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

Vaishnavi Thesma, Akhila Devabhaktuni, and Zihao Wu

January 24, 2023

Outline

- Introduction
- Motivation
- Approach
- Model & Architectures
- Results
- Conclusion

Introduction

- Pre-trained language representations in NLP systems have been researched for various tasks
  - Single-layer representations learned word-vectors
  - Multi-layer RNNs to form stronger representations
  - Pre-trained recurrent and transformer language models are fine-tuned directly
    - Reading comprehension, question answering, textual entailment

Motivation

- Limitations of pre-trained recurrent/transformer language models
  - Need task-specific datasets and task-specific fine-tuning
  - Requires fine-tuning on extremely large datasets
- Large datasets have several limitations
  - Limits applicability/generalizability of language models for various language tasks
  - Potential of narrow training distribution
  - Not always necessary for many language tasks

Motivation (cont.)

- Increase in capacity of transformer language models shows improvements in NLP task performance
  - From 100 million parameter- to 17 billion parameter- models
  - Each increase improved downstream NLP tasks
- In this paper, GPT-3 is introduced
  - 175 billion parameter autoregressive language model
  - Previous paper presentation addressed increasing the size of the datasets for better model performance
  - Now, this paper focuses on increasing the size of the language model for better performance

Larger models and performance

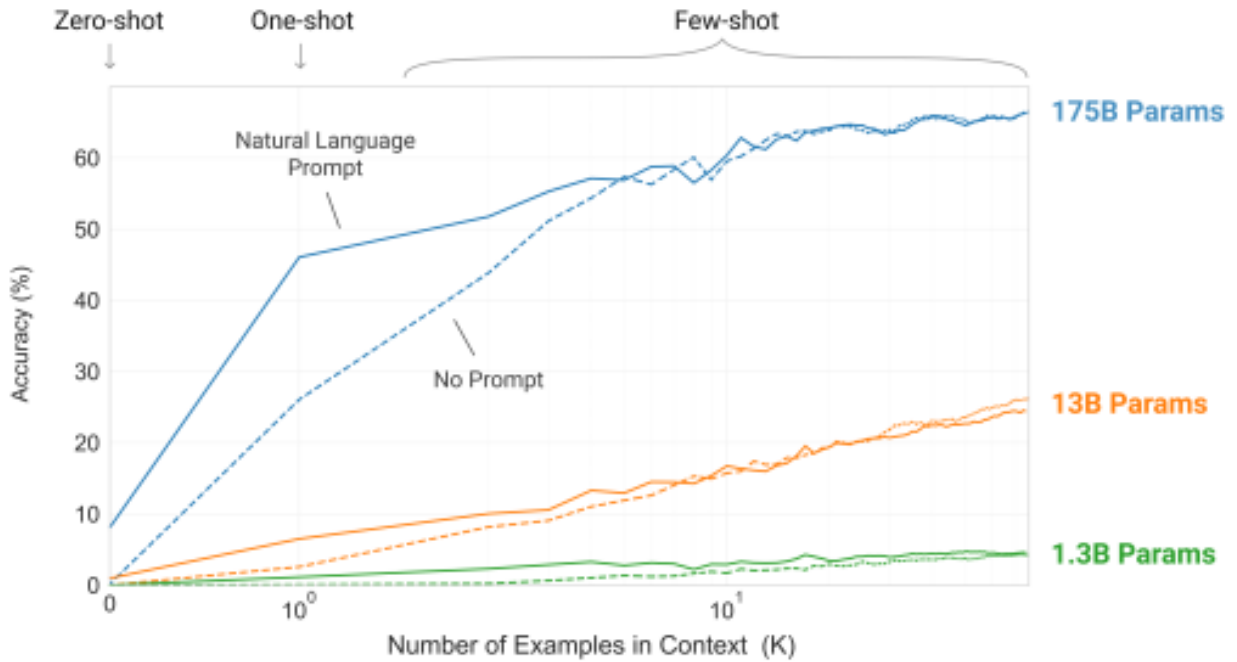


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

Approach

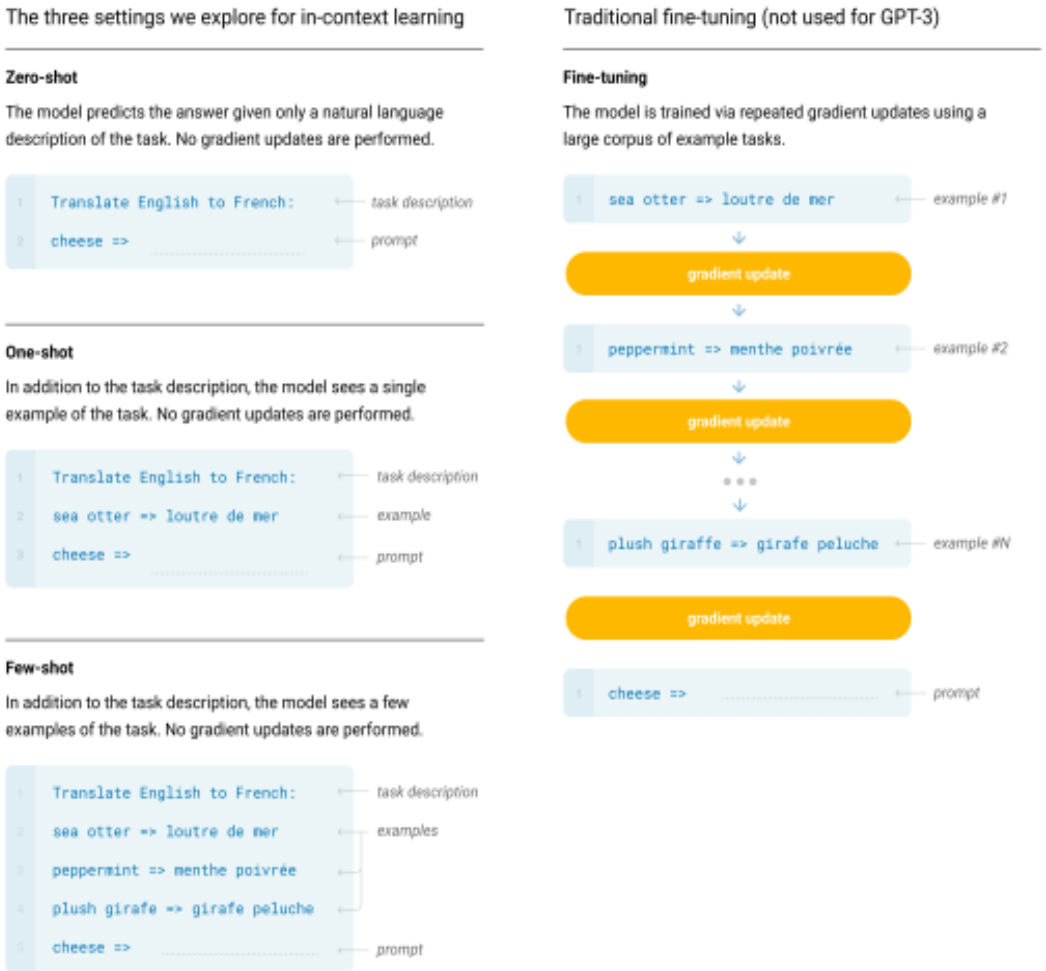


Figure 2.1: Zero-shot, one-shot and few-shot, contrasted with traditional fine-tuning. The panels above show four methods for performing a task with a language model – fine-tuning is the traditional method, whereas zero-, one-, and few-shot, which we study in this work, require the model to perform the task with only forward passes at test time. We typically present the model with a few dozen examples in the few shot setting. Exact phrasings for all task descriptions, examples and prompts can be found in Appendix G.

Approach (cont.)

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.

Model and Architectures

Generative Pre-trained models

OpenAI GPT1, GPT2, GPT3, LaMDA

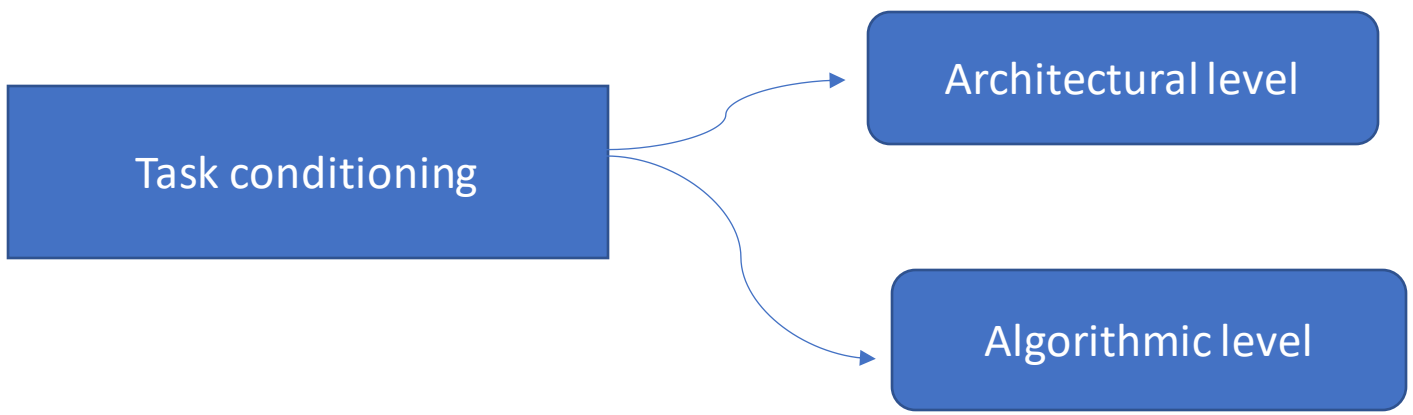
Language Models

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, \ldots, s_{n-1})$$

Language Models

$$p(output|input)$$

$$p(output|input, task)$$



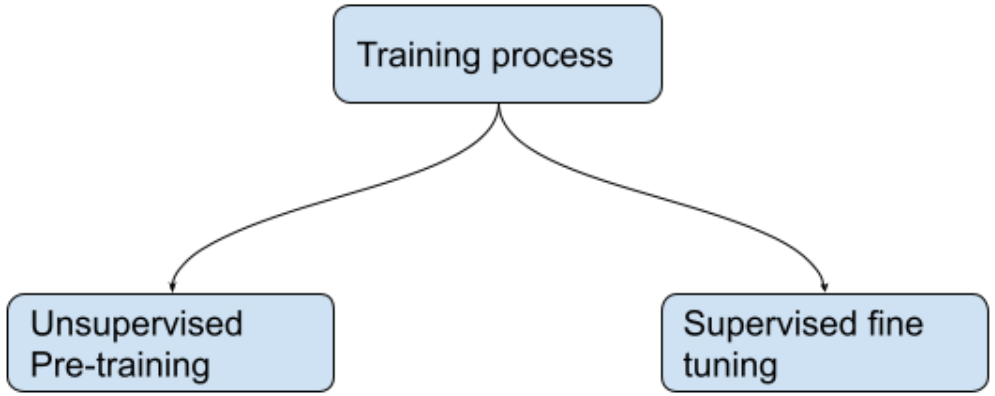
Language Modeling

- Multitask language model – MQAN
  - Translation training: (`translate to french, english text, french text`)
  - Reading comprehension: (`answer the question, document, question, answer`)
- Supervised Vs Unsupervised objective
- Objective: convergence of unsupervised objective

Training

WebText

OpenAI GPT/ GPT2/ GPT3



Unsupervised pre-training

$$\mathcal{U} = \{u_1, \ldots, u_n\}$$

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

T-DMCA: Transformer Decoder with Memory Compressed Attention

T-DMCA

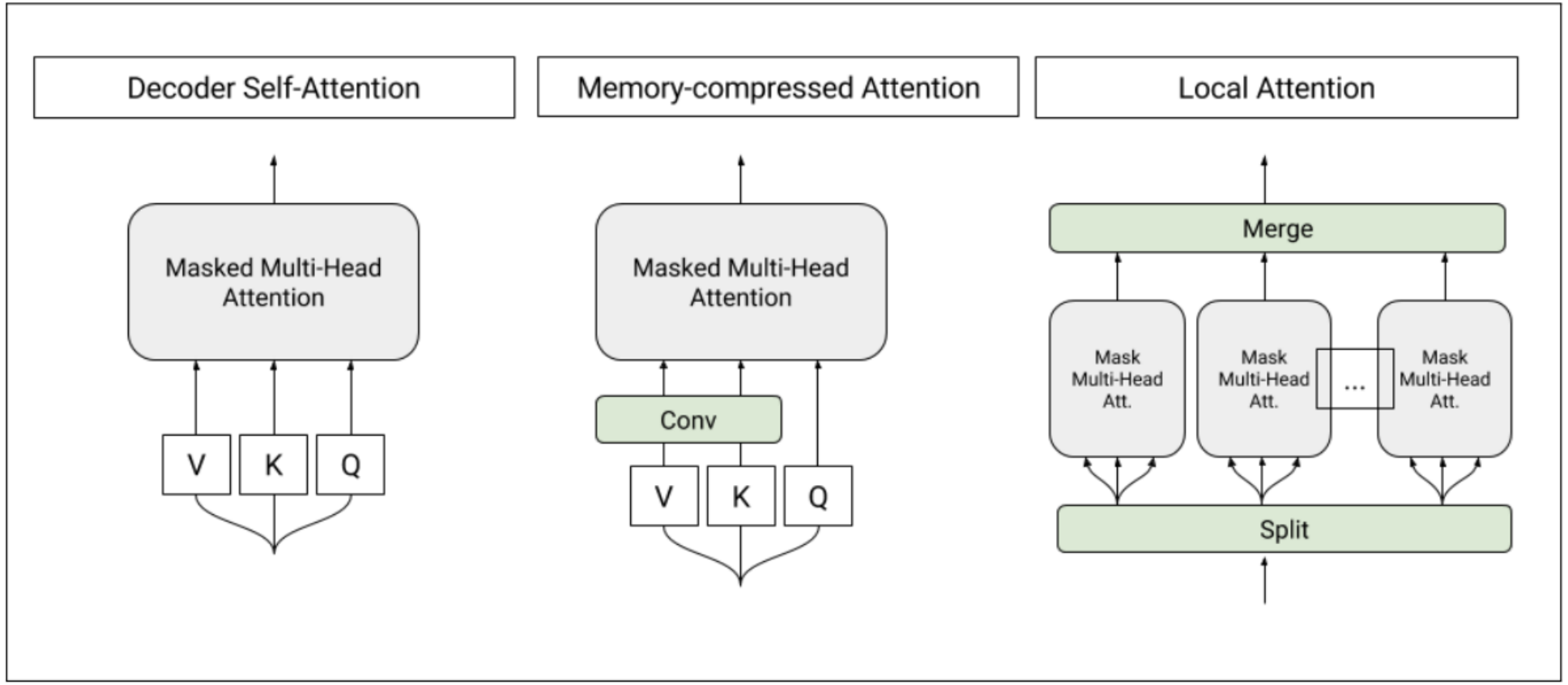
$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Local attention: 256 tokens

Memory compressed attention: global context capture

Final Architecture : LMLML

Self-attention in T-DMCA



Unsupervised pre-training

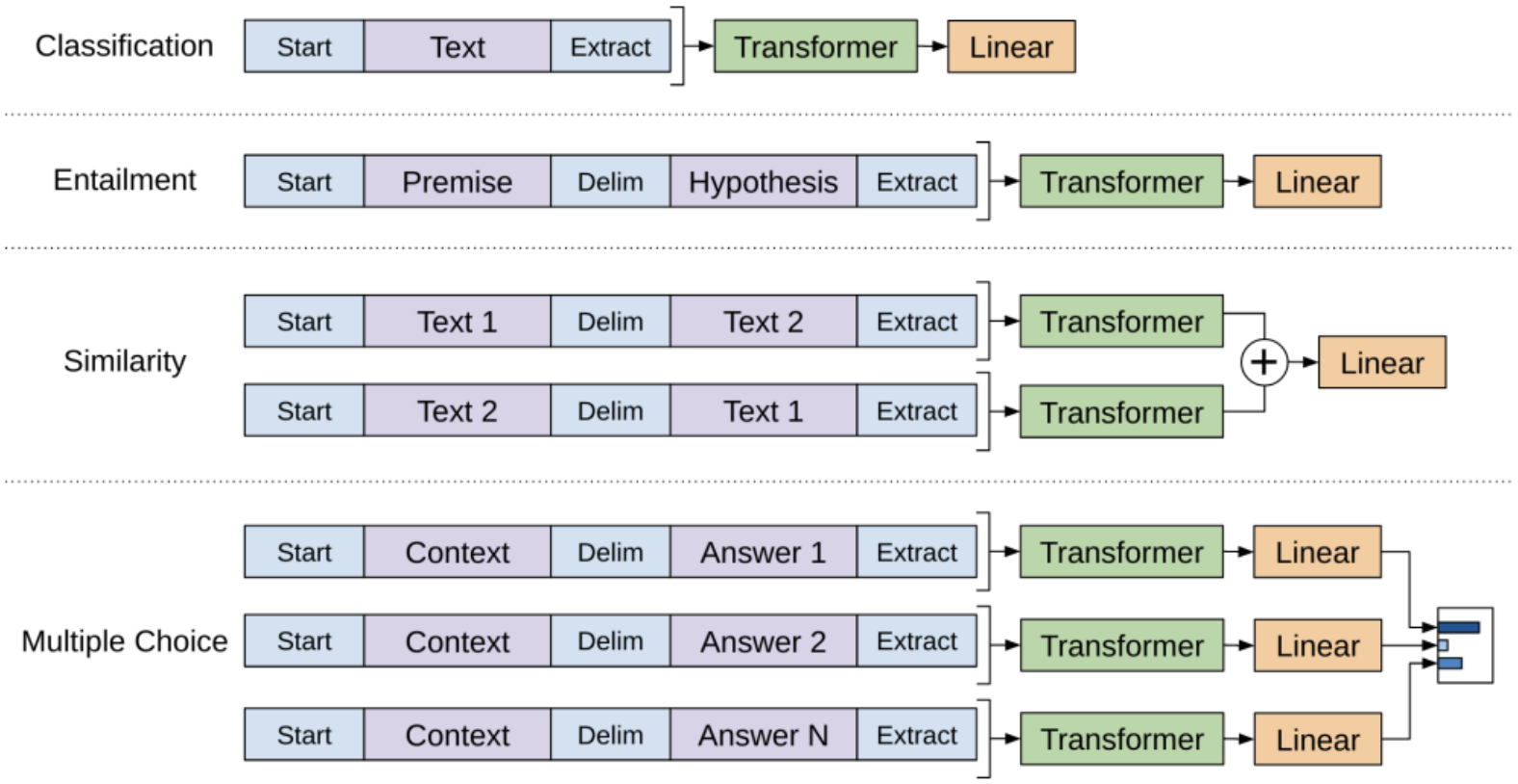
$$\mathcal{U} = \{u_1, \ldots, u_n\}$$

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

T-DMCA: Transformer Decoder with Memory Compressed Attention

$$\begin{aligned} h_0 &= UW_e + W_p \\ h_l &= \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n] \\ P(u) &= \texttt{softmax}(h_n W_e^T) \end{aligned}$$

Supervised fine-tuning



GPT2/ GPT3

- Layer normalization
- Modified initialization: Residual weights initialized by a factor of $1/\sqrt{N}$
- Reversible tokenization

GPT3

- Exception: Alternating dense and sparse attention patterns
- Goal: Performance Vs Model size

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Training Dataset

- Steps to improve quality of datasets
  1. Filtered version of CommonCrawl based on similarity
  2. Fuzzy deduplication at document level
  3. Added reference corpora.

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

Training Process

- Steps to improve quality of datasets
  1. Filtered version of CommonCrawl based on similarity
  2. Fuzzy deduplication at document level
  3. Added reference corpora.

- Contamination

- Process = Large batch size + small learning rate

Evaluation

- Few-shot learning: evaluation of K examples from evaluation set
- Picking K when separate development and test sets are available.
- Multiple choice task
- Binary classification
- Free form completion

Results

- language modeling
- question answering
- translate between languages
- Winograd Schema-like tasks
- commonsense reasoning
- reading comprehension tasks
- SuperGLUE benchmark suite
- NLI (Natural Language Inference)

Training Curves

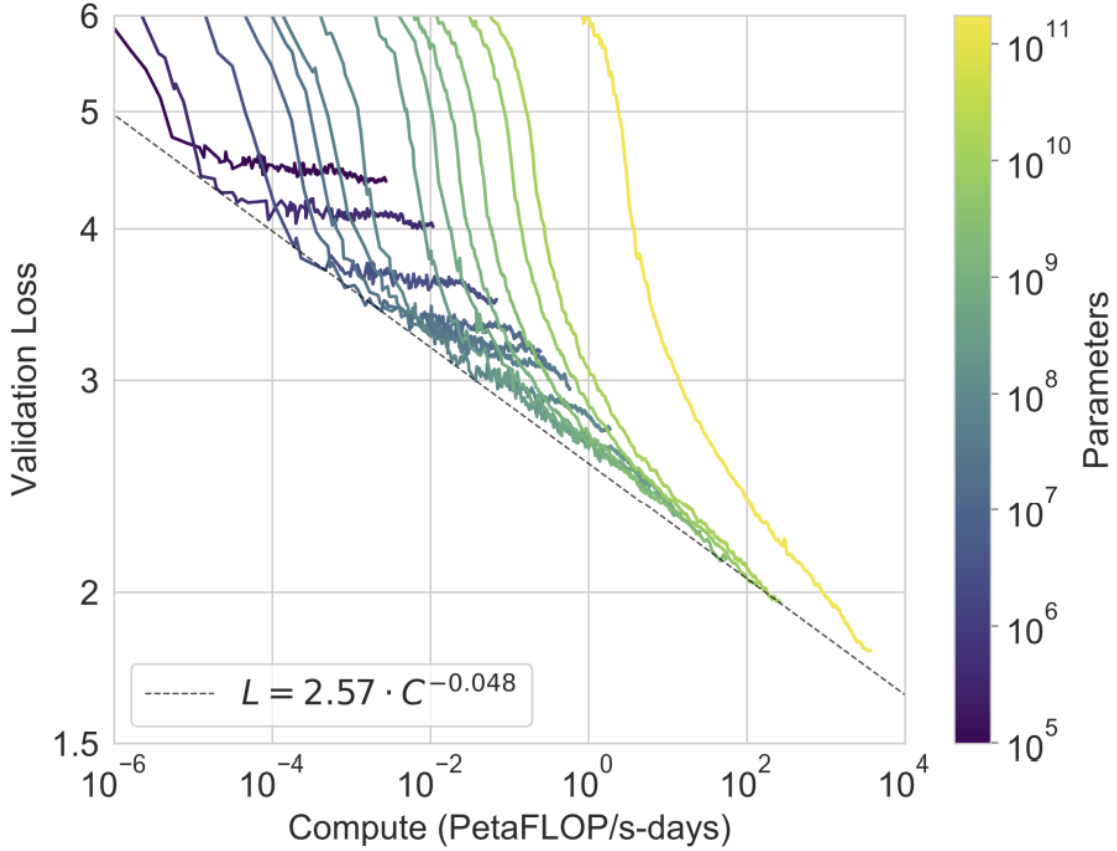


Figure 3.1: Smooth scaling of performance with compute. Performance (measured in terms of cross-entropy validation loss) follows a power-law trend with the amount of compute used for training. The power-law behavior observed in [KMH+20] continues for an additional two orders of magnitude with only small deviations from the predicted curve. For this figure, we exclude embedding parameters from compute and parameter counts.

Language modeling

Setting	PTB
SOTA (Zero-Shot)	35.8[a]
GPT-3 Zero-Shot	**20.5**

Table 3.1: Zero-shot results on PTB language modeling dataset. Many other common language modeling datasets are omitted because they are derived from Wikipedia or other sources which are included in GPT-3's training data.
[a] [RWC+19]

Language modeling

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

Table 3.2: Performance on cloze and completion tasks. GPT-3 significantly improves SOTA on LAMBADA while achieving respectable performance on two difficult completion prediction datasets. [a][Tur20] [b][RWC+19] [c][LDL19] [d][LCH+20]

Language modeling

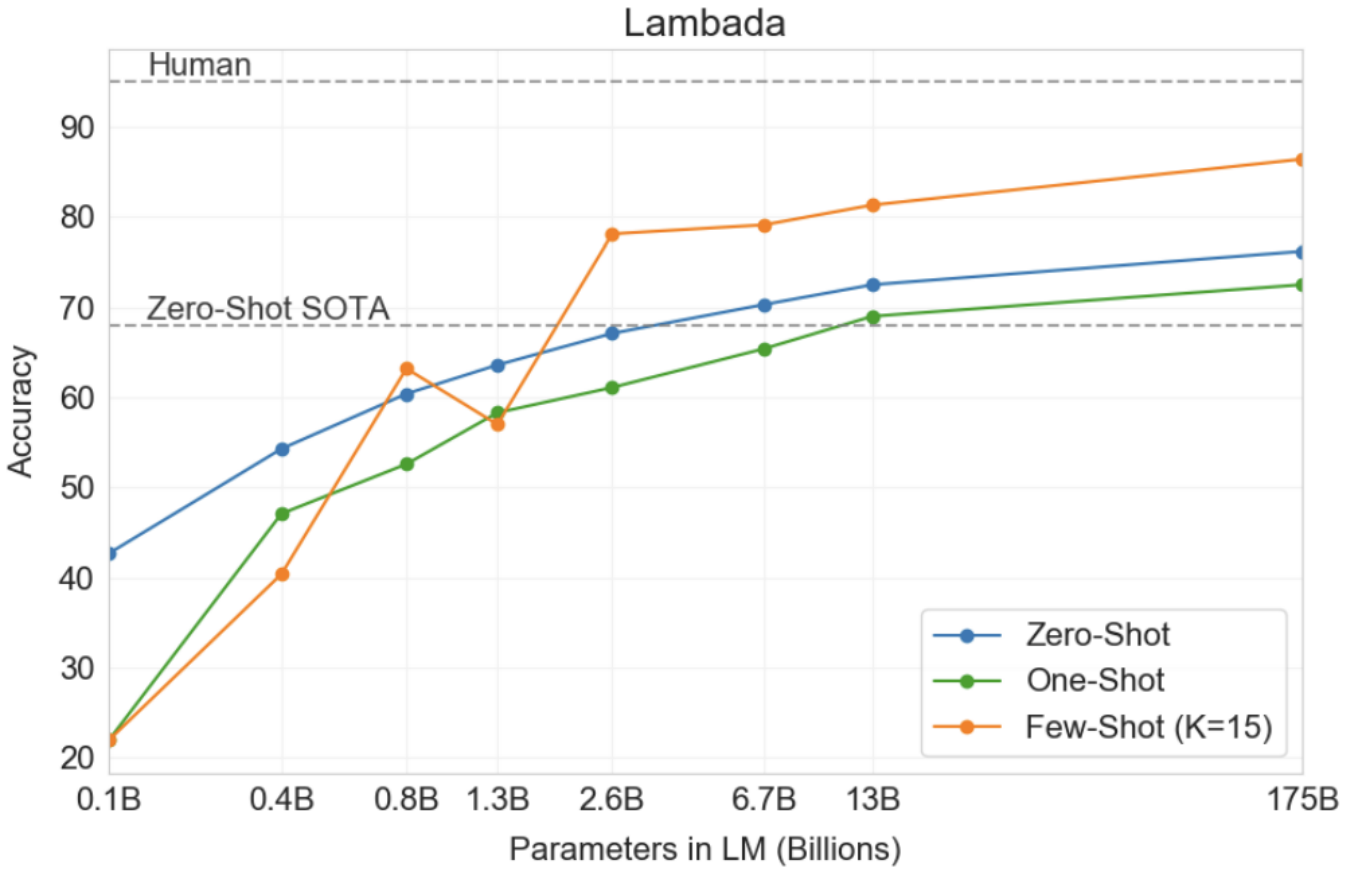


Figure 3.2: On LAMBADA, the few-shot capability of language models results in a strong boost to accuracy. GPT-3 2.7B outperforms the SOTA 17B parameter Turing-NLG [Tur20] in this setting, and GPT-3 175B advances the state of the art by 18%. Note zero-shot uses a different format from one-shot and few-shot as described in the text.

QA

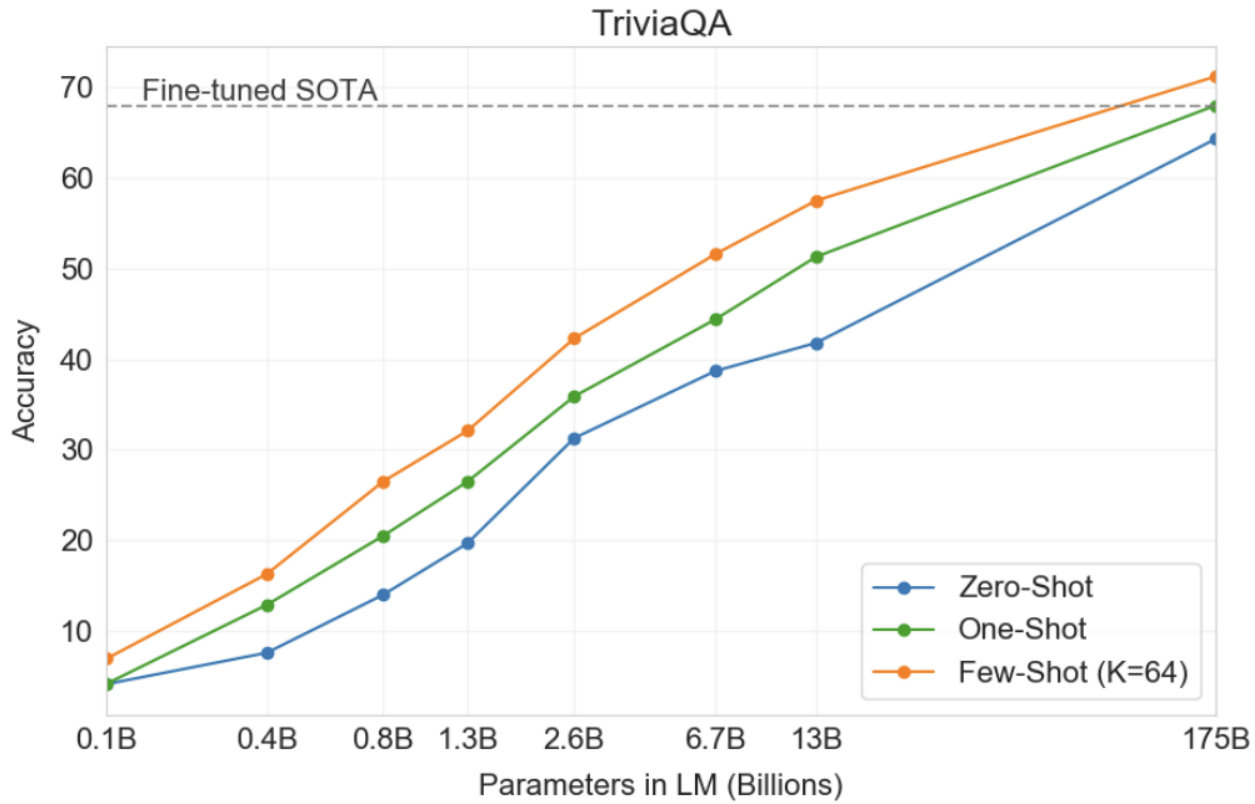


Figure 3.3: On TriviaQA GPT3's performance grows smoothly with model size, suggesting that language models continue to absorb knowledge as their capacity increases. One-shot and few-shot performance make significant gains over zero-shot behavior, matching and exceeding the performance of the SOTA fine-tuned open-domain model, RAG [LPP+20]

Winograd-Style Tasks

Figure 3.5: Zero-, one-, and few-shot performance on the adversarial Winogrande dataset as model capacity scales. Scaling is relatively smooth with the gains to few-shot learning increasing with model size, and few-shot GPT-3 175B is competitive with a fine-tuned RoBERTA-large.

Common Sense Reasoning

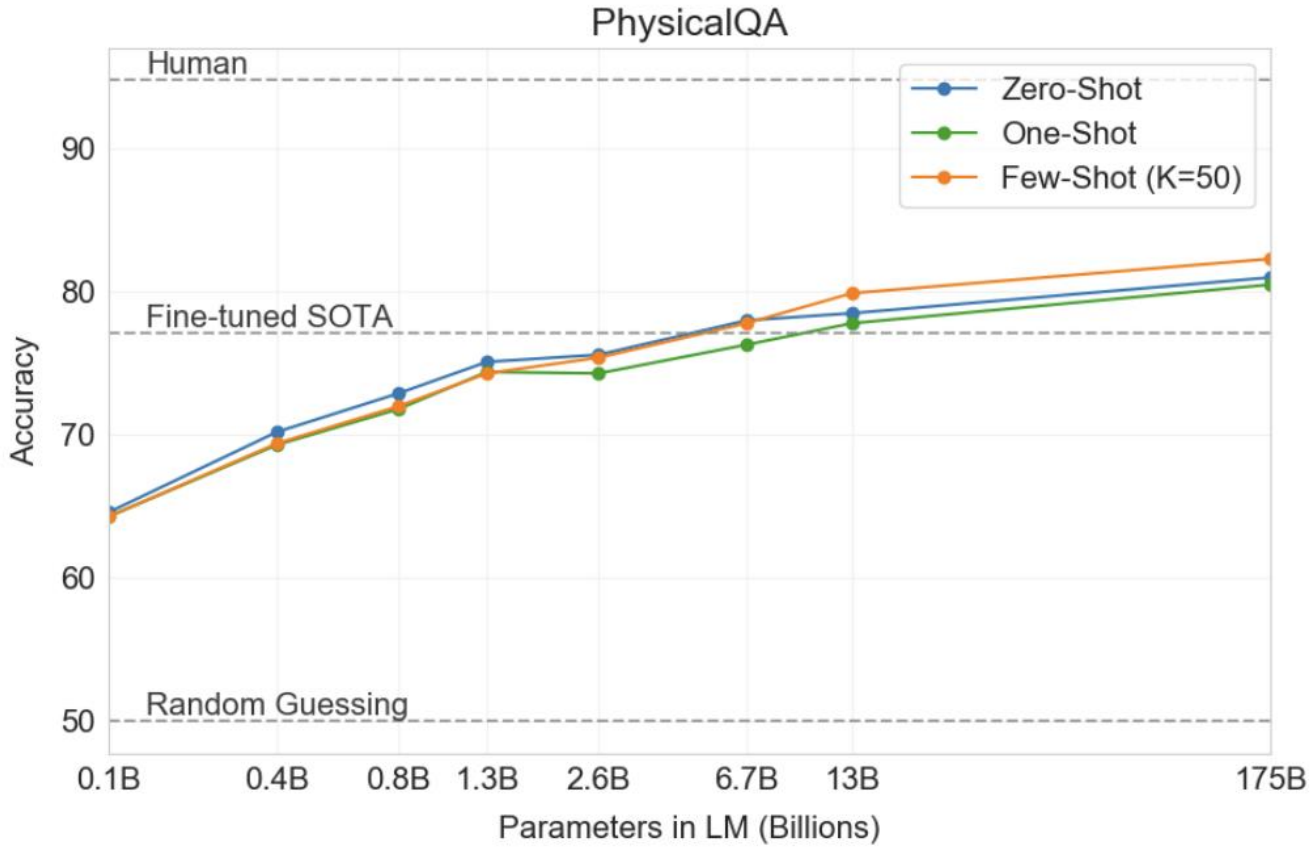


Figure 3.6: GPT-3 results on PIQA in the zero-shot, one-shot, and few-shot settings. The largest model achieves a score on the development set in all three conditions that exceeds the best recorded score on the task.

Reading Comprehension

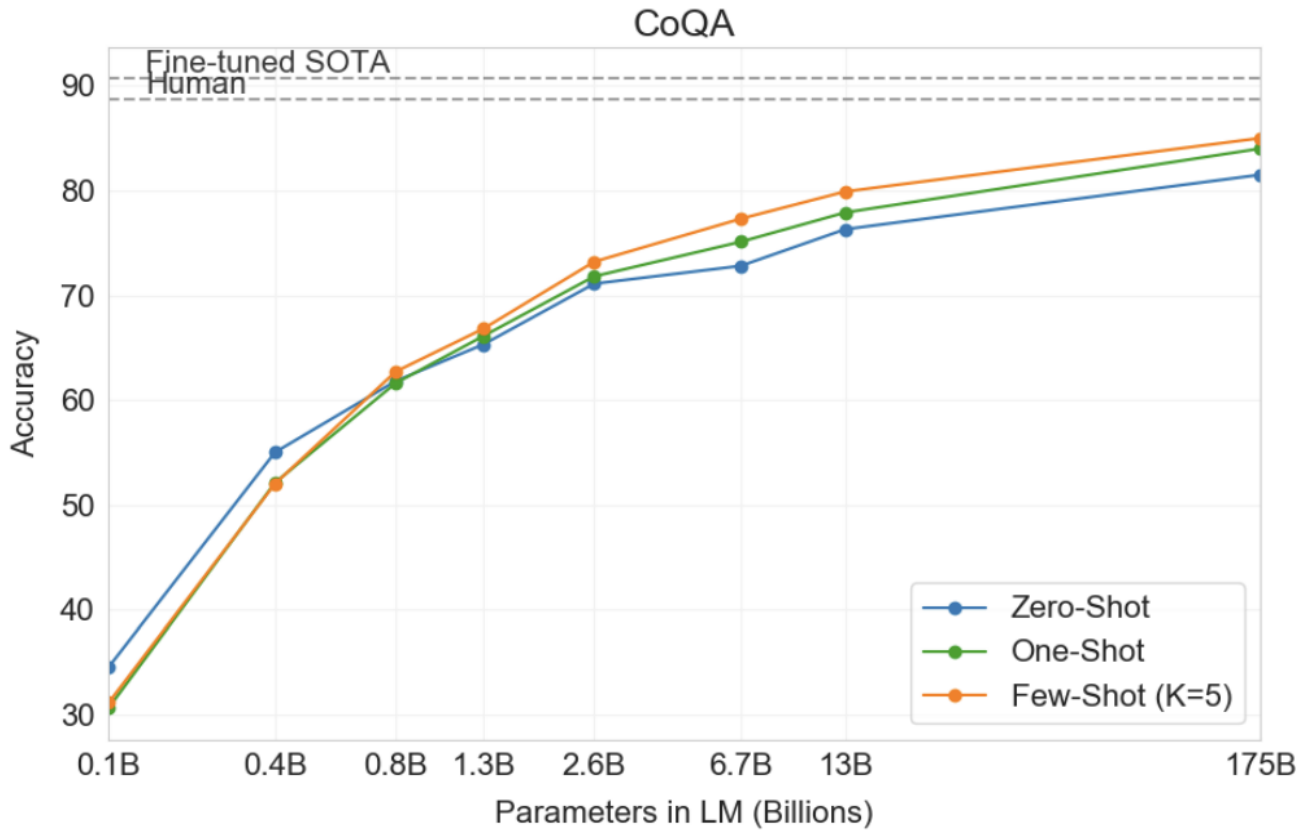


Figure 3.7: GPT-3 results on CoQA reading comprehension task. GPT-3 175B achieves 85 F1 in the few-shot setting, only a few points behind measured human performance and state-of-the-art fine-tuned models. Zero-shot and one-shot performance is a few points behind, with the gains to few-shot being largest for bigger models.

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

Table 3.8: Performance of GPT-3 on SuperGLUE compared to fine-tuned baselines and SOTA. All results are reported on the test set. GPT-3 few-shot is given a total of 32 examples within the context of each task and performs no gradient updates.

NLI

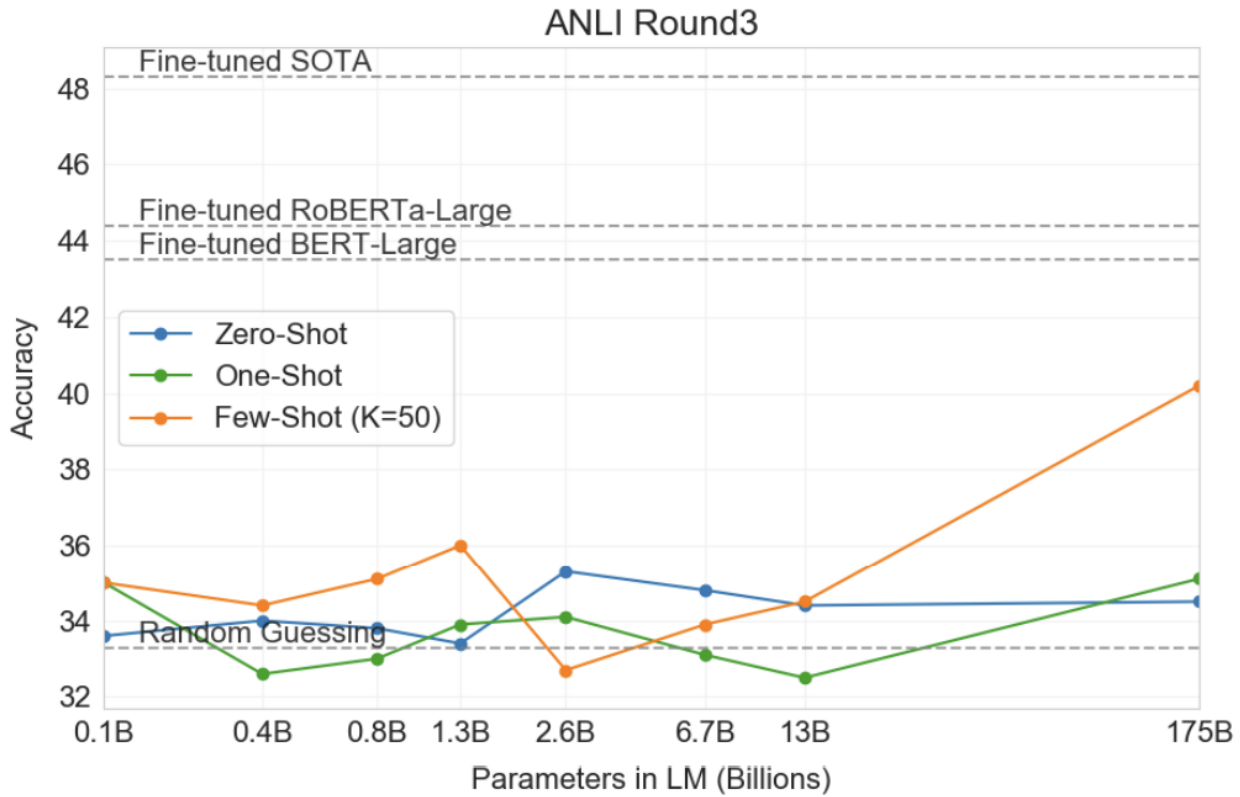


Figure 3.9: Performance of GPT-3 on ANLI Round 3. Results are on the dev-set, which has only 1500 examples and therefore has high variance (we estimate a standard deviation of 1.2%). We find that smaller models hover around random chance, while few-shot GPT-3 175B closes almost half the gap from random chance to SOTA. Results for ANLI rounds 1 and 2 are shown in the appendix.

Examples

- https://beta.openai.com/examples/

Limitation

- Do not include any bidirectional architectures or other training objectives such as denoising
- Poor sample efficiency during pre-training
- Expensive and inconvenient to perform inference
- Retains the biases of the data it has been trained on

Conclusion

- 175 billion parameter language model
- Strong performance on many NLP tasks
- Zero-shot, one-shot, and few-shot setting

References

- Generating Wikipedia by Summarizing Long Sequences, T-DMCA[link]
- Improving language understanding by Generative Pre-training [link]

Thank you!