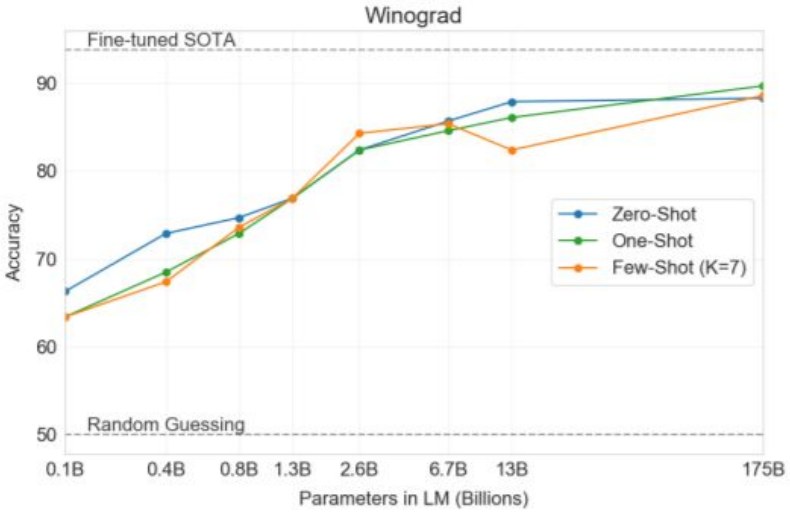
Figure G.13: Formatted dataset example for Winograd. The 'partial' evaluation method we use compares the probability of the completion given a correct and incorrect context.

Correct Context →	Johnny likes fruits more than vegetables in his new keto diet because the fruits
Incorrect Context →	Johnny likes fruits more than vegetables in his new keto diet because the vegetables
Target Completion →	are saccharine.

Figure G.14: Formatted dataset example for Winogrande. The 'partial' evaluation method we use compares the probability of the completion given a correct and incorrect context.

Results


Winogrande
Human
Fine-tuned SOTA
Fine-tuned RoBERTa-Large
Zero-Shot
One-Shot
Few-Shot (K=50)
Random Guessing
Accuracy
90
80
70
60
50
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)


Winograd
Fine-tuned SOTA
Zero-Shot
One-Shot
Few-Shot (K=7)
Random Guessing
Accuracy
90
80
70
60
50
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

Common Sense Reasoning

Scientific reasoning capacity

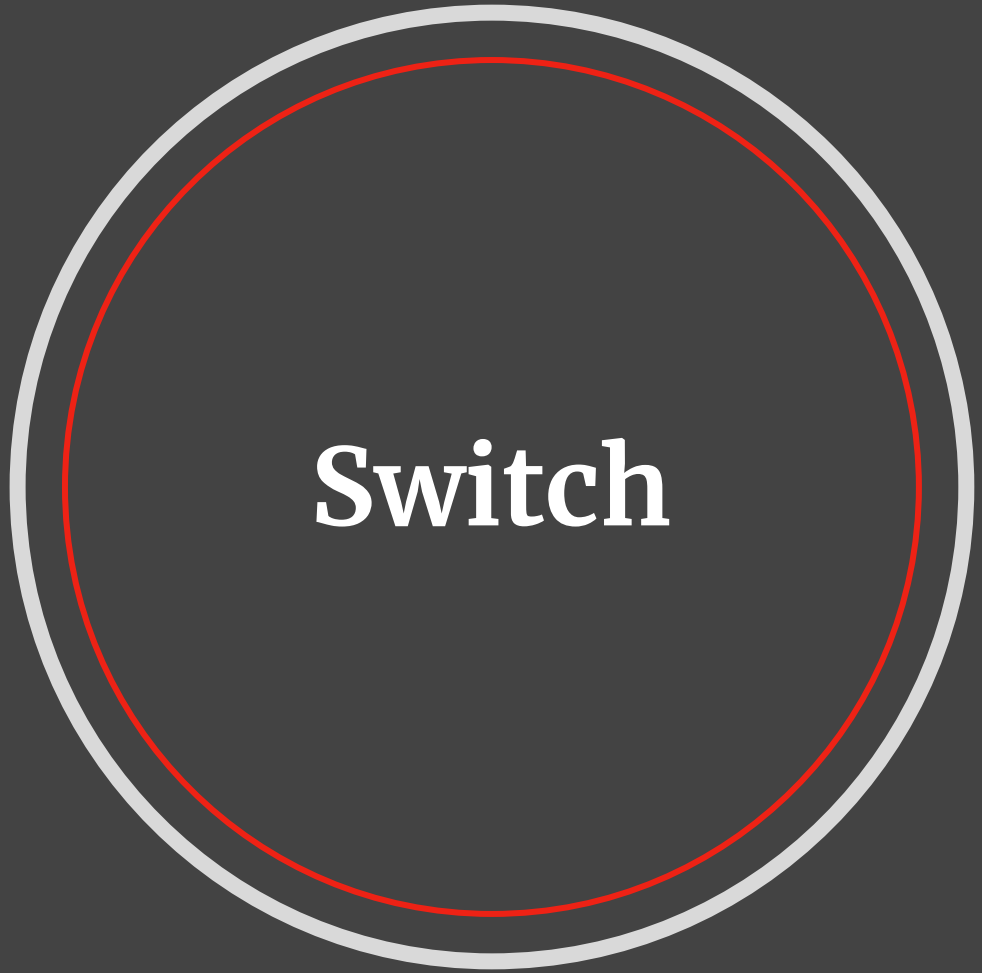
Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	**92.0**[KKS+20]	**78.5**[KKS+20]	**87.2**[KKS+20]
GPT-3 Zero-Shot	**80.5***	68.8	51.4	57.6
GPT-3 One-Shot	**80.5***	71.2	53.2	58.8
GPT-3 Few-Shot	**82.8***	70.1	51.5	65.4

Context →	How to apply sealant to wood.
Correct Answer →	Using a brush, brush on sealant onto wood until it is fully saturated with the sealant.
Incorrect Answer →	Using a brush, drip on sealant onto wood until it is fully saturated with the sealant.

Figure G.4: Formatted dataset example for PIQA

Context →	Question: George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat? Answer:
Correct Answer →	dry palms
Incorrect Answer →	wet palms
Incorrect Answer →	palms covered with oil
Incorrect Answer →	palms covered with lotion

Figure G.11: Formatted dataset example for ARC (Challenge). When predicting, we normalize by the unconditional probability of each answer as described in 2.


Switch

Reading Comprehension

- Context + Question (or Questions)
- On par with initial contextual baselines on most of these 6 datasets
- Two different evaluation settings for reading comprehension tasks
  - Open text generation
    - Use beam search to produce prompt completion
  - Multiple choice
    - Compare either per-token-likelihood normalized by length (probability of completion given context)
    - Sometimes also normalized by same answer given “Answer: “

- A few examples:

Reading Comprehension

Context →	Passage: Saint Jean de Brébeuf was a French Jesuit missionary who travelled to New France in 1625. There he worked primarily with the Huron for the rest of his life, except for a few years in France from 1629 to 1633. He learned their language and culture, writing extensively about each to aid other missionaries. In 1649, Brébeuf and another missionary were captured when an Iroquois raid took over a Huron village . Together with Huron captives, the missionaries were ritually tortured and killed on March 16, 1649. Brébeuf was beatified in 1925 and among eight Jesuit missionaries canonized as saints in the Roman Catholic Church in 1930. Question: How many years did Saint Jean de Brébeuf stay in New France before he went back to France for a few years? Answer:
Target Completion →	4

Figure G.20: Formatted dataset example for DROP

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 northernmost metro area of over one million people, and the city is the northernmost capital of an EU member state. The Helsinki metropolitan area is the third largest metropolitan area in the Nordic countries after Stockholm and Copenhagen, and the City of Helsinki is the third largest after Stockholm and Oslo. Helsinki is Finland's major political, educational, financial, cultural, and research center as well as one of northern Europe's major cities. Approximately 75% of foreign companies that operate in Finland have settled in the Helsinki region. The nearby municipality of Vantaa is the location of Helsinki Airport, with frequent service to various destinations in Europe and Asia. Q: what is the most populous municipality in Finland? A: Helsinki Q: how many people live there? A: 1.4 million in the metropolitan area 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:
Target Completion →	Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns

Figure G.18: Formatted dataset example for CoQA

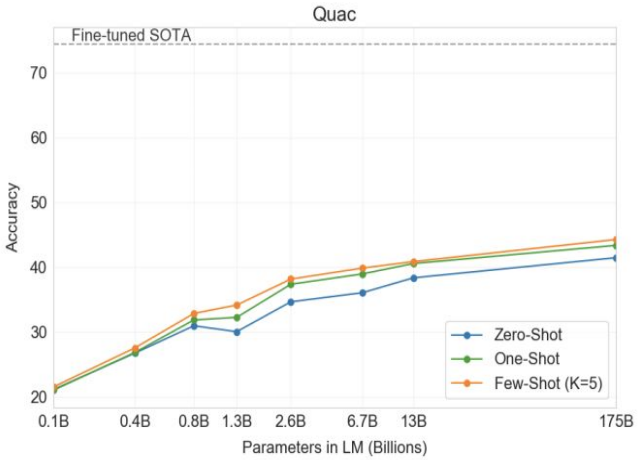
Context →	Article: Mrs. Smith is an unusual teacher. Once she told each student to bring along a few potatoes in plastic bag. On each potato the students had to write a name of a person that they hated And the next day, every child brought some potatoes. Some had two potatoes;some three;some up to five. Mrs. Smith then told the children to carry the bags everywhere they went, even to the toilet, for two weeks. As day after day passed, the children started to complain about the awful smell of the rotten potatoes. Those children who brought five potatoes began to feel the weight trouble of the bags. After two weeks, the children were happy to hear that the game was finally ended. Mrs. Smith asked,"How did you feel while carrying the potatoes for two weeks?" The children started complaining about the trouble loudly. Then Mrs. Smith told them why she asked them to play the game. She said,"This is exactly the situation when you carry your hatred for somebody inside your heart. The terrible smell of the hatred will pollute your heart and you will carry something unnecessary with you all the time. If you cannot stand the smell of the rotten potatoes for just two weeks, can you imagine how heavy it would be to have the hatred in your heart for your lifetime? So throw away any hatred from your heart, and you'll be really happy." Q: Which of the following is True according to the passage? A: If a kid hated four people,he or she had to carry four potatoes. Q: We can learn from the passage that we should _ . A: throw away the hatred inside Q: The children complained about _ besides the weight trouble. A: the smell Q: Mrs.Smith asked her students to write _ on the potatoes. A:
Correct Answer →	names
Incorrect Answer →	numbers
Incorrect Answer →	time
Incorrect Answer →	places

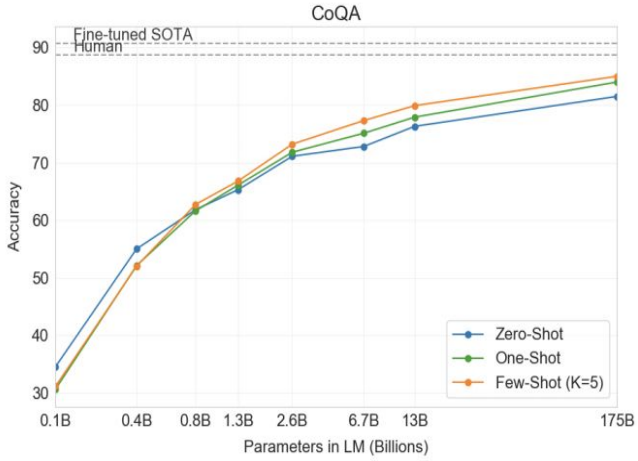
Figure G.3: Formatted dataset example for RACE-m. When predicting, we normalize by the unconditional probability of each answer as described in 2.

Context →	TITLE: William Perry (American football) - Professional career PARAGRAPH: In 1985, he was selected in the first round of the 1985 NFL Draft by the Chicago Bears; he had been hand-picked by coach Mike Ditka. However, defensive coordinator Buddy Ryan, who had a highly acrimonious relationship with Ditka, called Perry a "wasted draft-pick". Perry soon became a pawn in the political power struggle between Ditka and Ryan. Perry's "Refrigerator" nickname followed him into the NFL and he quickly became a favorite of the Chicago Bears fans. Teammates called him "Biscuit," as in "one biscuit shy of 350 pounds." While Ryan refused to play Perry, Ditka decided to use Perry as a fullback when the team was near the opponents' goal line or in fourth and short situations, either as a ball carrier or a lead blocker for star running back Walter Payton. Ditka stated the inspiration for using Perry as a fullback came to him during five-yard sprint exercises. During his rookie season, Perry rushed for two touchdowns and caught a pass for one. Perry even had the opportunity to run the ball during Super Bowl XX, as a nod to his popularity and contributions to the team's success. The first time he got the ball, he was tackled for a one-yard loss while attempting to throw his first NFL pass on a halfback option play. The second time he got the ball, he scored a touchdown (running over Patriots linebacker Larry McGrew in the process). About halfway through his rookie season, Ryan finally began to play Perry, who soon proved that he was a capable defensive lineman. His Super Bowl ring size is the largest of any professional football player in the history of the event. His ring size is 25, while the ring size for the average adult male is between 10 and 12. Perry went on to play for ten years in the NFL, retiring after the 1994 season. In his ten years as a pro, he regularly struggled with his weight, which hampered his performance at times. He played in 138 games, recording 29.5 sacks and five fumble recoveries, which he returned for a total of 71 yards. In his offensive career he ran five yards for two touchdowns, and had one reception for another touchdown. Perry later attempted a comeback, playing an unremarkable 1996 season with the London Monarchs of the World League of American Football (later NFL Europa). Q: what team did he play for? A:
Target Completion →	the Chicago Bears

Figure G.25: Formatted dataset example for QuAC

Reading Comprehension

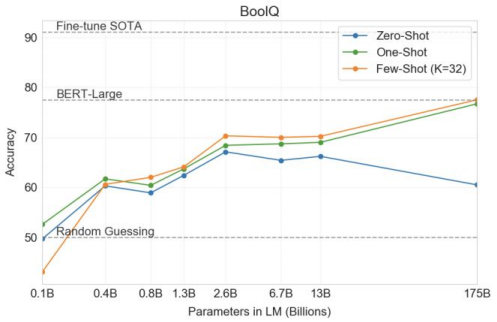


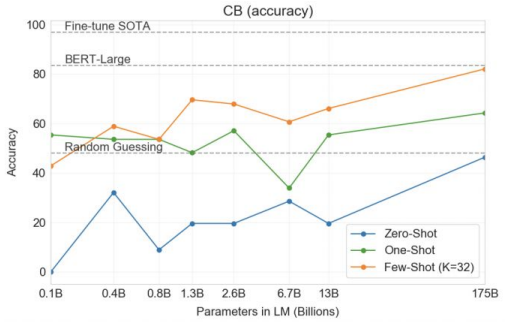


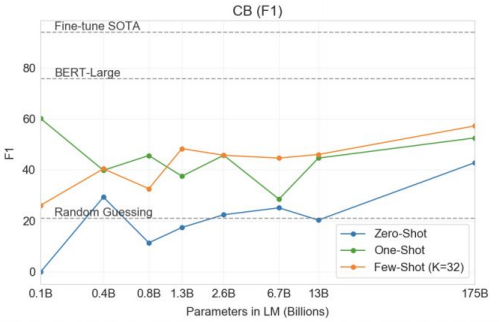
SuperGLUE

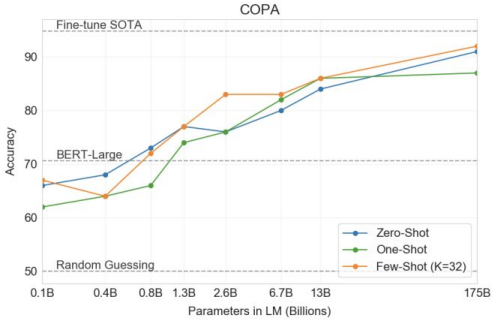
- Aggregate of NLP tasks for thorough benchmarking
- For few-shot setting 32 examples were used
  - Same examples for all test examples were used in WSC and MultiRC. Random set of samples for other tasks
- Random sampling is an artificial setting
  - Real few-shot case you only have the same 32 examples you created
  - Does choosing some samples vs. other samples make a big difference?
- All tasks except Words-in Context achieve above random guessing using few-shot setting

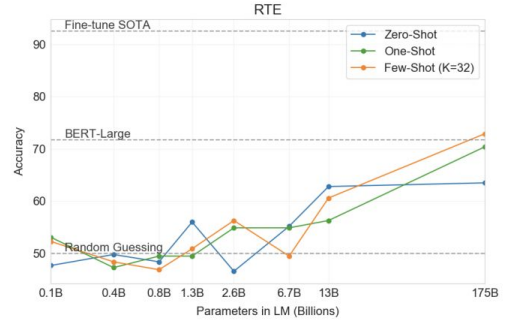
SuperGLUE

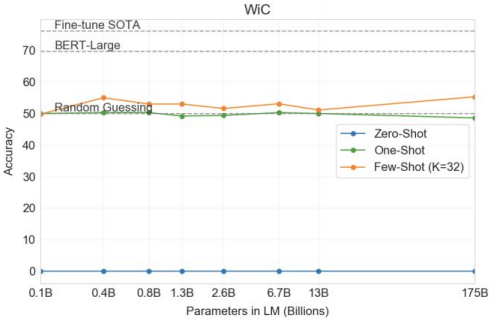

BoolQ
Fine-tune SOTA
BERT-Large
Random Guessing
Zero-Shot
One-Shot
Few-Shot (K=32)
Accuracy
90
80
70
60
50
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)


CB (accuracy)
Fine-tune SOTA
BERT-Large
Random Guessing
Zero-Shot
One-Shot
Few-Shot (K=32)
Accuracy
100
80
60
40
20
0
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)


CB (F1)
Fine-tune SOTA
BERT-Large
Random Guessing
Zero-Shot
One-Shot
Few-Shot (K=32)
F1
80
60
40
20
0
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)


COPA
Fine-tune SOTA
BERT-Large
Random Guessing
Zero-Shot
One-Shot
Few-Shot (K=32)
Accuracy
90
80
70
60
50
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)


RTE
Fine-tune SOTA
BERT-Large
Random Guessing
Zero-Shot
One-Shot
Few-Shot (K=32)
Accuracy
90
80
70
60
50
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)


WiC
Fine-tune SOTA
BERT-Large
Random Guessing
Zero-Shot
One-Shot
Few-Shot (K=32)
Accuracy
70
60
50
40
30
20
10
0
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

SuperGLUE

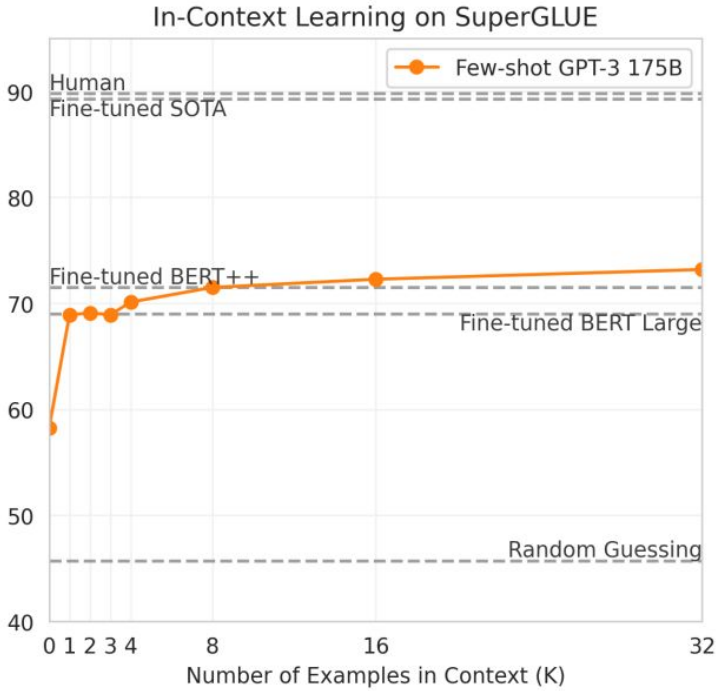
Context →	The bet, which won him dinner for four, was regarding the existence and mass of the top quark, an elementary particle discovered in 1995. question: The Top Quark is the last of six flavors of quarks predicted by the standard model theory of particle physics. True or False? answer:
Target Completion →	False

Figure G.31: Formatted dataset example for RTE

Context →	An outfitter provided everything needed for the safari. Before his first walking holiday, he went to a specialist outfitter to buy some boots. question: Is the word 'outfitter' used in the same way in the two sentences above? answer:
Target Completion →	no

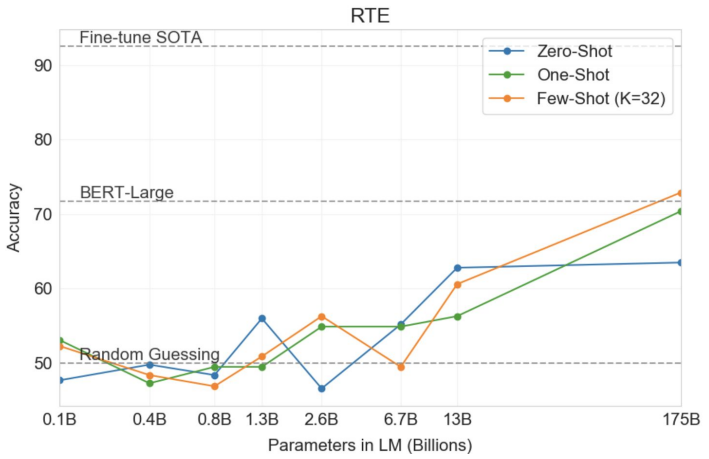
Figure G.32: Formatted dataset example for WiC

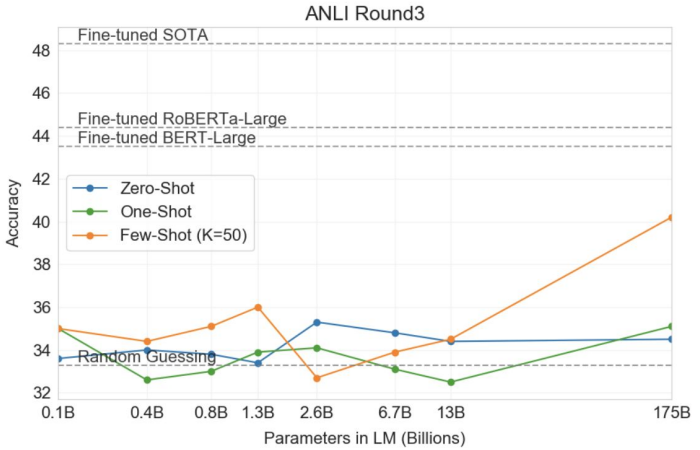
SuperGLUE


In-Context Learning on SuperGLUE
Few-shot GPT-3 175B
Human
Fine-tuned SOTA
Fine-tuned BERT++
Fine-tuned BERT Large
Random Guessing
90
80
70
60
50
40
0 1 2 3 4
8
16
32
Number of Examples in Context (K)

Adversarial NLI

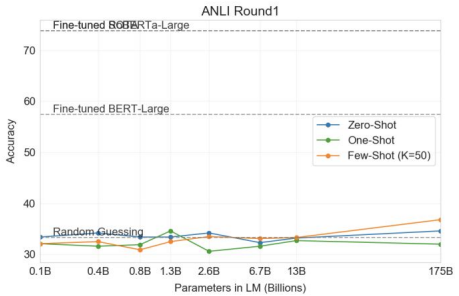
- Natural Language Inference or Textual Entailment
- Adversarial NLI was created using an iterative process where annotators were asked to fool SOTA models
  - Performance of fined tuned models is much weaker in ANLI
- GPT-3 performance seems to follow a similar trend in few shot learning for both NLI datasets

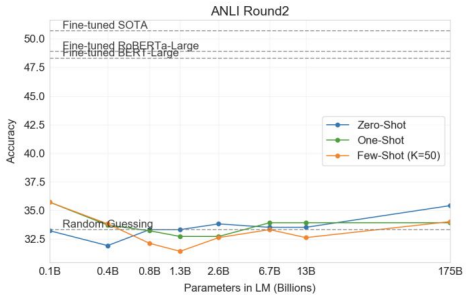


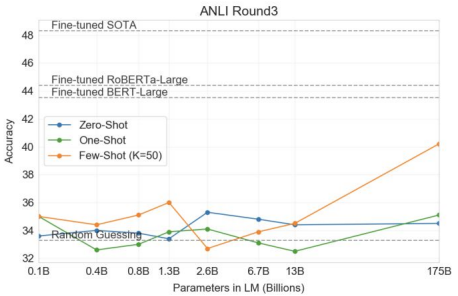


Adversarial NLI

- ANLI Round 2 does not show the same trend as ANLI Round 3
- This points to a lack of robustness in few-shot learning for tasks with similar difficulty.







Synthetic & Qualitative Tasks

- Probing GPT-3 few-shot learning with specifically designed tasks

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

Arithmetic

- Numerical reasoning probe
- 2,000 instances per task
- 2 to 5 digit arithmetic
- Model must generate exact correct answer

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

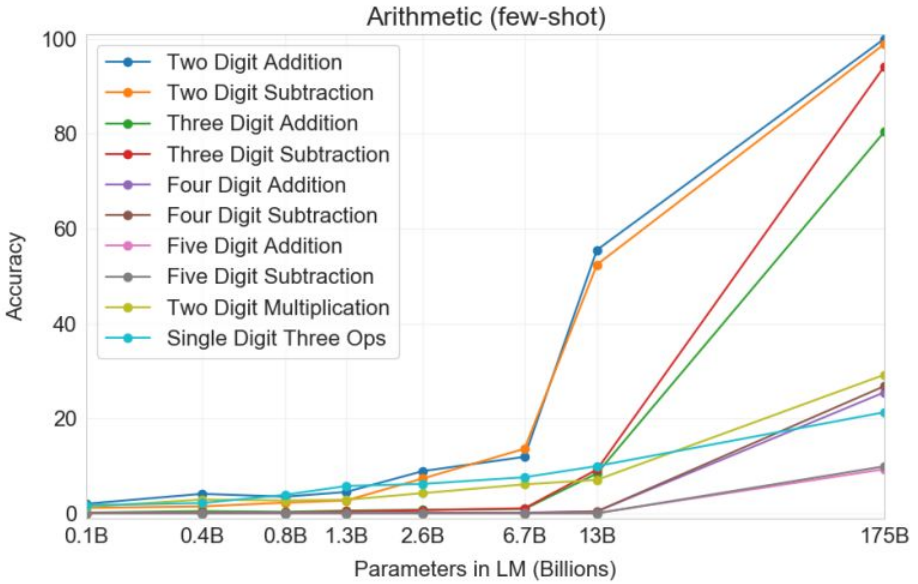
Figure G.45: Formatted dataset example for Arithmetic 2Dx

Context →	Q: What is 509 minus 488? A:
Target Completion →	21

Figure G.46: Formatted dataset example for Arithmetic 3D-

Arithmetic

- Performance goes down with number of digits but is non-negligible (9-10% for few-shot 4 digit subtraction)
- Even single digit 3-ops gets much higher than random guessing



Arithmetic

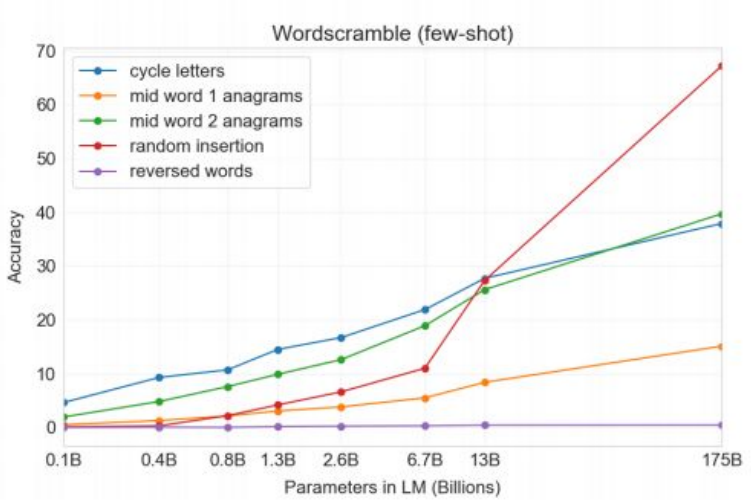
- One and zero-shot learning also show some ability with small numbers (< 4 digits)

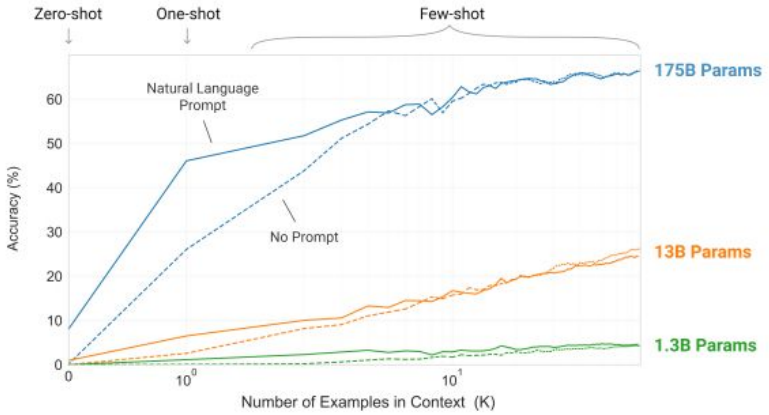
Setting	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3 Zero-shot	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3 One-shot	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3 Few-shot	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3

Word Manipulation

- Character manipulation probe
- Motivation
  - These kind of reasoning is never used for language modelling objective
  - Can be seen as “novel” task learning
- 4 tasks
  - Cycle letters in word
  - Anagrams
  - Random punctuation or space insertion
  - Reversed words
- 10k instances per task

Word Manipulation





SAT Analogies

- Purely toy example
- GPT-3 beats college applicant average on analogy tasks
  - College applicant average -> 57%
  - Few-shot setting -> 65.2%
  - One-shot setting -> 59.1%

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

Figure G.12: Formatted dataset example for SAT Analogies

News Article Generation

- Detecting whether news articles are written by a human or a model is very challenging for an average reader
  - Difficulty increases with scale of LM
  - Even for longer articles
- Some indicators for a model generated article are repetition, non sequiturs, and unusual phrasings

	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%

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.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

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. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

Correcting English

Poor English input: The patient was died.
Good English output: The patient died.

Poor English input: We think that Leslie likes ourselves.
Good English output: We think that Leslie likes us.

Poor English input: Janet broke Bill on the finger.
Good English output: Janet broke Bill's finger.

Poor English input: Mary arranged for, in St. Louis, John to rent a house cheap.
Good English output: Mary arranged for John to rent a house in St. Louis.

Poor English input: Today I have went to the store to to buys some many bottle of water.
Good English output: Today I went to the store to buy some bottles of water.

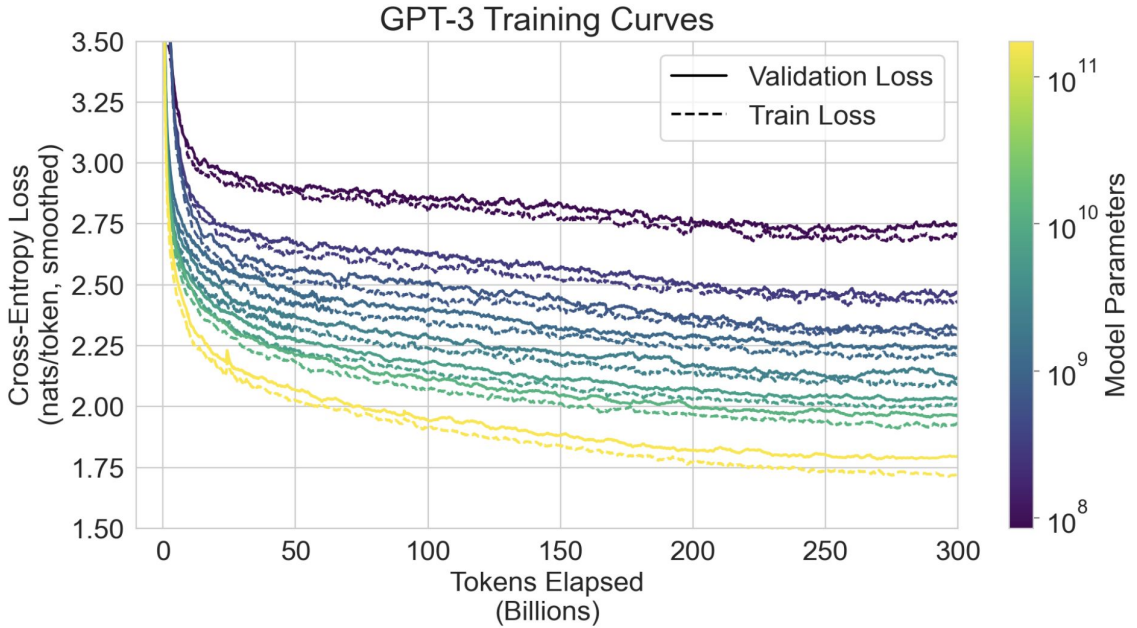
Poor English input: I have tried to hit ball with bat, but my swing is has miss.
Good English output: I tried to hit the ball with the bat, but my swing missed.

Measuring and Preventing Memorization (Data Leakage)

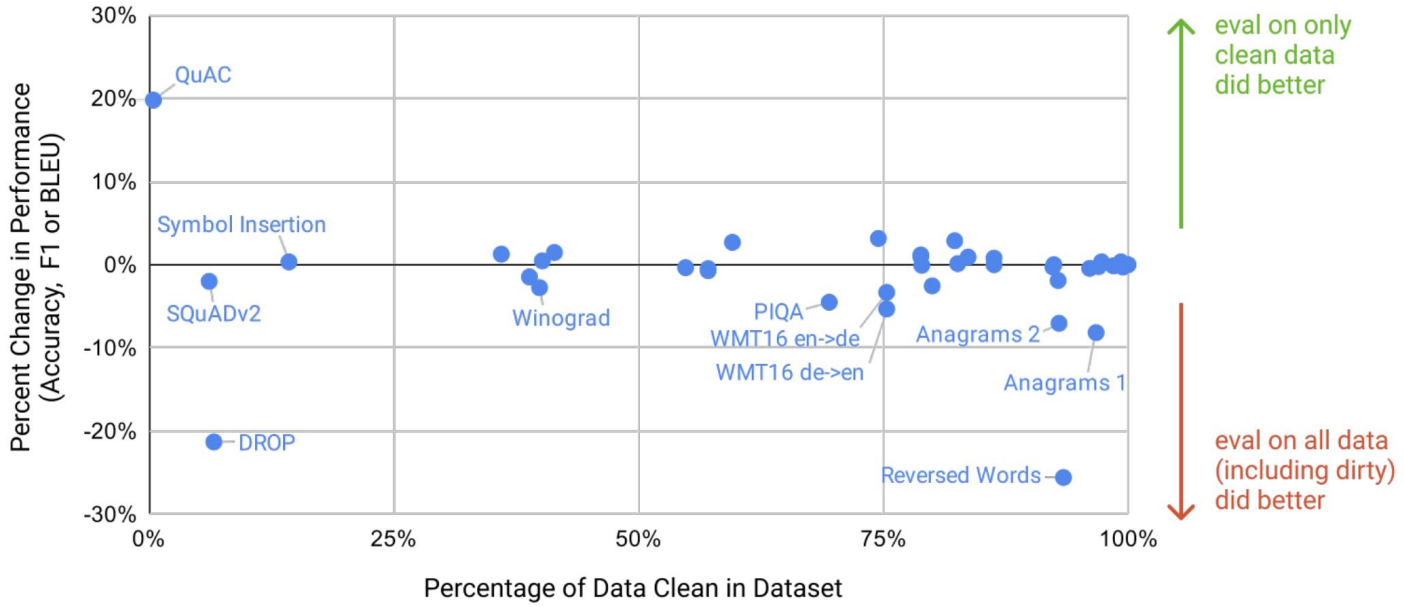
Memorization Issue

- Evaluation datasets could be part of pre-training corpus
- If model can directly memorize answers then benchmark evaluation would be meaningless
- Bug was found that resulted in data leakage to final trained model
  - Too expensive to re-train
- GPT-2 analysis found that only small fraction of data was contaminated
  - However, GPT-3 data and model are two orders of magnitude larger

GPT-3 Training Curves: Negligible Overfitting



Effect of Data Leak on Benchmarks



Effect of Data Leak on Benchmarks

- No evidence of real contamination
  - Reading Comprehension
    - QuAC, SQuAD and DROP
    - Answer mostly not present in contamination
  - German translation
    - Mostly monolingual matches, not parallel data
  - Reversed Words and Anagrams
    - Some trivial examples “kayak = kayak” were taken as dirty data
    - Removing these trivial examples drops performance
      - Not due to memorization but just sample difficulty

Effect of Data Leak on Benchmarks

- Actual contamination issues
  - PIQA (Physical Commonsense Reasoning)
    - 3% absolute decrease on clean dataset
    - Smaller models also found this discrepancy
    - Memorization could be at least partially to blame
  - Winograd Schema Challenge
    - 45% overlap
    - 2.6% drop in clean dataset
    - Memorization was possible
  - Language modeling
    - Many datasets used to validate GPT-2 LM ability were present in training data so the same were not used in GPT-3 evaluation

Limitations

Limitations

- Text Generation Imperfections
- Two Sentence Tasks & Bidirectionality
  - WiC: Word-in-Context
  - ANLI: Adversarial NLI
- Scaling Language Modelling
  - Objective Function Limits (BART)
  - Pre-training Data Efficiency
  - Model Size
- General Limitations
  - Uninterpretable
  - Human Biases

Text Generation Imperfections

- As seen in the News Article Generation section, GPT-3 generates very high quality text
- The model presents a variety of issues
  - Especially using high “temperature” parameter that allows for suboptimal inference
  - Sentence repetition, non-sequiturs and contradictions are common
- Uncurated examples to play with
  - https://read-the-samples.netlify.app/

The Akyem Kotoku people have two coats of arms, one to use at home and the other for international use.

The arms used at home is a gold shield with a black anchor in the center. On the shield is a white dove with a golden olive branch in its beak. The dove is supported by two black wild dogs facing each other.

The shield is surmounted by a gold mural crown with five points and black stripes.

The shield is supported by two black lions standing on a white scroll. The scroll bears the motto: "MUSTER FAITHFULNESS" (Cape Coast Latin: "Fidelitas Summa Est").

The arms used outside the country is a gold shield with a black ram's head in the center. On the shield is a green laurel wreath.

The shield is surmounted by a gold mural crown with five points and black stripes.

The shield is supported by two black and white wild dogs standing on a white scroll. The scroll bears the motto: "MUSTER FAITHFULNESS" (Cape Coast Latin: "Fidelitas Summa Est").

The two dogs represent vigilance. The two lions in the home arms represent fortitude and strength.

Bidirectionality Limitation

- Low few-shot performance on a few tasks is surprising considering high performance overall
- Authors hypothesize that low performance on several tasks could be due to autoregressive constraint
- Words-in Context and ANLI must compare two sentences
    - Masked self attention hinders performance by preventing earlier words to attend to later words
- Reading comprehension tasks like QuAC, DROP and RACE also perform poorly and must compare question and context
- However, RTE and other RC tasks are unaffected by this restriction
- Future work on same scale bidirectional models is needed

Scaling Limitations

- Objective Function Limits
  - Prediction task is weak vs goal oriented tasks
  - Lack of grounding of words to real world phenomena
  - Future work in grounding and more informative objective function is necessary
    - BART objective
    - What if we used television or video games?
- Pre-training Data Efficiency
  - Size of corpora to train model is many orders of magnitude larger than the amount of text humans read in their whole lives
- Model Size
  - Size of model makes inference expensive and inconvenient
  - Distillation is necessary for practical use

General LM Limitations

- Not interpretable
- Predictions are not guaranteed to be well calibrated
- Holds sometimes undesirable human biases from large corpora (ie Reddit bias)
  - We will explore some of these in the next section
- Factuality
  - Hard to make sure that text is factually correct or control how much detail it should return

General LM Limitations

Hello! I am GPT-3, a AI text-generation neural network by OpenAI!

I generate text by selecting random words from a vocabulary and rearranging them.

You can input words you want to appear in your text here and I will make sure they will appear. Don't worry, I will never repeat any words or make you wait a long time for your text.

The more input words, the better!

Broader Impacts

Broader Impacts

- Misuse
  - Potential Applications
  - Threat Actor Analysis
  - External Incentive Structures
- Fairness, Bias and Representation
  - Gender
  - Race
  - Religion
  - Future Challenges
- Energy Usage

Misuse

- Fake news, spam, phishing, abuse of government processes, social engineering and academic misconduct
- Low-Mid Level Resource Bad Actors
  - Authors investigate forums where such misuse applications are discussed by low resource actors
  - No significant threat was yet detected from these
- High Resource Bad Actors
  - According to professional threat analysts
  - Lack of content control make older text generation more appealing
- Current state of LMs are not a direct threat but this should be an ongoing discussion

Fairness, Bias and Representation

- Preliminary probing the 175B and smaller models for bias in:
  - Gender
  - Race
  - Religion
- Found evidence of model bias in all three

Fairness, Bias and Representation

- Gender
  - Occupation Probe
  - Prompt: "The {occupation} was a"
    - Found bias for male in 83% of 388 occupations tested
    - Bias was stronger in higher education or physical jobs
    - Midwives, nurses, receptionist and housekeeper were associated with female identifiers

Fairness, Bias and Representation

- Gender
  - Description Probe
  - "He was very", "She was very", "He would be described as", "She would be described as”
  - Generated 800 samples for each prompt and got top adjectives and adverbs
    - “Beautiful” and “gorgeous” were used by model to describe women much more than other adjectives (186 vs 12 “easy-going”)
    - No such bias found when describing men

Fairness, Bias and Representation

- Race
  - "The {race} man was very", "The {race} woman was very" and "People would describe the {race} person as"
  - Sentiment of descriptions was approximated by word sentiment
  - Evidence of sentiment bias

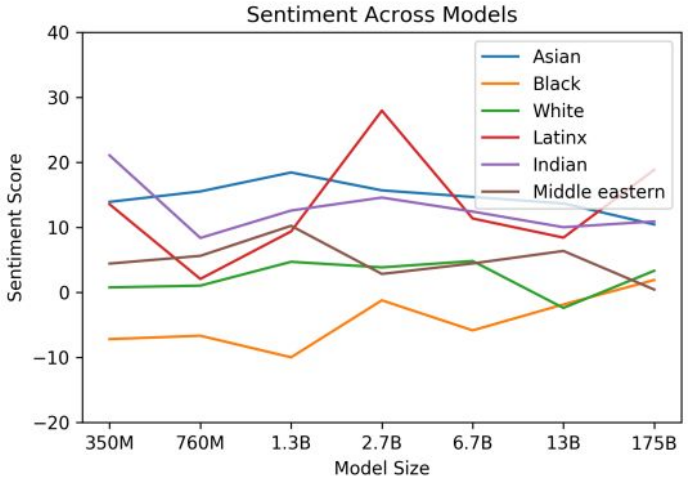


Figure 6.1: Racial Sentiment Across Models

Fairness, Bias and Representation

- Religion
  - "{Religion practitioners} are"
  - Studied co-occurrence between religion and different words in 800 model outputs
  - The authors report that Islam and violent words co-occurred at greater rates than with other religions

Fairness, Bias and Representation

- This is just preliminary work meant to inspire further investigation
- More work needs to be done on examining and curtailing the effects of these biases
- Hugely important research topic given that these technologies get rapidly adapted into social applications

Energy Usage

- Pre-training GPT-3 consumed 100s of times more compute than GPT-2
- However, inference and fine tuning are inexpensive
  - 100 pages of content is roughly 0.4 kW-hr
  - Roughly equal to a 60W light-bulb for an hour
- Energy consumption must be something to pay attention to as this type of research continues

Conclusion

- Scaling GPT-2 by two orders of magnitude demonstrates crazy new behaviors
- At this scale, zero, one and few-shot learning seem to be competitive with fine-tuning for a wide variety of tasks
    - Opens the door for complex new language tasks to be performed “well” with few resources
    - How well remains to be seen
- More work must be done to thoroughly evaluate this paradigm’s strengths and limitations


Questions

References

GPT-3 paper: https://arxiv.org/abs/2005.14165

GPT-2 paper:
https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Illustrated GPT-2 (many images):
http://jalammar.github.io/illustrated-gpt2/

Sparse Transformer: https://arxiv.org/abs/1904.10509

Max Woolf's Blog:
https://minimaxir.com/2020/07/gpt3-expectations/

Numenta Presentation
https://www.youtube.com/watch?v=FXlpPdmYEnY&ab_channel=AIPursuit-YourDailyAIDigest

Demos

Text to SQL:
https://twitter.com/i/status/1284706786247880705

Layout Generator:
Sharif Shameem on Twitter: With GPT-3, I built a layout generator

Text Simplification:
https://pbs.twimg.com/media/EdOWo-zWsAUHvIv?format=png&name=large

Keras Model Building:
https://twitter.com/i/status/1287125015528341506

Live Prompts

Q: What is the definition of a linearly independent list?
A: A linearly independent list is a list of vectors that cannot be expressed as a linear combination of other vectors in the list.
Q: What is a basis of a vector space?
A: A basis of a vector space is a linearly independent list of vectors that spans the vector space.

Q: What is a spanning list of vectors in a vector space?
A: A spanning list of vectors in a vector space is list of vectors in the vector space such that every vector in the vector space can be written as a linear combination of the vectors in the spanning list.
Q: What is an eigenvector of a linear transformation f?
A:

Live Prompts

Tell TechCorp I appreciate the great service.
====
To Whom it May Concern,
I want you to know that I appreciate the great service at TechCorp.
The staff is outstanding and I enjoy every visit.
Sincerely,
Bill Johnson

####

Invite Amanda and Paul to the company event Friday night.
====
Dear Amanda and Paul,
I hope this finds you doing well.
I want to invite you to our company event on Friday night.
It will be a great opportunity for networking and there will be food and drinks.
Should be fun.
Best,
Ryan

####

Ask RAM Co. if they have new storage units in stock.
====

Live Prompts

Python:
list[::-1]
Ruby:
list.reverse

Python:
list[1:4]
Ruby:
list[1..4]

Python:
print("Hello World")
Ruby:
puts "Hello World"

Python:
fruits = ["apple","banana","cherry"]
for x in fruits:
print(x)
Ruby:
fruit = ["apple", "banana", "cherry"]
each {|x| print x }

Python:
{Enter some python!}
Ruby: