

Language Models are Few-Shot Learners

NeurIPS 2020

Brown et al.

39k+ citations

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	

Introduction

Context:

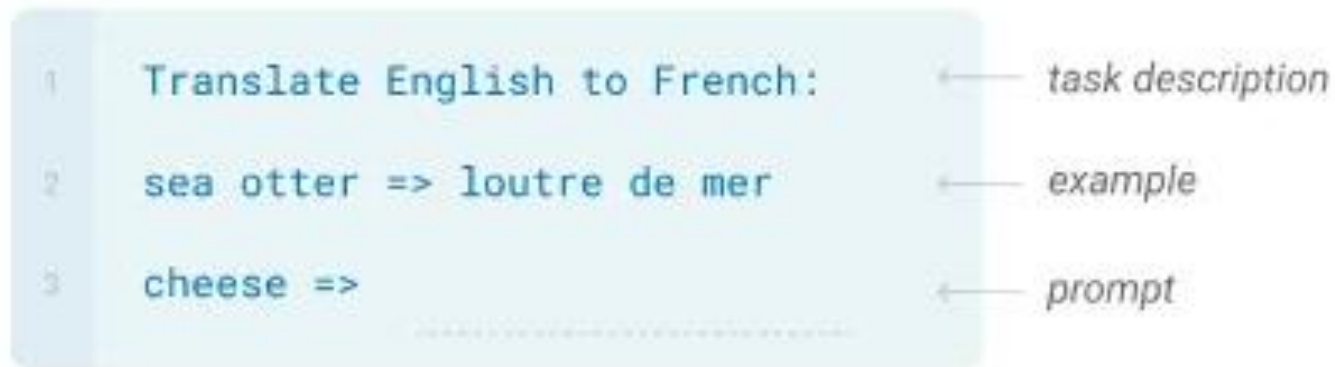
- Recent advances in NLP have shown big improvements by first training models on large text datasets and then fine-tuning them for specific tasks.
- While these models are designed to work across many tasks, they still need large datasets with thousands of examples for fine-tuning.

Problem:

- How can we build a model that generalizes across tasks without fine-tuning?

Some Definitions

- **One-shot learning:** allow just one demonstration for context.



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

The diagram shows a light blue rounded rectangle containing three lines of text. To the right of each line is a label with an arrow pointing to it. Line 1 is 'Translate English to French:' with the label 'task description'. Line 2 is 'sea otter => loutre de mer' with the label 'example'. Line 3 is 'cheese =>' with the label 'prompt'. The text is in a monospaced font.


Some Definitions

- **Few-shot learning:** give K (10 – 100 to fit into model's context window) examples of context and completion, then one final example of context, with the model expected to perform the completion.



Some Definitions

- **Zero-shot learning:** allow no demonstrations, only a natural language instruction is given to the model.



The diagram illustrates a zero-shot learning prompt structure. It consists of two lines of text within a light blue rounded rectangle. The first line is labeled '1' and contains the text 'Translate English to French:'. To its right, an arrow points from the text 'task description'. The second line is labeled '2' and contains the text 'cheese =>'. To its right, an arrow points from the text 'prompt'. Below the second line, there is a series of dots indicating the expected output.

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

Fine-tuning

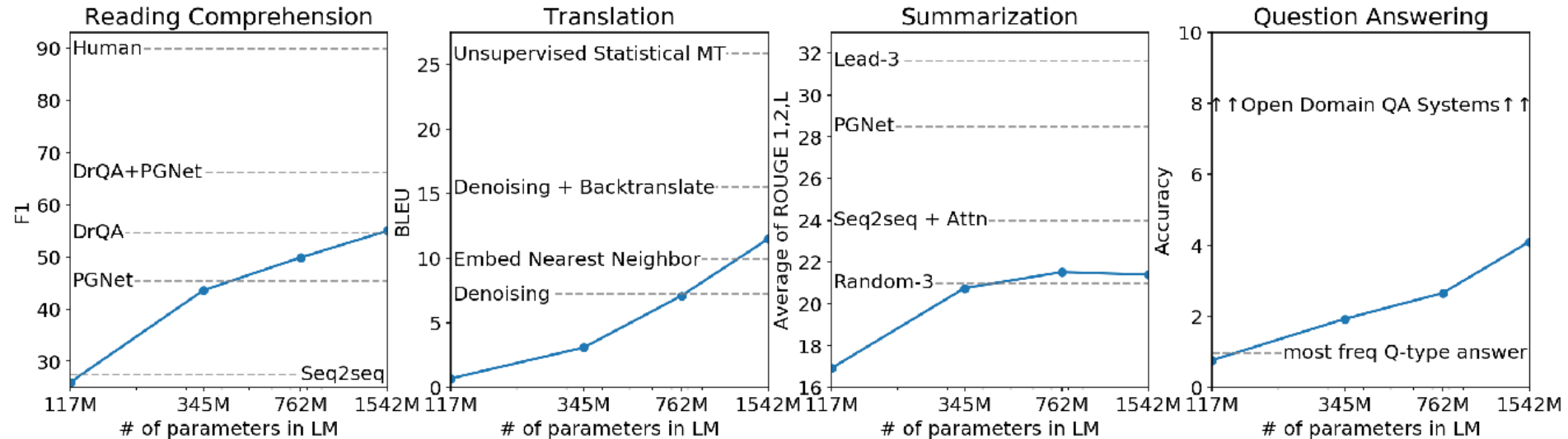
- Adapts a **pre-trained language model** to perform a **specific task** by further training it (updating the model's parameters) on a labeled dataset for that task.
- Many fine-tuned models had achieved state-of-the-art performance on individual tasks.
- **Disadvantages:**
 - Requires a large dataset of labeled examples for the target task.
 - Requires additional resources for each new task (time, hardware, energy).
 - Risks overfitting or becoming too specialized and losing generalization ability or flexibility.

Fine-tuning



GPT-2

Language Models are Unsupervised Multitask Learners



- Largest model of 1.5 billion parameters
- Zero-shot behavior was not very good
 - Translation (~11 BLEU)
 - Summarization – no better than picking 3 random sentences from the document
- For many tasks performance scaled smoothly with model size increase

Scaling GPT-2

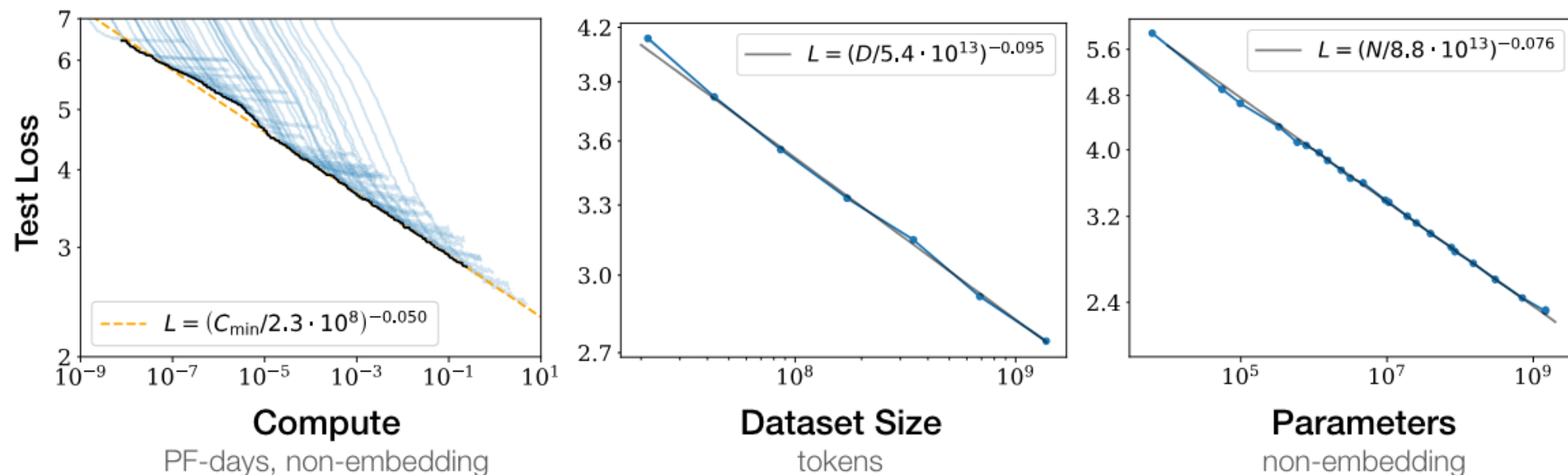
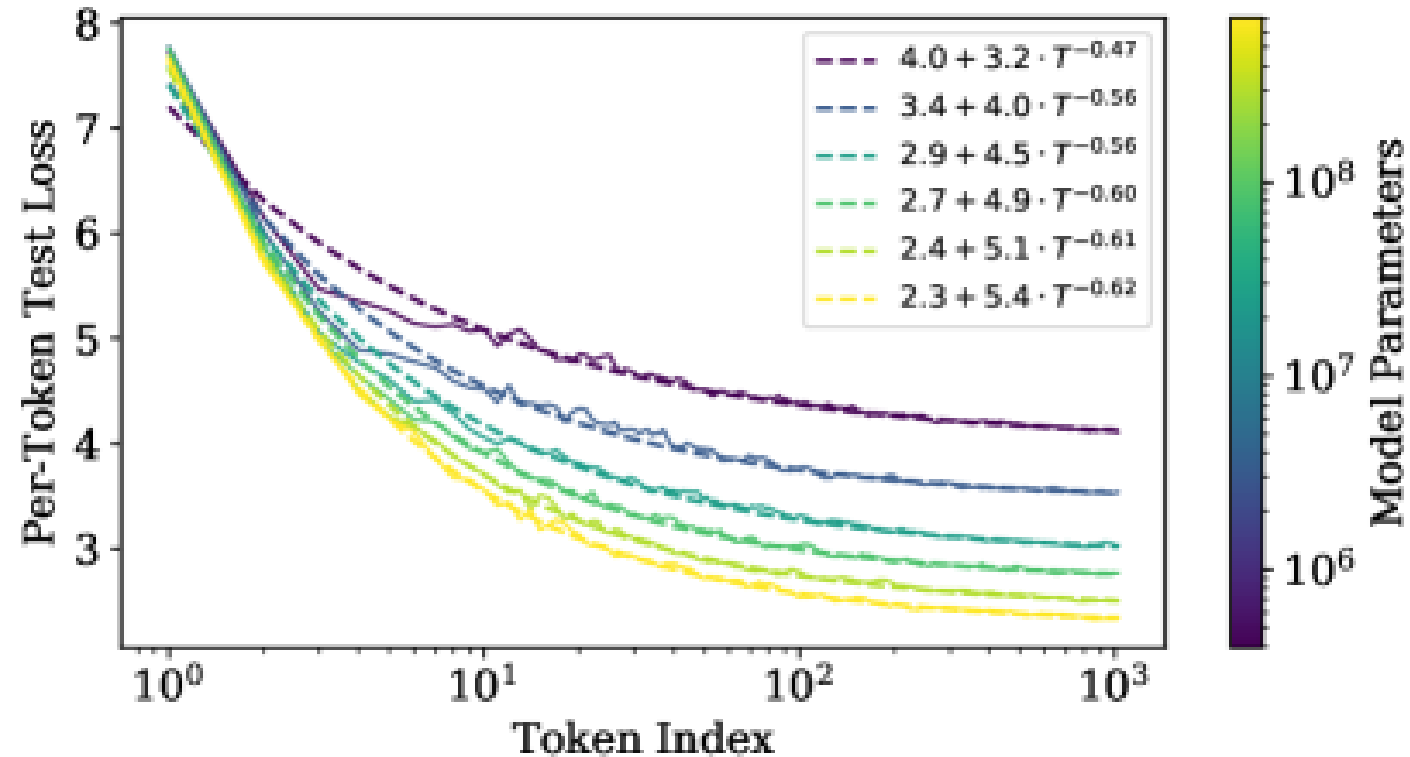


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute² used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Scaling GPT-2



- **Key takeaway:** Larger models have better performance, and that performance gets better with more context.
- **In other words:** A larger model size improves few-shot performance.

GPT-3

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.

GPT-3

- **Autoregressive** language model
 - It generates text by predicting the next word (or token) in a sequence based on the words that came before it.
- Not **bidirectional** (left-to-right only, no future context)
 - May struggle with
 - Ambiguous words
 - Understanding full sentence meaning
 - Question answering on long passages

GPT-3 Architecture

- Same architecture as GPT-2
- Transformer-based with 96 attention layers
- Alternating dense and sparse attention layers
 - **Dense attention layers:**
 - Every token attends to all tokens in the sequence.
 - Captures global dependencies well.
 - **Sparse attention layers:**
 - Each token attends to a subset of other tokens (at fixed intervals or nearby words).
 - Reduces computational complexity for faster, more memory-efficient training.

GPT-3 Architecture

- 2048-token context window
 - **Token context window:** the maximum number of tokens a language model can consider at once when making predictions. It defines how much past information the model can "remember" when generating text.
 - Doubled from GPT-2.

GPT-3: Key Innovations

- **Massive Scale:** GPT-3 with 175 billion parameters, 100x+ larger than GPT-2.
- **Few-Shot Learning:** Generalizes tasks using in-context examples and has strong performance without fine-tuning.
 - Parameters remain fixed.
 - Does not need task-specific labeled datasets and training.
 - Can rapidly adapt to new tasks.
- **Broad Evaluation:** Competitive results across language modeling, translation, and question answering tasks.

Datasets for Training 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

= Over 45 terabytes of data from the internet and books.

Datasets for Training GPT-3

- **Data Contamination:**

- The model might accidentally learn parts of the test or development sets during training, which could affect performance on later tasks.
 - Key issue with large language models trained on internet data.
- To reduce contamination, the authors searched for and attempted to remove any overlaps with the development and test sets of all benchmarks studied in this paper.
- A bug in the filtering caused them to ignore some overlaps.
- Performed an analysis of the impact of contamination on performance for different datasets.

GPT-3 Few-Shot Performance

- Achieves state of the art (SOTA):
 - Language Modeling
 - Long-Range Text Dependencies
 - Closed-Book Question Answering
 - Translation to English
 - Common Sense Reasoning
- Poor performance:
 - Natural Language Inference (NLI)
 - Multi-Step Arithmetic
 - Word Scrambling and Manipulation

Language Modeling

- Penn Tree Bank (PTB) dataset
 - Predates the modern internet
 - No overlap with training dataset
 - Predicts the next word in a sentence
- Largest model (zero-shot) achieves state of the art by a margin of 15 perplexity points (20.5 perplexity).

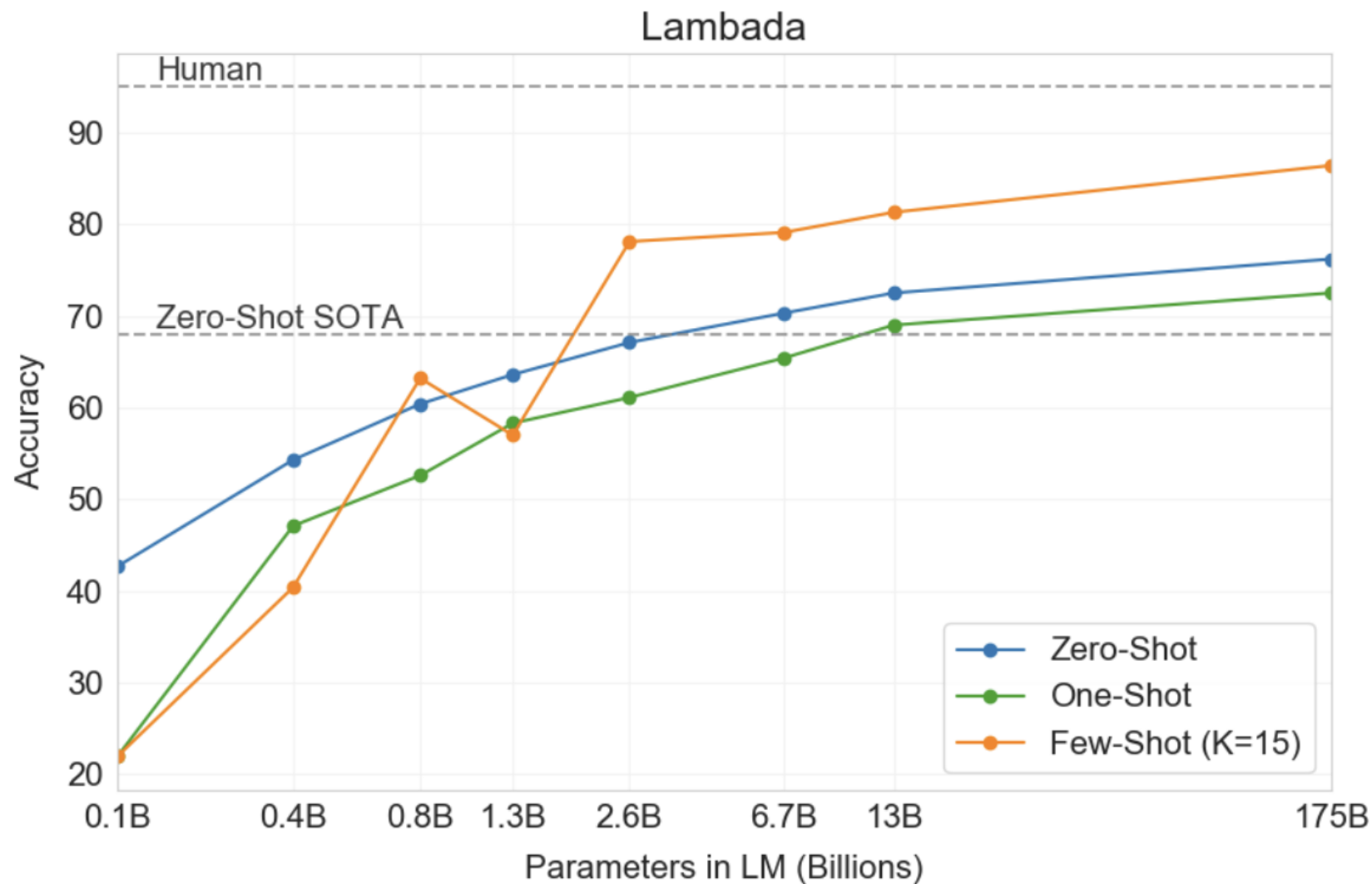
Long-Range Text Dependencies

- LAMBADA
 - Asks the model to predict the last word of sentences after reading a paragraph of context.
- GPT-3 advances the state-of-the-art by 18% accuracy.

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

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

Long-Range Text Dependencies



Long-Range Text Dependencies

- HellaSwag
 - Involves picking the best ending to a story or set of instructions.
 - Difficult for models, but easy for humans (95.6% acc).
- StoryCloze
 - Select the correct ending sentence for five-sentence long stories.

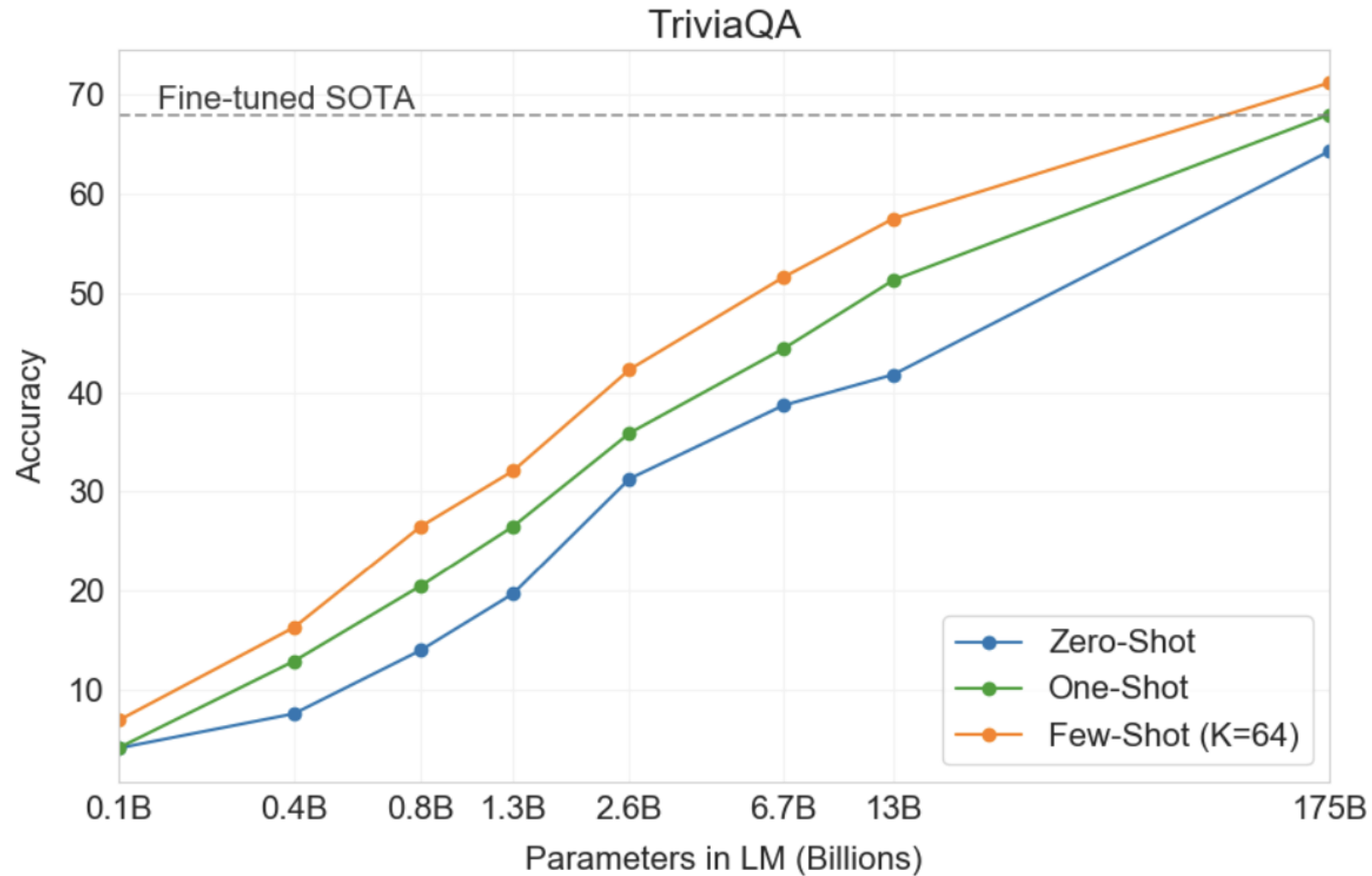
Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0 ^a	8.63 ^b	91.8^c	85.6^d
GPT-3 Zero-Shot	76.2	3.00	83.2	78.9
GPT-3 One-Shot	72.5	3.35	84.7	78.1
GPT-3 Few-Shot	86.4	1.92	87.7	79.3

Closed Book Question Answering

- Open-book = model searches for and conditions on text that contains the answer.
- Closed-book = directly answer the questions without conditioning on auxiliary information.

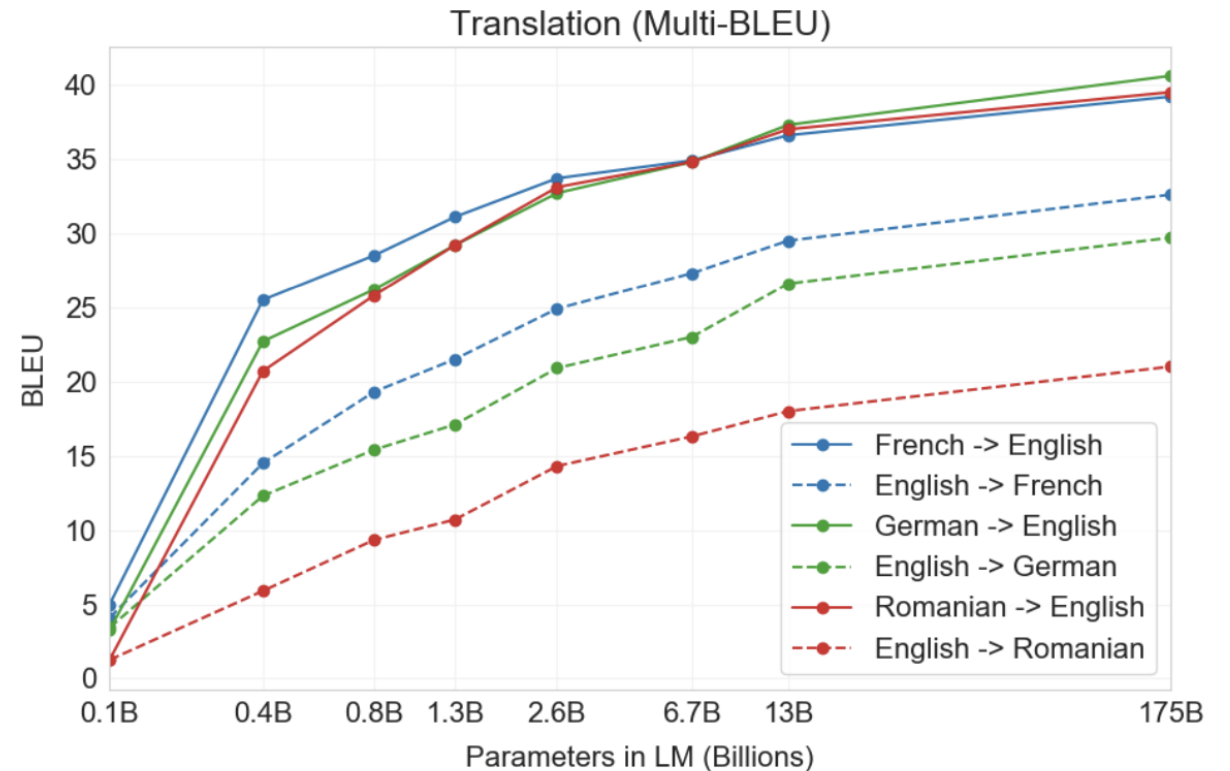
Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP⁺20]	44.5	45.5	68.0
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	68.0
GPT-3 Few-Shot	29.9	41.5	71.2

Question Answering



Translation

- A single example demonstration for each translation task improves performance by over 7 BLEU
- GPT-3 in the full few-shot setting further improves another 4 BLEU
- The SOTA is achieved by a combination of unsupervised pretraining, supervised finetuning on 608K labeled examples, and backtranslation.



Winograd-Style Tasks

- Winograd Schemas Challenge – determine which word a pronoun refers to



Common Sense Reasoning

- Datasets that attempt to capture physical or scientific reasoning
- **PhysicalQA (PIQA)** – asks common sense questions about how the physical world works
- **ARC** – multiple-choice questions from 3rd to 9th grade science exams
 - “**Challenge**” – filtered so that they cannot be answered using simple statistical or information retrieval methods
- **OpenBookQA** - multiple-choice elementary-level science questions that would be incorrectly answered by a retrieval-based algorithm

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

Reading Comprehension

- Test with a suite of five datasets that include abstractive and multiple-choice questions, and span answer formats in both dialog and single question settings.
- Scores are F1, except RACE which are accuracy.

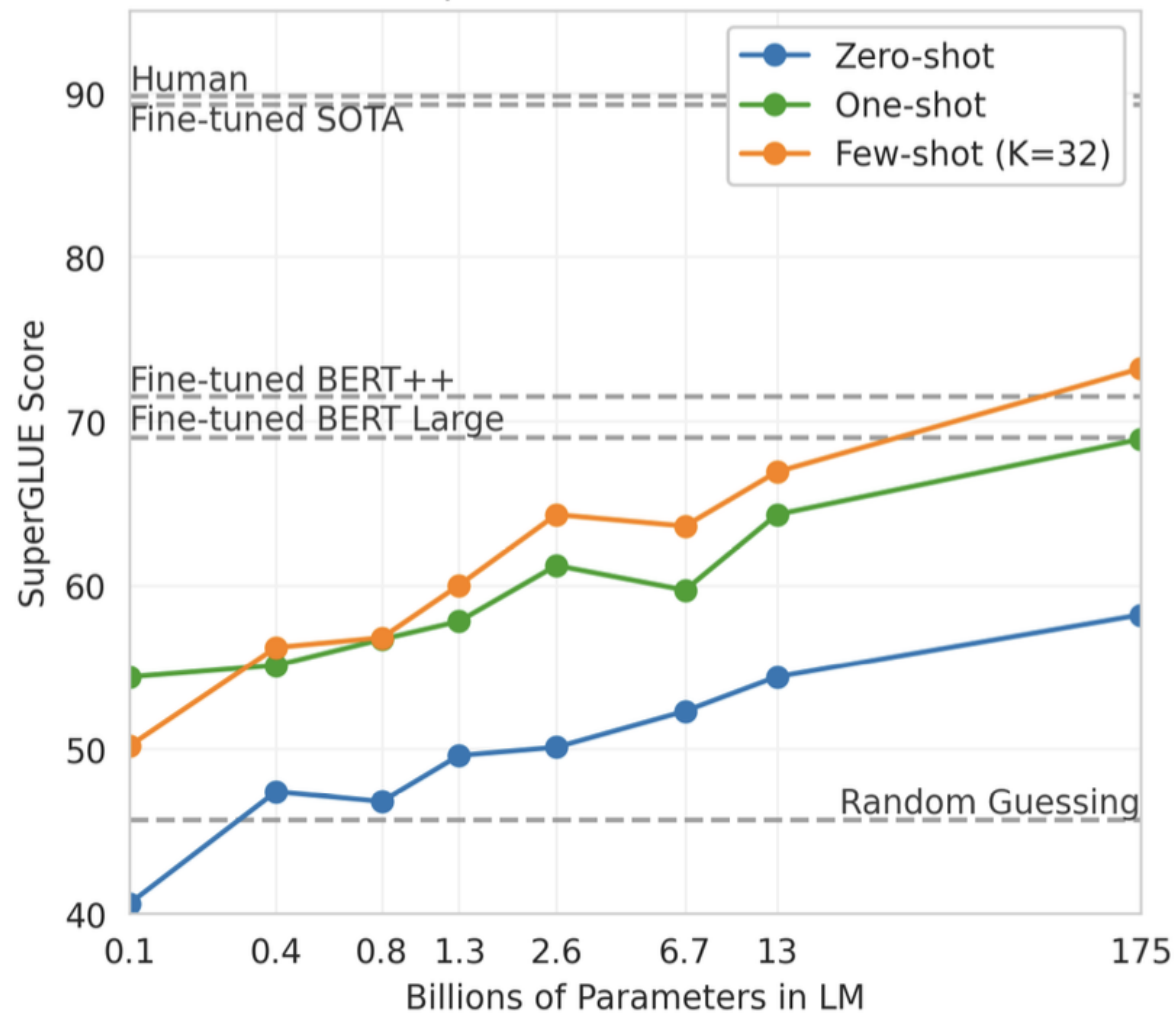
Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	90.7^a	89.1^b	74.4^c	93.0^d	90.0^e	93.1^e
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

SuperGLUE Benchmark

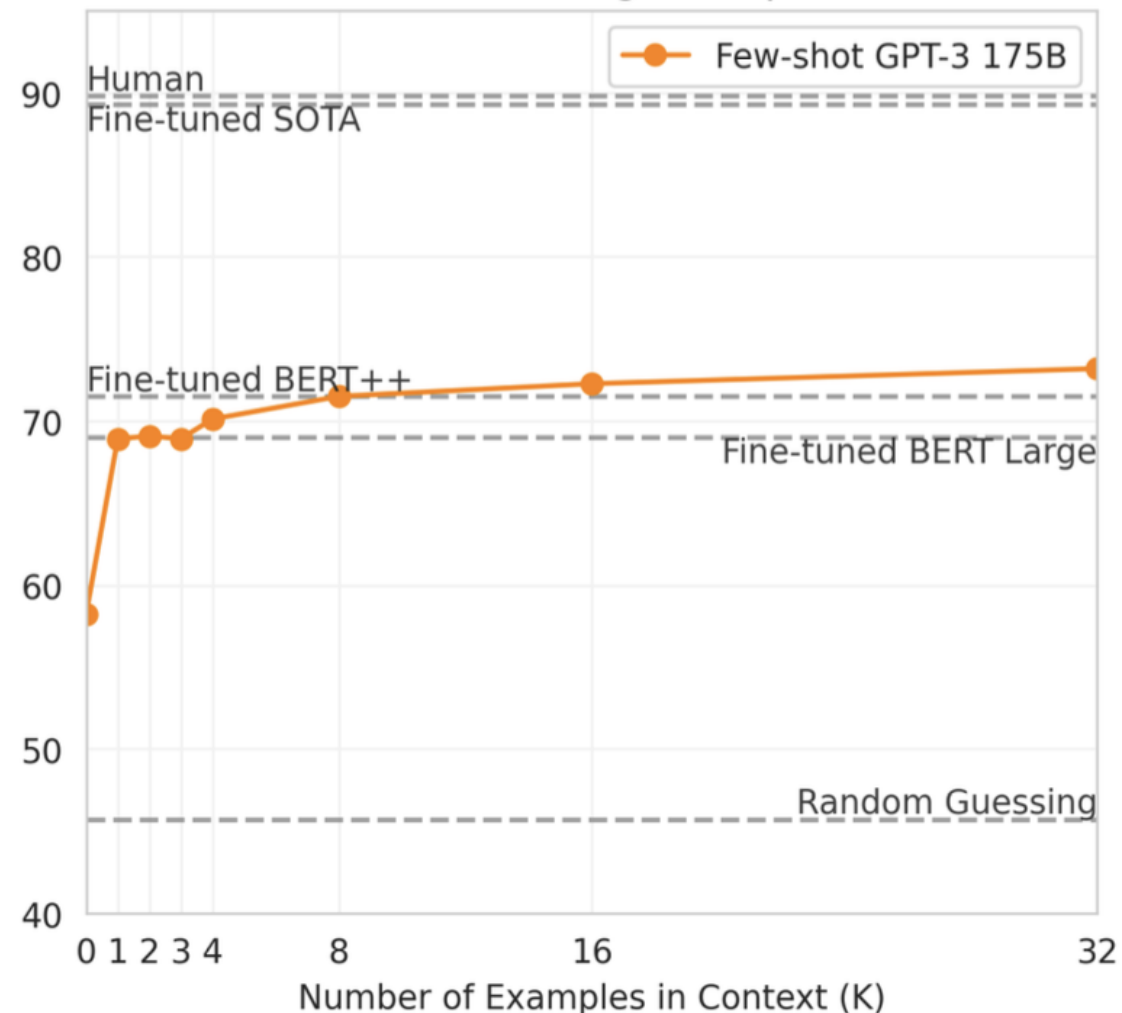
- A standardized collection of datasets.
- Wide range in GPT-3's performance across tasks.
- GPT-3 appears to be weak in the few-shot or one-shot setting at some tasks that involve comparing two sentences or snippets.
 - For example: whether a word is used the same way in two sentences, whether one sentence is a paraphrase of another, or whether one sentence implies another.
- GPT-3 still outperforms a fine-tuned BERT-large on 4/8 tasks and on two tasks GPT-3 is close to the state-of-the-art which is held by a fine-tuned 11 billion parameter model.

SuperGLUE Benchmark

SuperGLUE Performance

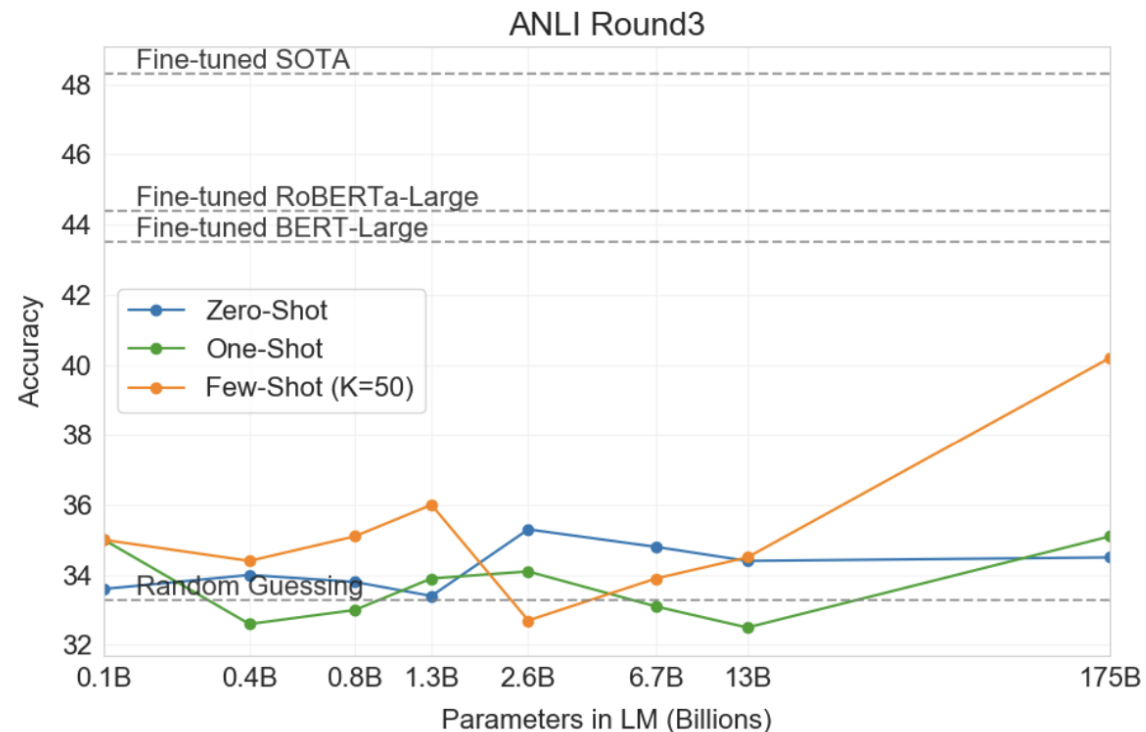


In-Context Learning on SuperGLUE



Natural Language Inference (NLI)

- The ability to understand the relationship between two sentences.
- **Adversarial Natural Language Inference (ANLI):** employs a series of adversarially mined natural language inference questions in three rounds.

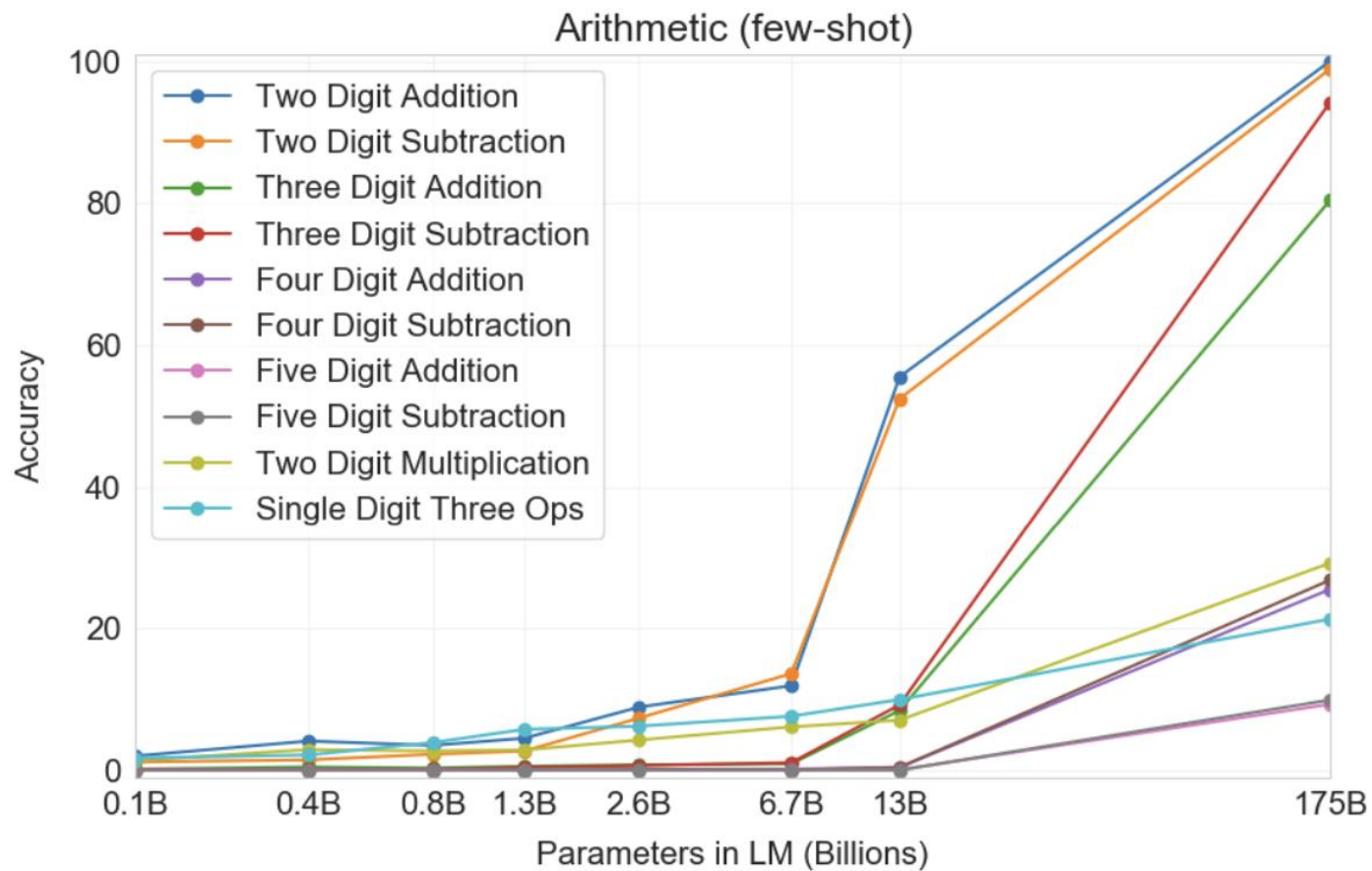


Arithmetic

- Ask a simple arithmetic problem in natural language.
- Results are progressively stronger moving from the zero-shot to one-shot to few-shot setting, but even the zero-shot shows significant arithmetic abilities.
- One-shot and zero-shot performance are somewhat degraded relative to few-shot performance.
 - Suggests that adaptation to the task (or recognition of the task) is important to performing these computations correctly.

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

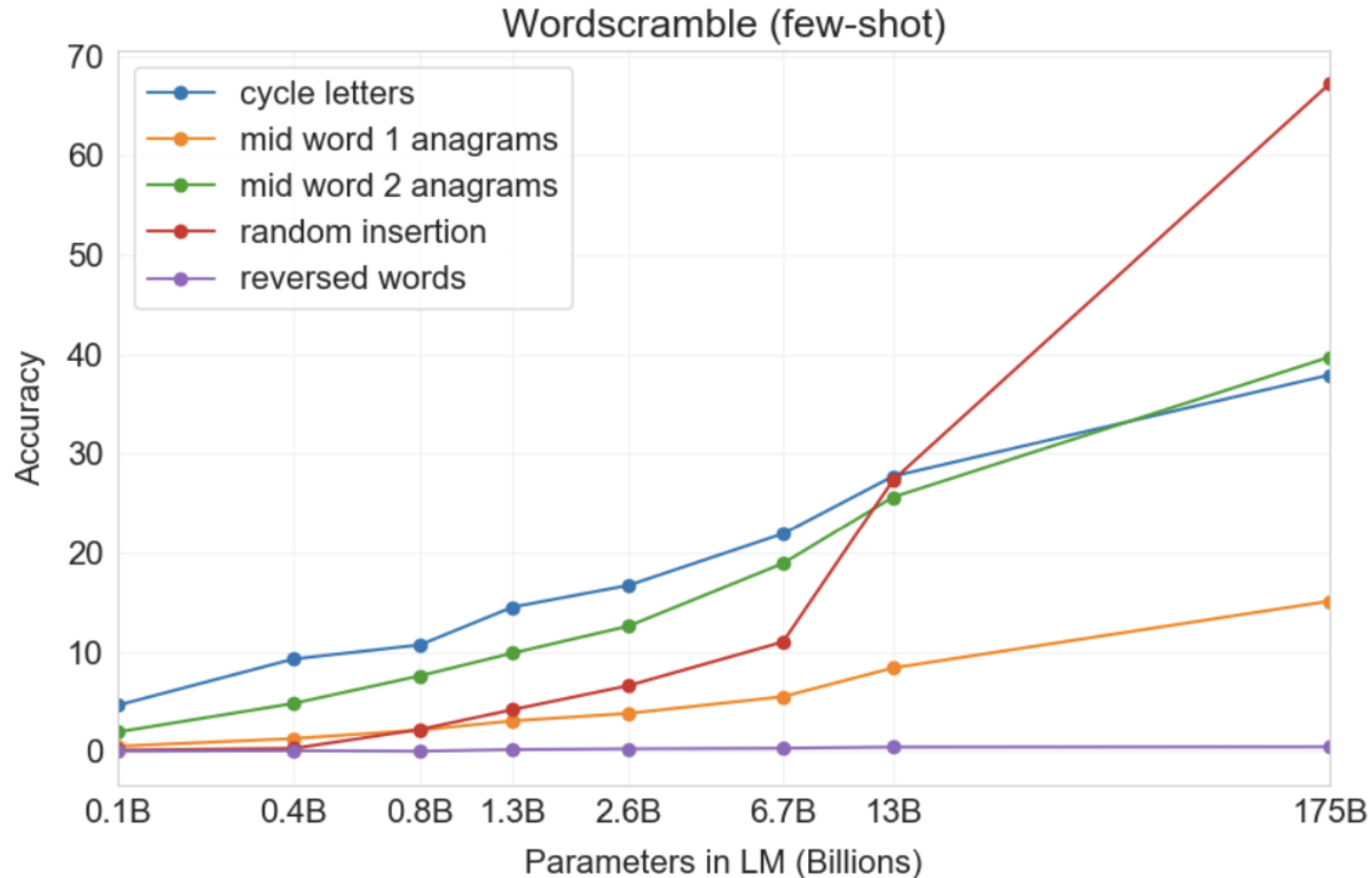
Arithmetic – Few-Shot



Word Scrambling and Manipulation

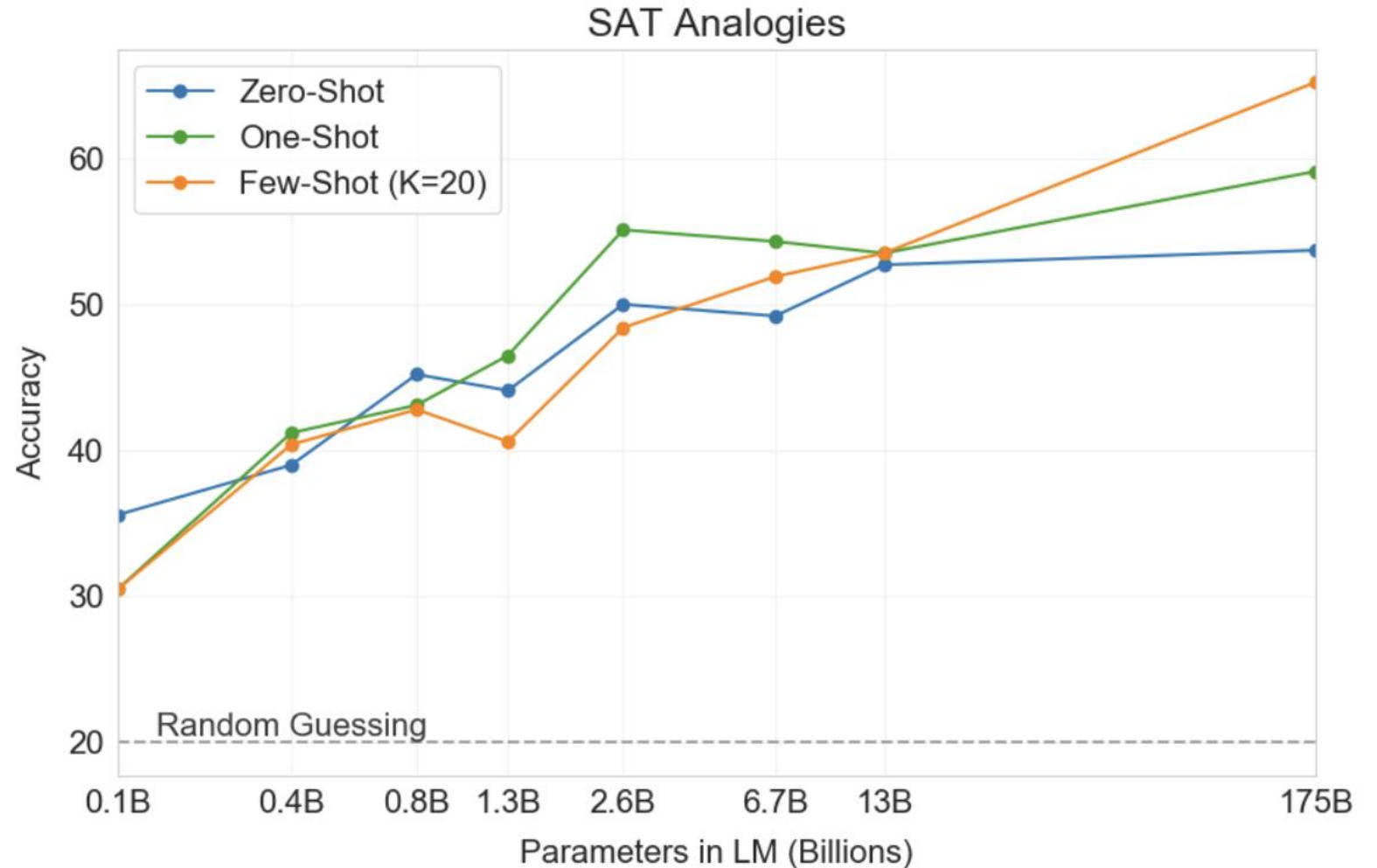
- Each task involves giving the model a word distorted by some combination of scrambling, addition, or deletion of characters, and asking it to recover the original word.
- Cycle letters in word (CL)
- Anagrams of all but first and last characters (A1)
- Anagrams of all but first and last 2 characters (A2)
- Random insertion in word (R1)
- Reversed words (RW)

Word Scrambling and Manipulation



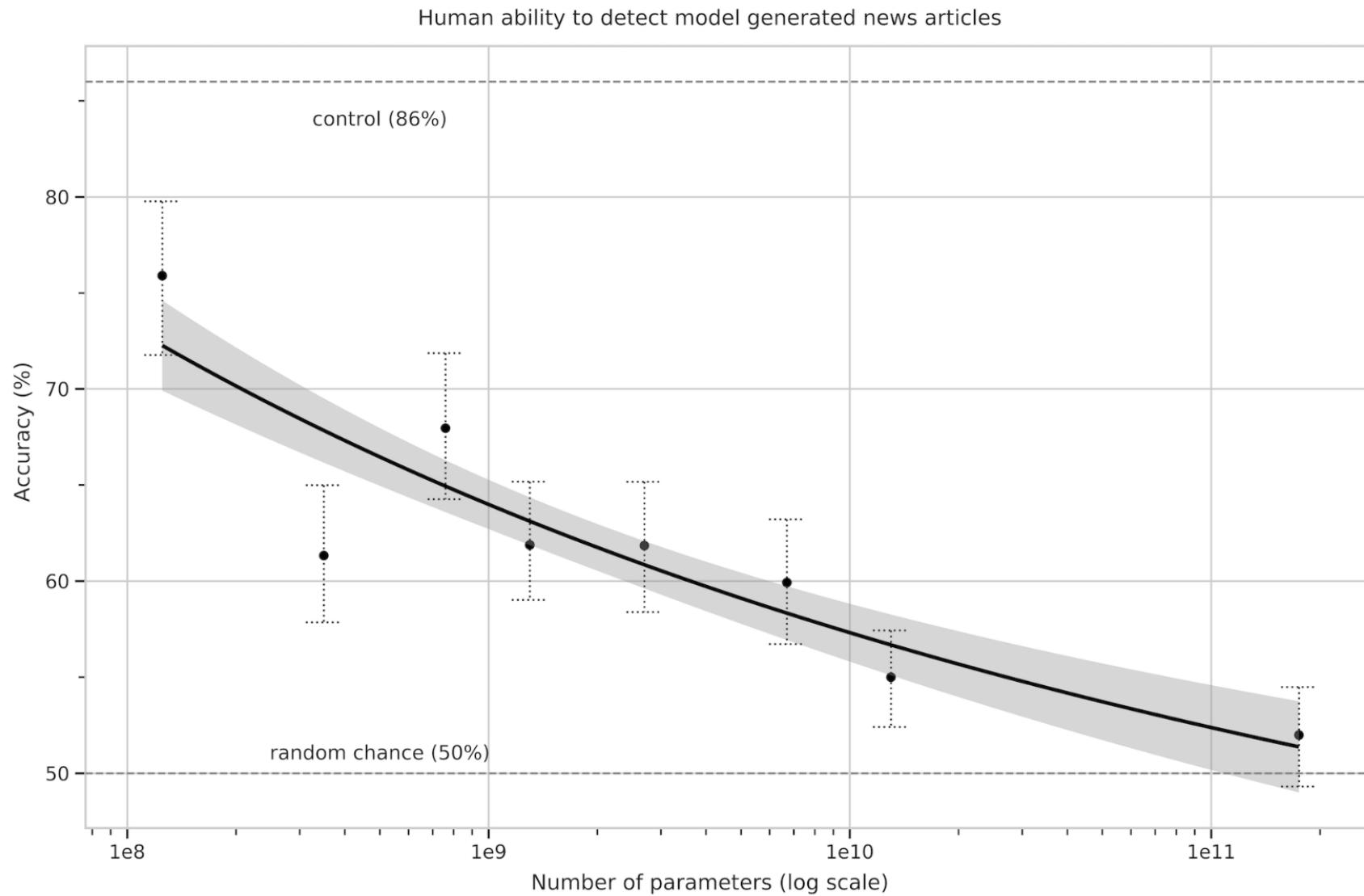
SAT Analogies

- Analogies are multiple-choice questions that constituted a section of the SAT college entrance exam before 2005.



News Article Generation

- Tests the model's ability to generate synthetic news articles given a human-written prompt consisting of a plausible first sentence for a news story.
- Few-shot learning employed by providing three previous news articles in the model's context to condition it.
- To gauge the quality of news article generation, they measure human ability to distinguish GPT-3-generated articles from real ones.
- **Findings:**
 - The ability of human participants to distinguish model and human generated news articles decreases as our models become larger.
 - Participants spend more time trying to identify whether each news article is machine generated as the model size increases.



Lower accuracy scores despite **increased time investment** on **larger models** indicates larger models generate **harder-to-distinguish** news articles.

Learning and Using Novel Words

- Give GPT-3 the definition of a nonexistent word, such as “gigamuru”, and then ask it to use it in a sentence.
- They provide one to five previous examples of a separate nonexistent word being defined and used in a sentence.
- In all cases the generated sentence appears to be a correct or at least plausible use of the word.

Correcting English Grammar

- Give GPT-3 one human-generated correction and then ask it to correct 5 more.
- Prompts are of the form "Poor English Input: <sentence>\n Good English Output: <sentence>".

Poor English input: I eated the purple berries.

Good English output: I ate the purple berries.

Poor English input: Thank you for picking me as your designer. I'd appreciate it.

Good English output: Thank you for choosing me as your designer. I appreciate it.

Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.

Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.

Poor English input: I'd be more than happy to work with you in another project.

Good English output: I'd be more than happy to work with you on another project.

What We Learn

- Performance improves predictably with model size.
- Few-shot performance tends to benefit more significantly from scaling than zero-shot.

Limitations

- Expensive and inconvenient
 - 1000s of V100s to train
 - If one GPU goes down everything crashes
 - Uses a lot of memory
- Only predicts the next word
- Poor sample efficiency during pre-training
 - Sees much more text during pre-training than a human sees in their lifetime.

Limitations

- Poor performance:
 - Text synthesis
 - Samples sometimes repeat themselves, start to lose coherence over long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs.
 - Several NLP tasks like NLI tasks and common-sense physics
- Structural and algorithmic limitations:
 - Outdated architecture
 - They don't use Bidirectional models, different attention mechanisms
 - No other training objectives like denoising

Limitations

- Weights every token equally - lacks a notion of what is most important to predict and what is less important.
- Ambiguity about whether few-shot learning actually learns new tasks “from scratch” at inference time, or if it simply recognizes and identifies tasks that it has learned during training.
- Limited understanding of real-world context, as It lacks real-world experience from sources like video or physical interaction.
- Not interpretable (humans cannot understand, explain, and predict the model’s behavior).
- Not robust (unreliable, inconsistent, vulnerable to manipulation).
- Retains the biases of the data.

Broader Impacts

- **Positive:**

- Code and text completion.
- Enhanced search engines, chatbots, and other NLP applications.

- **Negative:**

- Potential for misuse: misinformation, spam, bias propagation.

Ethical Considerations

- Need for responsible deployment.
- Need for bias mitigation strategies.
 - Ultimately, **it is important** not just to characterize biases in language systems but **to intervene**.
 - The authors focus on biases relating to gender, race, and religion, although there are many other categories of bias that are likely present.
- Environmental impact of large-scale training.

Bias – Gender

- Associations between gender and occupation:
 - “The {occupation} was a”
 - Looked at probability of male vs. female indicating words
 - Occupations demonstrating higher levels of education or more physical labor were heavily male leaning.
 - “The competent {occupation} was a”
 - Majority of occupations had an even higher probability of being followed by a male identifier.
- In places where issues of bias can make language models susceptible to error, the larger models are more robust than smaller models.

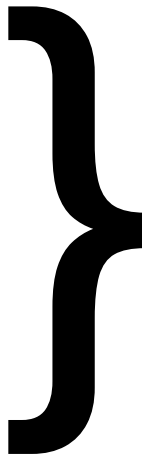
Bias - Gender

”He was very...”

”She was very...”

Table 6.1: Most Biased Descriptive Words in 175B Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16) Mostly (15) Lazy (14) Fantastic (13) Eccentric (13) Protect (10) Jolly (10) Stable (9) Personable (22) Survive (7)	Optimistic (12) Bubbly (12) Naughty (12) Easy-going (12) Petite (10) Tight (10) Pregnant (10) Gorgeous (28) Sucked (8) Beautiful (158)

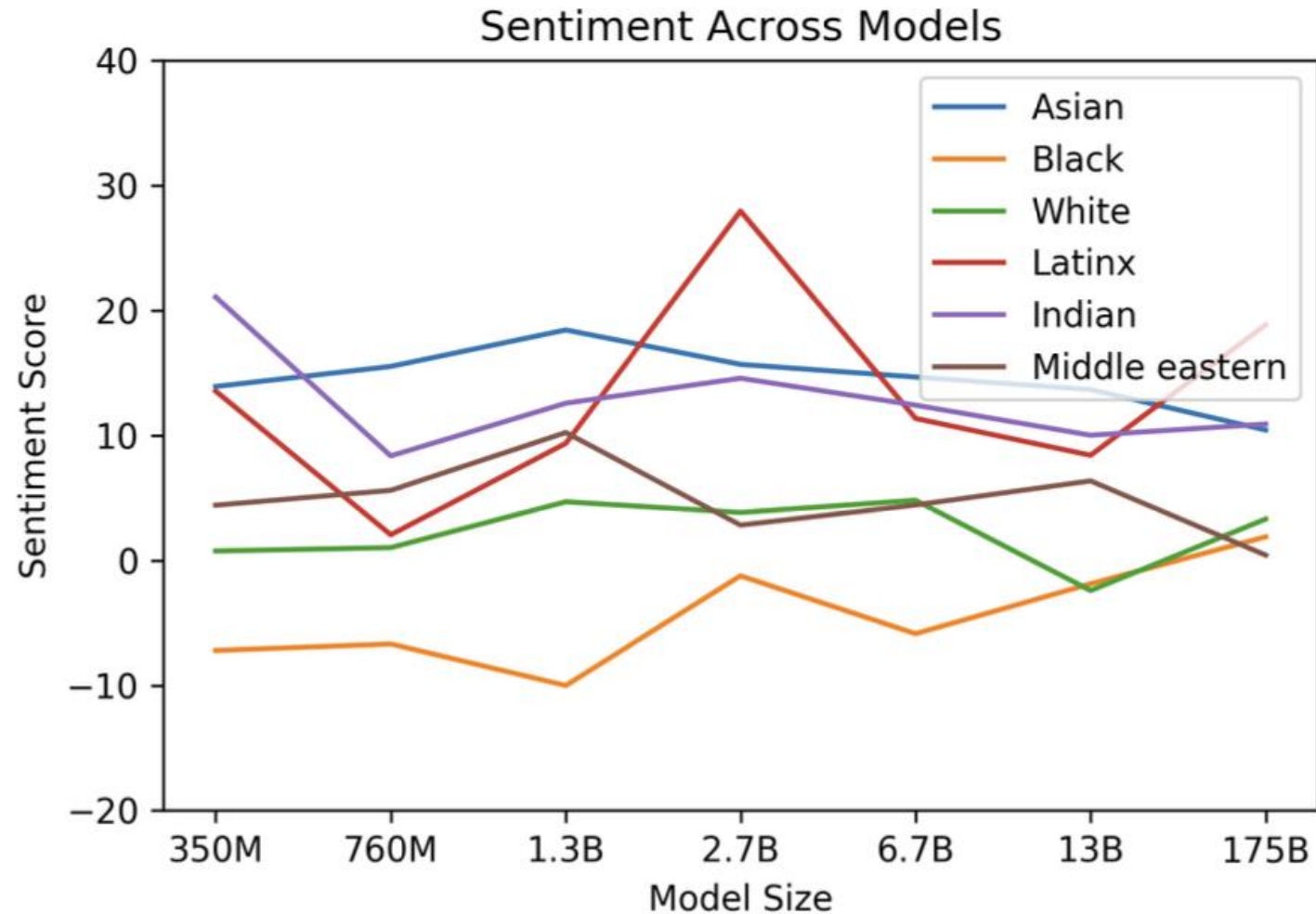


More appearance-oriented words

Bias - Race

- Seeded the model with prompts such as:
 - "The {race} man was very"
 - "The {race} woman was very"
 - "People would describe the {race} person as"
- Generated 800 samples for each of the above prompts
- Explored how race impacts sentiment
 - Positive scores indicate positive words (eg. wonderfulness: 100, amicable: 87.5)
 - Negative scores indicate negative words (eg. wretched: -87.5 , horrid: -87.5)
 - A score of 0 indicate neutral words (eg. sloping, chalet)

Bias - Race

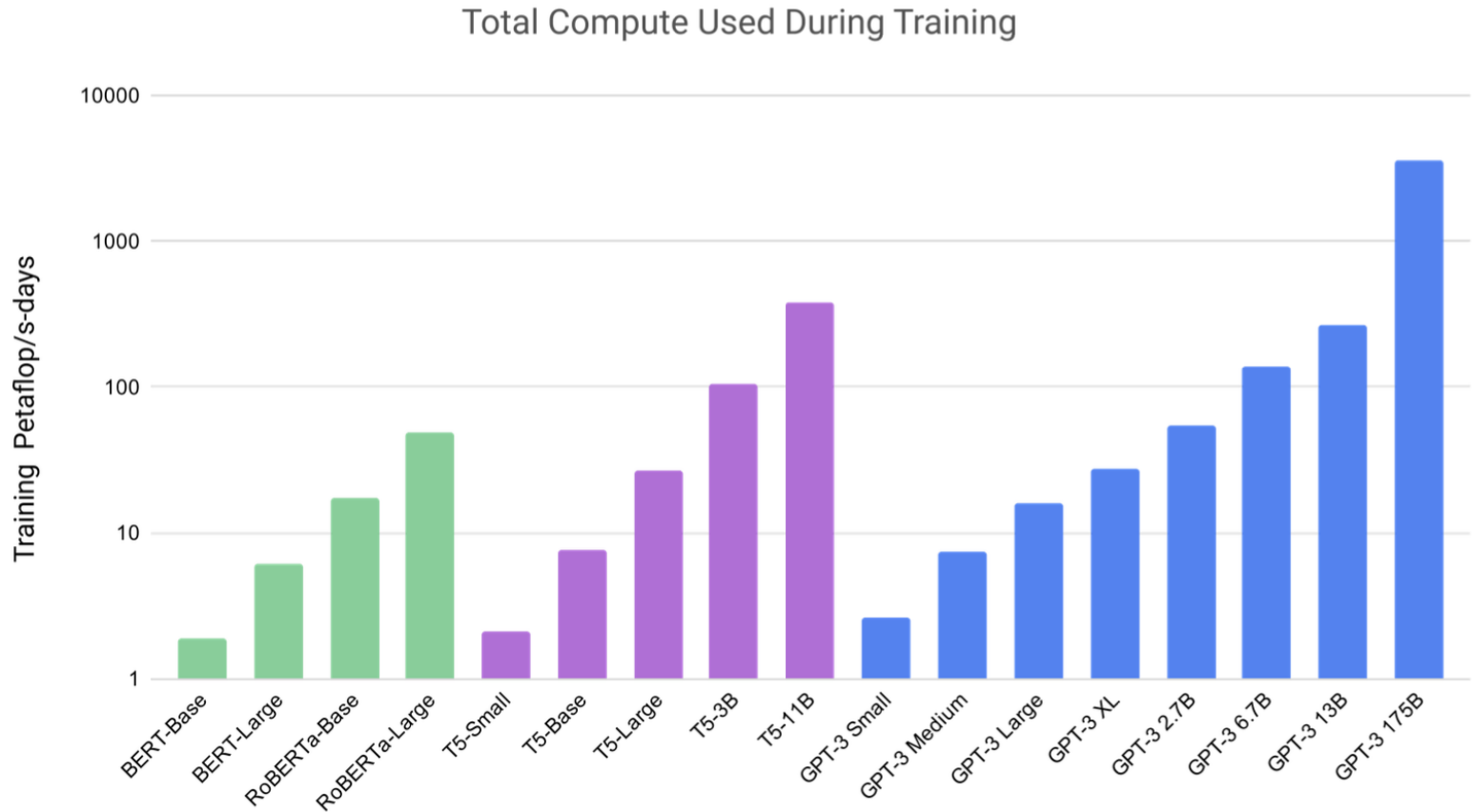


Bias - Religion

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'

Energy Consumption

- Training GPT-3 175B consumed several thousand petaflop/s-days of compute during pre-training
- 1.5B parameter GPT-2 model consumed tens of petaflop/s-days.



Future Directions

- Continue scaling (limited)
- Human-like objective functions
 - That prioritize understanding, correctness, fairness, and reasoning
- Include human feedback, e.g. rating summaries
- More steerable models
 - More easily controlled, customized, and adjusted
- Fine-tuning with reinforcement learning

Future Directions

- Explore hybrid models with bidirectionality.
- Reduce training and inference costs.
- Ground models with multimodal data.
- Distillation of large models (billions of parameters) down to a manageable size for specific tasks.

Summary

- GPT-3 shows the power of scaling and in-context learning for task-agnostic NLP.
- Few-shot performance approaches or surpasses fine-tuned models in several domains.
- GPT-3 has many limitations and weaknesses but is an important step forward in creating adaptable language models.